Principles of Planetary Climate

R. T. Pierrehumbert

July 25, 2009

Preface

When it comes to understanding the why's and wherefores of climate, there is an infinite amount one needs to know, but life affords only a finite time in which to learn it; the time available before one's fellowship runs out and a PhD thesis must be produced affords still less. Inevitably, the student who wishes to get launched on significant interdisciplinary problems must begin with a somewhat hazy sketch of the relevant physics, and fill in the gaps as time goes on. It is a lifelong process. This book is an attempt to provide the student with a sturdy scaffolding upon which a deeper understanding may be hung later.

The climate system is made up of building blocks which in themselves are based on elementary physical principles, but which have surprising and profound collective behavior when allowed to interact on the planetary scale. In this sense, the "climate game" is rather like the game of *Go*, where interesting structure emerges from the interaction of simple rules on a big playing field, rather than complexity in the rules themselves. This book is intended to provide a rapid entrée into this fascinating universe of problems for the student who is already somewhat literate in physics and mathematics, but who has not had any previous experience with climate problems. The subject matter of each individual chapter could easily fill a textbook many times over, but even the abbreviated treatment given here provides enough core material for the student to begin treating original questions in the physics of climate.

The Earth provides our best-observed example of a planetary climate, and so it is inevitable that any discussion of planetary climate will draw heavily on things that can be learned from study of the Earth's climate system. Nonetheless, the central organizing principle is the manner in which the interplay of the same basic set of physical building-blocks gives rise to the diverse climates of present, past and future Earth, of the other planets in the Solar system, of the rapidly growing catalog of extrasolar planets, and of hypothetical planets yet to be discovered. A guiding principle is that new ideas come from profound analysis of simple models – thinking deeply of simple things. The goal is to teach the student how to build simple models of diverse planetary phenomena, and to provide the tools necessary to analyze their behavior.

This is very much a how-to book. The guiding principle is that the student should be able to reproduce every single result shown in the book, and should be able to use those skills as a basis for explorations that go beyond the rather limited display of results that can be presented in a printed tome of reasonable size. Similarly, the student should have access to every data set used to produce the figures in the book, and ideally to more comprehensive data sets that draw the student into further and even original analyses. To this end, I have set as a ground rule that I would not use reproductions of figures from other works, nor would I show any results which the student would not be able to reproduce. With the exception of a very few maps and images, every single figure and calculation in this book has been produced from scratch, using software written expressly for the purposes of this book and provided as an online software supplement. The computer

implementation have pedagogy as their guiding principle, and readability of the implementation has been given priority over computational efficiency . A companion to this philosophy is what I call "freedom to tinker." The code should all be in a form that can easily be modified for other purposes. The goal is to allow the student to first reproduce the results in the book, and then use the tools immediately as the basis for original research. In this, I have been much inspired by what the book *Numerical Recipes* did for numerical analysis. This book does not sell fish. Instead, it teaches students how to catch fish, and how to cook them. As gastronomical literature goes, the book before the reader is somewhat in the spirit of one of Elizabeth David's extended pedagogical discourses on food (with recipes interspersed), whereas *Numerical Recipes* is somewhat more in the spirit of a traditional cookbook like *Joy of Cooking*.

The software underlying this book was implemented in the open-source interpreted language `Python`, because it lends itself best to the design principles annunciated above. It has a versatile and powerful syntax but nonetheless is easy to learn. In my experience, students with no previous familiarity with the language can learn enough to make a substantial start on the computational problems in this book in only two weeks of self-study or computer labs. `Python` also teaches good programming style, and is a language the student will not outgrow, since it is easily extensible and provides a good basis for serious research computations. It will work on virtually any kind of computer, and being open-source the instructor does not have the bother and expense of dealing with licensing fees. I do hope that the student and instructor will fall for `Python` as madly as I have, but I emphasize that this book is not `Python`-specific. The text focuses on ideas that are independent of implementation. Specific reference to Python is confined to the online supplement and to the Workbook section of each chapter, where `Python`-specific advice is isolated in clearly demarcated *Python tips*. The instructor who wishes to make use of some other computer language in teaching the course will find few obstacles. The transparency and readability of `Python` is such that the `Python` implementations should provide a convenient aid to re-implementation in other languages. It is envisioned that `Matlab` versions of most of the software will ultimately be made available.

In this book I have chosen to deal only with aspects of climate that can be treated without consideration of the fluid dynamics of the Atmosphere or Ocean. Many successful scientists have spent their entire careers productively in this sphere. The days are long gone when leading-edge problems could be found in planetary fluid dynamics alone, so even the student whose primary interests lie in atmosphere/ocean dynamics will need to know a considerable amount about the other bits of physics that make up the climate system. There are many excellent textbooks on what is rather parochially known as "geophysical fluid dynamics," from which the student can learn the fluid dynamics needed to address that aspect of planetary climate. That does not prevent me from entertaining a vision of adding one more at some point, as a sequel to the present volume. This sequel, entitled *Things that Flow* would treat the additional phenomena that emerge when fluid dynamics is introduced. It would continue the theme of taking a broad planetary view of phenomena, and of providing students with the computational tools needed to build models of their own. It would take a rather broad view of what counts as a "flow," including such things as glaciers and sea ice as well as the more traditional atmospheres and oceans. We shall see; for the moment, this is just a vision.

Remarks on notation and terminology

Since I have in mind the full variety of planets in our Solar System and in extrasolar systems, there is the question of what kind of terminology to use to emphasize the generality of the phenomena. Should we create new terminology that emphasizes that we are talking about an arbitrary system, at the risk of creating confusion by introducing new jargon? Or should we adopt terminology that emphasizes the analogy with familiar concepts from Earth and our own Solar

System? For the most part, I have adopted the latter approach, which leads to a certain amount of Earth-centric terminology. For example, if I sometimes refer to "the sun" or "solar radiation," it is to be thought of as referring to whatever star the planet under discussion is orbiting, and not necessarily Earth's Sun or even a star like it. In the same spirit, the term *solar constant* will be used to refer to the rate at which a planet receives energy from its star (as defined precisely in Chapter **??**), regardless of what that star may be and where the planet may be located. One may thus talk about the solar constant for Mars, for Earth, for Gliese 581d, or for that matter the difference between Earth's solar constant in June and July; from this remark, it is clear that the solar "constant" is a rather inconstant constant, but I will stick to the terminology since it has considerable familiarity within the field of climate physics. I will use the notation $L_{\circledast}$ to refer to the value of this quantity for the planet under consideration, and use the same $\circledast$ subscript to refer to all properties of a planet's star and the electromagnetic radiation which emanates from it. The new notation is necessary in order to distinguish the value pertinent to a given planet from the specific *number* $L_{\odot}$ which refers to the corresponding mean flux from the Sun itself, measured in the present era at a standard distance of 1 astronomical unit (roughly the Earth's mean orbit). The reader should also be cautioned that astronomers usually use the symbol L to refer to a star's *luminosity*, or net power output, rather than the flux as measured at some particular orbit. Since the luminosity of a star affects the climate of a planet only through the energy flux at the planet's own orbit, however, I have taken the liberty of co-opting the symbol for this flux. Astronomers are free to think of the quantity as a form of apparent luminosity, seen from the planet's orbit.

More proper terms would be *stellar radiation* and *stellar constant*, but those unfortunately call to mind starlight from the night sky and seem potentially confusing (though I will gradually break in the use of the terms to help the reader get used to the idea that there are a lot of stars out there, with a lot of planets with a lot of climates). The radiation from a planet's star will also sometimes be referred to as *shortwave radiation*, to emphasize that it is almost invariably of considerably shorter wavelength than the thermal radiation by which a planet cools to space.

In a similar vein "air" will mean whatever gas the atmosphere is composed of on the planet in question – after all, if you grew up there, you'd just call it "air." When I need to refer to the specific substances that make up our own atmosphere, it will be called "Earth air." All this is a bit like the way one refers to Martian "geology" and "geophysics," so we don't need to refer to Areophysics on Mars and Venerophyics on Venus when we are really talking about the same kind of physics in all these cases. Eventually, we will all need to learn to get used to terms like "periastron" as a generalization of "perihelion," as the focus of the field shifts more to the generality of phenomena amongst planetary systems.

To improve the readability of inline equations, I will usually leave out parentheses in the denominator. For example, $a/2\pi$ is the same thing as $\frac{a}{2\pi}$, whereas I would write $(a/2)\pi$ or $\pi a/2$ if meant $\frac{a}{2}\pi$ was intended.

With few exceptions, SI units (based on kilograms, meters and seconds) are used throughout this book. To avoid the baggage of miscellaneous factors of 1000 floating around, when counting molecules kilogram-moles are used, denoted with a capital, i.e. $Mole$. Thus 1 $Mole$ of a substance is the number of molecules needed to make a number of *kilograms* equal to the molecular weight – one thousand times Avogadro's number. There are a few cases where common practice dictates deviations from SI units, as in the use of millibars (mb) or bars for pressure when Pascals (Pa) involve unwieldy numbers, or the use of cm^{-1} for wavenumbers in infrared spectroscopy.

How to use this book

The short exercises embedded in the text are meant to be done "on the spot," as an immediate check of comprehension. More involved and thought-provoking problems may be found in the

accompanying Workbook section at the end of each chapter. The Workbook provides an integral part of the course. Using the techniques and tools developed in the Workbook sections, the student will be able to reproduce every single computational and data analysis result included in the text. The Workbook also offers considerable opportunities for independent inquiry launching off from the results shown in the text.

There are four basic kinds of problems in the Workbook. Some calculations are analytic, and require nothing more than pen and paper (or at most a decent pocket calculator). Others involve simple computations , data analysis, or plotting of a sort that can be done in a spreadsheet or even many commercially available graphing programs, without the need for any actual computer programming. Many of these problems involve analysis of datasets from observations or laboratory experiments, and all critical datasets are provided in the online supplement to this book in a tab-delimited text format which can be easily read into software of any type. Students who have competence in a programming language, either from prior courses or because the instructor has integrated programming instruction into the climate sequence (as is done at The University of Chicago) have the option of doing these problems in the programming language of their choice. They should be encouraged to do so, since these simpler problems make good warm-up exercises allowing students to consolidate and hone their programming skills. The third class of problems requires actual programming, but can easily be carried out from scratch by the student in the instructor's language of choice (perhaps with the assistance of some standard numerical analysis routines). While just about any language would do, I have found that interactive interpreted languages such as `Python` and `Matlab` offer considerable advantages, since they provide instant feedback and encourage exploration and experimentation.

The fourth class of problem consists of major projects which would take a lot of time for the student to implement from scratch. Some of these become relatively straightforward, though, if the basic tools such as the a moist adiabat routine of Chapter 2 or the "homebrew" exponential sum real gas radiation code of Chapter 4 are provided as tools the student can use in doing the problems. I have provided Python implementations of all such tools, but this is the class of problems that poses the greatest challenge for the instructor who has not adopted Python as the language of choice for instruction. It is highly recommended that the instructor take the time to re-code at least some of the critical tools in the language of choice if Python is not being used. I have provided algorithm descriptions in the text that are independent of the computer implementation, and examination of the Python code should also help. The object-oriented features and powerful list handling capabilities of Python mean that translations into languages that do not support these language features are apt to be more complicated and unwieldy, but still it is only the exponential sums radiation code that is likely to pose any real challenge to the instructor. It is an excellent training exercise to pursue the necessary code development as a team-project effort with a few enthusiastic graduate students, who can then serve as teaching assistants for the course.

There are just a handful of basic computational methods and computer skills needed to do the Workbook problems and to reproduce all the calculations in this book. None of the calculations require any more computer power than is available on any decent laptop computer. The required numerical skills are outlined and exercised in the Chapter 1 Workbook section, which the student should master before proceeding to the rest of the book. I have not provided detailed discussions of basic algorithms like ordinary differential equation integration, interpolation, or numerical quadrature, since they are well described in the book *Numerical Recipes*, available from Cambridge University Press. *Numerical Recipes* should be viewed as an essential companion to this book, though only a small part of the material in that opus is actually required for the problems that concern us here.

With very few exceptions, all the data sets needed to produce the figures in the text, or

needed do do the data analysis problems in the Workbook sections, have been provided in the online supplement in plain text tabular form. These are organized into subdirectories according to the chapter to which the data is pertinent. These data sets can be plotted and analyzed using virtually any software. The only exceptions to the text format are a very few data sets used in making temperature and humidity maps in Chapters 7 and 9, which are too large to handle conveniently in text form. These data sets are in the machine-independent `netCDF` format, but they are not heavily used and it is not necessary to learn how to deal with this file format for the purposes of working through this book. The format is widely used for archiving numerical simulations and observations, so effort expended in this direction will be well repaid. It is very easy to read `netCDF` data using various packages written for use with `Python`, and not very hard to do it within `MATLAB` either.

There are three groups of Python courseware that are provided as part of the online supplement. Even non-Pythonista should be aware of what is there, since it is recommended that the same basic organization be adopted regardless of the computer language used for instruction.

- First, there are the basic courseware modules `ClimateUtilities.py`, `phys.py` and `planets.py`. They are intended to be placed in a publicly available directory which the students' Python interpreter will look for when loading add-on software (called `modules` in Python). The student should be able to read these files so as to get a better understanding of what they are doing, but it is not expected that the student will need to modify them.

  `ClimateUtilities.py` provides basic graphics, input/output, data manipulation and numerical analysis utilities. The modules `phys.py` and `planets.py` are intended to replace the data tables that typically appear on the endpapers and in the appendices of books such as this one. Being software rather than paper, they can include a more versatile level of data organization and can include functions as well as static data. However, since they are also human-readable text files, even the non-Pythonista can consult them in order to find needed data.

- Second, there are the *Chapter Scripts*, organized in subdirectories according to the chapter to which they pertain. These reproduce all the computations and figures appearing in the respective chapters, provide the means of further explorations, and also illustrate techniques needed to solve the Workbook problems. Some instructors will want to have the student refrain from examining the chapter scripts until they have had a go at implementing the ideas on their own, while others may want to make them immediately available as a study aid. Whenever they are made available, students are expected to have their own individual copies of these, as the basic intended use of these scripts is that the student will modify them and customize them, and re-run them as needed.

- Third, there are the *Solution Scripts*, which carry out solutions to selected Workbook problems. Access to these requires an instructor password, and they are intended to be doled out to students after they have turned in their own work.

The online supplement also includes Python tutorials, and various sample scripts illustrating numerical techniques and basic data analysis tasks. Software requirements and tips on installation are provided here as well. Solution write-ups for selected Workbook problems are also provided (instructor password required); these are for the most part language-independent, even where the calculations themselves were done in Python.

Although I have tried to rely primarily on calculations done with software that was written expressly for this book – and which the student can read, understand, and customize – at a

few points I have found it necessary to make use of calculations carried out with a full-featured terrestrial radiation model. For this purpose I have used the `ccm` column radiation model from the National Center for Atmospheric Research. For the most part, I have designed the problems so that they can be done using various polynomial fits to calculations with this model; it is not strictly necessary for the student to have access to the model. Nonetheless, it is desirable that the student be able to reproduce the results on his or her own. A stand-alone `FORTRAN` version of the `ccm` radiation model can be downloaded from public sources and used to do the needed calculations, but to make life easier for the Pythonista, I have included as part of the courseware a user-friendly Python interface to the `ccm` model, which makes it easy to use the model in conjunction with other Python computations. More details are provided in the online supplement.

Every author like to think that their book will occasionally be perused from time to time even in a century or two. I think of this myself, when reading through the crumbling though still-illuminating pages of Arrhenius' work, or the less famous though equally crumbling (and still-illuminating) papers of Frank Very. Whether the reader of the future (should I be so lucky) is absorbing this material through crumbling paper or neural implant, the text and equations will still in some sense be readable. One cannot dare to hope, however that the associated software used in the computations for this book will still run on the hardware prevailing in the future, though one can hope that the underlying algorithms are eternal. Even the Python of 2050 is unlikely to look like the Python of 2010, if indeed the language still exists at all. For this reason, I have minimized the discussion of the detailed sofware implementation within the text, and left that to the supplementary online material. Surely, while Frank Very's ideas on radiative transfer are still of interest, any description of the charts and graphical techniques for doing the calculations at the time would have at most historical interest. The associated Python software is meant not to be a static thing, but a living entity, which will be adopted, ported and mutated by the community of users as necessary. It could well be that quantum computers of the next century will allow direct line-by-line calculations to replace all the approximate radiative fanciness developed in this book, but even then people will need a good set of example programs in order to help build an understanding of the underlying physics. Computation is not understanding. The calculations embodied in the suite of software upon which this book is built are intended to provide the nucleation point for understanding.

Prerequisites and suggested syllabi

Before tackling the material in this book, the student should have attained a good mastery of classical mechanics, such as is typically provided by a first-year college physics course or an advanced secondary school course. This is not so much because classical mechanics itself is heavily exercised in this book, as because classical mechanics introduces the student to the necessary kind of problem solving skills, as well as building foundational concepts such as energy conservation and its use in problem-solving. It would also be useful for the student to have some familiarity with the basics of electromagnetism, including the concept of electric and magnetic fields and the forces they exert on charged particles; it is certainly not necessary for the student to have dealt with Maxwell's Equations in their full glory, though. The treatment of thermodynamics in this book is designed to be self-contained, though a student with some prior exposure to the subject in a physics course will be able to approach the material on a deeper level. There is some lightweight use of chemical kinetics and equilibrium in Chapter 8, but a self-contained, if minimal, introduction to the subject has been provided.

As for mathematics, all that is really required is a thorough understanding of single-variable differential and integral calculus, including first order ordinary differential equations. There is some use of second order differential equations in Chapters 5 and 7, but with help most students will be able to grasp the generalization even if they haven't formally studied the subject. The discussion

of the diffusion equation in Chapter 7 makes use of a one-dimensional partial differential equation, but this material does not require a very deep understanding of the subject and for the most part can be grasped intuitively from a physical basis. In fact, for students who haven't before dealt with PDE's, or who are rusty on the subject, this material and the associated problems serve as a good refresher. There is some optional material in the final chapter which exercises multivariable calculus more heavily, but the discussion is designed so that the details can be skipped if necessary.

It takes several courses to cover the material in this book, where by a "course" I mean 30 hours of lectures in a typical 10-week quarter, or 45 hours of lectures in a typical 15-week semester. Europeans using different quanta of instruction can calibrate the following accordingly.

For complete beginners, Chapters 2 and 3, with just a brief dip into the material of Chapter 1, plus all the necessary computer and algorithm preliminaries, fit in a one-quarter introductory course, though with little room for digressions. It is essential to introduce the students to some programming environment sophisticated enough to handle numerical differential equation integration. Even if most of the planetary history information in Chapter 1 is skipped, or left to self-study, the material in the Computational Toolkit section of the Chapter 1 Workbook should be covered, since it will be needed to do the rest of the problems. It usually works best to save lecture time by having the students learn programming and algorithms in lab or section meetings, supplemented by copious use of computational examples in the course of the lectures. When using Python, I find that students usually have enough basic skills to do the computational problems after about two weeks of such training; more advanced programming and numerical analysis skills can be introduced as needed. Besides covering the Computational Toolkit problems, it is a good idea to have the students work through the problems covering basic physics and chemistry, to get them up to speed on the fundamental concepts they will need to proceed; the "chemistry" covered in these problems is mostly a matter of understanding how to do problems involving molecular weight.

In a semester, or for students who already possess a good knowledge of either thermodynamics or basic computer skills, a bit more material can be fit in. This could be a full coverage of the Earth and planetary history material of Chapter 1, a more thorough treatment of programming and numerical methods, or inclusion of the grey-gas portion of Chapter 4.

Chapters 4 and 5 can form the basis of a one-semester course for advanced undergraduates or beginning graduate students, but I have sometimes taken as much as a full quarter just to cover the material in 4 in depth, with plenty of time allowed for simulation projects using the material.

The surface energy balance material in Chapter 6 is less fundamental to planetary climate than some of the other topics addressed, but it is something that everybody needs to know eventually. It could be paired naturally with Chapter 4 in place of scattering. My own preference is to truncate the material somewhat and teach it together with the material on the seasonal cycle and geographic variations in Chapter 7, which is truly fundamental and essential. One cannot begin to understand the issues surrounding Pleistocene ice ages without the material in this chapter.

Chapter 8, which deals with planetary formation and evolution of atmospheric composition – including feedbacks between climate and composition – can comfortably fit into a one-quarter course. It depends somewhat on material from previous chapters as a prerequisite, but the nature of the material is independent enough that with suitable presentation of background, it could be used in a stand-alone course. In a semester, the material could be supplemented with additional instruction on atmospheric chemistry, oceanic carbonate chemistry, or silicate weathering.

Chapter 9, which points the student towards an appreciation of the importance of fluid dynamical effects not conidered as part of this book, can be worked in towards the end to fill out any of the above courses. It is a particularly good complement to the material in Chapter

7. Alternately, it can be left for the student to peruse at leisure, given that study of the other material has now doubt awakened considerable curiosity about what comes next.

The material is laid out in what seems to me to be the natural didactic order, but in view of the realities of teaching and the necessity of dividing the material up amongst multiple courses the instructor may wish to jump around a bit, so as to retain the students' interest. For example, Chapters 2 and 3 fit comfortably in a one quarter course and provide the student with an introduction to thermodynamics and radiation, but adhering to that syllabus would leave the student without an appreciation of the important real-gas aspects of CO_2 and water vapor. It is thus desirable to work in some of the more qualitative real-gas material from Chapter 4, including the use of polynomial OLR fits in solving climate problems. These can be introduced as simply a drop-in replacement for σT^4, once the qualitative underlying physics is explained. Similar opportunities abound, and I have tried to organize things so as to help out the instructor who wishes to wander nonlinearly through the subject matter.

Acknowledgements

Contents

Chapter 1

The Big Questions

1.1 Overview

This chapter will survey a few of the major questions raised by observed features of present and past Earth and planetary climate. Some of these questions have been answered to one extent or another, but many remain largely unresolved. This will not be a comprehensive synopsis of Earth and planetary climate evolution; we will be content to point out a few striking facts about climate that demand a physical explanation. Then, in subsequent chapters, we'll develop the physics necessary to think about these problems. Although we hope not to be too Earth-centric in this book, in the present chapter we will perforce talk at greater length about Earth's climate than about those of other planets, because so much more is known about Earth's past climate than is known about the past climates of other planets. A careful study of Earth history suggests generalities that may apply to other planets, and also raises interesting questions about how things might have happened differently elsewhere, and it is with this goal in mind that we begin our journey.

1.2 Close to home

When the young Carl Linneaus set off on his journey of botanical discovery to Lappland in 1732, he left on foot from his home in Uppsala. He didn't wait until he reached his destination to start making observations, but found interesting things to think about all along the way, even in the plant life at his doorstep. So it is with climate as well.

To the discerning and sufficiently curious observer, a glance out the window, a walk through the woods or town, a short sail on the ocean, all raise profound questions about the physics of climate. Even without a thermometer, we have a perception of "heat" or "temperature" by examining the physical and chemical transitions of the matter around us. In the summertime, ice cream will melt when left out in the sun, but steel cooking pots don't. Trees and grass do not spontaneously burst into flame every afternoon, and a glass of water left outdoors in the summer does not boil. Away from the tropical regions, it often gets cold enough for water to freeze in the wintertime, but hardly ever cold enough for alcohol to freeze. What is it that heats the Earth? Is it really the Sun, as seems intuitive from the perception of warmth on a sunny day? In that case, what keeps the Earth from just accumulating more and more energy from the Sun each day, heating up until it melts? For that matter, why don't temperatures plummet to frigid wintery

values every night when the Sun goes down? Similarly, what limits how cold it gets during the winter?

With the aid of a thermometer, such questions can be expressed quantitatively. The first, and still most familiar, kinds of thermometers were based on on particular reproducible and measurable effect of temperature on matter – the expansion of matter as it heats up. Because living things are composed largely of liquid water, the states of water provide a natural reference on which to build a temperature scale. The *Celsius* temperature scale divides the range of temperature between the freezing point of pure water and the boiling point at sea level into 100 equal steps, with zero being at the freezing point and $100C$ at the boiling point [1].

Through observations of fire and forge, even the ancients were aware that conditions could be much hotter than the range of temperatures experienced in the normal course of climate. However, they could have had no real awareness of how much *colder* things could get. That had to await the theoretical insights provided by the development of thermodynamics in the nineteenth century, followed by the invention of the refrigeration by Carl von Linde not long afterwards. By the close of the century, temperatures low enough to liquify air had been achieved. This was still not as low as temperatures could go. The theoretical and experimental developments of the nineteenth century consolidated earlier speculations that there is an *absolute zero* of temperature, at which random molecular motions cease and the volume of an ideal gas would collapse to zero; no temperature could go below this absolute zero. On the Celsius scale, absolute zero occurs at $-273.15C$. Most of thermodynamics and radiation physics can be expressed more cleanly if temperatures are given relative to absolute zero, which led to the formulation of the *Kelvin* temperature scale, which shifts the zero of the scale while keeping the size of the degrees the same as on the Celsius scale. On the Kelvin scale, absolute zero is at zero degrees, the freezing point of water is at $273.15K$, and the sea-level boiling point of water is at $373.15K$. Viewed on the Kelvin scale, the temperature range of Earth's climate seems quite impressively narrow. It amounts to approximately a $\pm 10\%$ variation about a typical temperature of $285K$. A 20% variation in the Earth's temperature (as viewed on the Kelvin scale) would be quite catastrophic for life as we know it. This remark can be encapsulated in a saying: "Physics may work in degrees Kelvin, but Earth life works in degrees Celsius,"

There is more to climate than temperature. Climate is also characterized by the amount and distribution of precipitation (rainfall and snowfall), as well as patterns of atmospheric winds and oceanic currents. However, temperature will do for starters. In this book we will discuss temperature at considerable length, and venture to a somewhat lesser degree into the factors governing the amount of precipitation. We will not say much about wind patterns, though some of their effects on the temperature distribution will be discussed in Chapter 9.

If you live outside the tropical zone, you will come to wonder why it is hotter in summer than in winter, and why the summer/winter temperature range has the value that it does (e.g. 30C in Chicago) and why the variation is generally lower over the oceans (e.g. 7C in the middle of the Pacific Ocean, at the same latitude as Chicago). If you communicate with friends living in the Arctic or Antarctic regions, and other friends living near the Equator, you will begin to wonder why, on average, it is warmer near the Equator than in the polar regions, and why the temperature difference has the value it does (e.g. 40C difference between the annual average around the Equator vs. the annual average at the North Pole). The physics underlying the seasonal cycle and the pole

[1] The scale is named for the Swedish astronomer Anders Celsius, who originally formulated a similar temperature scale in 1742. Celsius' scale was reversed relative to the modern one, putting 100 at the freezing point and zero at the boiling point. The Celsius scale is sometimes called *centigrade*, but Celsius is considered to be the preferred term. The official definition of the temperature scale is now based on standards that are more precise and unambiguous than the freezing and boiling point of water.

to equator temperature gradient is discussed in Chapters 7 and 9. If you climb a mountain (or even observe the snow-capped peaks of a mountain from the valley floor on a hot summer day), or if you go up in a hot-air balloon, or fly in an airliner which informs you of the outdoor temperature – you will notice that the air gets colder as one goes higher in altitude? Why should this be? This turns out to be a general feature of planetary atmospheres, and the basic physics underlying the phenomenon is discussed in Chapter 2.

The air that surrounds us is itself a matter of interest. We know that it is there because it has a temperature, exerts pressure, and because it is necessary that we breathe it in order to remain alive. But what is the air made of, and why does it have the composition it does? We can see water condense out of the air, but why don't other components condense in the course of natural weather and climate variations? How much air is there? And has it always been there with its present composition, or has it changed over time? If so, how much and how quickly?

We know that our planet journeys through the hard vacuum of outer space, clothed in a thin blanket of air – our atmosphere. It is natural to wonder how our atmosphere affects the Earth's climate. The airless moon shares the same orbit of the Earth, at the same distance from the Sun, so one can look to the Moon to get an idea of what the Earth's climate would be like if it had no atmosphere. We know the moon is airless because a reasonably thick atmosphere would bend the light rays from the Sun and stars, just as objects appear displaced when viewed through the surface of a swimming pool. But how to measure its temperature?

Of course, one could go there with a thermometer (and this did eventually happen) but people became curious about Lunar conditions long before it seemed likely that anybody would ever get there. Dante Alighieri himself, in the *Paridiso* written between 1308 and 1321, devoted fully one hundred cantos to a learned discussion between himself and Beatrice concerning the source of Lunar light and the solidity of the Lunar surface. By the mid nineteenth century, science had progressed to the point that the questions could be formulated more sharply, and the means for an answer had begun to emerge. With the discovery of infrared light by Sir William Herschel in 1800, astronomy opened a new window into the properties of planets and stars. Over the coming decades, it gradually became clear that all bodies emit radiation according to their temperature. This is known as *blackbody radiation* and will discussed in detail in Chapter 3. Infrared light from the Moon was detected by Charles Piazzi Smyth in 1856, and the first attempt to use it to estimate temperature was by the Fourth Earle of Rosse in 1870. The instruments available at the time were not up to the task. In 1878, Langley invented the bolometer, which made good observations of Lunar infrared possible. However, while Langley made the first accurate observations of Lunar infrared, theory was not quite up to the task of interpreting the observations. These issues were largely sorted out by 1913, though Langley gave up on his earlier estimates rather reluctantly. By 1913 it was pretty clear that the daytime temperature of the Moon at the point where the Sun is directly overhead is well in excess of $373K$ (the sea-level temperature of boiling water on Earth). Night-time temperatures were harder to determine accurately, since the infrared emission from cold objects is weak; however it was clear that temperatures at night dropped by well over $140K$ relative to the daytime peak. Pettit and Nicholson observed the temperature of the Moon during the Lunar eclipse of 1927, using the Mt. Wilson telescope. They found something even more remarkable: over the span of the few hours of the eclipse, the Lunar temperature fell from $342K$ at the point of observation to $175K$. Modern measurements show the daily average temperature at the Lunar equator to be around $220K$, while the mean temperature at 85N latitude is $130K$

It appears that without an atmosphere or ocean, the Earth would be subject to extreme swings of temperature between day and night. The Moon's "day" is 28 Earth days, since it always shows the same face to the Earth; on that basis, one could imagine that the day/night extremes were due to a longer night offering more time to cool down, but the rapid cooling during an eclipse

gives the lie to this idea. Given the rapid cooling of an airless body at night, it is likely that the Earth's summer/winter temperature difference would be far more extreme in the absence of an atmosphere. Further, a comparison of the pole to equator gradient in daily mean temperature with that on Earth suggests that the atmosphere significantly moderates this gradient, too. What is it about the atmosphere or ocean that damps down day/night or summer/winter swings in temperature? This subject will be taken up in Chapter 7, where we'll also learn why summer is warmer than winter and why the poles are on the average colder than the Equator. How does an atmosphere or ocean moderate the temperature difference between pole and equator? We'll learn something of that in Chapter 9.

At its hottest the Moon gets much hotter than Earth, and at its coldest it gets much colder. But how does the Moon's mean temperature stack up against that of Earth? The $220K$ mean equatorial temperature of the Moon is very much colder than the observed mean tropical temperature on Earth, which is on the order of $300K$. If the Earth's mean temperature were as low as that of the Moon, the oceans would be solidly frozen over. The cold mean temperature of the Moon does not come about because the Moon reflects more sunlight than the Earth; the Moon looks silvery but measurements show that it actually reflects *less* than Earth. Why is the Earth, on average, so much warmer than the Moon? Does this have something to do with our atmosphere, or is it the case that Earth is warmed by some internal heat source that the Moon lacks?

The search for the first stirrings of an answer to this problem takes us back to 1827, when Fourier published his seminal treatise on the temperature of the Earth. Fourier could not have known anything about the temperature of the Moon, but he did know a great deal about heat transfer – having in fact largely invented the subject. Using his new theory of heat conduction in solids, Fourier analyzed data on the rate at which average temperature increases as one descends deeper below the Earth's surface ; he also analyzed the attenuation of day/night or summer/winter temperature fluctuations with depth. (Fourier's solution for the latter problem will be derived in Chapter 7.) Based on these analyses, Fourier concluded that the flow of heat outward from the interior of the Earth was utterly insignificant in comparison to the heat received from the Sun. We'll see shortly that this situation applies to other rocky planets as well: dry rock is a good insulator, and doesn't let internal heat out very easily.

If the Earth is continually absorbing solar energy, it must also have some way of getting rid of it. Otherwise the energy would have accumulated over the past eons, leading to a molten, incandescent uninhabitable planet (see Problem **??**) – which is manifestly not the case. Fourier seems to have known that there was little or no matter in the space through which planets plied their orbits, and so he posited that planets lose heat almost exclusively through emission of infrared radiation (called "dark heat" at the time). [2] He also knew that the rate of emission of "dark heat" increased with temperature, which provided a means for an equilibrium temperature to be achieved: a planet would simply heat up until it radiated infrared energy at the same rate as it received energy from the Sun. Finally, Fourier refers to experiments showing that something in the atmosphere emits infrared radiation downward toward the ground, and seems to have been aware also of the fact that something in the atmosphere absorbs infrared. Based on these somewhat sketchy observations, Fourier inferred that the Earth's atmosphere retards the emission of infrared to space, allowing it to be warmer than it would be if it were airless.

Fourier's treatise made it clear the the thermal emission of infrared light was not just useful for astronomical observations – it was in fact part and parcel of the operation of planetary climate.

[2] Fourier also refers to the importance of heating from what he calls the "temperature of space." It is unclear whether he thought there was some substance in space that could conduct heat to the atmosphere, or whether he was referring to some invisible radiation which pervades space. His inferences regarding the importance of this factor were erroneous – the only real error in an otherwise remarkable paper.

At Fourier's time the state of understanding of infrared radiation emission was not sufficiently developed as to allow him to complete the calculation he set up. Nonetheless, he correctly formulated the problem of terrestrial temperature as one of achieving a balance between the rate at which solar radiation is absorbed and the rate at which infrared is emitted. With this great insight, the modern era of study of planetary temperature had begun. Fleshing out the "details," however, required major advances in several areas of fundamental physics. The basic principles of planetary energy balance, and of the manner in which an atmosphere increases planetary temperature, are introduced in Chapter 3 and elaborated on in the earlier parts of Chapter 4.

One of the many details that needed to be settled was the question of which components of the atmosphere affected the transmission of infrared radiation. In 1859 Tyndall found that the dominant components of the Earth's atmosphere – nitrogen and oxygen – are very nearly transparent to infrared radiation. He found instead that it was two relatively minor constituents – water vapor and CO_2 – which accounted for most of the infrared absorption and emission by Earth's air. Gases of this sort, which let solar energy through virtually unimpeded but strongly retard the outward loss of infrared radiation, are known as "greenhouse gases." Their warming effect on the lower portions of a planet's atmosphere, and on its surface (if it has one) is called the "greenhouse effect." The term was not coined by Fourier, and in some ways is misleading, since real greenhouses do not work by blocking infrared emission. However, the glass or plastic enclosure of a real greenhouse does warm the interior by reducing heat loss to the environment while allowing solar heating, and in that sense – viewed as a broader metaphor for the implications of energy balance – the analogy is apt. Besides CO_2 and water vapor, we now know of a number of additional greenhouse gases, including CH_4 (methane), which may have played a very important role on the Early Earth, and plays some role even today. In fact, it turns out that in some very dense atmospheres such as that of Titan, even nitrogen can become a greenhouse gas. What determines whether a molecule is or is not a good greenhouse gas, and how do we characterize the effects of individual gases, and thus the influence of atmospheric composition on climate? These questions will be taken up in the latter half of Chapter 4.

In thinking about the effect of greenhouse gases on climate, it is important to distinguish between *long-lived greenhouse gases* which are removed slowly from the atmosphere on a time scale of thousands of years or more, and *short-lived greenhouse gases* which are removed on a time scale of weeks to years by condensation or rapid chemical reactions. The short-lived greenhouse gases act primarily as a feedback mechanism. Their concentration adjusts rapidly to other changes in the climate, serving to amplify or offset climate changes caused by other factors – including changes due to long-lived greenhouse gases. Long-lived greenhouse gases can also participate in feedbacks, but only on time scales longer than their typical atmospheric adjustment time. Whether a greenhouse gas is long-lived or short-lived depends on environmental conditions. On the Earth, CO_2 is a long-lived greenhouse gas but water vapor is a short-lived greenhouse gas; however, on Mars, which gets cold enough for CO_2 to condense, that gas can be considered short-lived.

Greenhouse gases are largely invisible, but the atmosphere also holds a readily visible component that exerts a profound influence over our planet's energy balance – the clouds. Clouds on Earth are composed of suspended droplets of condensed water, in the form of liquid or ice. Clouds, like water vapor, act as a short-lived greenhouse gas affecting the rate at which infrared can escape to space. The infrared opacity of clouds is used routinely in weather satellites, since this property makes cloud patterns visible from space even on the night side of the Earth. However, clouds affect the other side of the energy balance as well, because cloud particles quite effectively reflect sunlight back to space. The two competing effects of clouds are individually large, but partly offset each other, so that small errors in one or the other term lead to large errors in the net effect of clouds on climate. Moreover, the effect of clouds on the energy budget depends on all the intricacies

of the physics that determine things like particle size and how much condensed water remains in suspension. For this reason, clouds pose a very severe challenge to the understanding of climate. This is the case not just for Earth, but for virtually any planet with an atmosphere. The physics underlying the effects of clouds on both sides of the radiation balance will be discussed in Chapters 4 and 5

1.3 Into deepest time: Faint Young Sun and habitability of the Earth

The Solar system was not always as we see it today. It formed from a nebula of material collapsing under the influence of its own gravitation, and once the nebula began to collapse, things happened very quickly. The initial stage of formation of the Solar system was complete by about 4.6 billion years ago. By this time, the Sun had begun producing energy by thermonuclear fusion; the formation of the outer gas giant planets and their icy moons by condensation, and the formation of the inner planets by collision of smaller rocky planetesimals, were essentially complete. The last major event in the formation of the Earth was collision with a Mars-sized body 4.5 billion years ago, which formed the Moon and may have melted the Earth's primitive crust in the process. All these collisions left behind a great deal of heat that had to be gotten rid of before the crust could stabilize. To determine how long it takes to get rid of this heat, we must learn about the mechanisms by which planets lose energy, and about how the rate of energy loss depends on temperature and atmospheric composition; this will happen in Chapters 3 and 4. It turns out that a planet loses energy almost exclusively by radiation of infrared light to space. While the precise rate of loss depends on the nature of the atmosphere, all estimates show that the surface of the Earth quickly cools to 2000K, at which point molten rock solidifies; in the absence of an atmosphere, this process takes a thousand years or less, while with a thick atmosphere it could take as long as two or three million years.

Once a solid crust forms, the flow of heat from the interior of the Earth to the surface is sharply curtailed, because heat diffuses very slowly through solid rock. In this situation, supply of heat from the interior becomes insignificant in comparison with the energy received from the Sun, and the Earth has settled into a state where the climate is determined by much the same processes that determine today's climate: a competition between the rate at which energy is received from the Sun against the rate at which energy is lost to space by radiation of infrared light. This is very likely to have been the case by 4.4 billion years ago, if not earlier. There are no actual rocks as old as this, but there are individual zircon crystals embedded in the Jack Hills formation of Western Australia which are 4.4 billion years old. Zircons of a similar age are also found within the 3.7 billion year old crustal rocks of the neighboring Narryer Gneiss Complex. These crystals provide indisputable evidence for the existence of at least some continental crust of a sort very like that we see today; they also provide convincing though less certain evidence of the existence of liquid water in contact with the early continental crust. The existence of liquid water does not in itself put much constraint on temperature, since water can be maintained in a liquid state even at temperatures in excess of 500 degrees Kelvin, provided the pressure exerted by water vapor in the atmosphere is high enough. The thermodynamics needed to address this issue will be introduced in Chapter 2. Certain aspects of the chemical composition of the zircons, however, suggest that they interacted with near-surface water having a temperature of 100C or less. By 4.4 billion years ago, it would appear, the Earth was no longer a molten volcanic inferno.

The precise nature of the climate evolution between 4.5 and 3.8 billion years ago is obscure at present. Depending on the composition of the atmosphere, the surface temperature could have

been as high as 200C or low enough to cause the ocean (if any) to freeze over completely, and the climate could well have swung wildly between the two extremes. In addition, the dates of Lunar craters indicate that the Earth very likely underwent a period of heavy bombardment by interplanetary debris between 4.1 and 3.8 billion years ago; it is generally supposed that this *late heavy bombardment* affected the rest of the inner Solar system as well, though that is far from certain. The energy brought in by impacts during this period could easily have been sufficient to bring surface temperatures episodically to values well in excess of 100C, sterilizing any nascent ecosystems. Life, if any, may have waged and won a battle for survival in deep ocean refugia.

By 3.8 billion years ago, the veil begins to lift. This is the age of the oldest intact rocks, found today in the Isua Greenstone Belt of Greenland. The appearance of these rocks marks the end of the *Hadean eon*, and the dawn of the *Archaean eon*. Remnants of 3.7 billion year old shales in the Isua formation show the unmistakable signs of deposition of sediments in open water. More intriguingly, these shales are rich in organic carbon, and this carbon preserves a chemical signature generally associated with microbial activity – life. The Barberton formation of South Africa and the Warrawoona formation of Australia, both about 3.5 billion years old, contain layered carbonate sedimentary structures known as *stromatolites*, which in later times are known to be laid down by microbial mats. This is not an unambiguous sign of life, since inorganic processes can also produce stromatolite-like features. Be that as it may, the early stromatolites certainly require ponds of open water evaporating into air. The Barberton and Warrawoona formations also contain microscopic features that are suggestive of bacterial fossils, though not unambiguously so.

The record of surface conditions during the subsequent billion years is hardly continuous, but preserved rocks dating to this period very commonly show a sedimentary character of a type most easily explained by deposition in an open, unfrozen ocean. The first truly unmistakable microbial fossils date to 2.6 billion years ago, where they are found in the Campbell formation of Cape Province, South Africa, and argue for open water conditions having a moderate temperature. At about this time, we bid farewell to the Archean eon, and enter the *Proterozoic eon*, which extends to the appearance of animal life 544 million years ago. Certain fine-grained silica based sedimentary rocks known as *cherts* preserve information about past temperatures, as well as a wealth of fossils. Very ancient cherts contain no unambiguous microbial fossils, but certain aspects of their chemical composition point to temperatures as high as 70C at 3.5 billion years ago, declining to 60C at 2 billion years ago, and declining further to 30C at 1 billion years ago. Well-preserved ancient cherts are rare, however, so this data by no means implies that temperatures were uniformly warm on the young Earth. It only indicates that the Earth attained high surface temperatures at least part of the time; there is ample room to hide lengthy cold periods within the gaps in the chert record, as we shall soon see.

The earliest geological indication of the presence of glaciers on Earth occurs in the upper part of the Pongola formation of South Africa, and dates to 2.9 billion years ago. The evidence consists of glacial sedimentary deposits called *diamictites*, material of a sort usually transported by floating ice, and even glacier-scratched rocks. This does not mean that there were no earlier glaciations, but in light of the chert record and widespread occurence of marine sedimentary rock it seems fairly certain that the Earth did not spend the bulk of its earlier history locked in a deep-freeze. Still, the Pongola glaciation seems to mark the beginning of Earth's long flirtation with ice. The Makganyene glaciation begining around 2.3 billion years ago, recorded in rocks of the Transvaal group of Southern Africa, was a big one, and may well have been global. We know this because a record of the Earth's magnetic field is preserved in the rocks, and this can be used to infer the latitude at which the rocks were located when the glacial deposits were laid down. This *paleomagnetic data* shows that there was ice within 12 degrees of the Equator, strongly suggestive of a global glaciation.

The first unambiguous bacterial microfossils (found in the Campbell group of South Africa) date to 2.6 billion years ago, shortly before the Makganyene glaciation. While earlier fossil and geochemical evidence is very strongly suggestive of life, the Campbell group fossils are the ocular proof that biology was well underway. These fossils mark a watershed in another important way, in that they are identifiable as *cyanobacteria* – the type of organisms that produce oxygen by photosynthesis. The issue of when cyanobacteria evolved is hotly debated, with some lines of indirect evidence putting their appearance early in the Archaean and others dating their onset to the time of the Campbell Group microfossils. Be that as it may, the appearance of these fossils speaks for a fairly benign environment, with open water and temperatures no more than about $40C$. After the Makganyene glaciation, microbial fossils become quite abundant. The two billion year old Gunflint Chert of Canada is one of many such marine sedimentary formations in which cyanobacterial microfossils are preserved.

So far, no glaciations have been reported in the period between two billion years ago and 800 million years ago, though there are abundant sedimentary rocks dating to this time. The record is far from continuous and the lack of glaciations in this period may be an artifact of preservation, but the evidence certainly indicates that icy climates were not dominant at this time, and were probably quite rare. The long hiatus in ice is terminated by the massive – and possibly global – Snowball Earth glaciations of the Neoproterozoic, about 700 million years ago. Thereafter, the climate alternated between fairly lengthy periods when the Earth was ice-free or nearly so, and periods when there was at least some ice in polar regions. The ice never again, however, reached the nearly global proportions it attained during the Neoproterozoic suggesting that the Earth passed some new threshold of climate stability in the Neoproterozoic. What might that be? This is one of the central questions of climate science.

Our overall picture of Earth history is that liquid water and moderate temperatures appeared at least episodically very shortly after the Moon-forming collision, and that the next three billion years had widespread open water with temperatures probably not exceeding $70C$ and generally much less. These conditions were punctuated by occasional glaciations, only a very few of which may have been global in extent. It was an environment that could support the evolution and survival of life, including (by 2.6 billion years ago, if not before) photosynthetic life requiring moderate temperature conditions. Let's keep this picture of relative stability in mind as we go on to discuss long-term changes in the atmosphere and the Sun – the two principle ingredients determining the Earth's climate, or indeed that of any planet.

There are many processes at play that cause the composition of a planet's atmosphere to evolve over time. In the earliest times, bombardments can help supply atmosphere-forming volatiles such as water, nitrogen and carbon dioxide. Equally, however, sufficiently energetic bombardments can cause loss of atmosphere through literally splashing it into orbit. On a volcanically active planet with a hot interior, such as the Earth or Venus, or the younger Mars, new atmosphere is continually being supplied by outgassing of volatile substances from the interior of the planet. The heat needed to keep the interior of the planet churning so it can recycle minerals formed at the surface and cook out volatile gases in the hot interior is supplied by leftover heat from formation of the planet and by radioactive decay. How long this process can continue before the planet freezes out becomes tectonically inactive depends on the size of the planet and the stuff it is made of; the nature of the gases which come out of volcanoes and other types of vents depends on the chemistry of the planet. For example, the early segregation of iron in the Earth's core made it harder to bind up oxygen in minerals, and therefore resulted in fairly oxidized gases like carbon dioxide (CO_2), and sulfur dioxide (SO_2) being released in preference to gases like methane (CH_4) and hydrogen (H_2) – though some of the latter two do nonetheless escape. The interior Earth also outgasses water vapor (H_2O), which is cooked out of hydrated minerals; the volume of the oceans appears to have

been in a steady state for a long time, though, indicating that the rate of release is balanced by the rate of formation and subduction of new hydrated minerals at the surface. Nitrogen (N_2) is fundamentally different from other current and past constituents of the Earth's atmosphere as it doesn't readily get incorporated into the minerals that form the bulk of Earth's crust and interior. Unlike, say, CO_2, nitrogen does not cycle through the Earth's interior. The bulk of the Earth's N_2 is in its atmosphere and stays there, where it has probably been for almost all of our planet's history. This is likely to be the case as well for any other rocky planet made of stuff similar to the Earth – Iron, oxygen (mostly bound up in minerals), silicon, magnesium and sulfur.

While atmosphere is being supplied by outgassing from the interior, other processes cause material to be lost from the atmosphere. Parts of a planet's atmosphere extend far out from the surface, where hot, fast-moving molecules can reach escape velocity and escape to space. Besides escape from random molecular motions, the solar-heated tenuous outer atmosphere can sustain fluid flows which cause atmospheric mass to fountain systematically into the void. In addition, the solar wind can literally blow way the outer portions of an atmosphere; the extent to which this happens is affected by the intensity of the planet's magnetic field, which shields the atmosphere from the solar wind. As outer parts of the atmosphere are eroded, new gases from lower altitudes well up to replace the lost material, sustaining the gradual loss of atmosphere. All three mass loss processes preferentially remove lighter molecules, either because lighter molecules move more swiftly for a given temperature, or because the outer atmosphere is enriched in gases having a lower molecular weight. For a given density, a smaller planet has lower surface gravity, and so binds its atmosphere less tightly; in consequence, escape of atmosphere to space proceeds more rapidly on a small planet. Impacts by large, swift bodies can impart sufficient energy to part of the atmosphere to blast it into space. This mechanism of atmosphere loss does not discriminate as to molecular weight, but as with the other mechanisms, it is easier for a small planet to lose atmosphere this way. Overall, the Moon or Mars is more prone to lose atmosphere than more massive bodies such as the Earth or Venus, to say nothing of Jupiter or Saturn. For Earth and Venus, escape to space is significant only for H_2 and He, and of these the latter is important only as an indicator of planetary history rather than as a physically or chemically active component of the atmosphere. Saturn's satellite Titan is an interesting case, as it maintains a mostly N_2 atmosphere more massive than that of Earth (per unit surface area) despite having a surface gravity lower than that of the Moon. The very cold temperature of Titan helps it retain its atmosphere, but it is nontheless likely that the persistence of the atmosphere requires a substantial rate of resupply from the interior of the planet.

Some components of the atmosphere can also be lost through chemical reactions with rocks at the Earth's surface. A particularly important example of this is the class of reactions commonly known today as *Urey reactions* [3] , which remove CO_2 from the atmosphere. When CO_2 dissolves in water, it forms a weak acid (carbonic acid), which reacts with silicate minerals (e.g. $CaSiO_3$) to form carbonate minerals (e.g. $CaCO_3$, or "limestone"). The reactions that form carbonate take place only in the presence of liquid water, so if a planet becomes so hot that liquid rain never reaches the surface, or if it somehow loses its water altogether, then CO_2 outgassed from the interior of the planet will accumulate in the atmosphere until the interior source is depleted or the rate of supply is balanced by loss to space. On Earth, all of the CO_2 presently in the atmosphere

[3]The reactions are named after the University of Chicago geochemist Harold Urey, who discussed them in a 1952 book called *The Planets: Their Origin and Development*. Although modern science was made aware of the importance of these reactions through Urey's work, the reactions were first introduced by the French chemist and metallurgist J.J. Ebelman more than a century earlier. Ebelman also introduced the notion that the silicate/carbonate reactions play an important role in determining atmospheric CO_2 and hence Earth's climate. Similar ideas were independently rediscovered by the Swedish geochemist A.G Högbom in 1894, and then finally by Urey. For more details of the history see Berner (1996) *Geochim Cosmochim Acta* **60**.

could be removed by the Urey reactions within 5000 years, forming a layer of limestone a mere 5 millimeters thick; if all the carbon stored in ocean water were to outgas as CO_2 and react to form limestone, the process would take a half million years and form a layer a half meter thick.

Life itself, once it appears, has a profound effect on atmospheric composition. While little methane escapes directly from the Earth's interior, bacteria known as *methanogens* can synthesize it from H_2 and CO_2 or from organic material produced by other organisms. Methanogens may well have dominated the ecosystems of the Earth's first two billion years, potentially allowing a methane-dominated atmosphere to build up. The advent of life also had a profound effect on nitrogen cycling. The bonds holding together N_2 are so strong that in the abiotic world only rare energetic events such as lightning strikes can form nitrogen compounds. In fact, though nitrogen is an essential ingredient of all living material, higher forms of life – including all plants and animals – are incapable of performing the chemical magic that makes N_2 available to organisms. This trick is accomplished by nitrogen-fixing bacteria, which can efficiently transform atmospheric N_2 into ammonium (NH_4), in turn transformed into nitrate (NO_3) which can be used by other organisms in the synthesis of living matter. Other bacteria, in oxygen-starved conditions, can make a living combining the oxygen in nitrate with carbon, returning N_2 to the atmosphere in the process. It was only in 1910, with the invention of the Haber Process for turning atmospheric N_2 into ammonia (NH_3) that humans caught up with the bacteria. This innovation has become essential to the human population, as the demands of industrial-style agriculture have far outstripped the ability of the natural bacterial ecosystem to supply nitrate (which is not to deny that other forms of agriculture might be able to live within the means provided by our bacterial friends). Nitrogen-fixing bacteria are still way ahead of industry in terms of chemical sophistication, though, since the Haber Process requires molecular hydrogen (made from fossil methane), an iron catalyst, temperatures exceeding 400C, and operates at a pressure over two hundred times that of air at the Earth's surface; Nitrogen-fixing bacteria can do the same trick, in contrast, in ambient temperature/pressure conditions and using materials found readily in their immediate environment.

In the absence of oxygen-producing photosynthetic life, only minuscule amounts of oxygen would be present in the atmosphere, since only a trickle could be produced through the breakdown of H_2O by exposure to sunlight. That there was very little oxygen in the early atmosphere (under 0.2%, compared to 20% today) is confirmed by the widespread presence of striking rock structures known as *banded iron formations* until 2.4 billion years ago. Banded iron formations can be laid down only when iron is very soluble in the ocean and can be transported long distances. This requires low oxygen, since in the presence of oxygen iron forms compounds that are not very soluble in water. Additional evidence stemming from the chemical composition of certain sulfur-containing minerals indicates that during at least some periods earlier than 2.6 billion years ago the atmospheric oxygen content might in fact have been orders of magnitude lower than 0.2%.

Note that the appearance of oxygen-producing photosynthesis is not synonymous with the rise of oxygen in the atmosphere. For oxygen to accumulate, a sufficient proportion of the organic matter must be buried before it is oxidized by other bacteria, taking the synthesized oxygen right back out of the atmosphere. Further, if the Earth had accumulated a great stock of available organic matter in the ocean during its anoxic phase, this backlog would have to be worked off before oxygen could begin to accumulate to any significant degree. Be that as it may, banded iron formations begin to falter somewhat after the time of the Campbell Group cyanobacterial microfossils, and disapear entirely by around 2 billion years ago. By this time, oxygen may have made up at least 3% of the atmosphere. Once oxygen made its appearance in the atmosphere at significant concentration, it changed all the rules of atmospheric chemistry, since it is so powerfully reactive. In particular, it made it much harder for CH_4 and H_2 to accumulate in the atmosphere, since the former oxidizes readily to CO_2 and the later to H_2O. The rise of oxygen may also have fostered

another great biological innovation – the *eukaryotic cell*, which has a complex internal organization with specialized structures, including a nucleus within which genetic material is segregated. We are made of eukaryotic cells, as are all animals and plants. Eukaryotic cells make their first unambiguous appearance in the fossil record about 1.5 billion years ago, in the Roper Group shales of Australia, though biochemical molecules preserved in ancient sediments suggest strongly that eukaryotic life may have evolved much earlier. However the answer to that issue may shake out, it is certain that eukaryotic life – even of the single-celled variety – did not proliferate and diversify until much later in the Proterozoic.

There was a sporadic reappearince of banded iron formations at the time of the Neoproterozoic Snowball events (600-700 million years ago), but by 600 million years ago oxygen was approaching its present concentration and banded iron formations disappeared for good (at least so far). This second pulse of oxygenation made the rise of multicellular animals possible, which happened in a great flurry of biological innovation known as the *Cambrian Explosion*, occupying a remarkably short span of years around 543 million years ago. It is even possible that the rise of animals helped to stabilize atmospheric oxygen levels, by providing a reliable means of transporting organic carbon to the sea floor where it can be buried and preserved. Simple multicellular ocean-dwelling plants appeared much earlier, as might be expected from the fact that photosynthetic organisms are not dependent on oxygen.

All of this atmospheric evolution takes place against a backdrop of a gradually brightening Sun. The energy produced within a star leaves the star almost exclusively in the form of electromagnetic radiation – loosely speaking, light of all wavelengths. The net power output is called the *luminosity*, and is measured in *Watts* (a measure of energy per unit time), just as if the star were a light bulb. Stars like the Sun get their energy by fusing hydrogen into helium, and as time goes on the proportion of helium in the Sun increases, thus increasing the mean molecular weight of the Sun. This in turn means that the core of the Sun needs to contract and heat up in order to maintain the pressure required to balance gravity. The increased density and temperature increases the rate of fusion more than the reduced availability of hydrogen reduces it, so the rate of production of energy – and hence the Solar luminosity – *increases* with time. The resulting evolution of luminosity over time rests on fundamental aspects of solar physics that are not seriously in question, and does not depend greatly on the finer points of solar modeling. The results of the standard solar evolution model can be fit by the expression

$$\mathcal{L}_\odot(t) = \frac{\mathcal{L}_\odot(t_\odot)}{1 + \frac{2}{5}(1 - t/t_\odot)} \tag{1.1}$$

where $\mathcal{L}_\odot(t)$ is the luminosity of the Sun at time t and $t_\odot$ is the current age of the Sun, usually taken as 4.6 billion years. This formula was originally proposed to describe the younger Sun, but it continues to be reasonably accurate out to times 4 billion years in the future as well.

It follows from Eq. 1.1 that 4 billion years ago the Earth received solar energy at only 75% of the rate it does today. All other things being equal – atmospheric composition in particular – that means the Earth would have been colder than it is today. How much colder? We will learn how to do this calculation in the simplest way in Chapter 3, and add sophistication to the calculation in Chapter 4. It turns out that *if the atmospheric composition were the same as today's atmosphere throughout Earth's history*, then Earth should have been completely frozen over 4 billion years ago, and given that ice reflects sunlight so well, it should still be completely frozen over today. However, we know that incidents of global ice coverage are rare in Earth history, if indeed they happened at all; we know with rather more certainty that the Earth is not solidly frozen over today. This contradiction is generally known as the *Faint Young Sun Paradox*, though of course, like most paradoxes it is not paradoxical at all once one understands what is going on. Calling it

a "paradox" is just a way of starkly bringing home the fact that to account for the basic facts of Earth's climate history, the atmosphere must indeed have undergone massive changes – changes of a sort that could substantially affect climate. How much would we need to increase CO_2 or CH_4 in order to make up for the faintness of the young Sun? This is another one of the big questions. It will be answered in Chapter 4. A related Big Question is the extent to which CH_4 (or some other long-lived greenhouse gas) substituted for CO_2 in maintaining the Earth's warmth during the Archaean. There is scattered evidence from mineral composition of fossil soils that during some periods of the Archaean the CO_2 concentration might not have been high enough to offset the Faint Young Sun. This has led some researchers to jump to the conclusion that CH_4 played the decisive role throughout the Archaean, but such a viewpoint rests on exceedingly shaky evidence. It matters a great deal whether CO_2 or CH_4 did the trick, since the long term control of the two gases is governed by very different processes, with very different time scales. The matter, at present, must be considered unresolved.

A corrollary to the above resolution of the Faint Young Sun Paradox is that any atmosphere that would be sufficient to keep the Early Earth unfrozen would make it uninhabitably hot with the present Solar output. The atmosphere must have somehow changed in lock step with the brightening Sun, in precisely such a manner as to keep the Earth in a habitable temperature range – one where liquid water exists at or near the surface, but where the water never gets hotter than about 100C – or indeed, even hotter than about 50C if cyanobacteria are to survive. It defies belief that the required co-evolution of atmosphere with solar output could maintain the required temperature range purely by coincidence, so it seems likely that some sort of temperature-regulating mechanism is in operation. In Chapter 8 we'll see how the Urey reaction can participate in a geochemical thermostat for the Earth and similar planets. The Snowball episodes represent temporary, and evidently rare, breakdowns of the temperature-regulating mechanism. Whatever the regulating mechanism may be, it must be sufficiently fail-safe to allow for recovery from global or near-global glaciations.

Given what we laid out earlier concerning the myriad processes churning out turmoil in atmospheric composition, the reader should quite rightly feel it a bit silly to make the stipulation "all other things being equal" in the statement of the Faint Young Sun Paradox. Indeed, even if the solar output were constant over time, any number of the aspects of atmospheric evolution we discussed earlier would have been sufficient to freeze or fry the Earth, so the gradual increase in solar output has no special claim on our attention in this regard. To be fair, when the "paradox" was first laid out, much less was known about the ways the Earth's atmosphere evolved, and with much less certainty. From the standpoint of current science, however, the traditional framing of the Faint Young Sun Paradox leaves a lot to be desired. It would be more satisfactory to refer to the Habitability Problem, which could be stated as follows: *How can the temperature of a planet be maintained in the habitable range for billions of years, in the face of geological, geochemical and biological turmoil in atmospheric greenhouse gas composition, and in the face of gradual increase in solar output?*. This is indeed one of the grandest of questions. The material discussed in Chapter 8 provides a plausible solution, but the book is far from closed on the Habitability Problem.

The history of the Faint Young Sun problem reveals something magnificent and deeply inspiring about the nature of discovery in Earth and planetary climate. The basic physics underpinning the energy source of stars was worked out by Hans Bethe by 1939, and the existence of benign conditions on the Early Earth was known (or at least inferred) even earlier. Yet, it was not until 1972 that Carl Sagan and George Mullen put the two bits together and inferred that there was indeed a big problem, requiring a profound answer [4]. This insight sparked a revolution

[4]Sagan and Mullen proposed accumulation of atmospheric ammonia as a resolution to the paradox, but later work on atmospheric chemistry showed that sunlight destroys ammonia too rapidly to allow it to build up to the

in thinking about planetary climate, that was in its own way as earthshaking as the discovery of DNA was in its own domain. This history highlights a lovely thing about our subject: The most profound new phenomena are often discovered by putting together a few bits of basic physics and chemistry in creative new ways. For the most part, new ideas come from playing with simple models, not from enormous incomprehensible computer simulations that take huge teams to put together. The entire goal of this book is to teach students to think the way Carl Sagan and other innovators did, and to provide the tools needed to build the simple models needed to turn a bright idea into real science.

The problem of maintaining long-term climate stability is not just an issue for Earth. Planets that could support some form of life are naturally of special interest, and while other forms of life than those we know of might prefer quite different conditions than prevail on Earth, the long-term maintainence of climate in a fairly narrow "habitable" range clearly would make it easier for life to evolve and persist. Any potentially life-bearing planet in any solar system anywhere must negotiate a way to maintain long-term habitability in the face of the gradual increase of the brightness of its sun and gradual or not-so-gradual changes in the composition of its atmosphere. Naturally, such considerations apply to the evolution of the climates of other planets in our very own solar system. Venus and Mars did not manage to maintain their habitability, or perhaps were never in a habitable state. How close did we come to the same fate?

1.4 Goldilocks in space: Earth,Mars and Venus

Until well into the 1960's, science fiction stories about Venus generally portrayed it as a steamy jungle planet, but one where intrepid explorers could perhaps survive unprotected on the surface. The idea of a jungle and breathable air was of course unfounded speculation, but the general picture of the climate was not wholly without merit. After all, the dense reflective cloud deck of Venus was readily observable – it is what makes Venus so bright as the "evening star" – and the reflection of sunlight could easily make up for the fact that Venus is closer to the Sun than is Earth. In fact in Chapter 3 we'll see that the reflectivity *more than* makes up for the proximity. In the late 1950's the picture of Venus as a habitable world began to unravel. Recall that the temperature of Earth's Moon was determined by examining infrared radiation from that body. Viewed in the infrared spectrum Venus appeared quite cool, but in the microwave ("radar") spectrum it was far too bright. In fact, seen in the microwave, Venus radiated like a body with a temperature well in excess of $600K$ ($327C$). A popular hypothesis at the time was that the anomalous microwave emission arose in the upper portions of the Venusian atmosphere. Another view held that the microwave emission came from the surface of the planet, and that the atmosphere was transparent to microwaves but relatively opaque to infrared. The latter idea suffered from the lack of a plausible mechanism to make the surface of Venus so hot. Then, in 1960 the young Carl Sagan proposed that Venus has a very thick atmosphere rich in greenhouse gases, which would heat up the surface to the required temperature. Little was known about the mass of the atmosphere or its composition at the time, but Sagan developed simple models of the greenhouse effect of a thick atmosphere, which showed that the trick could be accomplished with an atmosphere consisting of mostly carbon dioxide with some water vapor mixed in, having a total mass three or four times that of the Earth's atmosphere. Sagan even recognized that since the planet was too hot for water vapor at the hypothesized concentration to condense and reach the surface as liquid, the Urey reaction (which removes carbon dioxide from the atmosphere and turns it into limestone) could not take place. This would make it easy for carbon dioxide to accumulate in the atmosphere, though even

required concentrations. Attention later shifted to CO_2 and CH_4.

Sagan did not envision just how far this would go.

A series of interplanetary probe missions over the next two decades – four US Mariner missions, two US Pioneer missions and sixteen Soviet Venera missions including eight Venera missions that returned data successfully from the surface – refined the estimates of surface temperature and substantially revised the conception of the atmospheric mass and composition. By the late 1970's, it was known that the surface temperature was nearly uniform at $737K$. The atmosphere was found to be much more massive than originally thought, in fact sufficiently massive to raise the surface pressure to 92 times that of Earth's atmosphere. And, it was found that the atmosphere consisted almost entirely of carbon dioxide, with only traces of water vapor remaining. The thick clouds that give Venus its high reflectivity were found to be made not of water, but of droplets of sulfur dioxide and concentrated sulfuric acid. It took the better part of another decade before the challenges of dealing with the effect of such an exotic atmosphere on climate were fully mastered and a fully satisfactory account of the high surface temperature could be given. Still, the initial exploration of the problem was carried out with simple models very like the ones we'll introduce in the first half of Chapter 4. The discovery of the true nature of Venus climate is another illustration of the tenet that big ideas grow from little models.

Mars yielded up its climatic secrets somewhat earlier, because it was not hidden behind the thick atmospheric veil that complicated observation of Venus. Mars was observed in the infrared using the Mt. Wilson telescope during the opposition (time of closest approach) of 1926 and 1927. Like all infrared observations, the interpretation was complicated by interference from the Earth's atmosphere. By 1947, the understanding of the effect of atmospheres on the emission of infrared light had progressed to the point that the mass and composition of the Martian atmosphere could be estimated. Based on these measurements, Kuiper estimated that Mars had an atmosphere that was almost entirely CO_2, with a sufficient mass that the surface pressure on Mars would be only .03% of the surface pressure on Earth. This turns out to be an underestimate of the true mass by about a factor of twenty, but even so, the picture of Mars as a nearly airless planet was not far wrong. By way of comparison, infrared observations of Venus interpreted in the 1940's using similar techniques suggested that the surface pressure of Venus was a fifth that of Earth – an underestimate by a factor of nearly 500. The Mt. Wilson infrared observations of Mars also indicated that the atmosphere was almost entirely CO_2, with almost no water – so little, in fact, that water vapor wouldn't condense until temperatures fell below $-60C$, and at those temperatures it would condense into frost, not liquid. Infrared observations showed further that the visible polar ice caps of Mars are most probably made of water ice rather than frozen CO_2. Temperature estimates based on the Mt. Wilson infrared observations were less informative. Nicholson and Pettit, in the same paper in which they discuss Lunar temperatures, noted a very large day/night cycle of Martian infrared, indicating extreme diurnal constrasts unlike those found on Earth. They concluded that this was due to the lack of water vapor in the Martian atmosphere, but we shall encounter the true reason in Chapter 7. Writing in 1947, Adel reported quantitative estimates of Martian surface temperature ranging from as low as $236K$ to as high as $318K$ (or even higher if the surface was assumed to emit infrared inefficiently, as does granite). Some of the variation in reported temperatures may have been due to the fact that these were not whole-disk observations, insofar as Mars exhibits extreme temperature contrasts. The higher end of the estimates based on Mt. Wilson turned out to be far greater than the actual maximum ground temperature encountered anywhere on the planet.

Given the thin atmosphere, what does theory lead us to expect about the Martian surface temperature? By 1947, it was a simple exercise to compute the expected temperature of airless bodies like the Moon, using arguments like those we'll discuss near the beginning of Chapter 3. Planets with atmosphere present more of a problem. The atmosphere of Mars is thin compared

to that of Earth, but how thin does an atmosphere have to be (particularly if it's pure CO_2) in order to have a minimal effect on planetary temperature? We'lll learn how to answer that question in the latter portions of Chapter 3. Based on similar reasoning, Kuiper, writing in 1947, inferred correctly that the atmosphere would have only minor effect on T, in particular allowing severe night-time temperature drops. By the 1940's Mars was already looking inhospitable – a mostly cold, dry nearly airless body where (it was still hoped) conceivably lichens might eke out a living but certainly not Thuvia, Maid of Mars.

Ground-based and theoretical estimates of the Martian climate improved gradually over the next decade, but the real breakthrough came with the Mariner flyby of 1965 and the two Viking orbiters and landers of 1976. Spaceborne infrared observations gave the first detailed picture of geographic variations of the ground and air temperature, and Viking provided *in situ* air temperature measurements of the ground. These observations consolidated the picture of Mars as a planet which (as we find it now) has more in common with the airless moon than with Earth. Even hopes of lichens were dashed, though it is too soon to give up on bacterial life, especially in view of the innovative chemical entrepreneurship shown by non-photosynthetic bacteria on Earth. In discussing Martian temperatures, one must take care to distinguish the air temperature from the temperature of the ground itself, as the two differ considerably. At high noon in the tropics, the ground can indeed briefly get as warm $300K$, though temperature drop by $100K$ or more at night. The air temperature also shows an extreme day/night cycle, but the peak daytime air temperatures are far less than the peak ground temperature; at the Viking Lander 1 site, in the tropics at $22N$ latitude, the daytime air temperature never exceeds $260K$, and plummets to under $200K$ at night. The reason the air temperatures are so much lower than the ground temperatures will become clear in Chapters 4 and 6.

There are considerable seasonal and latitudinal variations in daytime temperature, and the Southern Hemisphere polar summer is notably warmer than the Northern Hemisphere polar summer. Night-time temperatures are comparatively uniform both geographically and seasonally; the Mars surface cools so fast that that once it is dark, it evidently doesn't matter much whether it has been dark for a few hours or a hundred days. The Southern Hemisphere winter pole does get notably colder than the rest of the planet, dropping to as low as $160K$. The Viking landers also provided the first clear picture of surface pressure variations on Mars. They showed that the Martian surface pressure varies from a high of about 1% of Earth's surface pressure to a low of about 0.6%, with the lowest values occurring in Southern Hemisphere winter. Since surface pressure gives a measure of the mass of air in the atmosphere (2), the large variation of pressure indicates that a considerable portion of Mars' CO_2 atmosphere snows out over the North pole in Northern winter and the South Pole in Southern winter, only to sublimate back into the atmosphere as spring approaches. A theory for the seasonal cycle of Martian temperature and pressure will be developed in Chapter 7.

So, Venus, like the porridge tasted by Goldilocks is too hot. Mars is too cold, and Earth is just right. One could quite reasonably object that this is a view prejudiced by our own status as a form of terrestrial life, and that conditions "too hot" by our standards could well be "just right" for somebody else. However, it appears that there's nobody home on either Venus or Mars (not even a microbial somebody), so if conditions there are indeed "just right" for somebody, it must not be very easy for such a somebody to evolve, given that it didn't happen in the past four billion years.

Presumably, Venus started out with a composition rather similar to Earth. What went wrong? Why did it keep most of it's CO_2 in its atmosphere, whereas most of Earth's CO_2 got bound up in carbonate rocks? Where did its water go? The answer came in 1967 with the theory of the *runaway greenhouse*, formulated first by M. Kombayashi and independently rediscovered

shortly thereafter by Andrew Ingersoll of Caltech. This theory puts together two simple bits of physics, the first being that water vapor content of a saturated atmosphere increases exponentially with temperature (Chapter 2), and the second being that water vapor is a greenhouse gas (Chapter 4). When the two are put together, it is found that a planet which receives sufficient solar radiation can get into a runaway cycle where the planet warms in response to absorbed sunlight, which causes more water vapor to enter the atmosphere, which causes more greenhouse effect, which leads to further warming in an unstable feedback loop that doesn't end until the entire ocean is evaporated into the atmosphere. At that point, the water vapor in the upper atmosphere breaks down into hydrogen and oxygen under the influence of high energy solar radiation, and the hydrogen escapes to space while the oxygen reacts with rocks. Without liquid water, the Urey reaction which turns CO_2 into limestone can't take place, so all the outgassed CO_2 stays in the atmosphere. The runaway greenhouse theory (explored in Chapter 4) gives rather precise predictions of the circumstances under which a runaway can occur, and explains why Earth did not undergo a runaway despite the fact that it has a water ocean. The work of Kombayashi and Ingersoll is another example of the general idea that big ideas come from simple models. Their reasoning was based on simple radiation models of the sort developed in the first half of Chapter 4. This work also illustrates another general principle of planetary climate: profound results can be obtained by combining a few bits of very basic physics in a novel way.

Radar mapping from the Magellan orbiter of 1990 revealed another remarkable fact about Venus: Unlike the Earth, with long-lived continents and a gradually subducted sea floor, Venus has a young-looking uncratered surface, suggesting that the crust may have been engulfed and resurfaced as recently as 500 million years ago. This has important implications for planetary habitability. Evidently, the formation of a planetary crust, as at end of Hadean, is not end of the peril from fires in the deep. For habitability, the crust has to be relatively stable, engulfed slowly in subduction zones (as is the Earth's sea floor) rather than being subject to episodic catastrophic volcanism as seems to have been the case on Venus. However, if the crust is engulfed *too slowly*, then limestone that forms by the Urey reaction is only sluggishly recycled, leading to a drawdown or atmospheric CO_2 and (under some circumstances) a very cold planet; since photosynthesis uses CO_2 as a feedstock, low CO_2 impedes habitability even if a planet is in an orbit where it doesn't need the greenhouse effect of atmospheric CO_2 in order to stay warm. The question of when a planet has *plate tectonics* like Earth or when it has episodic catastrophic resurfacing like Venus, is another one of the great questions of planetary science, though one we will not take up at any great length in this book.

Mars may be impoverished in atmosphere compared to Venus, but it has something Venus lacks: a geological record of the distant past preserved in its crust. In fact, the ancient features on the surface of Mars are far better preserved than is the case on Earth. Mars appears to have lost most of its atmosphere quite early on, leading to a near-halt in the rate of erosion of surface features. The first high resolution data of Martian surface features, returned by the Mariner mission, revealed a startling fact. Evidently, Mars was not always the dry, frigid planet cloaked in a tenuous atmosphere that we see today. The Mariner photographs revealed dry river-like channels for which the only likely explanation is flowing surface water, which would be impossible under the conditions of the present Martian climate. A more recent image of this type of feature is shown in Figure 1.1. The rate of cratering of a planet goes down with time, so the features can be dated by counting superposed craters; many of the major river-like features date to very early in Mars history, perhaps 4 billion years ago. This led to the concept of a "warm,wet Early Mars." But how could Mars have been so much warmer than it is at present, at a time when the Sun was so much fainter. Mars presents an even more extreme version of the Faint Young Sun paradox than does the Earth. It will probably come as no surprise that it was Carl Sagan who first pointed out the implications of Martian dry river networks. It is still somewhat disputed

Figure 1.1: Nanedi Vallis on Mars, observed by the Mars Orbiter Camera on the Mars Global Surveyor Mission. The image covers an area 9.8km wide and 18.5km tall.

whether the surface features really demand that Early Mars be warm and wet, but adopting the warm-wet view, the resolution of the Faint Young Sun problem, as was the case for Earth, lies in the supposition that the Early Mars atmosphere had a substantially stronger greenhouse effect. What kind of atmosphere could warm Mars to the point that liquid water could flow long distances at the surface? That question is taken up in Chapters 4 and 5.

If Mars started out with such a dense atmosphere, where did it go? Some possible answers are suggested in Chapter 8. Modern high-resolution images suggest other forms of massive climate change on Mars. In particular, there are tropical landform features suggesting that at some time in the past, glaciers formed in the Martian tropics, whereas virtually all the ice is today sequestered at the cold polar regions. The tropical glacier landforms suggest that at some time in the past, the equatorial regions of Mars were colder than the poles? How could that be? The answer is provided in Chapter 7.

The fact that Earth maintained habitable conditions while Venus succumbed to a runaway greenhouse and got too hot while Mars lost its atmosphere and got too cold raises the question of just how narrowly Earth escaped the fate of Mars and Venus. How much could Earth's orbital distance be changed before it turned into Mars or Venus, and how would the answer to this question change if Earth were more massive (making it easier to hold onto atmosphere) or less massive (making it easier to lose atmosphere? If Mars were as large as Earth, would it still be habitable today? What if Venus were as small as Mars? Perhaps if the orbits of Mars and Venus were exchanged, our solar system would have three habitable planets, instead of just one.

The range of orbital distances for which a planet retains Earthlike habitability over billions

of years is known as the *habitable zone*. Determing the habitable zone, and how it is affected by planetary size and composition as well as the properties of the parent star, is one of the central problems of planetary climate.

1.5 Other Solar System planets and satellites

For the *gas giant* planets – Jupiter and Saturn – many of the most striking questions that arise are fluid dynamical in nature. These questions include the origin of the banded multiple-jet structure of the atmospheric flow, and the dynamics of long-lived atmospheric vortices, most famously Jupiter's Great Red Spot. We shall have little to say about such fluid dynamical questions in this book. However, the thermal structure of the atmosphere provides an essential underpinning for any dynamical inquiry. Moreover, the thermal structure determines the rate of heat loss from the planet, and therefore plays a crucial role in the long-term evolution of the gas giants. The thermal structure also affects atmospheric chemistry and the nature of the colorful clouds that allow us to visualize the spectacular fluid patterns on these planets.

The gas giants present an interesting contrast to rocky terrestrial-type planets, because of the lack of a solid surface. Instead of solar radiation having the possibility of penetrating to the ground and being absorbed there, thus heating the atmosphere from below, solar radiation on the gas giants is deposited continuously throughout the upper portion of the atmosphere as the solar beam propagates downward and attenuates. Further, unlike the case of the rocky planets where heat flux from the interior is an insignificant player in climate, the fluid nature of the gas giants allows considerable heat flux from the interior to escape to space. For both Jupiter and Saturn, this heat flux is comparable to the flux of energy received from the Sun. One of our objectives in subsequent chapters will be to learn how the distinct nature of atmospheric driving on the gas giants affects the thermal structure. The gas giants also offer an interesting opportunity to test ideas about how climate is affected by atmospheric composition. These planets are mostly made of H_2, with a lesser amount of He and trace amounts of a range of other substances, including ammonia (NH_3), methane (CH_4) and water, the latter three of which exist in both gaseous and condensed forms. The composition affects the thermodynamics of the atmosphere, as well as the optical properties for both infrared and visible light.

Uranus and Neptune are like the gas giants in that they have no distinct solid surface at any depth that could significantly affect the atmosphere. However, they are usually classified separately as *ice giants* because they contain a much higher proportion of ice-forming substances such as water, ammonia and methane. The composition of the outer portions of the atmospheres can be determined by spectral observations, and contain a high proportion of hydrogen and helium. The overall density of the planets, however, constrains them to be composed primarily of an *ice mantle*. In the case of Uranus, the ice mantle must make up between 9.3 and 13.4 Earth masses worth of the total mass of the planet, which is 14.5 Earth masses. Similar proportions apply to Neptune. The commonly used term "ice mantle" is somewhat misleading, since the substance is actually a hot, slushy mixture that would be more aptly described as a water-ammonia ocean. Whatever term is used,the very thermal structure that determines the nature of the transition between the ice mantle and the more gaseous outer atmosphere engages all the same issues of atmospheric energy balance as one encounters on other planets. A novel feature of Uranus is its axial tilt. Its axis of rotation is nearly perpendicular to the normal to the plane of the orbit. In other words, the axis lies almost in the plane of the orbit. That means that in the Uranian Northern Hemisphere summer, the North pole is pointing directly at the Sun and the entire Southern Hemisphere is in darkness. By way of contrast, the Earth's axis is only tilted by 23.4^o relative to the normal at present. The

high axial tilt of Uranus potentially gives that planet an extreme seasonal cycle, though it will take a long time to observe it since Uranus' year lasts 84 Earth years. The effect of axial tilt on seasonal cycles of planets are discussed in general terms in Chapter 7. The very low solar radiation received at the distant orbits of Uranus and Neptune leads to extremely cold outer atmospheres, particularly in the case of Neptune. These planets provide an opportunity to examine the novel features of an atmosphere driven by an exceedingly weak trickle of solar energy, supplemented by an equally feeble trickle of heat from the interior. Despite the weak thermal driving, Neptune has by far the strongest winds in the Solar system, as well as a variety of interesting meteorological features. We will not say much about planetary winds, but as in the case of the gas giants, a good understanding of the thermal structure is a prerequisite for any attack on the meteorology.

The gas and ice giants also challenge our notion of *habitable zones* – orbits where a planet has some region where there are Earthlike temperatures allowing for liquid water. The gas and ice giants have no distinct surface, but there is some depth on each of them where the temperature is Earthlike and liquid water can exist. The atmosphere also have plenty of chemical feedstocks for organic molecules, including ammonia and methane. The pressures are no greater than those seen at the bottom of the Earth's ocean. One may have some prejudice in favor of surfaces for life to live on, but it must be recalled that on Earth life first arose in the oceans and indeed stayed there for many billions of years before venturing onto land. The gas and ice giants could just as well be thought of as being "all ocean" rather than "all atmosphere" so it is far from clear that they are inhospitable, at least for chemosynthetic forms of life that don't need much sunlight. Our thinking about habitable zones is overly prejudiced toward life that carries out its existence on a rocky surface.

From the standpoint of planetary climate, one of the most interesting Solar System bodies is not a planet at all, but a satellite. Titan, which orbits Saturn, is a fairly large icy body with a radius of that is 76% of that of Mars. Because it is composed of ice rather than rock, the surface gravity is low: 1.35 m/s^2, which is actually lower than the Moon's surface gravity, though the Moon is smaller than Titan. What makes Titan interesting, however is its dense atmosphere. The atmosphere of Titan consists mainly of nitrogen, with a surface pressure about 1.5 times that of Earth. What is even more interesting is that the lower portion of the atmosphere is about 30% methane. At the cold temperatures of Titan (about $95K$) methane can rain out, and participates in a hydrological cycle analogous to that of water on Earth – but operating at a much colder temperature. In subsequent chapters we will develop the physics to examine the similarities and differences between the role of methane on Titan vs. water on Earth. Titan's atmosphere is also a seething organic chemical factory, with complex long-chain hydrocarbon hazes being manufactured from methane in the upper atmosphere. These hazes absorb solar radiation, shade the surface, and are a key player in Titan's climate. Such organic hazes were first discovered on Titan, but there are speculations that similar hazes could have been present in methane-dominated atmospheres of the Early Earth.

A major question about Titan is why it has an atmosphere left at all . Given the low gravity, the N_2 atmosphere would be expected to escape fairly quickly (we'll have a look at the relevant physics in Chapter 8). Moreover, the chemical reactions in the atmosphere should gradually convert all the methane into a tarry sludge sequestered at the surface. In some way or another, the atmosphere of Titan must be dynamically maintained by recycling of chemicals deposted on the surface, and by outgassing of N_2 (probably in the form of ammonia) and CH_4 from the interior. Precisely how this happens is one of the Big Questions of Titan.

Even icy moons without an appreciable atmosphere can manifest features of considerable interest. Jupiter's moon Europa is a case in point. This satellite has a water ice crust between 10 and 50 km thick, but beneath the crust there lies a liquid water ocean. Europa shows an intriguing

range of crustal features, including some that suggest melt-through of the ocean. Of course, the existence of the ocean has attracted attention as possible habitat for life. The icy moons challenge the epistemological boundaries of planetary science. At the cold temperatures of Europa's surface, as on Titan, water ice is basically a rock, just as sand can be considered an "ice" of SiO_2 on Earth. The ice-rock forms minerals with ammonia and methane and other compounds, and when warm enough the ice-minerals can flow or melt and lead to cryovolcanism. When studying the crust and interior of Europa or Titan or other icy moons are we doing geology or oceanography or glaciology? Whatever one wants to call it, these moons are, as has been said, "always icy, never dull."

1.6 Farther afield: Extrasolar planets

Up until 1988, the Solar system was the only field of play for students of planetary climate, and it provided the only example against which theories of planetary formation could be tested. Revolutionary improvements in detection methods led to the first confirmed detection of a planet orbiting another star in that year, and instrumentation for planetary detection has continued to improve by leaps and bounds. As of the time of writing, over 228 planets orbiting stars within 200 parsecs (652.3 light years) of Earth have been detected, and the rate of detection of new planets is if anything accelerating. Certainly, much of the excitement surrounding the new zoology of planets has revolved around the prospect for detecting a planet that is habitable for life as we know it – or have known it to be in the past few billion years of Earth history. Perfectly aside from the habitability question, though, the rich variety of new planets discovered offers the student of planetary climate stimulus for thinking well outside the box of how the climates of known Solar System planets operate and have evolved over time.

Planets have been detected in orbit about a variety of different kinds of stars, so it is necessary to learn something of how stars are characterized. At the most basic level, stars are classified according to their luminosity (i.e. their net power output) and the temperature of the star at the surface from which the starlight escapes to space. The luminosity is determined by measuring the brightness of the star as seen from Earth's position, and then correcting for the distance between the star and the Earth. For relatively nearby stars, the distance can be measured directly by looking at the tiny shift in angular position of the star as seen from opposite ends of the Earth's orbit, but for stars farther than 500 parsecs (1630.8 light years), more indirect inferences are required. The stellar effective temperature is determined by measuring the spectrum of the starlight – how the brightness changes when the star is observed through different filters. Hotter stars have colors towards the blue end of the spectrum, while cooler stars tend toward the red. Hot stars emit more energy per unit surface area, but a reddish cool star can still have very high luminosity if it is very large, since it then has more surface area that is emitting. The energy sustaining the emission of starlight comes from the fusion of lighter elements into heavier elements. Since hydrogen is by far the most common element in the Universe that can participate in thermonuclear fusion, the overwhelming majority of stars ignite by fusing hydrogen into helium. These stars are known as *Main Sequence* stars, and stellar structure models predict that there is a distinct relation between the luminosity and emission temperature for stars that get their energy in this way. The stellar structure theories imply that the position of a star on the Main Sequence is determined primarily by its mass, with more massive stars being both hotter and more luminous. Moreover, stars spend most of their lifetimes on the main sequence, so that a scatter plot of luminosity vs. temperature of stars – a *Hertzsprung- Russell diagram* shows most stars to be clustered along the main-sequence curve. Indeed, the Main Sequence was discovered by plotting catalogs of stars in this way long before the energy source of stars was discovered.

Stars do not evolve *along* the main sequence. Rather, they enter it at a certain position when fusion ignites, remain near the same point for a certain amount of time while gradually brightening, and then leave the main sequence for a comparatively short afterlife as a brighter star with a more rapidly evolving spectrum. What happens when a star leaves the Main Sequence depends on the star's mass. The Sun will spend a billion years or so as a Red Giant, before collapsing into a gradually fading white dwarf. Main Sequence stars are thought to be the best candidates for hosts of habitable planets, since they provide relatively long-term stable stellar environments. Once a star leaves the main sequence, the climates of any unfortunate planets the star may have had will have been radically disrupted, if indeed the planets continue to exist at all; there will be relatively little time for any new form of life to establish itself. Nonetheless, planetary systems do exist off the main sequence, and these planets, too have their points of interest. The first confirmed detection of a planet orbiting a Main Sequence star did not occur until 1995.

Just as the luminosity of the Sun increases over time, the luminosity of other Main Sequence stars increases during their time on the Main Sequence. However, the lifetime of a star on the Main Sequence varies greatly with the mass of the star. The mass of the star determines the amount of nuclear fuel available to sustain the star's life on the main sequence, while the luminosity give the rate at which this fuel is consumed. The Main-Sequence lifetime of a star with mass $M_\circledast$ and luminosity $\mathcal{L}_\circledast$, scaled to values for the Sun, is estimated by

$$\tau_\circledast = \tau_\odot \frac{M_\circledast}{M_\odot} \frac{\mathcal{L}_\odot}{\mathcal{L}_\circledast} \tag{1.2}$$

Main sequence stars have a power law mass-luminosity relationship $\mathcal{L}_\circledast \propto M_\circledast^{3.5}$, so on the Main Sequence the lifetime scales with luminosity according to the law

$$\tau_\circledast = \tau_\odot (\frac{\mathcal{L}_\odot}{\mathcal{L}_\circledast})^{1.29} \tag{1.3}$$

Bright, hot, blue massive stars thus have a much shorter lifetime on the Main Sequence than dim, cool reddish dwarf stars. The Sun has a Main Sequence lifetime of about 10 billion years, and is nearly halfway through its time on the Main Sequence.

Figure 1.2 shows the Hertzsprung-Russell diagram for stars that are known to host one or more planets. Astronomers designate the color (equivalently surface temperature) of stars according to *spectral classes* given by the letters O, B,A,F,G,K,M extending from hottest to coldest, with numbers appended to indicate subdivisions within a spectral class. Our Sun is a class G2 Main Sequence star. The diagram shown in the Figure represents a tiny subset of the many millions of stars that have been catalogued, and none of these stars (so far) have spectral classes hotter than F. The stars cluster along the Main Sequence simply because there are more stars in general along the Main Sequence than elsewhere. There is also a selection bias due to the technologies available at present for detecting planets, which work better for some kinds of stars than for others. Thus, the gap in detections between K and M0 stars may be an artifact of detection bias rather than a reflection of some basic feature of planetary formation.

There is a rich supply of planets orbiting stars between spectral classes G0 and K, as well as a good handful of planets around bright red giant stars off the main sequence. The cluster of detections around M5 dim red-dwarf stars are particularly interesting, as many of these turn out to represent the most Earthlike planets to date – again a detection bias, because it is easier to detect low-mass Earthlike planets around low mass stars using present technology. These M-dwarf stars are very dim, so a planet has to be in a very close orbit about its star in order to be as warm as Earth. In compensation, these systems have very long lifetimes compared to the Sun and other brighter stars. According to Eq. 1.3, an M5 dwarf spends about 100 times as long on the main

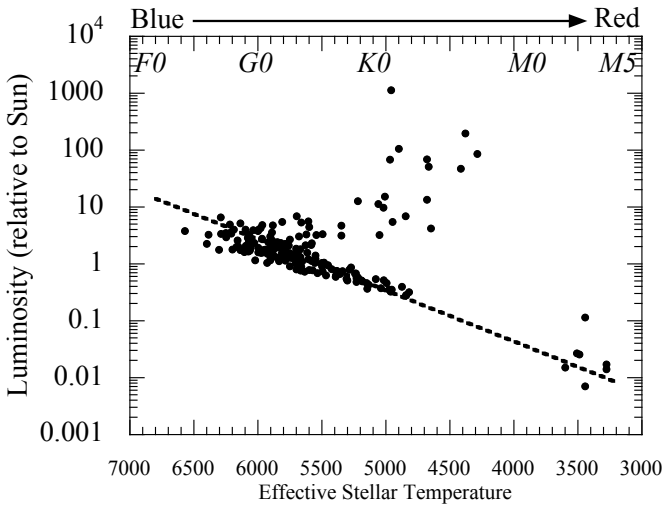

Figure 1.2: Scatter plot of luminosity vs. effective surface temperature for stars about which at least one orbiting planet has been detected as of 2007. Luminosity is given as multiples of luminosity of the Sun. The colder stars are redder while the hotter stars are more blue in color. The letters at the top indicate the standard spectral classification of stars in this temperature range, and the dashed line approximately locates the Main Sequence. The Sun is a class G2 Main Sequence star.

sequence as the Sun does, and so this kind of star will brighten only very slowly over time. Such a star provides a very stable climate to its planet, and requires much less adjustment of atmospheric conditions than the Earth had to accomplish to resolve the Faint Young Sun problem. In contrast, an F0 star would spend under a billion years on the Main Sequence, and if the history of life on Earth is anything to go by, life around an F0 star would be snuffed out at the prokaryotic stage, before it could even begin to think of making oxygen. Aside from affecting the lifetime, the spectral class of the star affects the degree of absorption of stellar radiation by whatever atmosphere the star's planet may have, and this too provides a lot of novel things to think about when pondering the climates of extrasolar planets.

That takes care of the stars, but what is known about the extrasolar planets themselves? Here, too, there is a detection bias, since it is much easier to detect massive planets comparable in mass to Jupiter than it is to detect more Earthlike planets. Most planets detected to date are very massive planets which, according to theories of planetary formation, are likely to be gas-giants like Jupiter or Saturn, or ice giants like Neptune or Saturn. The full variety of planetary climates offered by the various combinations of planetary mass, orbital characteristics and stellar characteristics is hard to convey by looking at just a few graphs. We'll give a small sampling of this variety below. In the course of the exercises in this book the student will have ample opportunities to explore the wider universe of exoplanets.

One of the key determinants of planetary climate is the rate at which the planet receives energy from its star. This is a function of the luminosity of the star and the distance of the planet from the star, which varies to some extent in the course of the planet's year (as discussed in Chapter 7). Planets that receive more stellar energy flux than the Earth will tend to be hotter, all other things being equal, whereas planets that receive less will tend to be colder. The left panel of Fig. 1.3 shows the mass of the planets discovered so far and plotted against the stellar flux heating the planet at the time of the planet's closest approach to its star. The masses are measured relative to the mass of Jupiter and the fluxes are measured relative to Earth's solar heating. One Earth mass is 0.00315 times the mass of Jupiter. On this diagram, a planet with a relative flux of unity would have an Earthlike temperature if its atmosphere were like Earth's. We see that a great many planets with a mass one tenth Jupiter mass or greater have been discovered; these are all likely to be gas giants or ice giants made mostly of hydrogen and helium. Though these are in some ways like Jupiters, their climate has no real analog in the Solar system, since most of them are in orbits where the planet receives at least as much stellar energy as the Earth receives from the Sun. These are all "hot Jupiters," and represent climates very different from anything in the Solar System. Could Jupiters receiving Earthlike stellar radiation be habitable for life? That is certainly a Big Question, requiring understanding of the climate of such planets. Even more exotic are the giant planets receiving vastly more stellar flux than the Earth – up to one thousand times as much, in fact. These are "roasters" – very hot gaseous planets in close orbits about their host stars. Planetary formation theory gave no inkling that such things should exist, and indeed the existence of roasters poses real challenges for the theory.

Another way that the new extrasolar planets offer novel climates is in the nature of their orbits. Solar System planets, except for Pluto, have fairly circular orbits. However, most exoplanets have highly elongated orbits with a considerable difference between the distance of closest approach to the star (the *periastron*) and the distance of farthest remove from the star (the *apastron*). The range of orbital elongation is shown in the right panel of Fig. 1.3. Since the stellar energy goes down like the square of the distance from the star, the planets with highly elongated orbits will have novel seasonal cycles unlike any encountered in the Solar System. They would tend to heat up to a great degree at periastron and cool down, perhaps freezing over any ocean, at apastron. Could such a planet be habitable? The answer depends a lot on the thermal response time of the

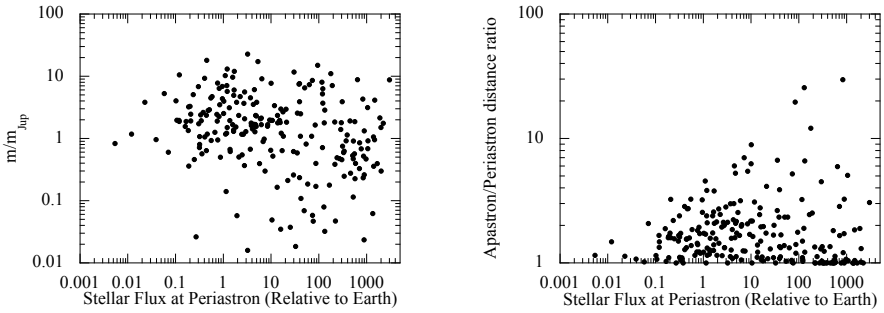

Figure 1.3: Left panel: Scatter plot of mass of extrasolar planets in units of Jupiter masses vs. the flux of stellar energy impinging on the planet at the time of closest approach (the periastron). Right panel: Scatter plot of the ratio of farthest (apastron) to closest (periastron) distance of the planet from its star in the course of the planets orbit vs. the stellar energy at the time of closest approach. The stellar energy flux is given as a multiple of the corresponding flux for the Earth.

planet's atmosphere and ocean, which could average out the orbital extremes. The relevant physics is discussed in Chapter 7.

The orbital period, or "year", of the extrasolar planets also varies widely, as shown in the left panel of Fig. 1.4. Planets have been discovered with a wide range of orbital periods, ranging from as little as two Earth days to as much as 6000 Earth days. Planets in close-orbits with short orbital periods are likely to be *tide locked* and always present the same face to the star, just as the Moon always presents the same face to the Earth. The Super Earths to be discussed shortly are mostly in this category. Tide-locked planets offer novel possibilities for planetary climate. The night-side could get very cold, and if the planet has an ocean, it might freeze over completely. The day side would be hot, but there could be a habitable zone near the ice margin. Further, the transport of moisture and heat from the day side to the night side poses interesting questions; the answers are important, since such transports in large measure will determine the nature of the planet's climate.

Low mass planets are of particular interest because according to theories of planetary formation they have the best chance to have a rocky composition similar to that of Earth, Mars or Venus. Relatively few planets with a mass of 10 Earth masses (0.03 Jupiter masses) or less have been discovered, but there has been recent progress in this area. The planets discovered so far in this range are all considerably more massive than Earth, and are therefore called *Super Earths*. The left panel of Fig. 1.3 shows that a handful of Super Earths have been discovered with stellar fluxes ranging from .25 that of Earth (yielding a too-cold planet) to 1000 times that of Earth (yielding a planet far too hot. The closest to being "just right" is Gliese 581c, with a flux of about three times that of Earth. Would such a planet be habitable, or would it turn into a Venus? The physics needed to answer such questions will be developed in Chapter 4.

The right panel of Fig. 1.4 gives an indication of the masses of planets that have been discovered about stars having various temperatures. The stellar temperature is of interest to climate since it determines the spectrum – the redness or blueness – of the starlight, which in turn affects the absorption of starlight by various atmospheric constituents. Hotter stars also put out more of the energic ultraviolet radiation, which can have a profound effect on atmospheric chemistry. We see that a few low mass Super Earths have been found around class G or K stars,

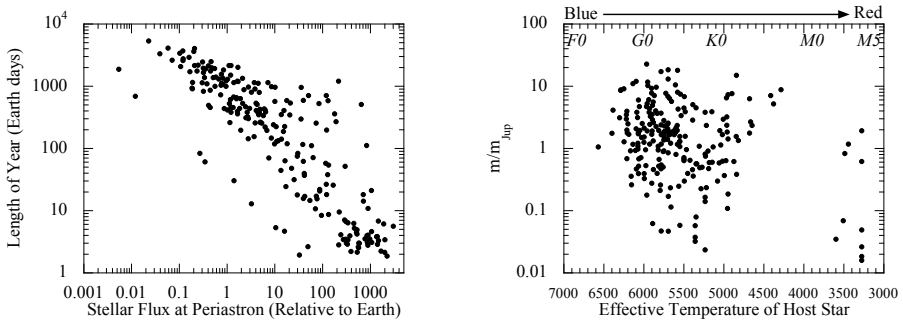

Figure 1.4: Left panel: Scatter plot of orbital period (in Earth days) vs. the flux of stellar energy impinging on the planet at the time of closest approach (the periastron). Right panel: Scatter plot of the mass of the planets (in units of Jupiter masses) vs. the effective radiating temperature of the host star.

but examination of the associated planetary data reveals that these planets are in close orbits and receive several hundred times as much stellar energy as the Earth receives from the Sun. They will unquestionably be extremely hot places, and unlikely to be able to sustain an atmosphere or liquid water. There is, however, a small cluster of Super Earths orbiting very cool stars with temperatures under 3700K. These stars are Main Sequence "M-dwarfs." They are very red, very small, and very dim, but by virtue of their dimness planets in close-orbits can still have a chance to be habitable. Moreover, dim stars like M-dwarfs have a long lifetime, and therefore provide a stable environment for their planets. Gliese 581 is such a system, but it appears that the two most Earthlike planets still miss being habitable – one likely to turn into a Venus if it has an ocean and the other likely to freeze into a Snowball. Part of the reason for interest in M-dwarfs comes purely from detection technology. It is comparatively easy to detect low-mass planets in close orbit around a low-mass star, and at the same time low-mass dim stars give a planet in a close orbit a chance to be habitable. As time goes on, it is likely that a habitable world around an M-dwarf will be discovered. As detection technology improves, the possibility for discovering Earthlike habitable planets will expand to other spectral classes of stars.

Naturally, one is at this point intensely curious as to the composition of these planets, whether they have water, and what their atmospheres (if any) are made of. Some information can be inferred from planetary formation theory and theories of atmospheric evolution , but there is as yet no great ability to determine atmospheric composition from observations. That will change in the next decade or two, as satellite-borne instruments come online which will be able to determine the spectrum of emission and reflection from extrasolar planets. The anticipated instruments will only return a single spectrum averaged over the whole visible surface of the planet, but a great deal can nonetheless be inferred about atmospheric composition from such data. Learning how to make the most of this single-pixel planetary astronomy, in which the Earth would appear as a pale blue dot (in Carl Sagan's words) opens a whole range of Big Questions. Much of the same radiative transfer used in calculating planetary climate bears on the interpretation of planetary spectrum as well.

1.7 Digression: About climate proxies

1.7.1 Overview of proxy data

Instrumental records of climate – that is, records of measurements of temperature and other quantities by scientific instruments – date back at most a few hundred years. The first accurate thermometer was invented in 1654 by Ferdinando II de'Medici, and two hundred years passed before anything like a global network of reliable temperature measuring stations began to become available. Written historical records of such events as frost dates, encounters with sea ice and depictions of mountain glacier length provide some information about the climate of the past few millennia, but for the most part one must rely on *climate proxies* for information on what climate was doing a century or more ago. A climate proxy is any measurable thing preserved in the geological record of Earth or other planets, from which some aspect of climate can be inferred; to be useful, a climate proxy must come with a *chronology*, that is, some means of telling what period the proxy dates to. There is a vast and ever-improving array of climate proxies. We have encountered a few already. For example, the existence of river networks on the surface of Mars tells us that at some time in the past the surface of Mars must have been warm enough to support a liquid (probably water) flowing for long distances along the surface; in this case, the chronology at the time of writing comes from counting the number of craters superimposed on the features. Similarly, the existence of marine sedimentary deposits and stromatolites during the Archean provides compelling evidence that the Earth was warm enough to support open ocean water through much of its early history.

Some of the more intuitive kinds of proxies derive from plants and animals that live on land, since physiology places certain constraints on the conditions in which various organisms can grow or thrive. The presence of cold-blooded animals like crocodiles is a sure sign that winters cannot have been much below freezing for extended periods of time. Some kind of plants require tropical conditions to survive, while others require cooler conditions; even the shape of leaves provides information about temperature. A great deal of this evidence is preserved in the fossil record. Where there are trees, the width of tree rings provides a record of annual variations in the temperature of the growing season (though it is sometimes hard to distinguish temperature from rainfall effects). Land-based proxies are only available after the time at which a fairly diverse ecosystem established itself on land. While the first primitive plants colonized land as early as 470 million years ago (in the *Ordovician period*), and the first plants with stem structures able to conduct water appeared perhaps 430 million years ago (the *Silurian period*, a diverse land ecosystem did not really get underway until the *Devonian period*, which began about 416 million years ago. Once plants colonized land, animals in search of a free lunch followed not long afterwards.

One of the richest sources of information about past climates of the Earth comes from material preserved in sea floor sediments. Sediment is laid down in layers like the pages of a book, in which the history of Earth's climate can be read. The Earth is a dynamic planet, and sea floor is constantly being re-created at mid-ocean ridges, and likewise pulled down into the Earth's interior for recycling at subduction zones. For this reason, the deep-ocean marine sediment record is mostly limited to the past 100 million years, and is rather sketchy for the first half of that period. The oldest remaining deep-sea floor is about 180 million years old, and there is precious little of that. Near-shore deposits on continental shelves, on the other hand, can be uplifted and preserved for hundreds of millions, or even billions, of years. Many of the key marine deposits that tell us about the climate of the Neoproterozoic (about 600 million years ago) are now high and dry in Namibia, while others are found in Arctic Canada. These various deposits have a lot of individual personality and are known to geologists by names such as the Acasta Gneiss, the Warrawoona Formation, the Akademikerbreen or the Isua Greenstone Belt; we've met many of these already in

the preceding sections.

Numerous aspects of the sedimentary record have been used to infer past conditions, and the ingenuity of paleoclimatologists is adding to the list all the time. Some sedimentary proxies involve the physical structure of the sediments, and are independent of chemistry or biology. *Diamictites* are a class of sediments known to be carried by land glaciers discharging into the ocean. They are a sure sign of very cold conditions, since it is only in such cases that a glacier can survive to sea level. *Ice-rafted debris* (IRD) is coarse material that can only be carried offshore by hitching a ride on icebergs or sea-ice. *Dropstones* – individual stones the size of pebbles and larger which fall with enough force to deform sediment layers – are a particularly striking form of ice-rafted debris. Inasmuch as stones do not float, they present very convincing evidence for ice. Continental dust in sediments provides an indication of the strength of wind, since larger particles require stronger winds to loft and transport them; the mineral composition of the dust can often mark where the dust came from, and hence the direction of the wind. Surveying the species of fossil algae found in sediments can provide information about the temperature of the layers in which the algae grew, since some organisms require colder temperatures while other require warmer temperatures.

A great deal of information can be gleaned from the chemical composition of the ocean, and of the microscopic creatures which dwell within it. Some chemical proxies do not rely on biology, and others make use of organisms mainly as nearly passive recorders of the composition of the ocean. Still other proxies are more intimately tied to biology, through the effect of temperature on the rate at which an organisms makes use of one element in preference to another. For example, the ratio of Magnesium (Mg) to Calcium (Ca) in corals and in the shells of certain micro-organisms is dependent on the temperature at which the organisms grew. The Strontium-Calcium substitution has been used similarly. Organic molecular proxies represent an important new source of data; it has been discovered that certain kinds of micro-organisms produce long-chain organic molecules with a somewhat variable chain-length which depends on the temperature at which the organism grows. When these molecules are robust enough to be preserved in the sedimentary record, the ratio of chain-lengths can be used to infer past temperatures. Alkenone and Tex-86 proxies fall into this class. They are useful only over time spans for which organisms producing the molecules exist, and can be presumed to have similar biochemistry to modern analogue organisms. Both alkenones and Tex-86 have been used to probe climates lying many tens of millions of years in the past, as well as more recent climates of the past few hundred thousand years.

If the chemical composition of a sedimentary sample has been altered by interaction with ocean water at some time after the sediment was first formed, then the information the sediment provides about past conditions is severely compromised. Post-depositional alteration is known as *diagenesis*, and it is a problem that plagues interpretation of geochemical sedimentary proxies. Paleoceanographers have had to become very clever sedimentary detectives in order to check for the nefarious effects of diagenesis, particularly when dealing with samples more than a few million years old. In some cases the existence of diagenesis has gone undetected for a decade or more, as will be discussed later in connection with the problem of hothouse climates.

1.7.2 Isotopic proxies

The chemical properties of an element are primarily determined by the number of protons in the nucleus (the *atomic number*, which also determines the configuration of the electron cloud. Nucleii also contain neutrons, and atoms having the same atomic number can appear in forms with different numbers of neutrons. These differing forms are known as *isotopes* of the element. Isotopic proxies have proved to be a versatile source of information about past climates. Some

isotopes are unstable, and decay into other elements; these can be used as "clocks," to determine when things happened. The original, and most famous, such application is radiocarbon dating, which makes use of the decay of the form of carbon having a molecular weight of 14 (carbon-14, or ^{14}C. Stable isotopes do not decay, and instead provide a tracer of past chemical reactions in which the substance participated. For elements heavier than Helium the stable isotopic composition of a planet is determined primarily by the synthesis of the elements in the supernova explosion which gave birth to the material eventually incorporated into the Solar system. The isotopic ratio can in some cases be further changed by the process of planetary formation. For example, oxygen has three stable isotopes: ^{18}O, ^{17}O and ^{16}O, the latter of which is by far the most common. About 1 in 500 atoms of oxygen on Earth are ^{18}O, which is nearly the same ratio as found in the Sun and which presumably represents the composition of the primordial Solar nebula. ^{17}O is less commonly used as a proxy because it is so much rarer than ^{18}O (only 1 in 2500 oxygen atoms on Earth), but variations in its composition are readily detectable with modern measurements, and have important specialized uses. Hydrogen has two stable forms: Deuterium (D) which has a proton and a neutron, and normal hydrogen (H) which has only a proton. About one hydrogen molecule in 6500 occurring in Earth's ocean water is D. This differs from the 1:1700 ratio in the outer Sun, because Deuterium is destroyed by nuclear fusion in the Sun and in the early Solar nebula; curiously, the composition of the outer Sun is nearly the same as Jupiter. The stable isotopes of carbon are ^{13}C and ^{12}C, of which the latter is by far the more common. About 1 in 100 carbon atoms on Earth are ^{13}C. The carbon isotopic system attracts particular attention because carbon is the basis of organic chemistry – the stuff of life as we know it. Stable isotopes of S, N, Ar and many other elements have proved useful as proxies. In fact, it is hard to think of any measureable isotopic ratio that hasn't proved useful as a proxy for some aspect of climate or atmospheric composition.

The utility of isotopes as climate proxies derives from what they tell us about processes that sort them out into different reservoirs. Isotopes of an element have *nearly* the same chemical and physical behavior, but not *exactly* the same behavior. The subtle differences come about in part because, at a given temperature, the molecules of the lighter isotopes have a higher typical velocity than those of the heavier isotopes, since all molecules have the same mean kinetic energy at any given temperature. Among other things, this means that lighter isotopic forms of a substance evaporate more readily than heavier ones, so that the vapor phase of a substance is typically depleted in heavy isotopic forms relative to the condensed phase with which it is in contact. We will find this effect particularly useful in the interpretation of water isotopes, but a very similar process applies to the gradual "evaporation" of a planet's atmosphere into outer space, and can be used to constrain the proportion of atmosphere that has been lost in such a fashion.

The rate at which a given element or a molecule containing that element undergoes chemical reactions also depends on the isotope involved. Other things being equal, heavier isotopes would tend to react more slowly, because they have lower speeds and therefore lower collision rates with other reactants than their lighter, nimbler cousins. However, there are more subtle effects involved that can either enhance or retard reaction rates. Specifically, the difference in molecular weight between isotopes can affect the characteristic vibration frequencies of molecule bonds, and this can in turn affect the probability that colliding reactants will stick together. Very often, the degree of preference for one isotopic form over another – the degree of fractionation occurring between reactant pool and reaction product – depends on the temperature at which the reaction is taking place. When this is so, the isotopic composition of the product provides a useful paleo-thermometer. In Section 1.3 we stated that "certain aspects" of the chemical composition of zircons constrained the temperature of Hadean water with which the zircons were in contact; with this cryptic phrase, we were in fact referring to the ratio of ^{16}O to ^{18}O in the Hadean zircons which, like all silicate minerals, contain oxygen (their chemical formula is $ZrSiO_4$). Similarly, it was the oxygen isotopic

ratio of cherts (later confirmed by silicon isotopic ratios) that was used to constrain Archaean temperatures in Section 1.3. A problem with all such paleothermometers is that fractionation is *relative to the isotopic composition of the reactant pool*, so that one needs to know the likely composition of the reactant pool in order to infer temperatures from the reaction product (e.g. the cherts or zircons), which are preserved long after the pool of reactants have been dissipated.

Biochemistry also has isotopic preferences. Photosynthetic Earth life (even of the sort that doesn't produce oxygen) prefers the lighter forms of carbon, so that organic material of photosynthetic origin is enriched in ^{12}C and depleted in ^{13}C relative to the inorganic carbon left behind. A record of the isotopic composition of inorganic carbon is preserved in carbonate minerals (e.g. $CaCO_3$) precipitated out and deposited on the ocean floor, with or without the help of shell-forming organisms. Organic material is directly preserved in sedimentary rocks such as shales. For example, the organic material in the Isua shales is presumed to be of biological origin because its carbon is isotopically light – enriched in ^{12}C relative to the average composition of carbon on Earth. Respiration – eating organic matter and combining it with oxygen to release energy – does not fractionate carbon to any significant extent, so the isotopic signature imprinted by photosynthesis is carried over to those of us creatures in the non-photosynthetic organic realm.

The simplest kind of isotopic fractionation to characterize is *equilibrium fractionation*. To keep things concrete, we'll illustrate the concept using oxygen isotopes. Consider two physically or chemically distinct substances, each of which contains oxygen atoms. For example, the two substances could consist of water in its vapor phase and water in its liquid phase. Alternately, the two substances might be different chemical compounds containing oxygen, such as calcium carbonate ($CaCO_3$) and water (H_2O), or silica (SiO_2) and water; the latter pair leads to the chert-based paleothermometer mentioned earlier. Each of the two substances will have some initial ratio of ^{18}O to ^{16}O. Now imagine bringing the two substances into contact, whereupon the two substances exchange heavier and lighter forms of oxygen, changing the initial ratios. After a very long time, the isotopic ratios in each substance will reach equilibrium and stop changing. At that point, we can define the equilibrium fractionation factor $f_{1,2}$ via the relation

$$r_1 = f_{1,2} r_2 \tag{1.4}$$

where r_1 is the ratio of ^{18}O to ^{16}O in the first substance after equilibrium has been reached, and r_2 is the corresponding ratio for the second substance. The fractionation factors differ for different pairs of substances, but typically (though not invariably) exhibit the following characteristics:

- The fractionation factor is typically quite close to unity
- The fractionation factor deviates most from unity at low temperatures, and approaches unity as temperature increases
- The deviation of the fractionation factor from unity increases as the contrast between the masses of the two isototopes increases (e.g. more fractionation for ^{18}O vs ^{16}O than for ^{17}O vs ^{16}O

For example, the oxygen in silica (SiO_2) is isotopically heavy in comparison with the oxygen in the water with which it is in equilibrium; the fractionation factor is 1.036 at $20C$ but falls to 1.018 at $100C$. Similarly, oxygen in liquid water is isotopically heavy in comparsion to that in the water vapor with which it is in equilbrium; in this case, the fractionation factor is 1.01 at $20C$ and falls to 1.005 at $100C$. It is the temperature dependence of the fractionation that makes it possible to use isotopic ratios as paleothermometers. Some kinds of fractionation appear to operate in nearly the same way whether the reactions happen within organisms or inorganically.

This appears to be the case for carbonate precipitation, which fractionates in much the same way regardless of whether it happens inorganically or in shell-forming organisms. Other forms of fractionation, such as the fractionation of carbon isotopes in photosynthesis, are more inherently biologically mediated, though even in such cases the fractionation factors tend to be similar across broad classes of organisms sharing the same biochemical pathways.

Isotopic composition is usually described using δ notation, which is defined as follows. Let r_A be the ratio of the number of molecules of isotope A in a sample to the number of molecules of the dominant isotope. Typically, r_A will be a rather small number. Next let $r_{A,S}$ be the isotopic ratio for a standard reference sample. Isotopic composition is invariably reported relative to a standard because the analytical instruments currently in use cannot measure the absolute composition to very high accuracy, but they can measure the difference relative to a standard very accurately; the standard is an actual physical substance, natural or manufactured, which can be put into the analytical instrument and serve as a basis for comparison. The choice of standard is a matter of convention, and there are various standards typically used in different contexts. For example, the standard for oxygen and hydrogen isotopes in ice or water is usually taken to be *VSMOW*, which stands for Vienna Standard Mean Ocean Water. The ratio of ^{18}O to ^{16}O in VSMOW is 1/498.7, and of D to H is 1/6420; this approximates the mean present composition of water in the ocean.

Once one agrees upon a standard, the isotopic composition of a sample can be described in terms of the quantity

$$\delta A \equiv \frac{r_A - r_{A,S}}{r_A, S} \tag{1.5}$$

Thus, negative values of δ indicate that the sample is depleted in isotope A relative to the standard, whereas positive values indicate that the sample is enriched. The δ value is usually expressed as a *per mil*, or parts-per-thousand, value. For example, a δ value of .001 would usually be expressed as 1 *per mil* or 1‰. A difference of 1‰is often equivalent to a miniscule variation in isotopic concentration, requiring high analytical precision to measure. For example, for the case of $\delta^{18}O$, a 1‰difference is equivalent to changing the ratio of ^{18}O molecules to ^{16}O molecules by $.001 \cdot \frac{1}{498.7}$, or a hair over 2 parts per million. Deuterium (D) is even tougher. For deuterium, a difference of 1‰amounts to a change in the D to H ratio of only 0.16 parts per million, though the challenge is offset by the fact that fractionation in the D/H system is considerably stronger than fractionation in the $^{18}O/^{16}O$ system owing to the lesser relative mass contrast in the latter case.

Carbonate minerals (e.g. $CaCO_3$ or $MgCO_3$) are important recorders of the oxygen and carbon isotopic composition of past environments. For carbonates, the isotopic composition is usually reported relative to the PDB standard, named after a mineral powder made from a naturally occurring fossil carbonate (the "Peedee Belemnite") The physical powder standard no longer exists, so it has been supplanted by a synthetic equivalent, known as $VPDB$. It is important to keep the standards in mind when interpreting the isotopic literature. Oxygen isotopes occur in both carbonate and water, and so can be reported relative to either the $VSMOW$ or $VPDB$ standard. The conversion between the two is given by

$$\delta^{18}O(VSMOW) = 1.03091\delta^{18}O(VPDB) + 30.91 \tag{1.6}$$

A useful rule of thumb to keep in mind when interpreting oxygen isotopes in marine carbonates is that a carbonate reading zero relative to $VPDB$ would be in equilibrium with water having zero $\delta^{18}O(VSMOW)$, if the carbonate formed at a temperature of around $18C$. This means that a carbonate $\delta^{18}O(PDB)$ of "around" zero goes along with the waters in which they formed having $\delta^{18}O(VSMOW)$ of "around" zero. Later, we'll clarify just what we mean by "around," and how that depends on temperature.

Having introduced the δ notation and the $VPDB$ standard, we are now in a position to get more quantitative about the things to be learned from the stable isotopes of carbon preserved in carbonates. The quantity of interest is $\delta^{13}C$, reported relative to the $VPDB$ reference. Carbon dioxide is cooked out of carbonates in the interior of the Earth, and outgases from volcanoes, subduction zones and the mid-ocean ridge with $\delta^{13}C \approx -6‰$. In a steady state, the flux of carbon in this carbon dioxide is balanced by the burial of carbon in the form of inorganic carbonates (e.g. $CaCO_3$ and organic carbon (schematically CH_2O). In the long run, most of this burial is in sea-floor sediments, since whatever forms on land tends to eventually get washed into the ocean. Note that even carbonate precipitated biologically in the form of shells of organisms is considered inorganic material, and acts pretty much (though not exactly) like inorganically precipitated carbonate. There are some kinds of recently evolved plants that use photosynthetic pathways that fractionate carbon very little, but for the most part, photosynthetically produced organic carbon has $\delta^{13}C$ values that are about 25‰lower than that of the carbon dioxide reservoir from which the photosynthetic organisms make their substance. For example, CO_2 in the atmosphere today has $\delta^{13}C \approx -8‰$, while land plants have $\delta^{13}C$ values running from -32‰to -25‰and marine organic carbon has $\delta^{13}C \approx -25‰$. Fossil fuels, which are made from ancient land plants (in the case of coal) or marine organisms (in the case of oil) have $\delta^{13}C \approx -25‰$.

If f_{org} is the fraction of carbon which is buried in the form of organic carbon, δ_o is the $\delta^{13}C$ of carbon outgassed from the Earth's interior, δ_{org} is the $\delta^{13}C$ of the organic carbon buried and δ_{carb} is the $\delta^{13}C$ of the carbonate carbon buried, then mass balance implies

$$\delta_o = f_{org}\delta_{org} + (1 - f_{org})\delta_{carb} \tag{1.7}$$

If data from both organic and carbonate sediments are available, this formula can be used directly to infer f_{org}. It is instructive, however, to make use of the intrinsic fractionation between inorganic carbon and photosynthetically produced organic carbon to infer the isotopic compositions of the two burial fluxes as a function of f_{org}. Note that because photosynthetic fractionation is relative to the composition of the inorganic carbon pool, it is incorrect to assume that $\delta_{org} \approx -25‰$. It could be considerably heavier, if the inorganic pool has very positive $\delta^{13}C$. Let's consider two limiting cases. Suppose that $f_{org} \approx 1$, so that nearly all of the carbon outgassed from the Earth's interior is intercepted by photosynthesis and buried as organic carbon. In that case, mass balance requires that $\delta_{org} \approx \delta_o \approx -6‰$. This can only happen if the inorganic carbon pool consists of isotopically heavy carbon with $\delta^{13}C \approx (-6 + 25)‰ = 19‰$. Since the carbonate that precipitates from an inorganic carbon pool tends to be isotopically heavier than the pool itself, the trickle of carbonate precipitated in this situation will have $\delta^{13}C$ in excess of 19‰. The precise value depends on aspects of ocean chemistry we will not pursue here. In the opposite limit, when $f_{org} \approx 0$ and there is little organic carbon burial, then $\delta_{carb} \approx \delta_o \approx -6‰$. The inorganic carbon pool this carbonate precipitates from is somewhat isotopically light compared to the carbonate itself, and the organic carbon fractionates relative to that value, yielding $\delta_{org} < (-6 - 25)‰ = -31‰$. The typical situation over the past two billion years has been for δ_{carb} to be somewhat positive, between 2‰and 5‰, while δ_{org} hovers around -22‰. There are periods, however, when carbon isotopes undergo considerable excursions from the typical situation. These *carbon isotope excursions* provide an important window into big events in the carbon cycle.

The above picture applies only when the carbon cycle is in a steady state. When any part of the carbon cycle is significantly out of equilibrium – for example, when one is building up a new pool of stored organic carbon in land plants and soils – the simple input-output isotopic calculation no longer works. One must then do a detailed accounting of the flows of carbon between the various reservoirs involved, and the attendent isotopic fractionations. This can be a very intricate process, especially since the fractionations involved are generally temperature dependent. There are other

important aspects of the isotopic carbon cycle we have swept under the rug, such as the important information that can be gained from vertical gradients of $\delta^{13}C$ in carbonates.

There is another biological process that can leave a distinct mark on the carbon isotopic record, namely *methanogenesis.* When there is oxygen around, organic matter generally gets decomposed into CO_2 by respiration. In anaerobic environments, methanogens get the goodies instead, and turn organic matter into CH_4. This is a multi-step process, each step of which fractionates carbon. The result is CH_4 which is much lighter isotopically than the organic feedstock from which it was produced. Biologically produced methane today has $\delta^{13}C$ values on the order of -50‰. When the atmosphere-ocean system is rich in oxygen, as has been the case for the past half billion years, methanogenesis plays a very minor role in the carbon cycle and usually leaves little imprint in the isotopic record. A possible exception to this general rule may occur as a result of gradual accumulation of large amounts of methane in the form of exotic ices called *clathrates*, which can form in ocean floor sediments and under arctic permafrost laters. If some event occurs which suddenly releases this stored methane into the atmosphere or ocean, the isotopically light methane quickly oxidizes into CO_2 which works its way into the carbonates. The net result is a negative carbonate carbon isotope excursion, the magnitude of which tells us something about the quantity of methane released relative to the net carbon in the ocean-atmosphere system.

At present, the arsenal of climate proxies is much more limited for other planets than it is for Earth. Biologically mediated proxies are obviously not in the cards for planets that seem to have no biology, but many of the abiological proxies used on Earth would be equally useful on other planets; cherts on Mars would provide much the same kind of information as cherts have provided on Earth. The use of these chemical proxies is limited primarily by the weight and power consumption of analytical instruments needed to carry out some of the analyses that are typically done on Earth materials. The same constraint applies to many of the means of determining chronology based on radioactive decay. Such techniques can be applied to other planets with a preserved geological record (notably Mars), but must await sample return missions. Meanwhile, a considerable amount has been accomplished by remote-sensing from planetary orbiters, and from low-power instrumentation on landers. The landforms of Mars have been imaged in great detail, and constitute a proxy for past climates with regard to occurence of water and glacial activity. The minerology of the surface can be determined by a range of remote-sensing techniques, so a fair amount is known about the occurence of clay minerals (a signature of water and weathering) in the ancient crust of Mars, and other minerals such as the iron compound hematite also tell us something about the aqueous environment of Early Mars. On all planets, a study of the isotopic composition of atmospheric gases (possible by *in situ* and remote spectroscopic means) provides valuable information on the source of the atmosphere and how much has been lost over time, insofar as lighter isotopes escape to space more readily than heavy isotopes. With sample return missions and improved robotic exploration, the future promises a rich expansion in planetary proxy studies, not least the prospect of drilling the Martian polar glaciers to see what climate mysteries they record.

1.7.3 Hydrogen and Oxygen isotopes in sea water and marine sediments

We will turn our attention now to the isotopes of hydrogen and oxygen contained in water and in sediments precipitated from the water column. We'll learn what the concentration of these isotopes tells us about the volume of glacier ice and the temperature of various parts of the ocean. "Normal" water is $H_2^{16}O$, but other isotopes of hydrogen or oxygen can substitute for the most prevalent isotopes, leading to various forms of heavy water, notably $HD^{16}O$ and $H_2^{18}O$.

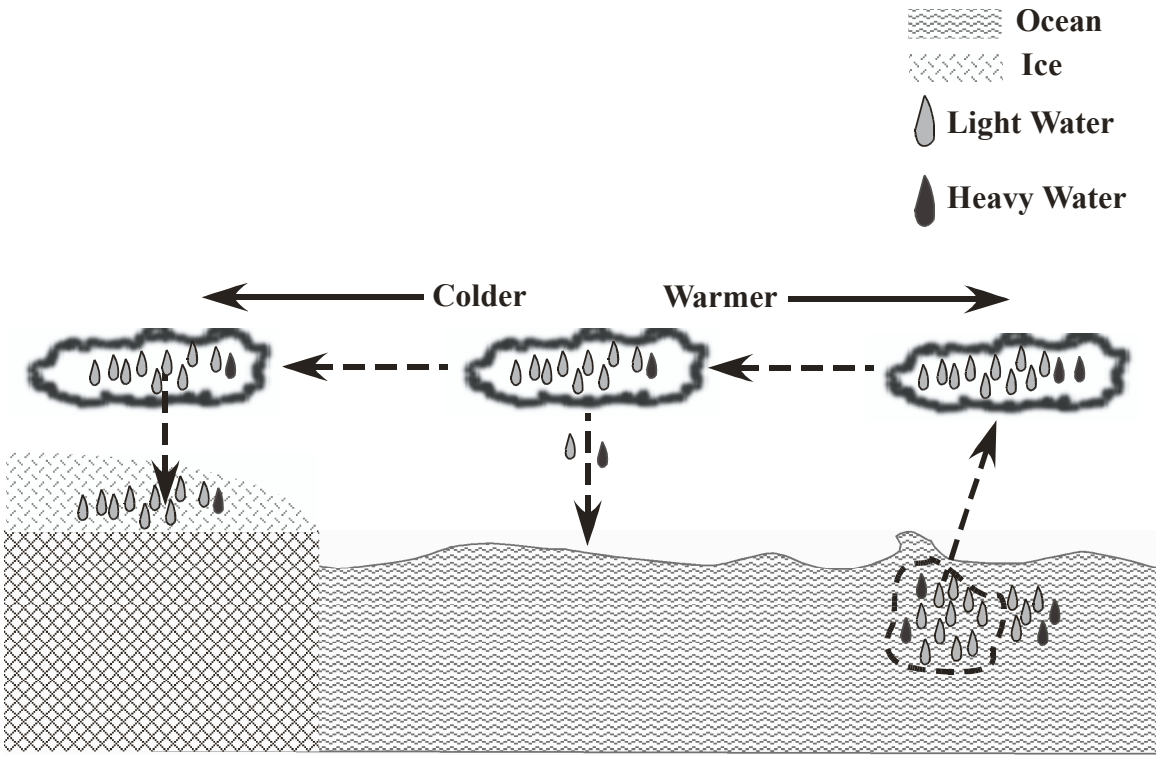

Figure 1.5: Sketch showing how the growth of ice sheets on land affects the isotopic composition of ocean water. The water vapor which evaporates from the ocean is enriched in lighter forms of water, and becomes more isotopically light as the heavy forms of water preferentially rain or snow out before the remainder is deposited on the glacier. This process systematically transfers isotopically light water to the glacier, leaving the ocean isotopically heavy.

At any given temperature light molecules, on average, travel with higher velocity than heavier molecules. This implies that during evaporation, the lighter isotopic forms of water evaporate more readily than the heavier forms, leading the vapor to be enriched in light water and depleted in heavy water relative to the liquid reservoir. Furthermore, when water vapor condenses into liquid or ice, the heavier forms condense more readily because the slower moving molecules can more easily stick together without bouncing off each other. In consequence, the rainfall is enriched in heavy forms relative to the vapor in the air, while the vapor left behind in the air is further enriched in the light forms and further depleted in the heavy forms. This is a form of distillation, very similar to the process by which one makes brandy from wine (or moonshine from fermented corn mash). Alcohol is more volatile than water, so the vapor in contact with heated wine is enriched in alcohol relative to the liquid; if part of the vapor cools and condenses, the water condenses out preferentially, leaving a potent essence at the end of the still if the remaining vapor is then condensed separately.

The way the distillation process affects the isotopic composition of sea water is sketched in Fig. 1.5. Let's suppose that the ocean starts with $\delta^{18}O$ of zero relative to VSMOW. Water will evaporate into the air until the air becomes saturated with water vapor (a concept that will be made precise in Chapter 2). Since heavy molecules evaporate less readily than light molecules, the

water vapor will be depleted in ^{18}O relative to the ocean – in other words, it will be *isotopically light*. More precisely, the ratio of ^{18}O to ^{16}O for the water vapor in the air will be less than the ratio of the original ocean water, leading to a negative $\delta^{18}O$ for the vapor. The amount of depletion depends weakly on temperature. At $273K$, the water vapor $\delta^{18}O$ is shifted by -11.7 ‰relative to the ocean. At $290K$ the shift is -10.1‰and at $350K$ the shift is -6.0‰[5]. This reduction in isotopic contrast between reservoirs as temperature increases is typical of almost all isotopic fractionation problems. Because the amount of water stored in the form of water vapor in the atmosphere is utterly dwarfed by the amount of water in the ocean, the selective removal of light isotopes makes the ocean only very, very slightly isotopically heavy. But what if we removed the atmosphere's water vapor, sequestered it in a glacier on land, and repeated the process many times over until a substantial fraction of the ocean water had been transformed into an isotopically light glacier? In that case, the systematic removal of large volumes of isotopically light water from the ocean would leave the ocean water isotopically heavy by a significant amount – it would have a significantly positive $\delta^{18}O$. Thus, the degree of to which the ocean is enriched in isotopically heavy forms of water tells us how much ice has built up on land. As ice volume becomes greater, the $\delta^{18}O$ (or similarly, δD) of the remaining ocean water becomes more positive. As an example, let's suppose we build an ice sheet by removing $200m$ depth of water from an ocean with a mean depth of $4km$, assuming the glacier to be built from vapor with $\delta^{18}O$ = -10‰. If δ_i is the $\delta^{18}O$ of the ice and δ_o is that of the ocean water, then conservation of molecules implies that $200\delta_i + (4000 - 200)\delta_o = 0$, if the ocean started with $\delta^{18}O = 0$. From this we conclude δ_o = .526‰.

In fact, for the reasons sketched in Fig. 1.5, the water that eventually snows out to form glaciers is much more isotopically light than the -10‰value one might expect from just looking at the vapor in equilibrium with ocean water. The initially evaporated water vapor may have $\delta^{18}O$ = -10‰, but on the way to the cold polar regions, some of that water will rain out back into the ocean, and the condensed water is isotopically heavy relative to the vapor, since heavy species condense more readily. That means that each time some atmospheric water vapor is lost to rainfall or snowfall back into the ocean, the vapor left behind becomes lighter. The precise extent of the additional lightening by the time the snow eventually falls out on a glacier depends on the amount of water lost on the way, which is in turn a function of the temperature difference between where the water was picked up from the ocean and where it was dropped on the glacier. Over the past 100,000 years, the $\delta^{18}O$ of the snow falling on Greenland has varied from -42 ‰in the coldest times to values around -35‰today. Antarctic ice is somewhat isotopically lighter than Greenland ice, and has $\delta^{18}O$ values ranging from -40‰to -55‰depending on location and age of the ice. Because of the additional fractionation on the way to the pole, the formation of the glacier would leave the ocean much more enriched in heavy isotopes than our previous estimate suggested. For glaciers having isotopic compositions comparable to the present ones, removing 200m of ocean to build glaciers would leave the ocean enriched by about +2‰, rather than a mere .526‰. The preceding discussion also shows that in order to translate the δ value of the ocean into an ice volume, one needs some estimate of the isotopic composition of the glaciers being formed. For the present glaciers, this can be determined by drilling into the ice, but for past ice that no longer exists one must rely on modelling.

So, if we could go back in time and grab a bucket of sea water, measuring its isotopic composition would tell us the volume of ice on the Earth, and this would tell us much about how cold the planet was. Wouldn't it be awfully nice if there were some way to do that?

[5] In reality, the isotopic composition of water vapor in the atmosphere just above the ocean deviates somewhat from the equilibrium value, because the water vapor is in a dynamic balance between evaporation from the ocean and mixing of dry air into the layer from aloft. It is only when the air is saturated with water vapor that the equilibrium fractionation applies exactly.

Figure 1.6: The shell of a departed benthic foram (*Cibicidoides Robertsonianus*. The specimen is about one half millimeter in diameter.

1.7.4 Forams to the rescue

As it happens, Nature has provided a handy way of determining the isotopic composition of past ocean waters, via the good works of single-celled shelly amoeba-like organisms known as *foramanifera*, nicknamed *forams* (see Fig. 1.6). These creatures build distinctive calcium carbonate ($CaCO_3$) shells which record the state of the water in which they grew. Because the shells have such diverse and unmistakable shapes, it is easy to recognize and select out the species which live at the the depth level one wishes to investigate. The two principle types of forams are *benthic* which live near the sea floor, and *planktonic* which require light and live near the ocean surface. The shells of both types wind up in tidy layers in the marine sediments, allowing one to read the state of the ocean in both depth and time, from a single sediment core. Benthic forams appear in the fossil record as early as 525 million years ago, but their use as paleoclimate indicators has been primarily restricted to the period over which significant portions of seafloor have survived – approximately the past 70 million years.

Forams are not just passive recorders of the isotopic composition of ocean waters, however. As is the case for any chemically distinct pair of reservoirs in contact, the oxygen in foram carbonate is systematically fractionated relative to the isotopic composition of sea water. There is some dispute as to the extent to which this fractionation can be thought of as equilibrium fractionation, but the fractionation factors behave much like inorganic equilibrium fractionation factors and do not differ greatly amongst species. Carbonate prefers the heavier forms of oxygen, and at a temperature of $18C$, the ^{18}O to ^{16}O of precipitated carbonate is greater than the ratio for the water in which it precipitates by a factor of about 1.03. In other words, carbonate is about 30‰enriched in ^{18}O compared to the water with which it is in equilibrium. As is typically the case for equilibrium fractionation, the degree of enrichment increases as temperature decreases. The change in fractionation with temperature is usually expressed as a *paleotemperature equation*. Many paleotemperature equations have been given, based on laboratory-cultured organisms, on field observations of recently living forams, and on laboratory measurements of inorganic carbonate

precipitation. All give similar results, though the differences are important if one is interested in high accuracy. A general feel for the numbers is adequately given by the following paleotemperature equation, which applies to the benthic foram *Uvigerina*.

$$T = 17.97 - 4.0 \cdot (\delta_c(VPDB) - \delta_w(VSMOW)) \tag{1.8}$$

where T is the temperature in degrees C at which the foram grew, $\delta_c(VPDB)$ is the measured $\delta^{18}O$ of the foram carbonate reported relative to $VPDB$, and $\delta_w(VSMOW)$ is the $\delta^{18}O$ of the water in which the foram grew, reported relative to $VSMOW$. Both δ values in the above equation are to be expressed in permil units.

The temperature dependence of foram fractionation is a two-edged sword. On the one hand, the temperature dependence allows forams to be used as paleothermometers. On the other hand, the temperature dependence means it is hard to disentangle ice-volume effects from temperature effects. According to Eq. 1.8, a $\delta^{18}O$ variation in foram carbonates of 2‰could represent an temperature change of $8K$ where the forams grew, or instead a change of 2‰in the water in which the forams grew – corresponding to an ice volume change equivalent to roughly 200m of sea level. The use of benthic forams mitigates this ambiguity to some extent, since the deep ocean temperature is much more uniform than surface temperature. This is so because the deep ocean is filled with waters that are created at the coldest parts of the surface ocean, typically located near the poles. When the climate is in a state having ice at one or both poles, this temperature hovers around the freezing point of sea water, whence the benthic oxygen isotopes primarily reflect ice volume rather than temperature – though the temperature effect is still by no means negligible. A $2K$ variation in deep ocean temperature, which is not implausible even in icy conditions, leads to about a 0.5‰variation in the $\delta^{18}O$ of carbonates, which if attributed instead to sea water composition would translate into an ice volume variation equivalent to about $50m$ of sea level. At the other extreme, when the climate is in a state without ice at either pole, the isotopic composition of ocean water itself can be considered nearly fixed, and the benthic foram isotopes provide an indication of the polar temperature, regardless of where the sediment core is actually drilled. In this case, a 2‰increase in benthic foram $\delta^{18}O$ would indicate an $8K$ warming of polar temperature.

Benthic forams thus provide a valuable overall indicator of the state of the climate, giving an indication of polar minimum temperature in ice-free climates, and ice volume in icy climates. Since greater ice volume generally goes with a colder climate, and both high ice volume and low temperature make the $\delta^{18}O$ of foram carbonate more positive, high benthic foram $\delta^{18}O$ indicates a cold climate while low benthic foram $\delta^{18}O$ goes along with a relatively warm climate. In ice-free climates, planktonic forams can be used to estimate surface temperature, but in icy climates the need to subtract out the ice volume effect makes it hard to get accurate surface temperature estimates by this means.

Forams also preserve other chemical signatures that are useful in reconstructing the past state of the climate system. Notably, they provide a record of the $\delta^{13}C$ of inorganic carbon of the ocean. Abiologically precipitated sea floor carbonates, or carbonates not associated with individual microfossils, also do this, but the additional depth information available by using benthic vs. planktonic forams provides valuable information about the state of the oceanic carbon cycle. The use of fractionation as a paleothermometer is not limited to isotopes. Notably, magnesium (Mg) substitutes for calcium (Ca) in foram shell carbonates, to an extent that depends on temperature; hence the Mg to Ca ratio in foram shells can be used as a paleothermometer [6]. As with oxygen isotopes, the fractionation is relative to the composition of sea water, but the Mg to Ca ratio of ocean water is not affected by formation of glaciers, and hence evolves relatively slowly over

[6] Magnesium-calcium paleothermometry can also be used with cores drilled into corals. The growth zone of corals has a distinct depth preference, which has proved particularly useful in estimating ocean surface temperature

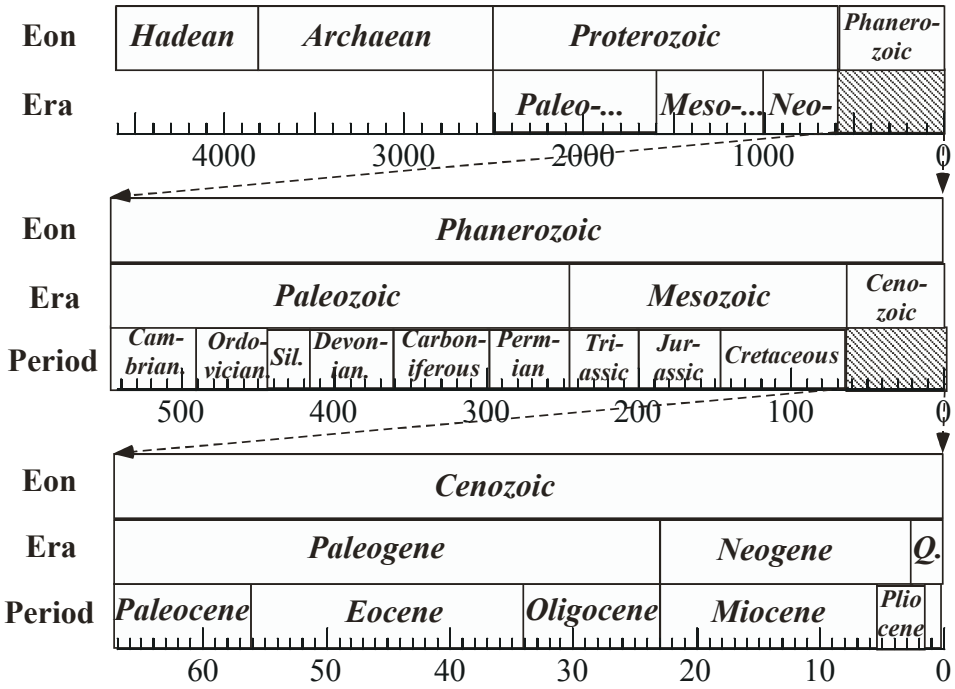

Figure 1.7: The geological time scale. Numbers on the scales represent millions of years before present. The entire span of time before the Phanerozoic is also known as the Precambrian.

geologic time. For this reason, magnesium-calcium paleothermometry has become a crucial tool for probing the past few million years of climate history, and the tool is being extended farther backward as understanding of the longer term evolution of the oceanic Mg to Ca ratio improves.

1.8 The Proterozoic climate revisited:Snowball Earth

With a few more tools at our disposal, we once more pick up the thread of Earth's climate history, starting with a more detailed look at certain aspects of the Proterozoic – the geological eon that extends from 2.5 billion years ago to 543 million years ago, and is subdivided into the Paleo- Meso- and Neo- Proterozoic, in order of decreasing age. From now on, we will have increasing need to refer to the various subdivisions of the geological time scale by name, so these are summarized in Fig. 1.7.

So far we have not said much about continents and continental drift. Continents consist of light material that has floated to the outer portions of the solid Earth and been incorporated into the crust. This material is too buoyant to be subducted into the interior of the planet to any significant degree, so once continental material had completely segregated, it remained at the surface as a kind of scum resting atop the churning cauldron of Earth's interior fluid motions. The continental material is constantly being pushed around, broken up and re-arranged in a process known as *continental drift*. The Earth is the only planet in the Solar System that has continents

in this sense, but it is likely that extrasolar rocky planets of a sufficient size to retain the heat that drives internal motions, and having a surface temperature similar to Earth, would also exhibit the dichotomy of drifting continents vs. areas of rapidly recycled mantle material (analogous to Earth's sea floors). It is at present unresolved whether a water ocean is necessary to maintaining this state of affairs. That is indeed one of the Big Questions of planetary science, but it is not one which we will take up in this book.

Continents are important to climate for three main reasons: They are a platform upon which polar glaciers can form; They are the primary sites of the silicate weathering reaction that governs atmospheric CO_2, and the amount of weathering is strongly affected by the continental configuration; They affect the geometry of ocean basins, and hence the ability of oceans to transport heat from one latitude to another. In addition, they provide distinct habitat for novel forms of life, though they were not colonized to any great degree (if at all) until land plants evolved in the late Ordovician and early Silurian.

There is not enough preserved continental crust to get a clear idea of the distribution of the continents until the very end of the Proterozoic. Geophysical modelling and indirect geochemical evidence has lead to a prevailing belief that the total volume of continental material was similar to the present volume throughout the Proterozoic, but this is not a very well settled area of geophysics. By the close of the Proterozoic, however, the picture of the distribution of continents begins to clarify; we will begin showing maps of paleogeography when we come to discuss that time.

The Proterozoic was the Age of Microbes. Indeed, in terms of the functioning of the biogeochemical cycles needed to sustain life, we are all still basically the guests of the prokaryotes, but up until the very end of the Proterozoic single-celled organisms had the world to themselves, with the more complex eukaryotic form of microbial life making a definite appearance roughly midway through the Proterozoic. The Proterozoic was above all a time of adjustment of biosphere and climate to the massive changes wrought by oxygenation of the atmosphere and ocean. It is a matter of current debate as to whether the oxygenation began in this eon because oxygenic photosynthesis only evolved at this time, or because other factors kept photosynthetic oxygen from accumulating earlier; be that as it may, the oxygenation occurred in fits and starts throughout the Proterozoic, and atmospheric oxygen levels probably reached values comparable to modern ones by the close of the eon. One effect of oxygenation we have already noted is that it is likely to have reduced the importance of CH_4 as a greenhouse gas, with CO_2 (aided by water vapor) becoming increasingly dominant. More speculatively, oxygenation could have changed the abundance of other greenhouse gases that could conceivably have been significant in the anoxic atmosphere, such as N_2O and SO_2. The nature of this greenhouse turnover, and the extent to which it constituted a habitability crisis for our planet, is another of the Big Questions. In Chapter 4 we'll learn how to evaluate the relative importance of the various greenhouse gases. Oxygenation would have also sharply limited the possibility for H_2 to accumulate in the atmosphere, which would have consequences for methanogenic ecosystems that used H_2 and CO_2 as a feedstock.

Oxygenic photosynthesis turns CO_2 into O_2 and organic carbon, represented schematically as CH_2O. In order for the oxygen to accumulate, the organic carbon produced by photosynthesis must be buried before it can be re-oxidized by other bacteria, which would just take the free oxygen back out of the system and turn it back into CO_2. For this reason, the evolution of oxygen is intimately tied in with the carbon cycle. Since organic carbon is isotopically light compared to the carbon in CO_2 outgassed from the interior of the Earth, the long term evolution of $\delta^{13}C$ in carbonates gives us a window into the carbon cycle, and a good overview of what is going on in the Proterozoic. As noted earlier, when the carbon cycle is approximately in a steady state, the $\delta^{13}C$ of carbonate is driven more positive as the proportion of organic carbon burial to total carbon burial increases. One must be cautious about this interpretation, however, since the carbon cycle

is thrown wildly out of equilibrium in the course of the Snowball episodes that constitute the most dramatic features of Proterozoic climate.

With regard to oxygenation, however, organic carbon burial is not the whole story. Various compounds of sulfur also play a key role in the cycling of oxygen through the Earth system; there is evidence that the role of the sulfur cycle is much more prominent in the Proterozoic than during later times. Bacteria are clever, and have ways of oxidizing organic matter that don't use up free O_2. In a process called *sulfate reduction*, certain bacteria can react the sulfate ion (SO_4^{--}) with organic carbon to produce bicarbonate (HCO_3^-) and the stinky rotten-egg gas hydrogen sulfide (H_2S), which reacts with iron oxides and a little bit of free oxygen to produce water and the mineral pyrite (FeS_2). If the pyrite is then buried without being oxidized further, the net process turns organic matter into mineral carbonate while leaving much of the O_2 liberated by oxygenic photosynthesis in the atmosphere/ocean system. Precisely how much is left in the atmosphere and ocean depends on where the oxygen in sulfate and iron oxides comes from, but the net result is that pyrite burial liberates oxygen in a way that doesn't show up in the carbon isotope record.

The participation of sulfur in a variety of reactions involving oxygen makes the stable isotopes of sulfur – ^{32}S, ^{33}S, ^{34}S and ^{36}S, of which the first is dominant – a key source of information about past behavior of oxygen. By comparing the degree of fractionation for the three minor isotopes, one can get additional information. One form of fractionation is *mass-independent fractionation* (MIF), which is nearly the same for all three minor isotopes. This kind of fractionation is believed to be produced only in photochemical reactions involving high energy ultraviolet light, and photochemical models of the atmosphere indicate that mass-independent sulfur fractionation can't be preserved in the sedimentary record unless the atmospheric oxygen concentration is extremely low – 10^{-5} of the present concentration or less, according to current estimates. It is the sulfur MIF proxy that tells us that Archaean oxygen levels are nearly zero; other oxygen proxies only require that Archaean oxygen be below 1% of present atmospheric concentration [7].

The more conventional mass-dependent sulfur fractionation is mediated by sulfate-reducing bacteria. The fractionation nearly disappears when sulfate concentration in ocean water is low, so a strong mass-dependent sulfur fractionation indicates both high sulfate concentration and strong productivity of sulfate-reducing bacteria. An increase in sulfate concentration, in turn, is generally taken as indicative of a rise in atmospheric oxygen, since that permits more oxidation of pyrite into sulfate on land. Once oxygen builds up to the point that at least near-shore bottom waters become oxygenated, a host of additional bacterially-mediated sulfur reactions, called *disproportionation reactions* become possible, and these provide additional means of producing sedimentary sulfides (e.g. pyrite) that are isotopically light in sulfur. The interpretation of mass-dependent sulfur isotope fractionation is an exceedingly complex subject, which is likely to remain in a considerable state of flux for some time to come.

The proxy record shows a great deal of activity towards the beginning of the Proterozoic (during the *Paleoproterozoic*), and also towards the end of the Proterozoic (in the later parts of the *Neoproterozoic*). In between lies a billion-year period that has sometimes been called "the most boring period in Earth history." During this Big Yawn, which stretched from about 1.8 billion years ago to 800 million years ago, carbonate $\delta^{13}C$ held steady near 0‰, indicating steady organic carbon burial. Mass dependent sulfur isotope fractionation suggests oxygen levels of around 10% modern concentrations during this period, though the upper bound on oxygen during this period is not well constrained. There is no indication of any significant glaciation. Eukaryotes appear in the microfossil record towards the beginning of the Big Yawn, but appeared to have little effect on

[7] The atmospheric chemistry models upon which this interpretation of the sulfur MIF is based rest on somewhat shaky assumptions, however.

biogeochemical cycling or climate evolution – unless perhaps they were somehow the cause of the long period of climate stability. It would not be surprising if closer study eventually revealed more features of interest in this period, but at this point we'll turn our attention to the more manifestly dramatic doings at the beginning and end of the Proterozoic.

All lines of evidence point to a Great Oxidation Event at the beginning of the Proterozoic. Preservation of mass-independent sulfur fractionation in sediments ceases abruptly between 2.45 and 2.3 billion years ago, and is never seen again throughout the rest of Earth history. A hiatus in banded iron formations begins at about this time. It is estimated that the oxygen content of the atmosphere soared to values well in excess of 1% of the present level but crashed back to lower levels afterwards, as witnessed by a transient reappearance of banded iron formations; the peak value in the event is at present not well constrained. A subsequent increase in mass-dependent sulfur isotope fractionation preserved in the sediments is indicative of an increase in sulfate concentration in the ocean, most likely associated with an increase of oxidation of pyrite on land. Based on this evidence and disappearance of banded iron formations, it is estimated that oxygen levels in the atmosphere recovered to somewhere around 10% of present atmospheric concentrations around 1.7 billion years ago, and stayed there until 700 million years ago when there was a further oxygenation event.

The Paleoproterozoic is characterized by wild swings in the $\delta^{13}C$ of carbonates. Around the time of the Great Oxygenation event, $\delta^{13}C$ has a major positive excursion, reaching values as high as 10 ‰before eventuallly subsiding to the lower values characteristic of the middle Proterozoic. This indicates a major transition in the carbon cycle, most likely an increase in the proportion of organic carbon burial. It is very suggestive of a take-off of oxygenic photosynthesis, but whether the cause is evolutionary, ecological or a matter of factors that allow better burial of carbon is a matter of dispute. There are several major glaciations within the Paleoproterozoic, of which one – the Makganyene alluded to earlier – was a Snowball event in which ice reached tropical latitudes. The Makganyene Snowball occurred within the interval between 2.32 and 2.22 billion years ago, and fine-scale examination of arguably synchronous glacial deposits in the Duitschland formation (Transvaal, S. Africa) indicates extreme carbon isotope excursions associated with major Paleoproterozoic glaciations: Carbonate $\delta^{13}C$ was around 5 ‰before the glaciation, then dropped to zero or even negative values as the glaciation progressed, recovering slowly afterwards. We'll be able to probe similar features in more detail in connection with the Neoproterozoic Snowballs. The Paleoproterozoic presents us with a puzzle whose pieces include oxygen, the effect of oxygen on greenhouse gases, the carbon cycle, and glaciation. Figuring out how these pieces fit together is one of the Big Questions.

The Neoproterozoic has many features in common with the Paleoproterozoic. The extreme carbonate carbon isotope excursions which had been dormant for so long resume in the Neoproterozoic. There are several major glaciations during the Neoproterozoic, and two of these were Snowball events in which ice reached tropical latitudes. The more recent of the two Snowballs is the Marinoan event, which occurred about 640 million years ago; the older is the Sturtian, centered on 710 million years ago. Neoproterozoic Snowball-related geological formations exhibit a distinctive sequence of events. The scene starts with high carbonate $\delta^{13}C$, up to 5‰, which is in fact higher than the modern value and indicative of a greater proportion of organic carbon burial than is the case at present. Then, the $\delta^{13}C$ drops, falling to zero or even negative values. At some point in this decline, one sees diamictites and other glacial deposits. The $\delta^{13}C$ continues to drop, and above the glacial deposits one finds *cap carbonates* – very unusual carbonate features that are believed to require very high deposition rates from waters highly supersaturated in carbonate. In the carbonates overlying the glacial deposit, the $\delta^{13}C$ becomes negative, typically around -5‰which is about the value for abiotic carbon outgassing from the Earth's interior. The $\delta^{13}C$ gradually

recovers to positive values over a long (but somewhat unconstrained) period of time afterwards. A particularly clear depiction of this sequence of events is given in the review by Hoffman and Schrag listed in the Further Readings for this chapter.

However, not all major carbon isotope excursions are associated with Snowball events. In fact, the greatest carbon isotope excursion in Earth history – the *Shuram excursion* – sets in gradually after a conventional glaciation which is thought to reach only to midlatitudes (the *Gaskiers* glaciation). The Shuram excursion brings the carbonate $\delta^{13}C$ all the way down to -12 ‰. There is no known process that could bring the carbonate $\delta^{13}C$ so far below the mantle outgassing value if the carbon cycle is in equilibrium. Indeed, the $\delta^{13}C$ is so implausibly low in the Shuram that it was long thought to be an artifact of diagenetic alteration. The Shuram is an enigmatic event – indeed one of the Big Questions. Current thinking has it that the Shuram is associated with a transient reorganization of the carbon cycle, in which a large isotopically light pool of suspended organic carbon in the ocean is oxidized and deposited as carbonate.

In fact, a lot is going on with oxygen across the Neoproterozoic, though it is a bit hard to determine what, where and when. What is certain is that oxygen must have been high – even near present levels – right down to the ocean bottom by the end of the Neoproterozoic, since bottom-dwelling animals appear in the fossil record by this time, and it is unquestionable that such creatures require a great deal of oxygen. At the other side of the Neoproterozoic, around 700 million years ago, there is further evidence of oxygenation, in that a sharp rise in mass dependent fractionation in sedimentary sulfides indicates the expression of sulfur disproportionation reactions, which indicates an oxygenation of at least some of the bottom waters. Another important clue as to what is going on is the reappearance of banded iron formations in connection with the Neoproterozoic Snowball events, suggesting that the ocean once more went anoxic, most plausibly as a result of global ice cover shutting down photosynthesis.

A very Big Question is why all this excitement suddenly resumed after nearly a billion years of stasis.

The Snowball events of the early and late Proterozoic are some of the most dramatic events of Earth history. We have used the term "Snowball" to refer to any glaciation where there is evidence of glaciation at tropical latitudes, but it is a matter of considerable debate whether the oceans were indeed nearly completely frozen over all the way to the equator during these events. Sometimes the term *Hard Snowball* is used to refer specifically to a state with near-total ice cover. A Big Question of climate physics is whether it is possible to cool down the planet enough to yield land-based ice sheets discharging into the tropics, without also freezing over the tropical ocean completely. This requres an understanding of ice-albedo feedback, which will be developed at several places throughout the book. However, it also involves ocean heat transports (which are good at melting ice) and glacier dynamics, which for the most part are subjects that will be left for another time and place.

The Snowball phenomenon is pregnant with Big Questions, the most obvious of which are: How do you get in? And how do you get out? And if your planet does succumb to a global Snowball, how long does it take to get out again? Is it a matter of centuries, millions of years or billions of years? On Earth, the upper limit set by the geological record for the duration of Neoproterozoic snowballs is about 20 million years, and the duration could well have been shorter. However, without a clear understanding of the nature of the event, it is hard to determine whether we just got lucky or whether the event could have lasted much longer.

Most theories for the entry into a Snowball involve the drawdown of whatever greenhouse gas had previously been maintaining the planet's warmth – usually CO_2 and CH_4 in some combination. Various hypotheses include methane destruction by oxygen, weathering enhancement

triggered by catastrophic methane release from sediments, weathering enhancement due to continental configuration or production of weatherable rock by massive volcanic eruptions, and (more speculatively) drawdown of CO_2 through runaway photosynthesis and oxygenation. Whatever the mechanism, a key requirement is that the mechanism be compatible with the observed reduction in $\delta^{13}C$ *before* the onset of glaciation. Not all of the relevant biogeochemistry will be treated in this book, but in order to evaluate the hypotheses, it is certainly necessary that one have the tools to assess how low any given greenhouse gas has to go in order to trigger a global glaciation. These tools will be provided in subsequent chapters.

Assuming for the moment that the cooling process caused a Hard Snowball, the next question is how to deglaciate the planet. We'll see in Chapter 3 that one would have to wait a billion years or more to exit from a Snowball if the exit were due to increase in solar output alone. Based on rather simple reasoning of the sort that will be covered in the remainder of this book, Kirschvink proposed that once the Earth freezes over, the weathering of silicate to carbonate (which requires liquid water washing over weatherable rocks) ceases, so that CO_2 outgassed from the Earth's interior accumulates in the atmosphere until it reaches concentrations sufficient to cause a deglaciation. This is another illustration of the principle that *Big Ideas come from Simple Models.* A Big Question (treated in subsequent chapters) is: how high does CO_2 have to go in order to trigger deglaciation of a globally ice-covered planet? Much hangs on the answer.

Both cap carbonates and the persistent negative carbon isotope excursion following the Snowball events are consistent with a massive buildup of CO_2 in the atmosphere during the frozen-over period. Once the planet gets warm enough to deglaciate, the powerful precipitation in the ensuing hothouse world would wash great quantities of land carbonates into the ocean, where they would precipitate to form cap carbonates. Further, if photosynthesis nearly shuts off during the glaciated phase, the inorganic carbon that accumulates in the ocean-atmosphere reservoir would have $\delta^{13}C$ comparable to the mantle outgassing value of about -6‰. As this reservoir is gradually transformed by silicate weathering into carbonate sediments, the isotopically light carbon works its way into the carbonates. If the reservoir is big enough at the termination of the snowball, it can keep the sedimentary carbonate $\delta^{13}C$ light even in the face of a resumption of photosynthesis. When interpreting the carbon isotope excursions in the course of a Snowball, it is essential to keep in mind that the carbon cycle is likely to be far out of equilibrium in the course of these events. A full understanding of the connection between the carbon isotope evolution and the sequence of events surrounding the Snowball requires a detailed accounting of flows of carbon between the many different carbon reservoirs in the Earth system – land carbonate rocks, marine carbonate sediments, atmospheric carbon dioxide, various species of dissolved inorganic carbon, and organic carbon.

Assuming that the the exit from a Snowballl state does indeed proceed from accumulation of a great deal of CO_2 in the atmosphere, several Big Questions arise in connection with the post-glacial climate. Is there a risk of triggering a runaway greenhouse? If not, how hot does the climate get in the tropics and polar regions? Would it be hot enough to sterilize the planet to any great degree? How long is the post-snowball recovery? In other words, how long does it take for weathering processes to draw the CO_2 back down to more normal levels?

Other Big Questions include: Why were there no Snowball event during the Big Yawn period of the middlle Proterozoic? Why did Snowballl events cease at the beginning of the Phanerozoic? Could they happen again, or is this particular threat behind us?

But let us not forget that another Big Question is whether there is a climate state with significant amounts of open water in the Tropics, which is nevertheless consistent with the full range of geological data accompanying the Marinoan and Sturtian events. One could pose the

same question for the Makganyene event, but there is less data to constrain the answer. The early and late Proterozoic manifests a lot of climate "weirdness" that is not seen elsewhere in Earth history, and such striking signatures would seem to call for an equally dramatic cause, rather than just a minor variation on the theme of ice ages. The global Snowball seems to fit the bill, but it remains to be seen whether other explanations are possible. The climate physics developed throughout this book will give the reader the underpinning needed to assess hypotheses as they develop, and even perhaps to formulate new ones.

Regardless of whether a true global Snowball glaciation ever happened on Earth, the Snowball certainly represents a state a water-covered planet could fall into if the right stellar and atmospheric conditions are encountered. Once a planet falls into a Snowball state,it is likely to stay there for a long time, and the consequences for existing life and evolution of new forms of life are profound. As such, Snowball states are a potential habitability crisis that extrasolar planets need to avoid or surmount. It is therefore worth understanding, in general terms, the physics of entry into, exit from, and duration of Snowball states, as well as the nature of the climate at various stages of the sequence and the effect of the sequence on life. This constitutes another Big Question, about which will have much to say.

1.9 The hothouse/icehouse dichotomy

The present climate has ice at both poles, and the ice volume has fluctuated episodically between the present amount and a considerably larger extent for the past two million years (about which more anon). However, the present icy climate is not at all typical of Earth history. A careful study of the climate evolution over the past seventy million years illustrates a transition between climate states archetypical of a theme that has been played over with variations during the 543 million years since the close of the Proterozoic. With this latest eon, known as the *Phanerozoic*, we complete the repertoire of the major divisions of geologic time, as summarized in Fig. 1.7. Though the very first preserved multicellular organisms appear at the close of the Proterozoic, the Phanerozoic is the eon in which multicellular organisms of a generally modern form become abundant and diversify, first in the ocean with colonization of land coming towards the middle of the eon. Though the Phanerozoic was not subjected to extreme variations of climate and atmospheric composition rivaling the Snowball or oxygenation transitions of earlier eons, the events of the Phanerozoic are by no means inconsequential.

1.9.1 The past 70 million years

Figure 1.8 shows the paleogeography at the end of the Cretaceous, 65 million years ago. The continent of Antarctica has approached the South Pole, and will continue to drift over the next 40 million years or so until it is more nearly centered on the pole. There is open water at the North Pole in the late Cretaceous, and the open Arctic Ocean continues throughout the subsequent time through the present. The modern continents of North and South America, Eurasia, and Africa are still early in their separation, leaving a narrow Atlantic ocean and a very broad Pacific. The continents will continue to drift apart as they approach their modern configuration; the Atlantic widens steadily throughout the span of time under discussion.

The record of benthic foram $\delta^{18}O$ in Fig. 1.9 provides a good overview of the climate evolution for the past 70 million years. Towards the beginning of the period, $\delta^{18}O$ is considerably lower than it is in the modern ocean. It reaches a minimum value of -0.1‰around 51 million

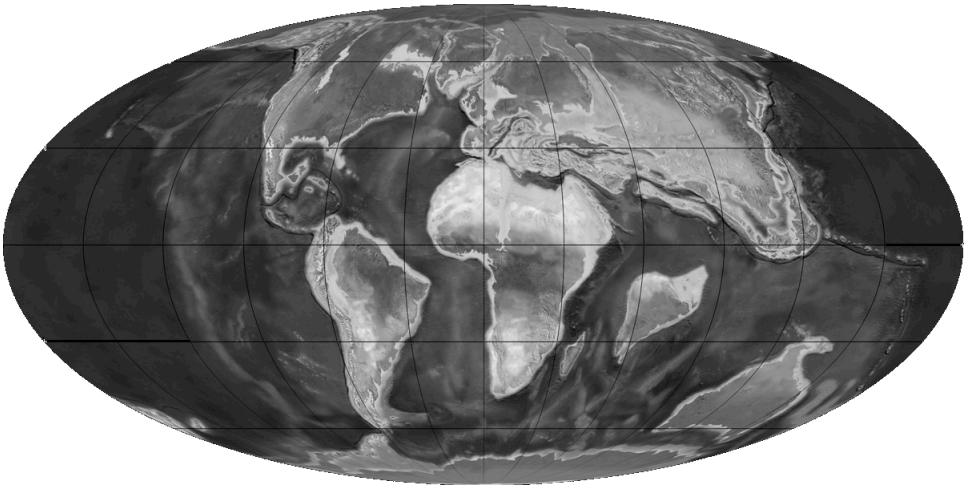

Figure 1.8: Position of the continents at the end of the Cretaceous, 65 million years ago. Continents are shown on a Mollweide map projection, with the North Pole at the top and the South Pole at the bottom. Light areas are continents while dark grey areas are ocean.

years ago. Independent geological evidence shows that there was no significant amount of ice sequestered in land glaciers until 36 million years ago, so the variations in $\delta^{18}O$ before this time can be interpreted as polar ocean temperature changes. Melting the present ice in Antarctica and Greenland would leave the ocean with a $\delta^{18}O$ of around -.7‰relative to $VSMOW$. Plugging this into the paleotemperature equation, we estimate a high-latitude ocean temperature of $15.5C$. This estimate applies to the coldest seasonal temperature attained either in the Arctic ocean or in the waters surrounding Antarctica. This is well above the freezing point of sea water, and so we conclude that the oceans were ice-free year round, even in the Arctic and Antarctic regions which are very cold in the modern climate. We'll refer to this kind of climate state as a *hothouse climate*. The late Cretaceous polar temperatures were about $4C$ cooler than those at the peak warmth, but still warm enough to guarantee ice-free conditions.

Other lines of evidence also support warm Northern high-latitude conditions and above-freezing winter conditions. In the early twenty-first century, the first useful deep-time Arctic marine cores were recovered, and TeX-86 proxies applied to these cores indicated Arctic ocean up to $22C$ during the time of the spike marked *PETM* in Fig. 1.9, with temperatures in the range of 17-18C before and after. Fossil vegetation from Arctic land also supports temperatures in this range. Moreover, the abundant evidence that lemurs and crocodiles were able to survive in high Northern latitudes points toward mild winters, since these creatures cannot survive sub-freezing temperatures for any significant length of time.

While evidence for warm, ice-free polar conditions in the Eocene and late Cretaceous is unambigous, the nature of the tropical climate is somewhat problematic. Up until the year 2001, most paleoceanographers would have said, based on planktonic foraminiferal ^{18}O that the Cretaceous tropical sea surface was no more than two or three degrees warmer than present, and the Eocene tropical sea surface was no warmer than today, and might even have been cooler. This posed the paradox of the "low gradient" climate – how to warm up the planet enough to prevent polar ice, without frying the tropics. In 2001, it was discovered that most of the evidence for a cool tropics was spurious, having been affected by diagenetic alteration of sediments. The surviving non-altered data indicated a warmer tropics, but there was precious little data left after the

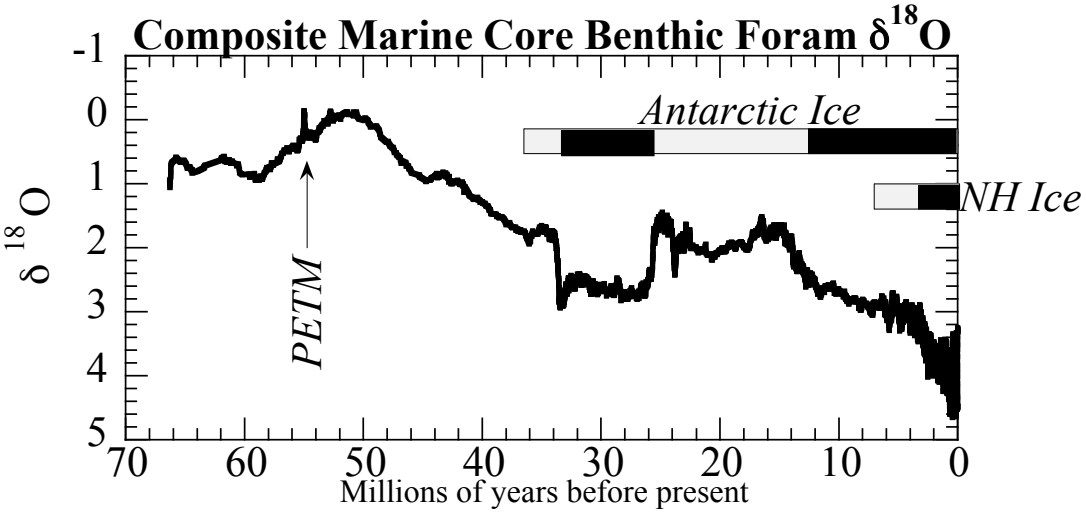

Figure 1.9: A composite record of benthic foram $\delta^{18}O$ (vs. PDB) over the past 70 million years, based on the average of several marine sediment cores. Data is from Zachos *et al.* 2001.

diagenetically contaminated data was discarded. Gradually, new sediments and new proxies have come to the fore, and the story continues to develop. Tex-86 proxies and Mg/Ca proxies now indicate tropical temperatures of up to $34C$ or even $37C$ in places, in the Eocene warm interval. Tropical surface temperatures were two or three degrees cooler in the late Cretaceous. This still gives a considerable reduction in the pole-equator temperature gradient as compared to the modern climate, but the problem is not as severe as it was. It is quite certain that the tropical sea surface temperatures could not have been as high as $40C$, since the planktonic forams that are seen in the sedimentary record could not have survived at such high temperatures. A striking feature of the data is that tropical surface temperatures seem to remain fairly constant throughout the Eocene, though polar temperatures (as indicated by the benthic forams) decrease towards the Oligocene.

Returning to Fig. 1.9 we see that following the peak Eocene warmth of 51 million years ago, the climate commenced a long slide toward the icehouse climate characterizing the latter part of the record. Between the peak warmth and the beginning of the Oligocene 34 million years ago, the minimum polar temperature dropped by $8C$ as indicated by benthic $\delta^{18}O$. At this point, small ephemeral ice sheets began to form on Antarctica, culminating in a more substantial glaciation of Antarctica that lasted until 26 million years ago, somewhat before the beginning of the Miocene. The Oligocene glaciation is visible as a pronounced ditch in the $\delta^{18}O$. This first attempt at glaciating Antarctica didn't last however,since the climate recovered and returned to a period of generally cold Antarctic conditions with sea ice but with land ice sheets having volume below 50% of the present volume. This situation lasted until 15 million years ago, when the slide towards icehouse conditions resumed. Antarctic ice sheets grew again, but the Northern hemisphere land glaciation had not yet been initiated by this time. The first abundant evidence of sea ice, based on ice-rafted debris in polar sediment cores, appears at 14 million years ago. The increase in $\delta^{18}O$ in the next several million years is due to a combination of continued cooling and Antarctic ice sheet growth, culminating in the initiation of the major Northern Hemisphere ice sheets around 6.5 million years ago, as we enter the Pliocene period. The oxygen isotopes begin to show substantial fluctuations at this time, which grow in amplitude as one enters the Pleistocene. These fluctuations are due to waxing and waning of ice sheets, predominantly in the Northern Hemisphere – in other

words, the coming and going of "ice ages." The nature of the fluctuations will be examined more closely in Section 1.10.

What accounts for the nature of the hothouse climate state, and for the subsequent hothouse/icehouse transition? This is another of the Big Questions. There is no support in astrophysics for solar variability of a magnitude and type that could explain the transition, so attention has settled primarily on long term fluctuations in the greenhouse gas content of the atmosphere. In an oxygenated atmosphere like that of the Phanerozoic, CO_2 is the only known long-lived greenhouse gas that can build up to concentrations sufficient to cause climate changes of a magnitude comparable to those seen over the Phanerozoic; to get fluctuations of the requisite magnitude requires amplification of the direct CO_2 effect by water vapor feedbacks, and cloud feedbacks can also substantially modify the response. Another thing that makes CO_2 a good suspect for the role of primary agent in Phanerozoic climate evolution is the fact that it is a central participant in all aspects of the inorganic and organic carbon cycle, which offers many possible mechanisms whereby CO_2 could evolve over the long term. There are a great many unresolved issues regarding the CO_2 theory of Phanerozoic climate evolution, but a central part of testing the theory is to understand the way CO_2 and water vapor act in concert to determine the temperature of a planet; the necessary estimates will be given in Chapter 4. One needs to understand how much CO_2 it would take to account for Eocene warmth, before one can decide whether there are plausible geochemical mechanisms that could lead to the required concentration.

The greatest impediment to testing the CO_2 theory of the hothouse/icehouse dichotomy is the difficulty of estimating past CO_2 levels. There are various geochemical and fossil proxies that can be brought to bear on the problem. For example, algae preferentially take up ^{13}C at a rate that depends somewhat on the CO_2 concentration in the water in which they grow. Carbon isotopes in fossil soil carbonate also preserves information about past CO_2, as does the density of pores (*stomata*) in fossil leaves. All estimates known to date are subject to considerable uncertainty. Nonetheless there is support for the idea that CO_2 concentrations around 70 million years ago could have been 6-10 times modern pre-industrial values. The evidence points to a general decline of CO_2 since that time,but there are also some indications that CO_2 may have already attained quite low levels during some periods well before the Pliocene icehouse climate set in. This is a rapidly evolving subject, however, and nothing definitive can be said at present. What is certain is that there are known geochemical mechanisms associated with the Urey reaction and silicate weathering, which have the potential for causing changes in atmospheric CO_2 of the required magnitude, and on a time scale consistent with the observations. These mechanisms will be discussed in Chapter 8. As with any climate problem, on Earth or elsewhere, uncertainties regarding cloud feedbacks complicate the problem of testing theories of climate response. It is not out of the question that part of the answer lies in modulation of cloud albedo by, say, biologically produced sulfur compounds that seed cloud formation. Further, if data should ultimately support the low-gradient picture of the Cretaceous and Eocene hothouse climates, some mechanism will be needed to keep the tropics from overheating while the poles are warmed by elevation of CO_2. This, too, may involve clouds, or it may involve changes in ocean circulation. It has even been suggested, with considerable physical support, that increases in hurricane intensity in a warmer world could provoke ocean circulation changes of the sort required to provoke a low-gradient climate.

Figure 1.9 exhibits a dramatic climate event of considerable importance. The spike in $\delta^{18}O$ at the Paleocene/Eocene boundary (marked "PETM" in the figure, for *Paleocene/Eocene Thermal Maximum*) is not a glitch in the data. It represents a real, abrupt and massive transient warm event. The spike looks small in comparison to the range of isotopic variation over the past 70 million years, but in fact it represents the planet accomplishing two million years worth of warming in a warm spike that (on closer examination) sets in within under 10,000 years and has a duration of

around 200,000 years. This isotopic excursion corresponds to a global warming of about 4C; other proxy records show that the warming had similar magnitude in the Arctic and at the Equator, and that it extended to the deep ocean. This climate event triggered a mass extinction of benthic species, probably due to a combination of warming, oxygen depletion, and ocean acidification. An important clue as to the cause of the warming is that the record of $\delta^{13}C$ from the same core (not shown) exhibits a major negative excursion at the same time, going from values of about $+1.2‰$down to about zero at the bottom of the excursion. This indicates a catastrophic release of large quantities of isotopically light carbon into the climate system, which presumably increased the atmospheric greenhouse effect and led to warming. One possibility is that the release came in the form of methane from destabilized clathrate ices in the ocean sediments; another is that the isotopically light carbon came from oxidation of suddenly exposed organic carbon pools on land, releasing large quantities of CO_2. Based on analysis of the carbon isotope record, it has been estimated that 4000-6000 gigatonnes of carbon were released into the ocean-atmosphere system, if the release were from organic matter. This compares with 700 gigatonnes of carbon in the form of CO_2 in the modern atmosphere. This would considerably enhance the atmospheric greenhouse effect, though the effect would largely wear off after a thousand years or so, over which time about 80% of the released carbon works its way into the ocean. It is far from clear that one can account for the observed magnitude and duration of *PETM* warming with the amount of carbon one has at ones' disposal. This is one of the Big Questions. We shall not answer it in this book, but the reader will be provided with the tools needed to assess the warming caused by various amounts of CO_2 or methane, and also (in Chapter 8) a bit of insight about the partitioning of carbon between atmosphere and ocean. These are tools one must have at hand in order to evaluate any theory of the *PETM*.

The Cretaceous is closed by the impact of a large asteroid or comet (known generically as a *bolide*). This is known as the *KT impact event* (for "Cretaceous/Tertiary," Tertiary being an obsolete term for the period following the Cretaceous). This event has little or no expression in the isotope record shown in Fig. 1.9, and is instead identified by global presence of a layer of the element iridium. The impact crater has also been identified, which allows an estimate of the energy of the impactor. The KT impactor had effects of extreme consequence despite the lack of an expression in the oxygen isotope record. Notably, it was the dinosaur killer – though many other species went extinct at the same time. Examination of the carbon isotope record shows also that the ocean carbon cycle remained highly perturbed for millions of years. There are many Big Questions associated with the consequences of a bolide impact. What is the mechanism by which the impact causes extinction? Is it direct blast and heat, or some longer-lasting change in the climate? It has been estimated that the impactor released about $5 \cdot 10^{23}$ Joules of energy. How does this energy compare to other energy sources in the climate system, and what effects should it have on the atmosphere and ocean? What are the broader effects of a bolide impact on climate, and how long do they last? Does the impact cause a warming (perhaps through release of greenhouse gases) or a cooling (through lofting of a dust and soot cloud)? Some of the climate questions will be taken up in Chapter 4.

The *KT* impact event is the archetype of impact events, which have been episodically important throughout Earth history. Similarly the Earth has experienced many other mass extinctions besides that at the *KT* boundary, not all of which are clearly associated with a bolide impact. All mass extinctions lead to Big Questions.

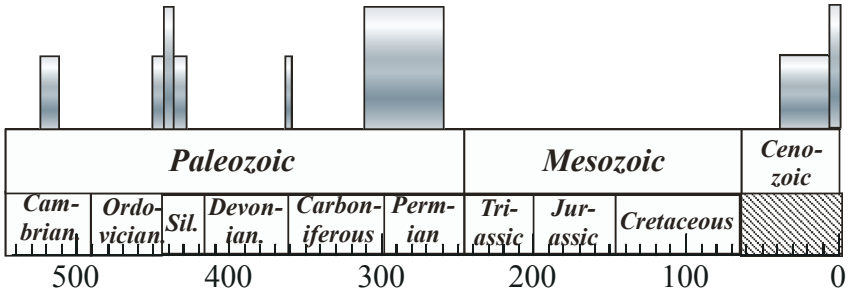

Figure 1.10: Known occurrences of land glaciation during the Phanerozoic. Tall shaded bars indicate periods of major glaciation in which ice reaches to latitudes of 50N or 50S, while short shaded bars indicate periods of more minor glaciation.

1.9.2 Hothouse and icehouse climates over the Phanerozoic

The Cretaceous hothouse climate and the Pleistocene icehouse climate represent opposite extremes of the Earth's typical climate state of the past half billion years. Going back further in time, the Snowball Earth represents an ultra-extreme on the cold end; going further afield in space, the runaway greenhouse represents an ultra-extreme on the hot end, though one that evidently never occurred on Earth. The Earth has experienced many individual hothouse and icehouse episodes in the past half billion years. For times earlier than 70 million years ago one cannot rely on $\delta^{18}O$ records from sea-floor sediments to provide a cross-check on ice volume. Instead, one must look for evidence of glaciation or temperate polar climate preserved as glacial features or fossil animal and plant occurrences on land and in near-coastal environments. This record is far less complete and well-preserved. An accurate long term sea level record would make it possible to estimate ice volume in the distant past, but recovering and interpreting sea level has proved difficult. It is certainly possible to detect the existence of major glaciations, but good estimates of ice volume are not available for pre-Mesozoic glaciations. The periods between major glaciations – tentatively labeled as hothouse climates below – could well have undetected episodes of polar ice embedded within them. The known episodes of major and minor land glaciation are summarized in Fig. 1.10.

The Cretaceous hothouse conditions extended back throughout the entire Mesozoic, and well into the late Permian. There is evidence that some of these periods, notably the early Jurassic, may have been even warmer than the Eocene. To find another era of glaciation to rival that of the Pliocene and Pleistocene icehouse, one needs to go back to late Carboniferous and early Permian. The sixty million year period centered on the Permo-Carboniferous boundary 300 million years ago was a time of very extensive glaciation, reaching to lower latitudes even than the Pleistocene ice ages, though not attaining Snowball proportions. This period is a crucial one for the CO_2 theory of Phanerozoic climate, since it is a time when there is quite strong evidence for low CO_2. The earlier Phanerozoic exhibits comparatively minor glaciations at the end of the Devonian, in the mid-Silurian and for a brief period in the mid-Cambrian, but so far as the Phanerozoic goes, the

Permo-Carboniferous glaciation is the big one to explain.

The changing paleogeography, shown in Fig. 1.11, is likely to have influenced climate. In particular, it is bound to be easier for a glacier to accumulate on land if there is land at or near one or both of the poles. Certainly there is land at the South Pole during the Carboniferous glaciation and the current glaciation that began in the mid-Cenozoic. However, this is clearly not the whole story, as there was plenty of land at the South Pole already 400 million years ago, but the Carboniferious glaciation didn't set in until nearly a hundred million years later. Likewise, Antarctica was already near the South Pole during the Cretaceous, but Antarctic glaciation only took off in the mid-Cenozoic. Most likely, fluctuations in CO_2 – probably itself affected by continental configuration – play a crucial role in the timing of glaciations.

A general theme in the evolution of paleogeography is the assembly and breakup of *supercontinents*. From 500 million years ago to 400 million years ago one can see the Southern supercontinent Gondwana near the South Pole, though there are a few leftover bits of land that are not part of Gondwana. By 300 million years ago, Gondwana has merged with these bits to form the global supercontinent Pangea, which then breaks up into the present continents over the course of the rest of the Phanerozoic. The interiors of supercontinents are isolated from the moderating effects of the oceans on climate, and so could be expected to experience harsh seasonal swings in temperature. Do we expect supercontinent interiors to be deserts or steaming, moist fern forests? This is another of the Big Questions.

Is there a clear dominance of hothouse or icehouse conditions over the past half billion years? The record of the past hundred million years certainly supports the notion that the largely ice-free hothouse is the preferred state of the Earth's climate, but going further back in time it is harder to say whether the apparent dominance of hothouse conditions is an artifact of poor preservation of the polar deposits where glaciers are most likely to have occurred. Some of the episodes we think of as hothouse climates could well have had significant amounts of ice.

In any event, the delineation of the circumstances which favor icehouse or hothouse climates, and the factors governing the transition between the two, constitutes one of the Big Questions of climate science. It seems likely that if the hothouse/icehouse transition of the past 70 million years can be understood, similar mechanisms could be applied to the rest of the Phanerozoic. Variations on the theme would include a greater range of different continental configurations – notably the breakup of supercontinents – as well as biological innovations such as the colonization of land and the evolution of deepwater carbonate shell-forming micro-organisms, both of which can affect the global carbon cycle.

During the Phanerozoic, life on Earth went through a number of mass extinctions rivaling or exceeding the end-Cretaceous event. The biggest mass extinction of all occurred at the end of the Permian, wiping out 96% of all marine species, 70% of all land vertebrates, and a large fraction of all land plants and invertebrates. It is the only mass extinction that included insects to any great extent. There is no clear evidence for a bolide impact at this time, though it remains possible that an impact occurred but failed to leave a trace in the fossil record. In any event, the cause of the end-Permian mass extinction ranks as one of the Big Questions.

1.10 Pleistocene Glacial-Interglacial cycles

Now we'll take a closer look at what's been going on in the past five million years. The earlier portion of this time is known as the *Pliocene* epoch, and the latter portion, beginning around 1.8 million years ago, is the *Pleistocene*. The choice of the Pliocene-Pleistocene boundary is based on

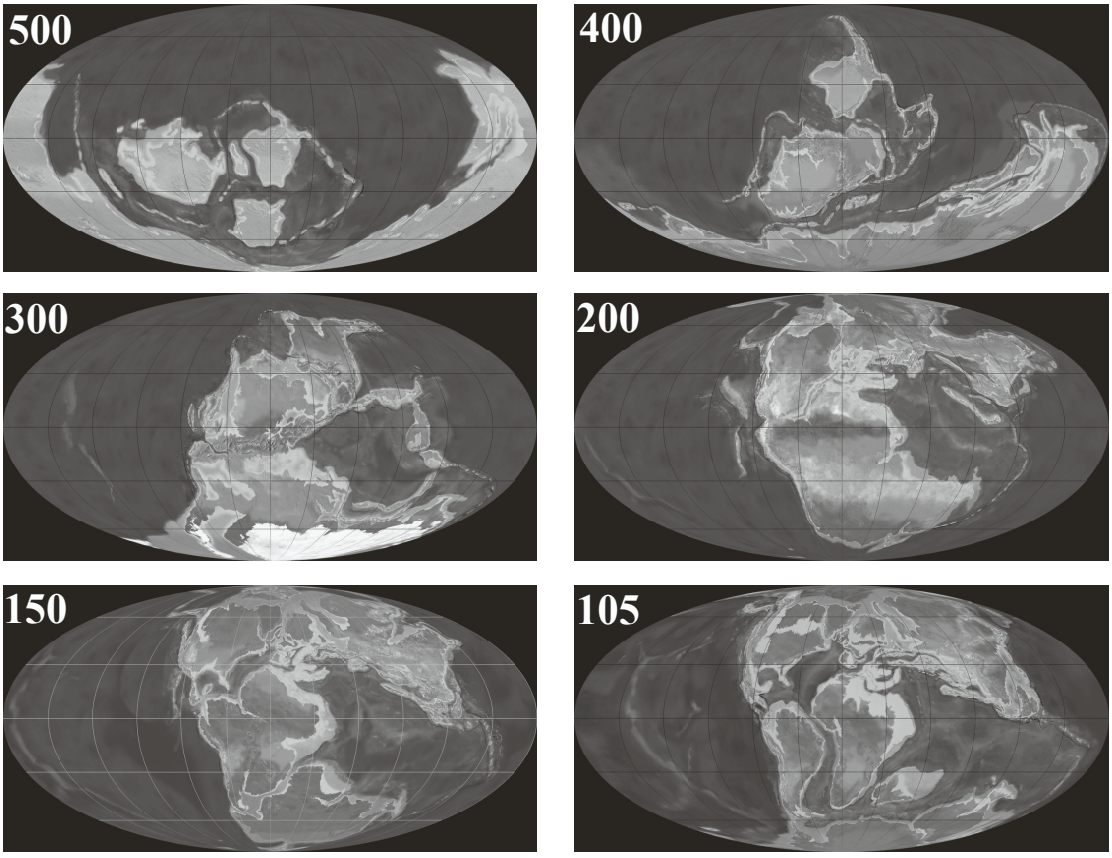

Figure 1.11: Evolution of geography over the Phanerozoic. Map projection and shading as for Fig. 1.8. The numbers give the time in millions of years ago.

an obsolete notion of the time when Northern Hemisphere ice sheets attained nearly their modern extent. At the level of geological periods, the Neogene-Quaternary more closely approximates this distinction. The Quaternary period extends all the way to the present, but the Pleistocene is terminated somewhat arbitrarily at the end of the last major ice sheet retreat around 10,000 years ago. The remainder of time up to the present is the *Holocene* epoch. This is a very human-centric division, since the time we are experiencing now is a fairly ordinary Pleistocene-type interglacial period – or at least it was up until the commencement of the industrial revolution. A more rational division would extend the Pleistocene up to about the year 1700, whereafter we would enter something with a name like "Anthropocene" (for reasons to be discussed in Section 1.12).

The Pliocene and Pleistocene are a time of establishment of the great Northern Hemisphere ice sheets. During this time, the ice sheets settled into a rhythm of expansion and retreat – the rhythm of the coming and going of *ice ages*. The notion of an "ice age" is distinct from that of the "icehouse climate state" introduced earlier. The latter term refers to a span of time (usually several million years) during which there is permanent ice at one or both poles. Within the time embraced by an icehouse climate state, ice volume is not constant, but fluctuates episodically with variations in ice volume that can be on the order of a factor of two (judging from the Pleistocene). Individual episodes of large ice volume within an icehouse climate are referred to as *ice ages*, with the warmer periods in between referred to as *interglacials*, though the ice doesn't come close to disappearing completely. In the Pleistocene, the fluctuation in ice volume is dominated by changes in Northern Hemisphere ice sheets, but as ice ages come and go, the entire globe becomes colder and warmer.

1.10.1 The Marine Sediment Record

Figure 1.12 shows the $\delta^{18}O$ of benthic forams in a tropical Pacific core over the past 4 million years. The short-period fluctuation represents fluctuations in both ice volume and benthic temperature, and in addition there is a downward trend in temperature and increasing trend in ice volume over the earlier two million years of the period. By 2 million years ago, the flucutations have settled into a fairly regular pattern, with a dominant period of about 40,000 years. The period may be crudely estimated by counting peaks in the isotope record. This says that major ice advances in the early Pleistocene occur roughly every 40,000 years. About 800,000 years ago, there is a major transition in which the amplitude of the glacial-interglacial cycle becomes markedly larger, and the periodicity lengthens to about 100,000 years. During this period, an asymmetry between glaciation and deglaciation becomes readily apparent: the climate cools and ice builds up over long periods of time, but deglaciation occurs rather precipitously.

The periodicity of the ice ages, and the reason for the transition to a dominant 100,000 year cycle later in the Pleistocene is another of the Big Questions of climate science. We will learn in Chapter 7 that the periodicities are almost certainly connected with the quasiperiodic variations of the Earth's orbital characteristics – namely the tilt of its rotation axis and the departure of the orbit from circularity. The cycles are known as *Milankovic cycles,* after the scientist who first formulated a detailed theory connecting orbital parameters with the coming and going of ice ages. Milankovic's theory was largely ignored for decades, because not enough was known about the pattern of ice ages to give the theory a fair test. It was only revived in the 1970's, when data of the sort given in Figure 1.12 first became available. Even today, the means by which cycles in orbital characteristics are expressed in the climate record are far from clear, and remain a subject of active research.

This is another case in which the lesson learned from Earth carries over to other planets,

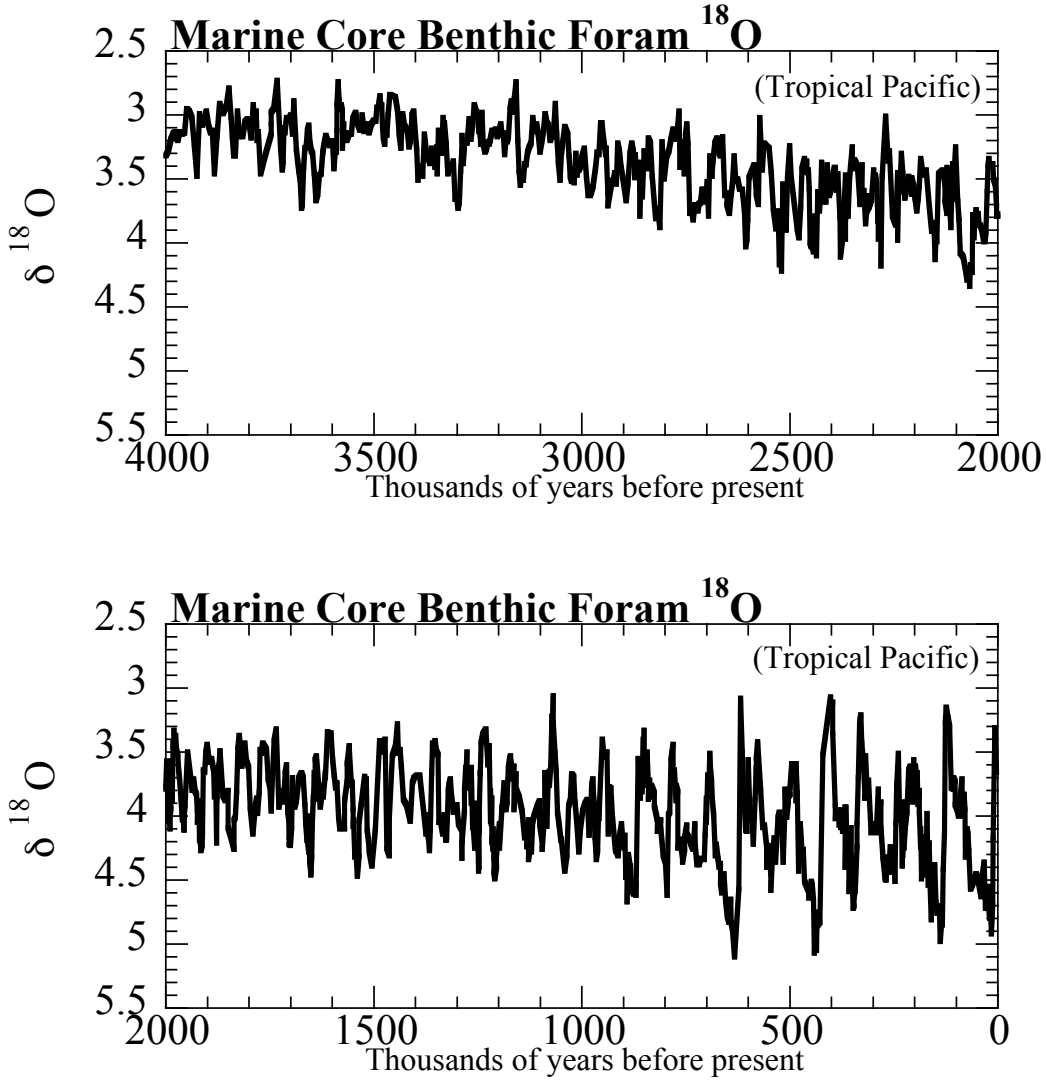

Figure 1.12: Benthic foram oxygen isotopes from the Ocean Drilling Program Core ODP 849 (see Mix *et al* in the Further Readings for this chapter). Values are reported relative to the *PDB* standard. The core is located in the tropical Pacific, but the benthic data is representative of the global climate state. Note that the vertical axis has been reversed, so that upward excursions represent warmer and less icy times.

as the orbital parameters of Mars also show Milankovic cycles. Compared to Earth, however, the study of how these cycles are expressed in the paleoclimate record of Mars is in its very infancy.

1.10.2 Ice core records

The ice at the base of the Antarctic ice sheet is about a million years old, so one can also retrieve a record of past climate by drilling cores into the ice. Many aspects of climate are recorded in the ice, but the ones that will concern us here are stable isotopes of water (δD and $\delta^{18}O$), recorded in the ice itself, and composition of past air preserved in bubbles within the ice. The stable isotopes are essentially recorders of temperature; the ice becomes more isotopically light as the polar temperature becomes colder, since more of the heavy isotopes have been distilled out in that case.

The upper panel of Fig. 1.13 shows the δD time series from the Vostok and Epica ice cores in the Antarctic. The Vostok data is systematically below the Epica data because Vostok is further inland, higher and colder, but otherwise tracks the Epica data. This record only covers the period within which the glacial-interglacial cycles have already settled into a 100,000 year cycle; older ice is too distorted to yield a useful record. This record confirms the 100,000 year cycle seen in the marine cores, and also confirms the asymmetry between slow buildup of glaciers and rapid deglaciation. It also shows that there is a strong Antarctic warming and cooling that occurs in association with the glacial/interglacial cycles.

The CO_2 data is shown in the lower panel of Fig. 1.13. In the course of the glacial/interglacial cycles, CO_2 fluctuates between 180 parts per million (by count of molecules) and 300 parts per million. Moreover, the fluctuation is very nearly synchronous with the warming and cooling. The correlation between CO_2 and temperature does not determine which causes which (or whether they mutually reinforce each other), but since CO_2 is a greenhouse gas, it is certain that the rise in CO_2 warms the planet and reinforces the termination of ice ages, whereas conversely the fall of CO_2 enhances cooling and reinforces the onset of glaciation. The origin of the glacial/interglacial CO_2 cycle is another of the Big Questions. Our understanding of glacial/interglacial cycles cannot be complete without resolving this question.

Greenland ice also records past climate, as seen in Fig. 1.14. The base of the Greenland ice cap is not as old as the Antarctic ice, so one can only go back in time about 100,000 years here. By way of compensation, though, the rate of snow accumulation in Greenland is much higher than in Antarctica, so one can see the past with much higher time resolution.

The Greenland record records a sharp deglaciation leading to the modern era, consistently with the Antarctic record. However, we can see in Greenland that there are many high frequency temperature fluctuations embedded within the glacial period –especially the period from about 60,000 years ago to 10,000 years ago. These don't have a strict periodicity, but have a time scale on the order of a thousand years, and hence are collectively referred to as *millennial variability*. The expressions of millennial variability seen in Greenland isotopes are called *Dansgaard-Oeschger events* after their discoverers. There are many other climate proxies that reflect millennial variability of the sort seen in Greenland isotopes, and these often represent precipitous switches in the climate state – referred to as *abrupt climate change.*. One of the most striking of these abrupt change events occurs as the planet is coming out of the most recent ice age – the *Last Glacial Maximum* (or *LGM*). The spike marked "B" in the figure is the *Bolling* warm period, and represents a recovery of the climate to full interglacial warmth. However, in the wake of the Bolling the climate abruptly reverted to full glacial temperatures, in an event known as the *Younger Dryas* (marked "YD" in the figure). The Younger Dryas is believed to have been triggered by a massive draining

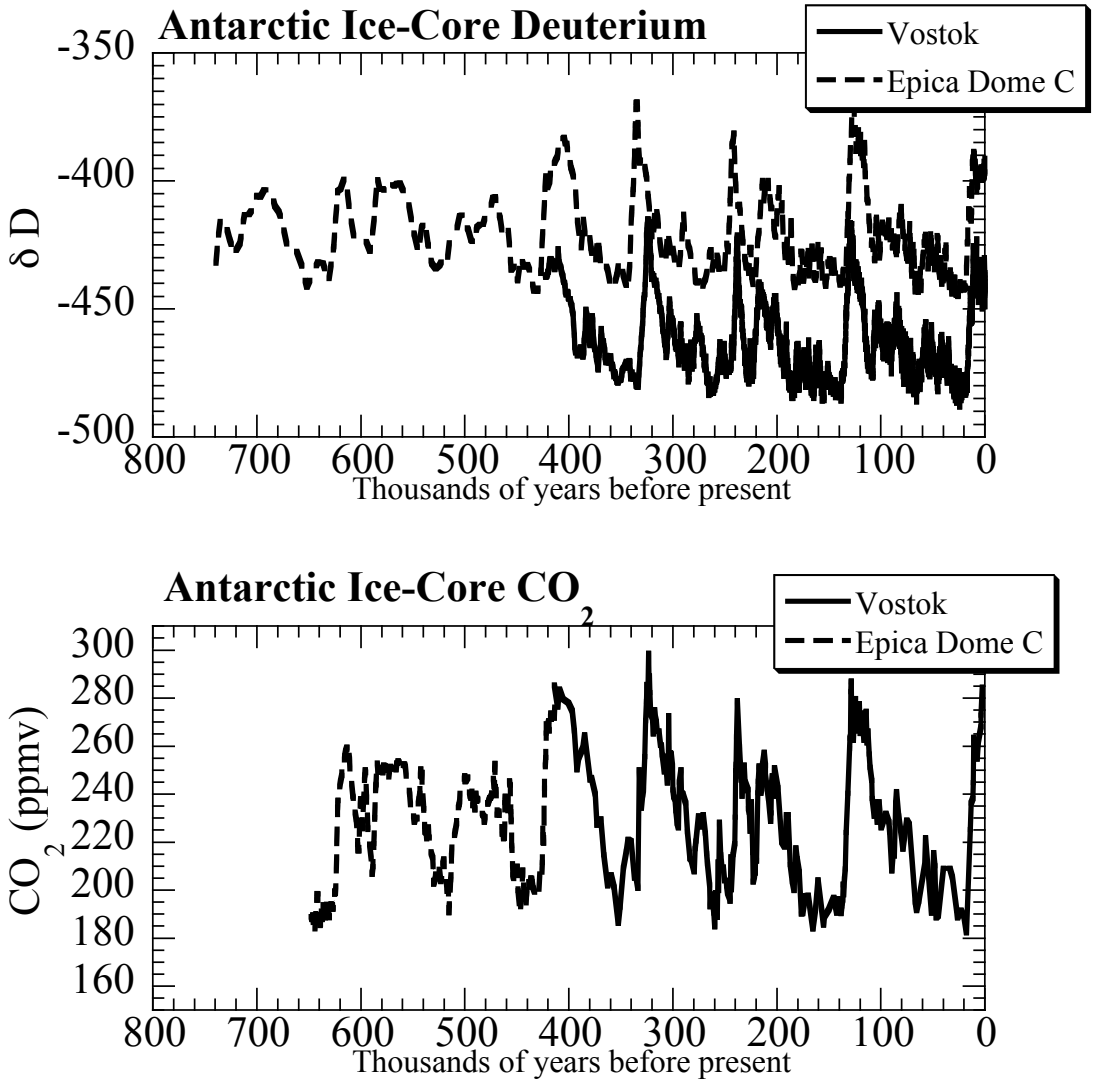

Figure 1.13: Data from the Vostok and Epica Antarctic ice cores. The upper panel shows the variation in deuterium depletion of the ice, which is a proxy for temperature. Higher (less negative) values indicate warmer conditions. The lower panel gives the variation in the CO_2 concentration in air bubbles trapped in the ice.

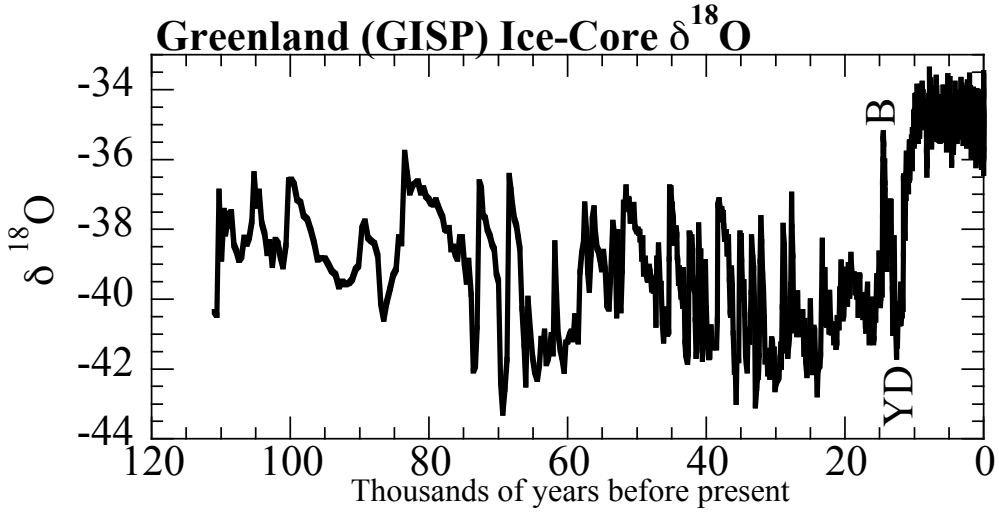

Figure 1.14: Oxygen isotope data from the GISP-2 Greenland ice core. Larger (less negative) values correspond to warmer temperatures.

of a glacial lake into the ocean, but generally speaking the mechanisms of both the Younger Dryas and of millennial variability remain as Big Questions that are yet to be resolved. This is an especially important question because the flat, quiescent period at the end of the Greenland record – the *Holocene* period we have lived in for all of the history of civilization – represents an abrupt cessation of high amplitude millennial variability. What accounts for the uncommonly stable climate of the Holocene, and what would it take to break this situation? This is a Big Question with considerable import indeed.

1.11 Holocene climate variation

The climate variations of the Holocene have been (at least so far) more subtle than the massive variations we have discussed previously. In part, this is simply because the Holocene is a short period of time, and there hasn't been enough time for something really dramatic to happen, given the relatively slow pace of many of the geological processes that modify climate. But the limited span of the Holocene is not the whole story. The Holocene has not witnessed extreme millennial scale variability of the sort seen in the preceding glacial time. Indeed, one of the Big Questions concerning the Holocene is the physical basis for the relatively stable Holocene climate. In particular, given the massive assault on climate by industrial society (see Section 1.12) it becomes all the more pressing to understand what it would take to break the equable and steady Holocene climate enjoyed during the rise of civilization.

Still, the Holocene has not been without its points of interest so far as climate variations go, the more so because there were human civilizations around to witness and be affected by these variations. A key driver of Holocene climate change is the *precessional cycle*, to be discussed in Chapter 7. It is the same precessional cycle that plays a role in the rhythm of the Pleistocene ice ages. The Earth's spin axis precesses like a top, so that the way the "tilt seasons" line up with the "distance seasons" associated with the varying distance of the Earth from the Sun goes through a cycle lasting about 22,000 years. A quarter cycle is only 5500 years, which brings these

cycles within the span of recorded history. At present, the Earth is farthest from the sun when the Northern Hemisphere points towards from the Sun (i.e. during Northern summer), and is closest to the Sun when the Northern Hemisphere points away (i.e. during Northern winter). This gives us relatively warm winters and relatively cool summers. 11000 years ago the situation was reversed, and Northern Hemisphere summers received considerably more sunlight than they do today, particularly at high latitudes. This should have made polar regions considerably warmer than today, but for reasons that are only partly understood, the time of warmest summers was delayed several thousand years, to a time called rather tendentiously the *Climatic Optimum* about 7000 years ago. It is not entirely clear what the climate was "optimal" for, and this is in any event a Northern-centric view as Southern hemisphere continents if anything experienced a weak cooling at this time (as one would expect from the nature of the precessional cycle). The alternate term *Altithermal* is preferred, as being less value-laden. In any event, at this time the Northern polar regions were getting up to 10% more solar radiation in summer than they are today, leading to warmer summers. The warming is expected to be greatest over land, since oceans average out the warm summers and cold winters. Tree ring records indicate that high-latitude land masses were between 2C and 4C warmer in the summer at the time of the Altithermal. These estimates are corroborated by an expansion of the tree line to higher altitude land in the European Arctic. Cold-tolerant trees are primarily sensitive to growing-season temperatures, and are little effected by a moderate decrease in winter temperatures; in fact, more severe winters can in some cases be favorable to tree growth, since cold winters interrupt the life cycle of various insect pests. While the Altithermal is certainly connected with the precessional cycle, the magnitude of the warming, the cause of the delay in warming (partly associated with leftover glaciers from the last ice age), and the role of ocean circulations and vegetation changes in the altithermal, are all active subjects of research.

The precessional cycle has also had a profound effect on the distribution of precipitation in low latitudes. The Sahara is a desert today, but the dry river features known as *wadis* were flowing with water six thousand years ago, at which time the desert was a savannah grassland. This wet period commenced about 14,500 years ago and the Sahara abruptly reverted to desert about 4700 years ago. There are also intriguing indications that the time of initiation of present tropical mountain glaciers follows a precessional cycle. For example, the Peruvian Andean glaciers of the Southern Hemisphere date back to the last ice age, while the Kilimanjaro ice fields were laid down during the African Humid Period about 10,000 years ago and Himalayan glaciers of the northern subtropics tend to be even younger. The connection between the seasonal cycle of solar radiation and the tropical precipitation distribution involves atmospheric circulations – monsoons and the Hadley circulation – that cannot be treated without a full understanding of atmospheric fluid dynamics. We will therefore have only limited opportunities to pursue the precessional precipitation cycles in the course of this book, though the treatment of the precessional cycle in solar forcing will provide the student with the necessary background for further study.

The *Little Ice Age* is another Holocene climate fluctuation of considerable interest. This term refers to a period of generally cool Northern Hemisphere extratropical land temperatures extending from approximately 1500 to 1800. Tree ring estimates suggest the Northern Hemisphere mean temperature dropped by something over $0.5C$ between the year 1400 and 1600. The cooling is corroborated by records of advances of mountain glaciers, sailors' observations of sea ice, and agricultural records. The Little Ice Age is too short and too recent to have anything to do with the precessional cycle, and while it is possible that fluctuations in the ocean circulation could have produced the cooling, the prime candidate for an explanation of the Little Ice age is a temporary slight dimming of the Sun. Sunspot observations do indicate a cessation in the normal solar sunspot cycle – called the *Maunder Minimum*– at about the time of the Little Ice Age. However, most estimates of the associate solar output change are far too small to yield a significant cooling.

Various mechanisms are under investigation which could amplify the response to the small solar fluctuation, but in the grand scheme of things the Little Ice Age is a rather subtle event, and accordingly hard to understand, particularly in terms of simple models.

To put the Holocene in perspective, it is salutory to note that the "abrupt" PETM event discussed in Section 1.9.1 lasted nearly 200,000 years, and set in over a period of around 10,000 years – as long as the full length of the Holocene. There really hasn't been much time for things to happen in the Holocene, and the time span of Figure 1.9 no doubt contains many 10,000 year periods as quiescent as the Holocene has been up until recently. Short as the Holocene is, we will see next that human activities have been able to cause some dramatic changes in the composition of the atmosphere and consequently in the Earth's climate. One wonders what it would take to trigger a hyperthermal event such as the PETM, which is over in the wink of an eye by the standards of Figure 1.9, but has a duration twenty times as long as the span of human civilization to date.

1.12 Back to home: Global Warming

We have seen that CO_2 has been a major factor in determining climate througout Earth's history, and that life, in turn, has greatly shaped the carbon cycle. Life is in the midst of disrupting the carbon cycle once more, but this time it's technological life that has provided the necessary innovations. Over billions of years, a great deal of organic carbon has been sequestered in the Earth's crust without oxidizing. Most of this carbon is in very dilute forms which cannot easily be tapped to provide economically useful amounts of energy. However, a very small portion winds up in nearly pure forms that are moreover chemically altered in a fashion that makes them especially convenient as fuels. These are the *fossil fuels* – coal formed from land plants and oil formed in marine environments. Natural gas can be produced by thermal alteration of either coal or oil. Fossil fuels represent concentrated solar energy stored in the form of organic carbon, which has been accumulating over hundreds of millions of years. This pool of readily oxidizable carbon exists precisely because it is in geological formations that have kept it apart from oxygen over the ages. It is only the evolution of technical civilization that is making it possible to dig up and oxidize hundreds of millions of years worth of stored fossil carbon within a few centuries.

In the year 2005, over 8 gigatonnes ($8 \cdot 10^{12} kg$) of carbon were released by fossil fuel burning, and annual emissions continue to grow rapidly. There are several ways to see that this is a very big number – a major upset to the natural carbon cycle. First, the pre-industrial atmosphere contained about 600 gigatonnes of carbon, so the 2005 annual emission is fully 1.3% of the undisturbed atmospheric content. If the same amount were released into the atmosphere each year, it would take only 75 years to double the atmospheric CO_2 content, provided all the released CO_2 stayed in the atmosphere. Alternately, one could compare the fossil fuel emissions to the volcanic outgassing which in the long term balances silicate weathering and sustains the carbon cycle. Precise estimates of volcanic outgassing are hard to come by, but generally are on the order of 0.1 gigatonnes of carbon per year or less. Thus, *fossil fuel carbon emissions are eighty times larger than background volcanic outgassing.* In fact, the very largest carbon flux number involved in the whole carbon cycle is the net CO_2 carbon fixed into organic carbon each year by worldwide photosynthesis, and fossil fuel emissions even look impressive when compared to this number. Based on satellite chlorophyll observations, it has been estimated that photosynthesis fixes 100 *gigatonnes* of carbon each year, about half on land and half in the oceans. The year 2005 fossil fuel emissions were fully 8% of this number. In other words, worldwide photosynthetic productivity would have to increase by 8% to take up the fossil fuel CO_2 *and 100% of that carbon would have to be buried as organic matter*

without being recycled by respiration. That, of course, would be a completely absurd situation, as virtually all of the photosynthetically fixed carbon is quickly respired back into the atmosphere, largely by bacteria who have had several billion years to become proficient at making use of organic carbon wherever they find it. As an example, land photosynthesis fixes about 50 *gigatonnes* of carbon each year, but the flux of organic carbon to the oceans in all the world's rivers is a mere 0.4 *gigatonnes* per year (one twentieth of fossil fuel carbon emissions). And there is no evidence that much of the remainder of the photosynthetically fixed carbon is remaining on land as soil organic carbon. To say that humans have become a force of geological proportions vastly understates the case, for by this measure human influences on the carbon cycle overwhelmingly dominate the natural sources.

The result of all our busy digging and burning has been a steady increase in atmospheric CO_2. Figure 1.15 shows the time series of atmospheric CO_2 concentration since 1750. CO_2 has a very long atmospheric lifetime, so it is well-mixed. In consequence, one finds nearly the same CO_2 concentration whereever one measures it, so long as the measurement is not in the immediate vicinity of major sources or sinks. The part of the record since 1950 comes from direct analyses of air samples at the Mauna Loa observatory, whereas the earlier part of the record comes from air trapped in bubbles in the ice of the Siple Dome site, Antarctica, but the two records match up well where they meet. At the dawn of the industrial era CO_2 concentrations are near 280 molecules per million (*ppmv* for short), right where they were left at the end of the most recent ice age. After 1750 the concentrations begin to rise, and by 2007 the concentrations have exceeded $380 ppmv$ – fully 35% above the pre-industrial value. Most of the increase has happened since the mid-twentieth century, and the rate of increase seems to be accelerating along with population and economic growth.

Not all of the carbon released by fossil fuel burning has remained in the atmosphere. Estimates based on careful historical inventories suggest that only about half of the total carbon released to date remains in the atmosphere as carbon dioxide. Most of the remainder has slowly infiltrated the ocean, with a lesser amount having been taken up by the terrestrial ecosystem (net of deforestation). In fact, it has been shown that the rate at which the ocean can take up the excess CO_2 is limited by the mixing between the upper ocean and the deep ocean. This is a slow process, and if all fossil fuel burning were to suddenly cease, it would take in excess of 600 years for 80% of the excess CO_2 to be taken out of the atmosphere. The remainder would stay in the atmosphere for millennia longer, owing to certain chemical processes (discussed briefly in Chapter 8) which limit the ability of the ocean to take up CO_2. The slow net removal rate of CO_2 allows fossil fuel emissions to accumulate in the atmosphere. Another consequence of the long lifetime of CO_2 in the atmosphere is that the climatic effects of elevated CO_2 will persist for centuries to millennia, even after any (much to be hoped-for) dramatic restriction of fossil-fuel burning. Allowing for uptake by the ocean, there are enough fossil fuel reserves– primarily in the form of coal – to ultimately increase the atmospheric CO_2 concentration to at least six times the pre-industrial value. The number could go much higher if the ocean sink were to become less efficient, or if land ecosystems were to turn around and become a CO_2 source rather than a sink.

This all leads us to a series of very Big Questions: if the rise in CO_2 is allowed to continue to a doubling of the pre-industrial value, how much will the Earth warm? How will the warming be distributed? How much will sea level rise as a result of melting land ice and thermal expansion of ocean water? What will happen to precipitation patterns? How will all of this affect human societies and natural ecosystems? The basic physics needed to treat these questions is identical to what is used to account for the influence of CO_2 and other long lived greenhouse gases on past climates. The problem in this instance has more immediacy as many generations of our descendents will be living with the consequences of our fossil fuel emissions in the next several

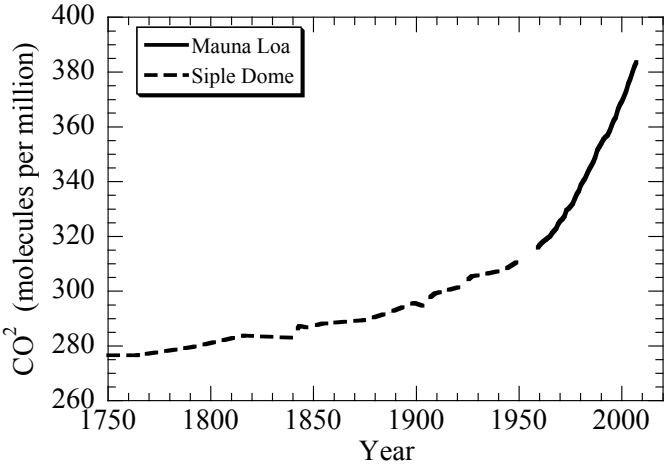

Figure 1.15: Annual mean CO_2 concentration from 1750 to the time of writing. The earlier part of the record is from air trapped in bubbles in the Siple Dome Antarctic ice core. The more recent part of the record is from instrumental measurements at the Mauna Loa observatory. The units are molecules of CO_2 per million molecules of air.

decades. In order to understand what kind of planet we are leaving these descendents, there is a demand for greater detail in the understanding of the climate changes to be wrought by these rapid increases in atmospheric CO_2.

Interest in the effect of CO_2 changes on climate long predates the kind of data shown in Figure 1.15 which showed that CO_2 was on the increase, and in fact predates the realization that human activities really could cause CO_2 to increase appreciably. Likewise, global warming was a concern long before it was confirmed that the Earth really was warming in response to increases of CO_2. These things were all anticipated theoretically a century or more before global warming burst onto the scene as an issue of political consequence, and the driving force was basic curiosity about the physics governing planetary temperature. It's a line of inquiry that extends right back to Fourier's pathbreaking inquiry into how an atmosphere affects the energy budget of a planet, and hence its temperature. The discovery of global warming is a great triumph of two centuries of developments in fundamental physics and chemistry. It is not a matter of people having noticed that both CO_2 and temperature were going up, and concluding that the first must be somehow causeing the second. Both the rise of CO_2 as a consequence of fossil fuel burning, and the consequent rise in temperature as a response to the Earth's perturbed energy balance, were anticipated long before either was observed.

After Fourier, the tale resumes with Tyndall, whose work on the infrared absorption of CO_2 and water vapor was mentioned near the beginning of this chapter. Tyndall was interested in these gases because of the questions raised by Fourier regarding the factors governing planetary temperature. He was also interested in the recently-discovered phenomenon of the ice-ages, and with several contemporaries thought perhaps ice ages could arise from a reduction in CO_2. In that, he was partly right; the Pleistocene ice ages are cold partly because of the glacial interglacial CO_2 cycle, even though the ultimate pacemaker of the ice ages is the rhythm of Earth's orbital parameters. Tyndall died, however, before he ever had the chance to translate his measurements into a computation of the Earth's temperature. That task was left to the Swedish physical chemist Svante

Arrhenius, who in 1896 performed the first self-consistent calculation of the Earth's temperature incorporating the greenhouse effect of water vapor and CO_2. Interestingly, Tyndall's measurements were not sufficient to provide the information about weak absorption over long path lengths, so for the absorption data he needed he turned to Langley's observations of infrared emitted by the Moon. It was a felicitous re-use of data intended originally for determination of the Moon's temperature, and indeed was a more correct use of the data than Langley was able to accomplish. This shows the benefit of curiosity-driven science: measurements taken to satisfy curiosity about lunar temperature wound up being instrumental in permitting an evaluation of the effect of the Earth's atmosphere on the Earth's temperature. Astronomers initiated the study of infrared as an observational technique, but the radiative transfer work stimulated by their needs soon provided the crucial tool needed to understand planetary climate. Arrhenius not only estimated the Earth's then-current temperature, but also estimated how much it would warm if the amount of CO_2 in the atmosphere were to double. Using clever scaling analyses from Langley's data, he was able to do this without a firm knowledge of just what the atmosphere's CO_2 content actually was. Not long afterwards, he realized that industrial burning of coal was dumping CO_2 into the atmosphere, and could eventually bring about a doubling; he described this process as "evaporating our coal mines into the atmosphere." At then-current rates of consumption, it appeared that a doubling would take up to a millennium, and Arrhenius would no doubt have been surprised to know that his own great-grandchildren could well live to witness the doubling. This takes our story to about 1900. What happened then?

A long hiatus. In part, there was little sense of urgency, because a failure to anticipate the explosive growth of fossil fuel use looming in the coming decades led to a belief that any problem was off in the far-distant future. Besides that, two unfortunate turns of events held back the study of global warming for decades. The first was a highly touted experimental study published in 1900 by the prominent physicist Knut Angstrom, which purported to show that the radiative effects of CO_2 are "saturated," i.e. that the gas already absorbs as much as it can at the atmosphere's then-present concentration, so increases would have no effect. A concomitant and closely associated (and equally wrong) idea was that the strong absorption of water vapor would completely swamp any effect CO_2 might have. The experiment turned out to be wrong, but such was Angstrom's reputation and such was the resistance to the idea that humans could change climate that it was decades before anybody definitively checked the result. Moreover, it turns out that even if Angstrom had been right, it would not have negated the greenhouse effect; this misunderstanding hinged on the poorly developed understanding of radiative transfer in a temperature stratified atmosphere. The "grey gases" we will study in the first half of Chapter 4 are "saturated" in the sense of Angstrom, but nonetheless allow for an increase in the greenhouse effect as more greenhouse gas is added to the atmosphere. The second barrier to progress was the belief that the huge carbon content of the ocean would buffer the atmosphere, overwhelming anything human industry could have thrown at it. The carbonate chemistry needed to defeat this idea was largely worked out by the 1930's;indeed, it requires nothing more than is taught routinely in high-school chemistry courses today. However, it had not been assimilated into a coherent and widely appreciated picture of the uptake rate of CO_2 by the oceans, in part because of lack of knowledge of the rate of mixing between the upper ocean and the deep ocean. A paper published by Revelle and Suess in 1957 is widely credited with having broken the logjam, but in fact mentions the essential carbonate buffering mechanism (fully worked out by earlier researchers, and cited as such) almost as an afterthought. The attempt of Revelle and Suess no doubt helped to revive interest in the question of oceanic CO_2 uptake rates, but in fact the paper came to exactly the wrong conclusion – that fossil fuel emissions were unlikely to lead to any significant increase in atmospheric CO_2 concentration. The true implications of the carbonate buffer for CO_2 increase due to fossil fuel burning was finally brought out clearly in a paper two years later, by Bert Bolin

and Erik Eriksson.

Despite Revelle and Suess's conclusion, the idea gained hold that somebody should actually systematically check and see what atmospheric CO_2 was doing. This program was initiated by Charles Keeling while at Caltech, and was subsequently encouraged by Revelle. Keeling's work culminated in the Mauna Loa data shown in Fig. 1.15. The techniques for recovering past CO_2 from air bubbles trapped in ice were not to be developed until the 1970's, so Keeling had to wait a decade or so before it was clear that CO_2 was really rising, and a bit more time after that before there was a clear idea of just how high CO_2 already was relative to the pre-industrial value. This work, together with developments in infrared radiative transfer stimulated by astronomical observation and military interest in infrared target detection lead to new breakthroughs in the formulation of radiative transfer. The work culminated in 1967 with the calculation by Manabe and Wetherald of the Earth's temperature using modern radiative physics. They were also able to calculate the warming due to a doubling of CO_2, allowing for expected changes in water vapor content as the planet warmed. This was not the end of the story, which indeed continues today, since there was much to be done in terms of embedding the radiative transfer in a fully consistent computation incorporating the fluid dynamics of the atmosphere – a *general circulation model.* It was, however, the beginning of the modern chapter of the study of global warming. With the publication of the Charney report by the US National Academy of Sciences in 1979, global warming began to be perceived as a real threat. The powers that be were slow to awaken to the magnitude of the problem, and several more years were to pass before the creation of the Intergovernmental Panel on Climate Change in 1988, which initiated regular, comprehensive surveys of the state of the science surrounding global warming. At the time of writing, the world still awaits substantive action to curb fossil fuel emissions.

All aspects of the essential chemistry, radiative physics and thermodynamics underlying the prediction of human-caused global warming have been verified in numerous laboratory experiments or observations of the Earth and other planets. Other aspects of the effect of increasing greenhouse gases rely on complex collective behavior of the interacting parts of the climate system; this includes behavior of clouds and water vapor, sea ice and snow, and redistribution of heat by atmospheric winds and ocean currents. Such things are impossible to test in laboratory experiments. To some extent, aspects of our theories of the collective behavior have been tested against the seasonal cycle of Earth, interannual variability, and past climates, as well as attempts to simulate other planetary climates. The ultimate test of the theory, though, is to verify it against the uncontrolled and inadvertent experiment we are conducting on Earth's own climate. Can we see the predicted warming in data? This is not an easy task. For one thing, the atmospheric CO_2 increase is only a small part of the way towards doubling, and the climate has not even fully adjusted to the effect of this amount of extra radiative forcing: oceans take time to warm up, and delay the effect for many years (for reasons to be discussed in Chapter 7. Thus, so far the signal of the human imprint on climate is fairly small. Set against that is a fair amount of noise complicating the detection of the signal. Climate, even unperturbed by human influence, is not steady from year to year, but is subject to a certain amount of natural variability. This can be due to volcanic eruptions and subtle variations in the brightness of the Sun. There are also various natural cycles in the ocean-atmosphere system that cause the planet to be a bit warmer or colder from one year to the next. Chief among these is the El Niño phenomenon of the tropical Pacific. During El Niño years, the coupled dynamics of the tropical ocean and atmosphere causes warm water to spread throughout the Pacific, leading to a warming of mean surface temperatures both in the tropics and further afield. La Niña years represent a bunching up of the warm water, and an accentuated upwelling of cold water, leading to cold years. The two phases alternate erratically, with a typical time scale of three to five years.

The fact that the signal is hard to detect does not mean that global warming is of little consequence. The difficulty arises precisely because we are trying to detect the signal before it becomes so overwhelmingly large as to be obvious. Given the long lifetime of CO_2 in the atmosphere, it would be highly desirable to keep the signal from ever getting that large, as if it ever does it will take many centuries to subside. Let's now take a look at some of the data, and see if there are any signs that the theoretically anticipated warming is really taking place.

Figure 1.16 shows a times series of estimated global mean surface temperature, based on recorded temperatures measured with thermometers. There is a lot of arduous statistics and data archaeology behind this simple little curve. Particularly for the data going into the early part of the curve, there has been a need to standardize measurements to allow for the various different ways of taking a temperature reading. Most of the oceanic measurements, for example, were taken by commercial or military ships of one sort or another, and some of the entries in ships' logs record things like "bait tank temperature" or "engine inlet temperature." There has also been a need to screen out stations that have been strongly affected by local land use changes such as urbanization, and to avoid spurious trends due to changes in the spatial distribution of temperature measurements (e.g. fewer Antarctic readings once Antarctic whaling essentially ceased). To help correct for biases in individual classes of temperature measurements, the long-term trends are presented as *anomalies* relative to the average of a station's long-term average standardized to a fixed base period (e.g. 1951-1980 for the data in Fig. 1.16).

There is little temperature trend between 1880 and 1920, but between 1920 and 2005 the temperature has risen by nearly $1C$. The rise hasn't been steady and uninterrupted, however. It takes the form of an early rise between 1920 and 1940, followed by a 30 year period when temperatures remained fairly flat, whereafter the temperature rise resumes and has continued to the present. Given that CO_2 has been rising at an ever-increasing rate over the industrial period, why was the warming interrupted between 1940 and 1970? The answer lies largely in another effect of human activities on climate. Burning of fossil fuels, and especially coal, releases sulfur compounds into the atmosphere which form tiny highly reflective droplets known as *sulfate aerosols*. By 1995,the effect was finally quantified with sufficient accuracy to permit reasonable estimates of the effect, and it began to appear that most of the evolution of climate of the twentieth and twenty-first centuries could be accounted for by a combination of rising greenhouse gases (mainly CO_2) due to human activity, with an offsetting cooling effect of sulfate aerosols. The reason small particles are so good at scattering light back to space is discussed in Chapter 5, where the optical properties of sulfate aerosols will be discussed in detail. By the year 2000, the greenhouse warming signal had unquestionably risen above both the noise of natural variability and the offsetting effect of aerosol cooling. Sulfur is an active element in many actual and hypothetical planetary atmospheres, and so the study of sulfate aerosols on Earth informs other planetary problems, including the clouds of Venus and Venus-like extrasolar planets.

The Earth's emissions in the microwave spectrum have been monitored continuously by satellite-borne instruments since 1979, and these observations make it possible in principle to obtain reconstructions of atmospheric temperature trends which are independent of the somewhat inhomogeneous surface station network. Processing the microwave data acccurately enough to obtain reliable temperature trends proved very difficult, and there were many false steps along the way. Nonetheless, the main problems were resolved by early in the twenty-first century. The microwave temperature retrievals give the temperature of the atmosphere averaged over fairly deep layers, in constrast to the surface stations which measure near-surface air temperature. The left panel of Figure 1.17 shows the satellite retrieval of temperature in layer of the atmosphere known as the *lower troposphere* – extending from sea level to roughly 5 km in altitude. The satellite record tracks the GISS surface station record very closely, with the exception of the very strong 1997 El

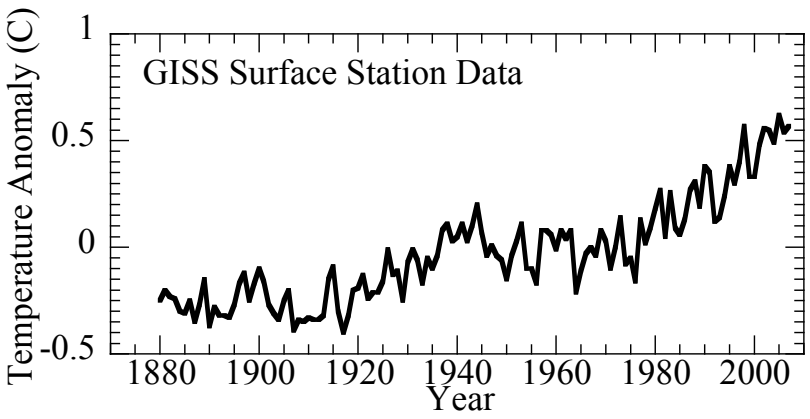

Figure 1.16: Global average annual mean surface temperature since 1870, estimated from surface temperature observations. Data source is the NASA GISS surface station analysis. The temperature is given as an anomaly relative to the mean temperature for the years 1951 to 1980. To turn these into actual global mean temperatures in degrees Celsius, add $14C$ to the anomaly.

Niño, during which the satellite indicates that the lower tropospheric layer warmed considerably more than the near-surface air. Both satellite and GISS records reproduce the cooling caused by the El Chicon eruption (which overwhelmed the 1982 El Niño) and the Pinatubo eruption (which accentuated the La Niña cooling following the 1991 El Niño, leading to a very cold year). The substantial agreement between the satellite and surface station record proves beyond doubt that the warming observed in recent times is not an artifact of any supposed inadequacies of the surface station record.

The situation looks quite different higher up in the atmosphere. The right panel of Fig. 1.17 shows the temperature trend in a portion of the atmosphere called the *lower stratosphere*, extending from about 15 to 25 km in altitude. Here, volcanic eruptions produce a pronounced warming, as opposed to the cooling seen at lower layers. This suggests that volcanic aerosols heat the upper atmosphere by absorbing sunlight. The pattern of upper level cooling and lower level warming produced by high altitude solar-absorbing layers will be discussed in Chapter 4, and the reflective effect of aerosols will be brought into the picture in Chapter 5. Leaving out the warm spikes associated with volcanic eruptions, the lower stratosphere appears to have undergone a pronounced cooling over the span of the satellite record. Is stratospheric cooling compatible with CO_2-induced warming in the lower troposphere? This is a Big Question that is resolved in Chapter 4. Ozone destruction also cools the stratosphere, since ozone absorbs sunlight. That portion of the cooling should go away as ozone recovers as a consequence of the Montreal Protocol banning ozone-destroying chlorofluorocarbons.

The Big Question of how much the Earth will warm upon a doubling or quadrupling of CO_2, and how fast it will do so, engages a number of associated questions. Insofar as water vapor is itself a powerful greenhouse gas, any tendency for water vapor content to increase with temperature will amplify the warming caused by CO_2. This is known as *water vapor feedback*. This feedback is now considered to be on quite secure ground, but the study of the behavior of water vapor in the atmosphere offers many challenges, and is a problem of considerable subtlety. In subsequent chapters, we'll provide the underpinnings needed for a study of this host of questions. Clouds present an entirely greater order of difficulty, as they warm the planet through their effect on

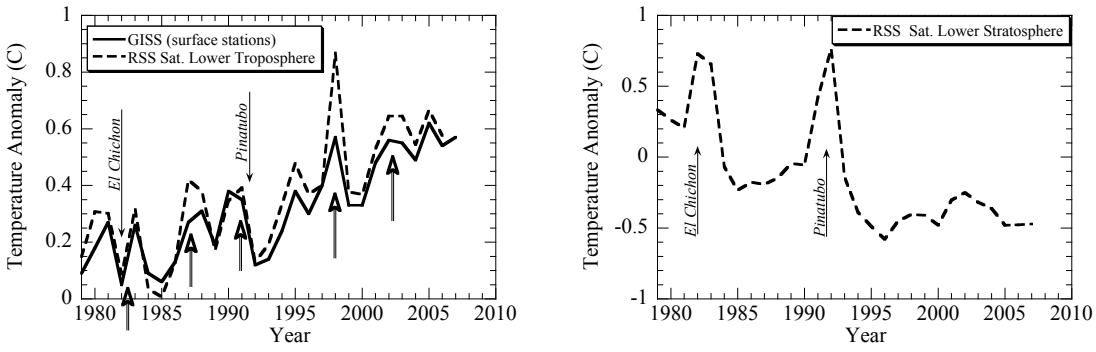

Figure 1.17: Atmospheric temperature time series derived from analysis of microwave satellite data. Left panel: Mean temperature anomaly for the layer of the atmosphere below about 5km, compared with the GISS instrumental record. Temperature is given as an anomaly relative to the same base value as used in the GISS instrumental record. Right panel: Temperature anomaly for the lower stratosphere (layer from about 15 km to 25 km). In both figures, the El Chichon and Pinatubo volcanic eruptions are marked. In the left panel, major El Niño events are indicated by upward open-shafted arrows.

outgoing infrared radiation, but cool the planet through their reflection of solar radiation. The net effect depends on the complex processes determining cloud height, cloud distribution, cloud particle size and cloud water or ice content. The infrared effects of clouds will be discussed in Chapter 4 and the reflective effects of clouds on sunlight will be discussed in Chapter 5. Uncertainties about the behavior of clouds are the main reason we do not know precisely how much warmer the planet will ultimately get if we double the CO_2 concentration. Typical predictions of equilibrium global average warming for a doubling of CO_2 range from a low of around $2C$ to a high of around $6C$, with some potential for even greater warming with a low (but presently unquantifiable) probability. Because of other uncertainties in the system (particularly the magnitude of the aerosol effect and especially the indirect aerosol effect on cloud brightness) simulations with a range of different cloud behaviors can all match the historical climate record so far, but nonetheless yield widely different forecasts for the future. There is no analysis at present that excludes the possibility of the higher end of the forecast range, for which the effects would likely be catastrophic. There are other feedbacks in the climate system that complicate the forecast. These include feedbacks from melting snow and ice, and from the dynamics of glaciers on land. They also include changes in vegetation, and changes in the ocean circulation which can affect the delay due to burial of heat in the deep ocean.

Global warming – perhaps more aptly called "global climate disruption" – is an event of geological proportions, but one which is caused by human activities. The natural range of CO_2 for the past 800,000 years, and almost certainly for the entire two million years of the Pleistocene, has been 180 to 280 molecules per million. Owing to human activities, the CO_2 concentration is already far above the top of the natural range that has prevailed for the entire lifetime of the human species, and without action will become much higher still. The human species and the natural ecosystems we share the Earth with have adapted over the Pliocene and Pleistocene to glacial-interglacial cycles, but a world with doubled CO_2 will subject them in the course of two centuries or less to a temperature jump to levels far warmer than the top of the range to which societies and organisms have adapted. Even if climate sensitivity is at the low end of the predicted range and if human

societies hold the line at a doubling of CO_2, the resulting $2C$ warming represents a substantial climate change; it takes a great deal to change the mean temperature of the entire globe, and a $2C$ global mean increase is a summary statistic that masks much higher regional changes and potentially quite massive effects on sea ice, glaciers and ecosystems. If climate sensitivity turns out to be at the high end, the warming could be $4C$ or more, and if that is compounded by an increase to four times pre-industrial CO_2 the global mean increase could reach $8C$. That is twice the degree of warming in the PETM, and though the PETM looks abrupt, it is very likely to have set in on a longer time scale than it would take human industrial society to burn the remaining reserves of fossil fuels. If this is allowed to happen, it will take thousands of years for the climate to recover to a normal state. Could global warming disrupt the natural glacial-interglacial cycle? What would the consequences of that be? Those are indeed Big Questions.

As seen by paleoclimatologists ten million years in the future, whatever species they may be, the present era of catastrophic release of fossil fuel carbon will appear as an enigmatic event which will have a name of its own, much as paleoclimatologists and paleobiologists refer today to the PETM or the K-T boundary event. The fossil carbon release event will show up in ^{13}C proxies of the carbon cycle, in dissolution of ocean carbonates through acidification of the ocean, through mass extinctions arising from rapid warming, and through the moraine record left by retreating mountain glaciers and land-based ice sheets. As an event, it is unlikely to permanently destroy the habitability of our planet, any more than did the K-T event or the PETM. Still, a hundred generations or more of our descendents will be condemned to live in a planetary climate far different from that which nurtured humanity, and in the company of a greatly impoverished biodiversity. Biodiversity does recover over the course of millions of years, but that is a very long time to wait, if indeed there are any of our species left around at the time to do the waiting. Extinction may not be precisely forever, but it is close enough.

1.13 The fate of the Earth,the lifetime of biospheres

Even if a planet enters a habitable phase at some stage in its life, it will not remain habitable forever; various kinds of crises can bring its habitability to an abrupt or gradual end. This brings us to the Big Question of lifetime of biospheres; the answer has implications for how likely it is that complex or intelligent life will have had time to evolve elsewhere in the Universe.

Certainly, the Earth's habitability will end when the Sun leaves the main sequence and expands into a Red Giant. Perhaps some of the outer planets or their satellites will enter a brief habitable phase at that time, but it will not be long lasting. That particular crisis is about four billion years in Earth's future, but other habitability crises are likely to set in long before then. In particular, as the Sun continues to brighten, at some point the brightness will outstrip the ability of the silicate weathering process to compensate by drawing down CO_2. At that point the Earth would succumb to a runaway greenhouse, become lethally hot, and eventually lose its water to space. When will that happen? That is a Big Question, and some current estimates put the remaining natual lifetime of Earth's biosphere at as little as a half billion years. Given that it took four billion years of Earth History before intelligent life emerged, that makes our existence look like quite a close call. Even before the runaway stage, silicate weathering will draw down CO_2 to the point where most forms of photosynthesis will no longer be able to operate. Can more efficient forms of photosynthesis fill in the gap? That's a Big Question as well, but one of a primarily biological nature that we will not attempt to answer.

As the Sun's luminosity increases, Earth may become uninhabitable, but other planets in the Solar System may become more hospitable; in any event they will go through interesting

transformations. Mars will warm up, but given that it has little or no active tectonics to generate a new atmosphere, it is unlikely to become Earthlike unless some artificial means is found to give it a more substantial atmosphere. Could Europa melt and become a waterworld? What will happen to Titan as the Sun gets brighter?

Alternately the end could come by ice rather than fire. Earth's life and climate are ultimately maintained by a brew consisting of solar energy and the CO_2 outgassing from the interior. The CO_2 has a warming effect of its own, which can be modified by organisms that intercept it and transform it into oxygen, methane, or other compounds. If the tectonic release of CO_2 ceases, as it will once the Earth exhausts its interior heat sources, all that will come to an end. There will be nothing to offset silicate weathering, and CO_2 will draw down until the Earth turns into a snowball – unless the runaway greenhouse from a brightening Sun gets us first.

This class of questions naturally generalizes to the question of how the time scales that limit the biosphere's lifetime would be different for planetary systems around other stars. We have already mentioned that hotter stars have a shorter life on the main sequence, while cooler stars last longer; the former will have planets with short-lived biospheres compared to the latter. The question of how long the silicate weathering thermostat can cope with changing stellar luminosity, and how long outgassing can sustain the climate, is far subtler, and will have interesting dependences on the planet's size, composition and orbit. There could well be other chemical cycles other than silicate weathering and CO_2 outgassing, which could provide climate regulation; the search for such possibilities is still in its infancy. There could also be novel habitability crises, associated with long term evolution of planetary systems with highly eccentric orbits or systems perturbed by binary star companions.

All this climate catastrophe presupposes no intervention by the inhabitants. In fact, there are quite realistic possibilities for technologically adept inhabitants to stave off the catastrophe at least until their star leaves the Main Sequence. A runaway greenhouse could be prevented by simply reducing the effective stellar brightness, through orbital sunshades or injection of reflecting aerosols into the upper atmosphere. Indeed, such *geoengineering* fixes have been proposed to offset the global warming effect of anthropogenic CO_2 increases. They are a rather desperate and alarming prospect as a solution to global warming, since they offset a climate forcing lasting a thousand years or more with a fix requiring more or less annual maintainence if catastrophe is not to strike; far better to keep CO_2 from getting dangerously high in the first place. However, if the alternative a half billion years out is a runaway greenhouse, the risk of maintaining sunshades will no doubt seem quite acceptable. The loss of CO_2 outgassing as Earth's tectonic cycle ceases also has a relatively easy technical fix. Inhabitants could use a small portion of the energy received from the star to cook CO_2 back out of carbonates, in a process nearly identical to that by which cement is manufactured. Given the slow rate of silicate weathering, only modest quantities of carbonate would have to be processed. All this can be done,but it would appear to require long term planning and intelligent intervention. A good understanding of the principles of planetary climate will be needed by any beings contemplating such interventions. This book, we hope, will be a good place to start.

1.14 For Further Reading

Many of the problems and explorations in the Workbook sections of this book require the use of some basic numerical analysis. The necesary algorithms are enumerated and exercised in the Workbook section of this chapter. The essential reference for the derivation and implementation of the algorithms is the *Numerical Recipes* series. The current edition is:

- Press WH, Teukolsky SA, Vetterling WT and Flannery BP 2007: *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press

The algorithm description is independent of programming language, but the implementations given in this particular edition are based on `c++`. Earlier editions are available for other programming languages. The `c` language implementations in the 1992 edition *Numerical Recipes in c* provide a clean basis for re-implementation in other programming languages, particularly convenient for the reader who wishes to avoid some of the intricacies of `c++`. The reader will need only a very small part of the material covered the *Numerical Recipes* opus; the relevant sections are pointed out in the Workbook for this chapter.

Extrasolar planet observations grow by leaps and bounds. New planets are added to the catalog on a practically monthly basis, and there are frequent updates to the database of planetary characteristics. For the most current data, consult the online *Extrasolar Planets Encyclopedia*, at `http://exoplanet.eu`. The data shown in this chapter is from the paper

- Butler RP *et al* 2006: Catalog of Nearby Exoplanets, *Astrophysical J.*, **646**.

Conditions during the earliest period of Earth's history, the time required for the crust to form and cool, and the time evolution of heat flux from the interior of the Earth to the surface are discussed in:

- Sleep NH, Zahnle K and Neuhoff PS 2001: Initiation of clement surface conditions on the earliest Earth, *Proc Nat Acad Sci*, **98**, 3666-3672.
- Turcotte DL 1980: On the thermal evolution of the Earth, *Earth Planet. Sci. Lett.*, **48**, 53-58.

The long-term evolution of the brightness of the Sun and similar stars is discussed in:

- Gough, DO 1981: Solar interior structure and luminosity variations, *Solar Physics*, **74**, 21-34.
- Sackmann I-J, Boothroyd AI, and Kraemer KE 1993: Our Sun. III. Present and Future, *Astrophysical Journal*, **418**, 457-468.

For a very engaging introduction to what we know about life on the Early Earth, the reader is directed to the book

- Knoll A 2004: *Life on a Young Planet*. Princeton University Press

The original paper on the Faint Young Sun problem is

- Sagan C and Mullen G 1972: Earth and Mars: Evolution of atmospheres and surface temperatures , *Science*, **177**, 52-56.

A good review of the history of oxygen on Earth and the proxy methods used to infer this history can be found in

- Canfield DE 2005: The Early History of Atmospheric Oxygen: Homage to Robert M. Garrels *Annu Rev Earth Planet Sci*, **33**, 1-36. doi:10.1146/annurev.earth.33.092203.122711

For a general introduction to the Snowball Earth problem, see

- Hoffman, PF and Schrag DP 2002: The snowball Earth hypothesis: testing the limits of global change. *Terra Nova* **14**, 129-155.

The discussion of late Cretaceous and Cenozoic paleoclimate drew largely on the following papers:

- Sluijs A *et al* 2006: Suptropical Arctic Ocean temperatures during the Palaeocene/Eocene thermal maximum. *Nature* **441** doi:10.1038/nature04668 .
- Moran K *et al* 2006: The Cenozoic palaeoenvironment of the Arctic Ocean. *Nature* **441** doi:10.1038/nature04800
- Pearson PN *et al* 2007: Stable warm tropical climate through the Eocene Epoch. *Geology* **35** doi: 10.1130/G23175A.
- Forster A *et al* 2007: *Geology* **35** doi:10.1130/G23874A.
- Mix, A.C., *et al.* 1995. Benthic foraminifera stable isotope record form Site 849, 0-5 Ma: Local and global climate changes. Pages 371-412 in N.G. Pisias *et al.* editors, Proceedings of the Ocean Drilling Program, Scientific Results 138, College Station, Texas, USA.

Evolution of Phanerozoic climate, occurrence of glaciations, and evolution of CO_2 content of the atmosphere are discussed in

- Crowley TJ and Berner RA 2001: CO_2 and climate change, *Science*, **292**, 870-872. DOI: 10.1126/science.1061664
- Zachos J *et al.* 2001: Trends, Rhythms and Aberrations in Global Climate 65 Ma to Present, *Science*, **292** , 686-693. DOI: 10.1126/science.1059412

Veizer's long term fossil ^{18}O tropical temperature record, discussed in Crowley and Berrner (2001), is not generally considered reliable.

The GISP, Vostok and EPICA ice core records are described in

- Grootes PM, and Stuiver M. 1997: Oxygen 18/16 variability in Greenland snow and ice with 10^3 to 10^5-year time resolution. *J. Geophys. Res.* **102** 26455-26470.
- Petit, JR *et al.* 2001: Vostok Ice Core Data for 420,000 Years, *IGBP PAGES/World Data Center for Paleoclimatology Data Contribution Series No. 2001-076.* NOAA/NGDC Paleoclimatology Program, Boulder CO, USA.
- Petit, JR *et al.*: 1999, Climate and Atmospheric History of the Past 420,000 years from the Vostok Ice Core, Antarctica, *Nature* **399** 429-436.
- Siegenthaler TF *et al.* 2005: Stable Carbon Cycle-Climate Relationship During the Late Pleistocene. *Science* **310** 1313-1317.

The intellectual history surrounding anthropogenic climate change ("global warming") is surveyed in the following book and acccompanying web site:

- Weart S 2008: *The Discovery of Global Warming*. Harvard University press
- `http://www.aip.org/history/climate/index.html`

Additional information can be found in *The Warming Papers*, a set of critical readings with essays by D. Archer and the author, forthcoming from Wiley/Blackwell circa 2009.

Chapter 2

Thermodynamics in a Nutshell

2.1 Overview

The atmospheres which are our principal objects of study are made of compressible gases. The compressibility has a profound effect on the vertical profile of temperature in these atmospheres. As things progress it will become clear that the vertical temperature variation in turn strongly influences the planet's climate. To deal with these effects it will be necessary to know some thermodynamics – though just a little. This chapter does not purport to be a complete course in thermodynamics. It can only provide a summary of the key thermodynamic concepts and formulae needed to treat the basic problems of planetary climate. It is assumed that the student has obtained (or will obtain) a more fundamental understanding of the general subject of thermodynamics elsewhere.

2.2 A few observations

The temperature profile in Figure 2.1, measured in the Earth's tropics introduces most of the features that are of interest in the study of general planetary atmospheres. It was obtained by releasing an instrumented balloon (radiosonde) which floats upwards from the ground, and sends back data on temperature and pressure as it rises. Pressure goes down monotonically with height, so the lower pressures represent greater altitudes. The units of pressure used in the figure are *millibars* (*mb*). One *bar* is very nearly the mean sea-level pressure on Earth, and there are 1000 *mb* in a *bar*.

Pressure is a very natural vertical coordinate to use. Many devices for measuring atmospheric profiles directly report pressure rather than altitude, since the former is generally easier to measure. More importantly, most problems in the physics of climate require knowledge only of the variation of temperature and other quantities with pressure; there are relatively few cases for which it is necessary to know the actual height corresponding to a given pressure. Pressure is also important because it is one of the fundamental thermodynamic variables determining the state of the gas making up the atmosphere. Atmospheres in essence present us with a thermodynamic diagram conveniently unfolded in height. Throughout, we will use pressure (or its logarithm) as our fundamental vertical coordinate.

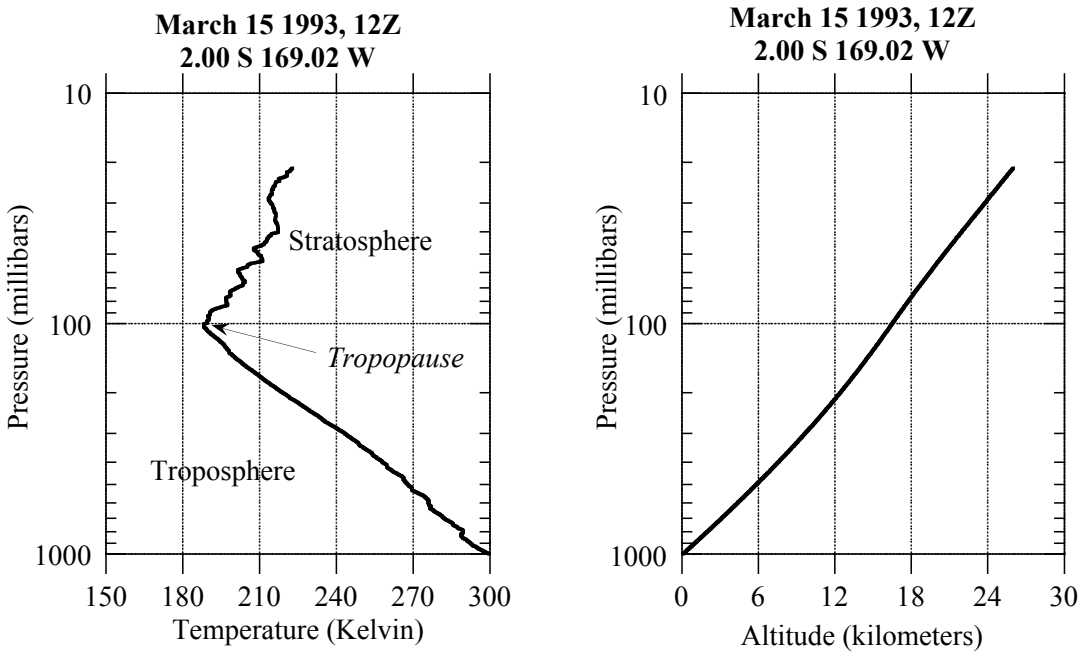

Figure 2.1: Left panel: Temperature profile measured at a point in the tropical Pacific. Right panel:The corresponding altitude. The measurements were obtained from a radiosonde ("weather balloon") launched at 12Z (an abbreviaton for Greenwich Mean Time) on March 15, 1993.

However, for various reasons one might nevertheless want to know at what altitude a given pressure level lies. By altitude tracking of the balloon, or using the methods to be described in Section 2.3, the height of the measurement can be obtained in terms of the pressure. The right panel of Figure 2.1 shows the relation between altitude and pressure for the sounding shown in Figure 2.1. One can see that the height is very nearly linearly related to the log of the pressure. This is the reason it is often convenient to plot quantities vs. pressure on a log plot. If p_o is representative of the largest pressure of interest, then $-\ln(p/p_o)$ is a nice height-like coordinate, since it is positive and increases with height.

We can now return to our discussion of the critical aspects of the temperature profile. The most striking feature of the temperature sounding is that the temperature goes down with altitude. This is a phenomenon familiar to those who have experienced weather in high mountains, but the sounding shows that the temperature drop continues to altitudes much higher than sampled at any mountain peak. This sounding was taken over the Pacific Ocean, so it also shows that the temperature drop has nothing to do with the presence of a mountain surface. The temperature drop continues until a critical height, known as the *tropopause*, and above that height (100mb, or 16 km in this sounding) begins to increase with height [1]. The portion of the atmosphere below the tropopause is known as the troposphere, whereas the portion immediately above is the stratosphere. "Tropo" comes from the Greek root for "turning" (as in "turning over"), while "Strato.." refers to stratification. The reasons for this terminology will become clear shortly. The stratosphere was discovered in 1900 by Lèon Phillipe Teisserenc de Bort, the French pioneer of instrumented balloon flights.

The sounding we have shown is typical. In fact, a similar pattern is encountered in the atmospheres of many other planets, as indicated in Figure 2.2 for Venus, Mars, Jupiter and Titan. In common with the Earth case, the lower portions of these atmospheres exhibits a sharp decrease of temperature with height, which gives way to a region of more gently decreasing, or even increasing, temperature at higher altitudes. In the case of Venus, it is striking that measurements taken with two completely different techniques, probing different locations of the atmosphere at different times of day and separated by a decade, nonetheless agree very well in the region of overlap of the measurements. This attests both to the accuracy of the measurement techniques, and the lack of what we would generally call "weather" on Venus, at least insofar as it is reflected in temperature variability. In the case of Mars, the temperature decrease shows most clearly in the summer afternoon when the surface is still warm from the Sun. As nignt approaches the upper level temperature decrease is still notable, but the lower atmosphere cools rapidly leading to a low level *inversion*, or region of temperature increase. In the Martian polar winter,the whole atmosphere cools markedly, and is much more isothermal than in the other cases.

The temperature decrease with height in the Earth's atmosphere has long been known from experience of mountain weather. It became a target of quantitative investigation not long after the invention of the thermometer, and was early recognized as a challenge to those seeking an understanding of the atmosphere. It was one of the central pre-occupations of the mountaineer and scientist Horace Bénédict de Saussure (1740-1799). In the quest for an explanation, many false steps were taken, even by greats such as Fourier, before the correct answer was unveiled. As will be shown in the remainder of this chapter, some simple ideas based on thermodynamics and vertical mixing provide at least the core of an explanation for the temperature decrease with

[1] Though the temperature minimum in vertical profiles of the Earth's atmosphere is the most obvious indicator of a transition between distinct layers, we'll eventually be able to provide a more dynamically based definition of the tropopause. In the definition we'll ultimately settle on, the tropopause need not be marked by a temperature minimum, though on Earth the temperature minimum lies only somewhat above the dynamically defined tropopause, and therefore serves as an approximate indicator of the tropopause location.

height. Towards the end of Chapter 3 we will introduce a theory of tropopause height that captures the essence of the problem; the theory of tropopause height will be revisited with increasing sophistication at various points in Chapters 4, 5 and 6. Nonetheless, some serious gaps remain in the state of understanding of the rate of decrease of temperature with height, and of the geographical distribution of tropopause height. In Chapters 3 and 4 we will see that the energy budget of a planet is crucially affected by the vertical structure of temperature; therefore, a thorough understanding of this feature is central to any theory of planetary climate.

2.3 Dry thermodynamics of an ideal gas

2.3.1 The equation of state for an ideal gas

The three thermodynamic variables with which we will mainly be concerned are: temperature (denoted by T), pressure (denoted by p) and density (denoted by ρ). Temperature is proportional to the average amount of kinetic energy per molecule in the molecules making up the gas. We will always measure temperature in degrees Kelvin, which are the same as degrees Celsius (or Centigrade), except offset so that absolute zero – the temperature at which molecular motion ceases – occurs at zero Kelvin. In Celsius degrees, absolute zero occurs at about -273.15C. Pressure is defined as the force per unit area exerted on a surface in contact with the gas, in the direction perpendicular to the surface [2]. It is independent of the orientation of the surface, and can be defined at a given location by making the surface increasingly small. In the mks units we employ throughout this book , pressure is measured in Pascals (Pa); 1 Pascal is 1 Newton of force per square meter of area, or equivalently 1 $kg/(ms^2)$. For historical reasons, atmospheric pressures are often measured in "bars" or "millibars." One bar, or equivalently 1000 millibars (mb) is approximately the mean sea-level pressure of the Earth's current atmosphere. We will often lapse into using mb as units of pressure, because the unit sounds comfortable to atmospheric scientists. For calculations, though, it is important to convert millibars to Pascals. This is easy, because 1 mb = 100 Pa. Hence, we should all learn to say "Hectopascal" in place of "millibar." It may take some time. When pressures are quoted in millibars or bars, one must make sure to convert them to Pascals before using the values in any thermodynamic calculations.

Density is simply the mass of the gas contained in a unit of volume. In mks units, it is measured in kg/m^3.

For a perfect gas, the three thermodynamic variables are related by the perfect gas equation of state, which can be written

$$p = knT \tag{2.1}$$

where p is the pressure, n is the number of molecules per unit volume (which is proportional to density) and T is the temperature. k is the *Boltzmann Thermodynamic Constant*, a universal constant having dimensions of energy per unit temperature. Its value depends only on the units in which the thermodynamic quantities are measured. n is the *particle number density* of the gas. To relate n to mass density, we multiply it by the mass of a single molecule of the gas. Almost all of this mass comes from the protons and neutrons in the molecule, since electrons weigh next to nothing in comparison. Moreover, the mass of a neutron differs very little from the mass of a

[2] Pressure can equivalently be defined as the amount of momentum per unit area per unit time which passes *in both directions* through a small hoop placed in the gas. This definition is equivalent to the exerted-force definition because when a molecule with velocity v and mass m bounces elastically off a surface, the momentum change is $2mv$, but only half of the molecules are moving toward the surface at any given time. The momentum flux definition, in contrast, counts molecules going through the hoop in both directions

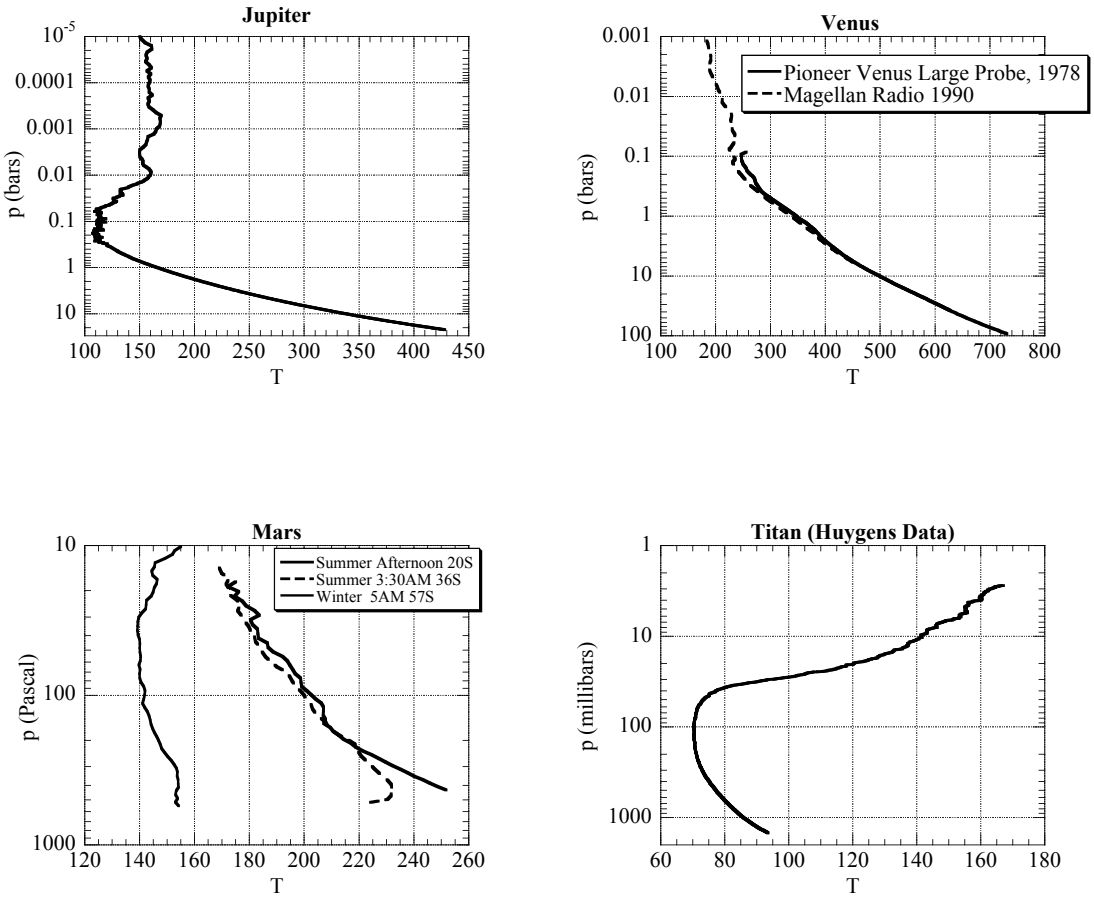

Figure 2.2: The vertical profile of temperature for a portion of the atmosphere of Jupiter (upper left panel), for the atmosphere of Venus (upper right panel) , and for the atmosphere of Mars (lower left panel) and that of Titan (lower right panel). The Venus Magellan and Mars data derive from observations of radio transmission through the atmosphere, taken by the Magellan (late 1980's) and Global Surveyor orbiters, respectively. The information on the lower portion of the Venus atmosphere comes from one of the four 1978 Pioneer Venus probes (the others show a similar pattern). The Jupiter data derives from *in situ* deceleration measurements of the Galileo probe. The full Mars profile dataset reveals considerable seasonal and geographical variation. The profiles shown here were taken in the Southern Hemisphere by Mars Global Surveyor. The warmest one is in the late afternoon of 1998 in the Summer subtropics while the next warmest is at night-time under otherwise similar conditions. The coldest Mars sounding shown in the plot is from the winter South Polar region. 100 $Pascal = 1\ mb$

proton, so for our purposes the mass of the molecule can be taken to be $M \cdot \mu$ where μ is the mass of a proton and M is the *molecular weight* – an integer giving the count of neutrons and protons in the molecule. (The equivalent count for an individual atom of an element is the *atomic weight*). The density is thus $\rho = n \cdot M \cdot \mu$). If we define the *Universal Gas Constant* as $R^* \equiv k/\mu$ the perfect gas equation of state can be rewritten

$$p = \frac{R^*}{M}\rho T \tag{2.2}$$

In *mks* units, $R^* = 8314.5(m/s)^2K^{-1}$ We can also define a gas constant $R = R^*/M$ particular to the gas in question. For example, dry Earth air has a mean molecular weight of 28.97, so R_{dryair} = 287 $(m/s)^2K^{-1}$, in mks units.

If μ is measured in kilograms, then $1/\mu$ is the number of protons needed to make up a kilogram. This large number is known as a *Mole*, and is commonly used as a unit of measurement of numbers of molecules, just as one commonly counts eggs by the dozen. For any substance, a quantity of that substance whose mass in kilograms is equal to the molecular weight of the substance will contain one Mole of molecules. For example, 2 kg of H_2 is a Mole of Hydrogen molecules, while 32 kg of the most common form of O_2 is a Mole of molecular Oxygen. If n were measured in $Moles/m^3$ instead of molecules per m^3, then density would be $\rho = n \cdot M$. One can also define the *gram-mole* (or *mole* for short), which is the number of protons needed to make a gram; this number is known as *Avogadro's number*, and is approximately $6.022 \cdot 10^{23}$.

Generally speaking, a gas obeys the perfect gas law when it is tenuous enough that the energy stored in forces between the molecules making up the gas is negligible. Deviations from the perfect gas law can be important for the dense atmosphere of Venus, but for the purposes of the current atmosphere of Earth or Mars, or the upper part of the Jovian or Venusian atmosphere, the perfect gas law can be regarded as an accurate model of the thermodynamics.

An extension of the concept of a perfect gas is the *law of partial pressures*. This states that, in a mixture of gases in a given volume, each component gas behaves just as it would if it occupied the volume alone. The pressure due to one component gas is called the *partial pressure* of that gas. Consider a gas which is a mixture of substance A (with molecular weight M_A) and substance B (with molecular weight M_B). The partial pressures of the two gases are

$$p_A = kn_AT, p_B = kn_BT \tag{2.3}$$

or equivalently,

$$p_A = R_A\rho_AT, p_B = R_B\rho_BT \tag{2.4}$$

where $R_A = R^*/M_A$ and $R_B = R^*/M_B$. The same temperature appears in both equations, since thermodynamic equilibrium dictates that all components of the system have the same temperature. The ratio of partial pressures of any two components of a gas is a convenient way to describe the composition of the gas. From Eq. 2.3, $p_A/p_B = n_A/n_B$, so the ratio of partial pressure of A to that of B is also the ratio of number of molecules of A to the number of molecules of B. This ratio is called the *molar mixing ratio*. When we refer to a mixing ratio without qualification, we will generally mean the molar mixing ratio. Alternately, one can describe the composition in terms of the ratio of partial pressure of one component to total presssure of the gas ($p_A/(p_A + p_B)$ in the two-component example). Summing the two partial pressure equations in Eq. 2.3, we see that this is also the ratio of number of molecules of A to total number of molecules; hence we will use the term *molar concentration* for this ratio [3]. If ξ_A is the molar mixing ratio of A to B, then the

[3]The term *volumetric* mixing ratio or concentration is often used interchangeably with the term *molar*, as in "ppmv" for "parts per million by volume." The reason for this nomenclature is that the volume occupied by a given quantity of gas at a fixed temperature and pressure is proportional to the number of molecules of the gas contained

molar concentration is $\xi_A/(1+\xi_A)$, from which we see that for the molar concentration and molar mixing ratio are nearly the same for substances which are very dilute (i.e. $\xi_A \ll 1$). We will use the notation η_A throughout the book to denote the molar concentration of substance A.

Exercise 2.3.1 Show that a mixture of gases with molar concentrations $\eta_A = n_A/(n_A + n_B)$ and $\eta_B = n_B/(n_A + n_B)$ behaves like a perfect gas with mean molecular weight $M = \eta_A M_A + \eta_B M_B$. (i.e. derive the expression relating total pressure $p_A + p_B$ to total density $\rho_A + \rho_B$ and identify the effective gas constant). Compute the mean molecular weight of dry Earth air. (Dry Earth air consists primarily of 78.084% N_2, 20.947% O_2, and .934% Ar, by count of molecules.)

The *mass mixing ratio* is the ratio of the mass of substance A to that of substance B in a given parcel of gas, i.e. ρ_A/ρ_B. From Eq. 2.4 it is related to the molar mixing ratio by

$$\frac{\rho_A}{\rho_B} = \frac{M_A}{M_B}\frac{p_A}{p_B} \tag{2.5}$$

Throughout this book,we will use the symbol r to denote mass mixing ratios and η for molar mixing ratios, with subscripts added as necessary to distinguish the species involved. Yet another measure of composition is *specific concentration*, defined as the ratio of the mass of a given substance to the total mass of the parcel (e.g. $\rho_A/(\rho_A + \rho_B)$ in the two-component case). We'll use the symbol q, with subscripts as necessary, to denote the specific concentration of a substance. Using the law of partial pressures, the specific concentration of substance A in a mixture is related to the molar concentration by

$$q_A \equiv \frac{\rho_A}{\rho_{tot}} = \frac{M_A}{\bar{M}}\frac{p_A}{p_{tot}} = \frac{M_A}{\bar{M}}\eta_A \tag{2.6}$$

where $\bar{M}$ is the mean molecular weight of the mixture, with the mean computed using weighting according to molar concentrations of the species, as in Exercise 2.3.1.

All of the ratios we have just defined are convenient to use because, unlike densities, they remain unchanged as a parcel of air expands or contracts, provided the constituents under consideration do not undergo condensation, chemical reaction or other forms of internal sources or sinks. Hence, for a compressible gas, two components A and B are well-mixed relative to each other if the mixing ratio between them is independant of position.

Constituents will tend to become well mixed over a great depth of the atmosphere if they are created or destroyed slowly, if at all, relative to the characteristic time required for mixing. In the Earth's atmosphere, the mixing ratio of oxygen to nitrogen is virtually constant up to about 80km above the surface. The mixing ratio of carbon dioxide in air can vary considerably in the vicinity of sources at the surface, such as urban areas where much fuel is burned, or under forest canopies when photosynthesis is active. Away from the surface, however, the carbon dioxide mixing ratio varies little. Variations of a few parts per million can be detected in the relatively slowly mixed stratosphere, associated with the industrial-era upward trend in fossil fuel carbon dioxide emissions. Small seasonal and interhemispheric fluctuations in the tropospheric mixing ratio, associated with variations in the surface sources, can also be detected. For most purposes, though, carbon dioxide can be regarded as well mixed throughout the atmosphere. In contrast, water vapor has a strong

in that quantity. To see this, write $n = N/V$, where N is the number of molecules and V is the volume they occupy. Then, the ideal gas law can be written in the alternate form $V = (kT/p)N$. Hence the ratio of standardized volumes is equal to the molar mixing ratio, and so forth. Abbreviations like "ppmv" for molar mixing ratios are common and convenient, because the "v" can unambiguously remind us that we are talking about a volumetric (i.e. molar) mixing ratio or concentration, whereas in an abbreviation like "ppmm" one is left wondering whether the second "m" means "mass" or "molar."

internal sink in Earth's atmosphere, because it is condensible there; hence its mixing ratio shows considerable vertical and horizontal variations. Carbon dioxide, methane and ammonia are not condensible on Earth at present, but their condensation can become significant in colder planetary atmospheres.

Exercise 2.3.2 (a) In the year 2000, the concentration of CO_2 in the atmosphere was about 370 parts per million molar. What is the ratio p_{CO_2}/p_{tot}? Estimate p_{CO_2} in mb at sea level. Does the molar concentration differ significantly from the molar mixing ratio? What is the mass mixing ratio of CO_2 in air? What is the mass mixing ratio of *carbon* (in the form of CO_2) in air – i.e. how many kilograms of carbon would have to be burned into CO_2 in order to produce the CO_2 in 1 kg of air? Note: The mean molecular weight of air is about 29. (b) The molar concentration of O_2 in Earth air is about 20%. How many grams of O_2 does a 1 liter breath of air contain at sea level (1000mb)? At the top of Qomolangma (a.k.a. "Mt. Everest," about 300mb)? Does the temperature of the air (within reasonable limits) affect your answer much?

2.3.2 Specific heat and conservation of energy

Conservation of energy is one of the three great pillars upon which the edifice of thermodynamics rests. When expressed in terms of changes in the state of matter, it is known as the First Law of Thermodynamics. When a gas expands or contracts, it does work by pushing against the environment as its boundaries move. Since pressure is force per unit area, and work is force times distance, the work done in the course of an expansion of volume dV is pdV. This is the amount of energy that must be added to the parcel of gas to allow the increase in volume to take place. For atmospheric purposes, it is more convenient to do write all thermodynamic relations on a per unit mass basis. Dividing V by the mass contained in the volume yields ρ^{-1}, whence the work per unit mass is $pd\rho^{-1}$. This is not the end of energy accounting. Changing the temperature of a unit mass of the substance while holding volume fixed changes the energy stored in the various motions of the molecules by an amount $c_v dT$, where c_v is a proportionality factor known as the *specific heat at constant volume*. For example it takes about 720 Joules of energy to raise the temperature of 1kg of air by 1K while holding the volume fixed. For ideal gases, the specific heat can depend on temperature, though the dependence is typically weak. For non-ideal gases, specific heat can depend on pressure as well.

Exercise 2.3.3 There are 20 students and one professor in a well-insulated classroom measuring 20 meters by 20 meters by 3 meters. Each person in the classroom puts out energy at a rate of 100 Watts (1 Watt = 1 Joule/second). The classroom is dark, except for a computer and LCD projector which together consume power at a rate of 200 Watts. The classroom is filled with air at a pressure of 1000mb (no extra charge). The room is sealed so no air can enter or leave, and has an initial temperature of $290K$. How much does the temperature of the classroom rise during the course of a 1 hour lecture?

Combining the two contributions to energy change we find the expression for the amount of energy that must be added per unit mass in order to accomplish a change of both temperature and volume:

$$\delta Q = c_v dT + pd\rho^{-1} \tag{2.7}$$

Using the perfect gas law, the heat balance can be re-written in the form

$$\delta Q = c_v dT + d(p\rho^{-1}) - \rho^{-1}dp = (c_v + R)dT - \rho^{-1}dp \tag{2.8}$$

From this relation, we can identify the *specific heat at constant pressure*, $c_p \equiv c_v + R$, which is the amount of energy needed to warm a unit mass by 1K while allowing it to expand enough to keep pressure constant.

The units in which we measure temperature are an artifact of the marks one researcher or other once decided to put on some device that responded to heat and cold. Since temperature is proportional to the energy per molecule of a substance, it would make sense to set the proportionality constant to unity and simply use energy as the measure of temperature. This not being common practice, one has occasion to make use of the *Boltzmann thermodynamic constant*, k, which expresses the proportionality between temperature and energy. More precisely, each degree of freedom in a system with temperature T has a mean energy $\frac{1}{2}kT$. For example, a gas made of rigid spherical atoms has three degrees of freedom per atom (one for each direction it can move), and therefore each atom has energy $\frac{3}{2}kT$ on average; a molecule which could store energy in the form of rotation or vibration would have more degrees of freedom, and therefore each molecule would have more energy at any given temperature. The energy-temperature relation is made possible by an important thermodynamic principle, the *equipartion principle*, which states that in equilibrium, each degree of freedom accessible to a system gets an equal share of the total energy of the system. In constrast to physical constants like the speed of light, the Boltzmann constant should not be considered a fundamental constant of the Universe. It is just a unit conversion factor.

2.3.3 Entropy, reversibility and Potential temperature; The Second Law

One cannot use Eqn 2.8 to define a "heat content" Q of a state (p, T) relative to a reference state (p_o, T_o), because the amount of heat needed to go from one state to another depends on the path in pressure-temperature space taken to get there; the right hand side of Eqn 2.8 is not an exact differential. However, it can be made into an exact differential by dividing the equation by T and using the perfect gas law as follows:

$$ds \equiv \frac{\delta Q}{T} = c_p \frac{dT}{T} - R\frac{dp}{p} = c_p d\ln(Tp^{-R/c_p}) \tag{2.9}$$

assuming c_p to be constant. This equation defines the *entropy*, $s \equiv c_p \ln(Tp^{-R/c_p})$. Entropy is a nice quantity to work with because it is a *state variable* – its change between two states is independent of the path taken to get from one to the other. A process affecting a parcel of matter is said to be *adiabatic* if it occurs without addition or loss of heat from the parcel. By definition, $\delta Q = 0$ for adiabatic processes. In consequence, adiabatic processes leave entropy unchanged, *provided that the changes in state of the system are slow enough that the system remains close to thermodynamic equilibrium at all times.* The latter condition is satisfied for all atmospheric phenomena that will concern us, but could be violated, for example, in the case of explosive adiabatic expansion of a formerly-confined gas into a vacuum. Entropy can also be defined for gases whose specific heat depends on temperature and pressure, for inhomogeneous mixtures of gases, and for non-ideal gases.

The *Second Law of Thermodynamics* states that entropy never decreases for energetically closed systems – systems to which energy is neither added nor subtracted in the course of their evolution. The formal derivation of the law from the microscopic properties of molecular interactions is in many ways an unfinished work of science, but the tendency towards an increase in entropy – an increase in disorder – seems to be a nearly universal property of systems consisting of a great many interacting components. A process during which the entropy remains constant is *reversible*, since it can be run both ways.

The Second Law is perhaps more intuitive when restated in the following way: *In an energetically closed system, heat flows from a hotter part of the system to a colder part of the system, causing the system to evolve toward a state of uniform temperature.* To see that this statement is equivalent to the entropy-increase principle, consider a thermally insulated box of gas having uniform pressure, but within which the left half of the mass is at temperature T_1 and the right half of the mass is at temperature $T_2 < T_1$. Now suppose that we transfer an amount of heat δQ from the left half of the box to the right half. This transfer leaves the net energy unchanged, but it changes the entropy. Specifically, according to Eq. 2.9, the entropy change summed over the two halves of the gas is $ds = (\frac{1}{T_2} - \frac{1}{T_1})\delta Q$. Since $T_2 < T_1$, this change is positive only if $\delta Q > 0$, representing a transfer from the hotter to the colder portion of the gas. Entropy can be increased by further heat transfers until $T_1 = T_2$, at which point the maximum entropy state has been attained.

The Second Law endows the Universe with an arrow of time. If one watches a movie of a closed system and sees that the system starts with large fluctuations of temperature (low entropy) and proceeds to a state of uniform temperature (high entropy), one knows that time is running forward. If one sees a thermally homogeneous object spontaneously generate large temperature inhomogeneities, then one knows that the movie is being run backwards. Note that the Second Law applies only to closed systems. The entropy of a subcomponent can decrease, if it exchanges energy with the outside world and increases the entropy of the rest of the Universe. This is how a refrigerator works.

Entropy is a very general concept, of which we have seen only the most basic instance. For a homogeneous ideal gas near thermodynamic equilibrium, the notions of reversible (i.e. isentropic) and adiabatic processes are equivalent, but caution must be exercised when extending this picture to more complex systems involving mixtures of gases. For example, if a box of gas at uniform temperature T contains pure N_2 in its left half and pure O_2 in its right half, then entropy will increase when the two gases spontaneously mix, even if no energy is let into the box. The entropy can still be defined in terms of changes in $\delta Q/T$, but it requires careful attention to what precisely is meant by δQ and to which subsystems the heat changes are being applied. The references given in the Further Readings section of this chapter provide a deeper and more general understanding of the use of entropy in solving thermodynamic problems.

Now let's get back to basics. Entropy can be used to determine how the temperature of an air parcel changes when it is compressed or expanded adiabatically. This is important because it tells us what happens to temperature is a bit of the atmosphere is lifted from low altitudes (where the pressure is high) to higher altitudes (where the pressure is lower), provided the lifting occurs so fast that the air parcel has little time to exchange heat with its surroundings. If the initial temperature and pressure are (T, p), then conservation of entropy tells us that the temperature T_o found upon adiabatically compressing or expanding to pressure p_o is given by $Tp^{-R/c_p} = T_o p_o^{-R/c_p}$ This leads us to define the *potential temperature*

$$\theta = T(\frac{p}{p_o})^{-R/c_p} \tag{2.10}$$

which is simply the temperature an air parcel would have if reduced adiabatically to a reference pressure p_o. Like entropy, potential temperature is conserved for adiabatic processes.

To understand why the presence of cold air above warm air in the sounding of Figure 2.1 does not succumb immediately to instability, we need only look at the corresponding profile of potential temperature, shown in Figure 2.3. This figure shows that potential temperature increases monotonically with height. This profile tells us that the air aloft is cold, but that if it were pushed down to lower altitudes, compression would warm it to the point that it is warmer than the

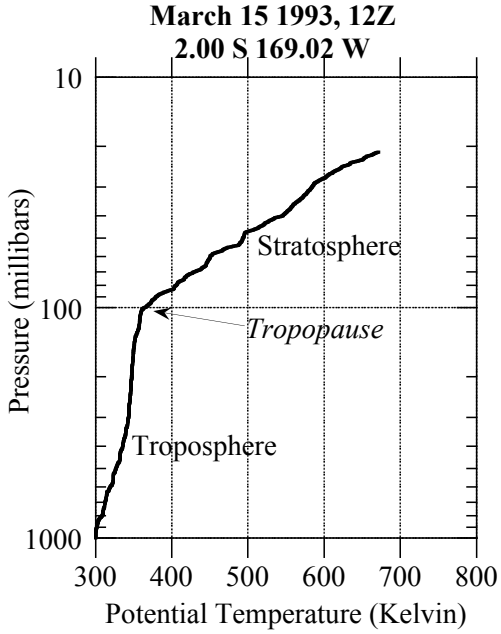

Figure 2.3: The dry potential temperature profile for the sounding in Figure 2.1

surrounding air, and thus being positively buoyant, will tend to float back up to its original level rather than continuing its descent. We see also where the stratosphere gets its name: potential temperature increases very strongly with height there, so air parcels are very resistant to vertical displacement. This part of the atmosphere is therefore strongly stratified.

The troposphere is stable, but has much weaker gradients of θ. In a compressible atmosphere, a well-stirred layer would have constant θ rather than constant T, since it is the former that is conserved for adiabatic processes such as would be caused by rapid vertical displacements. This is the essence of the explanation for why temperature decreases with height: turbulent stirring relaxes the troposphere towards constant θ, yielding the *dry adiabat*

$$T(p) = \theta \cdot (\frac{p}{p_o})^{R/c_p} \tag{2.11}$$

In this formula, θ has the constant value $T(p_o)$.

On the dry adiabat, the slope $d\ln T/d\ln p$ has the value R/c_p. From the first equality in Eq. 2.9,it can be seen that this result remains valid even if R/c_p depends on temperature and pressure. In general, the slope $d\ln T/d\ln p$ provides a convenient measure of how sharply temperature decreases in the vertical; positive values correspond to decrease of temperature with height, since pressure decreases with altitude. The *dry adiabatic slope* is then R/c_p. In atmospheric science, it is common to characterize the temperature structure by $-dT/dz$ – the *lapse rate* – but there are very few circumstances in which the altitude z is a convenient coordinate for climate calculations. Unless noted otherwise, we will (somewhat unconventionally) use the term *lapse rate* to refer to the slope $d\ln T/d\ln p$. With this definition,the dry adiabatic lapse rate would be R/c_p. (see Problem **??** for an exploration of the more conventional use of the term.)

It is evident from Figure 2.3 that something prevents θ from becoming completely well mixed. An equivalent way of seeing this is to compare the observed temperature profile with the dry adiabat. For example, if the air at 1000mb in Figure 2.3, having temperature 298K, were lifted dry-adiabatically to the tropopause, where the pressure is 100mb, then the temperature would be $298.(\frac{100}{1000})^{2/7}$, i.e. 154.3K (using the value $R/c_p = 2/7$ for Earth air). This is much colder than the observed temperature, which is 188K. We will see shortly that in the Earth's atmosphere, condensation of water vapor is one of the factors in play, though it is not the only one affecting the tropospheric temperature profile. The question of what determines the tropospheric θ gradient is at present still largely unsettled, particularly outside the Tropics.

It is no accident that the value of R/c_p for air lies close to the ratio of two small integers. It is a consequence of the equipartition principle. Using methods of statistical thermodynamics, it can be shown that a gas made up of molecules with n degrees of freedom has $R/c_p = 2/(n+2)$. Using the expression for the gas constant in terms of the specific heats, the adiabatic coefficient can also be written as $R/c_p = 1 - 1/\gamma$, where $\gamma = c_p/c_v$; for exact equipartion, $\gamma = 1 + 2/n$. The measured values of γ for a few common atmospheric gases are shown in Table 2.1. Helium comes close to the theoretical value for a molecule with no internal degrees of freedom, underscoring that excitation of electron motions plays little role in heat storage for typical planetary temperatures. The diatomic molecules have values closest to the theoretical value for $n = 5$, one short of what one would expect from adding two rotational and one vibrational internal degrees of freedom. Among the triatomic molecules, water acts roughly as if it had $n = 6$ while carbon dioxide is closer to $n = 7$. The two most complex molecules, methane and ammonia, are also characterized by $n = 7$. The failure of thermodynamics to access all the degrees of freedom classically available to a molecule is a consequence of quantum theory. Since the energy stored in states of motion of a molecule in fact comes in discrete-sized chunks, or "quanta," one can have a situation where a molecule hardly ever gets enough energy from a collision to excite even a single vibrational degree of freedom, for example, leading to the phenomenon of partial excitation or even non-excitation of certain classical degrees of freedom. This is one of many ways that the quantum theory, operating on exceedingly tiny spatial scales, exerts a crucial control over macroscopic properties of matter that can effect the very habitability of the Universe. Generally speaking, the higher the temperature gets, the more easy it is to excite internal degrees of freedom, leading to a decrease in γ. This quantum effect is the chief reason that specific heats vary somewhat with temperature.

Exercise 2.3.4 (a) A commercial jet airliner cruises at an altitude of 300mb. The air outside has a temperature of 240K. To enable the passengers to breathe, the ambient air is compressed to a cabin pressure of 1000mb. What would the cabin temperature be if the air were compressed adiabatically? How do you think airlines deal with this problem? (b) Discuss whether the lower portion of the Venus temperature profile shown in Figure 2.2 is on the dry CO_2 adiabat. Do the same for the Summer afternoon Mars sounding. (c) Assume that the Jupiter sounding is on a dry adiabat, and estimate the value of R/c_p for the atmosphere. Based on your result, what is the dominant constituent of the Jovian atmosphere likely to be? What other gas might be mixed with the dominant one?

2.4 Static stability of inhomogeneous mixtures

An atmosphere is *statically unstable* if an air parcel displaced from its original position tends to continue rising or sinking instead of returning to its original position. Such a state will tend to mix itself until it becomes stable. Static stability is important to planetary climate because it affects the vertical mixing which creates a planet's troposphere.

	H_2O	CH_4	CO_2	N_2	O_2	H_2	He	NH_3
Crit. point T	647.1	190.44	304.2	126.2	154.54	33.2	5.1	405.5
Crit. point p	221.e5	45.96e5	73.825e5	34.0e5	50.43e5	12.98e5	2.28e5	112.8
Triple point T	273.15	90.67	216.54	63.14	54.3	13.95	2.17	195.4
Triple point p	611.	.117e5	5.185e5	.1253e5	.0015e5	.072e5	.0507e5	.061e5
L vap(b.p.)	22.55e5	5.1e5	–	1.98e5	2.13e5	4.54e5	.203e5	13.71e5
L vap(t.p.)	24.93e5	5.36e5	3.97e5	2.18e5	2.42e5	??	??	16.58e5
L fusion	3.34e5	.5868e5	1.96e5	.2573e5	.139e5	.582e5	??	3.314e5
L sublimation	28.4e5	5.95e5	5.93e5	2.437e5	2.56e5	??	??	19.89e5
ρ liq(b.p.)	958.4	450.2	1032.	808.6	1141.	70.97	124.96	682.
ρ liq(t.p.)	999.87	??	1110.	??	1307.	??	??	734.2
ρ solid	917.	509.3	1562.	1026.	1351.	88.	200.	822.6
c_p(0C/1bar)	1847.	2195.	820.	1037.	916.	14230.	5196.	2060.
$\gamma(c_p/c_v)$	1.331	1.305	1.294	1.403	1.393	1.384	1.664	1.309

Table 2.1: Thermodynamic properties of selected gases. Latent heats of vaporization are given at both the boiling point (the point where saturation vapor pressure reaches $1bar$) and the triple point. Liquid densities are given at the boiling point and the triple point. For CO_2 the 'boiling point' is undefined, so the liquid density is given at $253K/20bar$ instead. Note that the maximum density of liquid water is $1000.00kg/m^3$ and occurs at $-4C$. Densities of solids are given at or near the triple point. All units are mks, so pressures are quoted as Pa with the appropriate exponent. Thus, $1bar$ is written as $1e5$ in the table.

For a well-mixed atmosphere, the potential temperature profile tells the whole story about static stability, since, according to the ideal gas law, the density of an air parcel with potential temperature θ_0 will be $\rho_0 = p_1/(R\theta_0 \cdot (p_1/p_0)^{R/c_p})$ upon being elevated to an altitude with pressure $p_1 < p_0$. The ambient density there is $\rho_1 = p_1/(R\theta_1 \cdot (p_1/p_0)^{R/c_p})$. The displaced parcel will be negatively buoyant and return toward its original position if $\rho_0 > \rho_1$, which is true if and only if $\theta_0 < \theta_1$, i.e. if the potential temperature increases with height.

For an inhomogeneous atmosphere, this is no longer the case, since the gas constant R depends on the mean molecular weight of the mixture, which varies from place to place. As an example, we may consider an atmosphere which has uniform θ computed on the basis of a reference pressure p_0, but which consists of pure N_2 for $p > p_0$ and pure CO_2 for $p < p_0$. One immediately notes that the system has an unstable density jump at p_o, since the density is $p_o/R_{N_2}\theta$ just below the interface and $p_o/R_{CO_2}\theta$ just above the interface. Since N_2 has lower molecular weight (28) than CO_2 (44), the gas constant for N_2 is considerably greater than the gas constant for CO_2. Given that $R_{N_2} > R_{CO_2}$, the density is greater just above the interface than it is just below the interface. This is an unstable situation, and the N_2 layer will tend to mix itself into the CO_2, despite the constancy of θ.

The phenomenon is very familiar: it is why helium balloons rise in air, even when they are at the same temperature as their surroundings. The low molecular weight of helium makes it lighter (i.e. lower density) than air having the same temperature and pressure.

Exercise 2.4.1 Make sense of the following statement: "For the Earth's atmosphere, moist air is lighter than dry air." Would this still be true for a planet whose atmosphere is mainly H_2?

Suppose now that a parcel of N_2 is lifted from just below the interface to an altitude where the pressure is some lower pressure p_1. Let the density of the parcel when it arrives at its destination

be ρ_{lift} and the density of the ambient CO_2 be ρ_{amb}. The density difference is then

$$\rho_{lift} - \rho_{amb} = \frac{p_1}{\theta}\left(\frac{1}{R_{N_2}(p_1/p_0)^{-(R/c_p)_{N_2}}} - \frac{1}{R_{CO_2}(p_1/p_0)^{-(R/c_p)_{CO_2}}}\right) \tag{2.12}$$

The biggest effect in this equation comes from the fact that $R_{N_2} > R_{CO_2}$, which assures that the N_2 retains its buoyancy as it is lifted. A much weaker effect comes from the fact that R/c_p is slightly greater for N_2 (.286) than for CO_2 (.230). This modulates the density difference as the parcel is lifted, but the effect is slight. Even lifting a great distance, so that $p_1/p_0 = .01$, the pressure factor in the first term is only modestly greater than the pressure factor in the second term (3.74 vs. 2.89), yielding a modest reduction in the buoyancy of the lifted parcel. For other pairs of gases, the difference in R/c_p could be more significant, and in principle it could also go in the opposite direction and increase buoyancy rather than reducing it.

For an arbitrary profile of composition and temperature, one can define a *potential density*, which is the density an air parcel would have if compressed or expanded adiabatically to a standard reference pressure p_o. Using the gas law, and the fact that mixing ratios are conserved (whence R/c_p is conserved on adiabatic compression or expansion of the parcel), the potential density relative to reference level p_o is

$$\begin{aligned}\tilde{\rho}(p|p_o) &= \frac{p_o}{R\theta} \\ &= \frac{p_o}{RT}\left(\frac{p}{p_o}\right)^{R/c_p} = \rho(p)\left(\frac{p}{p_o}\right)^{R/c_p - 1}\end{aligned} \tag{2.13}$$

The R and c_p in this equation must be taken from the values prevailing at pressure p, since these values are determined by composition, which in this calculation is presumed to remain fixed as the parcel is displaced to the reference pressure p_o. When $\tilde{\rho}(p|p_o)$ increases with p, then the system will be stable in the sense that a parcel from $p < p_o$ will be positively buoyant when pushed down to p_o and thus tend to return to its original level; similarly a parcel lifted to p_o from some pressure $p > p_o$ will be negatively buoyant and also tend to return to where it came from. When the atmosphere is homogeneous, R/c_p is constant and one can determine whether a parcel at pressure p_1 will be buoyant when displaced adiabatically to p_2 by adiabatically reducing both to the standard pressure p_o and comparing the densities there. In other words, for a homogeneous atmosphere, one can tell immediately from a single plot of potential density relative to p_o which regions are stable and which regions will tend to overturn. When the atmosphere is inhomogeneous, this is no longer the case. When a parcel with density $\rho(p_1)$ is displaced to p_2, it will be positively buoyant if

$$\rho(p_1)\left(\frac{p_1}{p_2}\right)^{R(p_1)/c_p(p_1)-1} < \rho(p_2) \tag{2.14}$$

However, if both parcels are reduced to a common pressure p_o, then the first parcel would be buoyant relative to the second if

$$\rho(p_1)\left(\frac{p_1}{p_o}\right)^{R(p_1)/c_p(p_1)-1} < \rho(p_2)\left(\frac{p_2}{p_o}\right)^{R(p_2)/c_p(p_2)-1} \tag{2.15}$$

i.e.

$$\rho(p_1)\left(\frac{p_1}{p_o}\right)^{R(p_1)/c_p(p_1)-1}\left(\frac{p_o}{p_2}\right)^{R(p_2)/c_p(p_2)-1} < \rho(p_2) \tag{2.16}$$

This yields the same criterion as the correct criterion given in Eq. 2.14 only if R/c_p is constant.

The lack of a globally valid potential density complicates the precise analysis of the static stability of inhomogeneous atmospheres. Strictly speaking, one needs to examine potential density

profiles for a range of different p_o covering the atmosphere. In practice, the potential density based on a single p_o can often provide a useful indication of static stability within a reasonably thick layer of the atmosphere because the variations of R/c_p are usually slight unless the compositional variations are extreme.

An alternate approach to dealing with the effect of composition on buoyancy is to define a modified potential temperature based on a *virtual temperature*. The virtual temperature is the temperature at which the gas law for a standard composition (e.g. dry air) would yield the same density as the true gas law taking into account the effect of the actual composition on density (e.g. moist air). This approach is a common way of dealing with the effect of water vapor on buoyancy in Earth's atmosphere. It is equivalent to the potential density approach and shares the same limitations. Making use of a virtual temperature can be convenient for some purposes, but it can also be confusing in that one needs to keep track of the contexts in which one uses the virtual temperature and the ones in which one must use the actual temperature.

In planetary atmospheres, compositional variations often arise as a result of condensation, evaporation or sublimation of one of the atmospheric components. For example, the Martian atmosphere contains a moderate amount of Ar, and when CO_2 condenses at the winter pole, it leaves an Ar-enriched layer near the surface. Because Ar has higher molecular weight than CO_2, this layer tends to be statically stable. When CO_2 sublimates from the polar glacier in the Spring, the pure CO_2 layer resulting has enhanced buoyancy relative to the Ar-rich atmosphere above. Similarly, the evaporation of light CH_4 from the surface of Titan makes the low level air positively buoyant with respect to the overlying N_2, favoring upward mixing. Water vapor has a small but significant effect on buoyancy on the present Earth; the importance of the effect would increase sharply under warmer conditions in which the water vapor content of the atmosphere is greater. Condensation can also lead to compositional variations in the interior of atmospheres, and in some circumstances it may be necessary to take these into account when determining where the vertical mixing that creates the troposphere takes place. In principle, chemical reactions could also create compositional gradients strong enough to create buoyancy, though there do not at present appear to be any atmospheres where this is known to be a significant effect.

2.5 The hydrostatic relation

The hydrostatic relation relates pressure to altitude and the mass distribution of the atmosphere, and provides the chief reason that pressure is the most natural vertical coordinate to use in most atmospheric problems. Consider a column of any substance at rest, and suppose that the density of the substance as a function of height z is given by $\rho(z)$. Suppose further that the range of altitudes being considered is small enough that the acceleration of gravity is essentially constant; The magnitude of this acceleration will be called g, and the force of gravity is taken to point along the direction of decreasing z. Now, consider a slice of the column with vertical thickness dz, having cross sectional area A in the horizontal direction. Since pressure is simply force per unit area, then the change in pressure from the base of this slice to the top of this slice is just the force exerted by the mass. By Newton's law, then, we have

$$Adp = -Agdm = -Ag\rho dz \tag{2.17}$$

where dm is the increment of mass in the column per unit area. An immediate consequence of this relation is that

$$dm = -\frac{dp}{g} \tag{2.18}$$

which states that the amount of mass in a slab of atmosphere is proportional to the thickness of that slab, measured in pressure coordinates. A further consequence, upon dividing by dz is the relation

$$\frac{dp}{dz} = -\rho g \tag{2.19}$$

This differential equation expresses the *hydrostatic relation*. It is exact if the substance is at rest (hence the "static"), but if the material of the column is in motion, the relation is still approximately satisfied provided the acceleration is sufficiently small, compared to the acceleration of gravity. In practice, the hydrostatic relation is very accurate for most problems involving large scale motions in planetary atmospheres. It would not be a good approximation within small scale intense updrafts or downdrafts where the acceleration of the fluid may be large. Derivation of the precise conditions under which the hydrostatic approximation holds requires consideration of the equations of fluid motion, which will be taken up in a sequel to the present book.

An important consequence of the hydrostatic relation is that it enables us to determine the total mass of an atmosphere through measurements of pressure taken at the surface alone. Integrating Eqn 2.18 from the ground ($p = p_s$) to space ($p = 0$) yields the relation

$$m = \frac{p_s}{g} \tag{2.20}$$

where m is the total mass of the atmosphere located over a unit area of the planet's surface. Note that this relation presumes that the depth of the layer containing almost all the mass of the atmosphere is sufficiently shallow that gravity can be considered constant throughout the layer. Given that gravity decays inversely with the square of distance from the planet's center, this is equivalent to saying that the atmosphere must be shallow compared to the radius of the planet. For a well mixed substance A with mass-specific concentration κ_A relative to the whole atmosphere, the mass of substance A per square meter of the planet's surface is just $m\kappa_A$

Using the perfect gas law to eliminate ρ from Eqn 2.19 yields

$$\frac{dp}{dz} = -\frac{g}{RT}p \tag{2.21}$$

where R is the gas constant for the mixture making up the atmosphere. This has the solution

$$p(z) = p_s \exp(-\frac{g}{R\bar{T}}z), \bar{T}(z) = (\frac{1}{z}\int_0^z T^{-1}dz)^{-1} \tag{2.22}$$

Here, $\bar{T}(z)$ is the harmonic mean of temperature in the layer between the ground and altitude z. If temperature is constant, then pressure decays exponentially with scale height RT/g. Because temperature is measured relative to absolute zero, the mean temperature $\bar{T}(z)$ can be relatively constant despite fairly large variations of temperature within the layer. In consequence, pressure typically decays roughly exponentially with height even when temperature is altitude-dependent.

Exercise 2.5.1 Compute the mass of the Earth's atmosphere, assuming a mean surface pressure of 1000mb. (The Earth's radius is 6378km, and the acceleration of its gravity is $9.8m/s^2$). Compute the mass of the Martian atmosphere, assuming a mean surface pressure of 6mb. (Mars' radius is 3390km, and the acceleration of its gravity is $3.7m/s^2$.)

Note that the hydrostatic relation applies only to the total pressure of all constituents; it does not apply to partial pressures individually. However, in the special case in which the gases are well mixed, the total mass of each well-mixed component can still be determined from surface data

alone. One simply multiplies the total mass obtained from surface pressure, by the appropriate (constant) mass-specific concentration. It is important to recognize, however, that even for a well mixed gas the mass per unit area of a constituent with partial pressure p_A is *not* simply p_A/g; it is $(M_A/\overline{M})(p_A/g)$, where M_a is the molecular weight of substance a and $\overline{M}$ is the mean molecular weight of the mixture, which varies according to how much of A is added to the mixture. This becomes especially important to keep straight when dealing with situations in which a certain amount of, say, CO_2 is added to an atmosphere. Sometimes when one talks about adding "1 *bar* of CO_2 to an atmosphere" it means adding an amount of CO_2 that would increase the partial pressure of CO_2 by 1 *bar*; however, a phrase like this is often used to mean adding an amount of CO_2 that would have a surface pressure of 1 *bar* if the CO_2 were alone in the atmosphere. The two things are not the same, as illustrated by the following exercise, which also makes the point that the partial pressure of the rest of the atmosphere does not remain fixed as one adds or takes away CO_2, or any other gas with a different molecular weight than the mean.

Exercise 2.5.2 Suppose an atmosphere initially consists of 1 *bar* of pure N_2. One then adds an amount of CO_2 to the atmosphere that would yield a surface pressure of 1 *bar* if it were all alone in an atmosphere. Show that at the end of this process the surface pressure is 2 *bar*, the partial pressure of CO_2 is 0.76 *bar*, the partial pressure of N_2 is 1.24 *bar*, and the molar concentration of CO_2 is 38%.

In the study of atmospheric dynamics, the hydrostatic equation is used to compute the pressure gradients which drive the great atmospheric circulations. Outside of dynamics, there are rather few problems in physics of climate that require one to know the altitude corresponding to a given pressure level. Our main use of the hydrostatic relation in this book will be in the form of Eqn 2.18, which tells us the mass between two pressure surfaces.

The hydrostatic relation also allows us to derive a useful alternate form of the heat budget, by re-writing the heat balance equation as follows:

$$\delta Q = c_p dT - \rho^{-1} dp = c_p dT - \rho^{-1} \frac{dp}{dz} dz = d(c_p T + gz) \tag{2.23}$$

assuming c_p to be constant. The quantity $c_p T + gz$ is known as the *dry static energy*. Dry static energy provides a more convenient basis for atmospheric energy budgets than entropy, since changes in dry static energy following an air parcel are equal to the net energy added to or removed from the parcel by heat sources such as solar radiation. For example, if there are no horizontal transports and if there is no net flux of energy between the atmosphere and the underlying planetary surface, then the rate of change of the net dry static energy in a (dry) atmospheric column is the difference between the rate at which solar energy flows into the top of the atmosphere and the rate at which infrared radiation leaves the top of the atmosphere; one needs to know nothing about how the heat is deposited within the atmosphere in order to determine how the net dry static energy changes. This is not the case for entropy.

Note that the dry static energy as defined above is actually the energy *per unit mass* of atmosphere. Thus, the total energy in a column of atmosphere, per unit surface area, is $\int (c_p T + gz)\rho dz$, which by the hydrostatic relation is equal to $\int (c_p T + gz) dp/g$ if the pressure integral is taken in the direction of increasing pressure.

2.6 Thermodynamics of phase change

When a substance changes from one form to another (e.g. water vapor condensing into liquid water or gaseous carbon dioxide condensing into dry ice) energy is released or absorbed even if the temperature of the mass is unchanged after the transformation has taken place. This happens because the amount of energy stored in the form of intermolecular interactions is generally different from one form, or *phase* to another. The amount of energy released when a unit of mass of a substance changes from one phase to another, holding temperature constant, is known as the *latent heat* associated with that phase change. By convention, latent heats are stated as positive numbers, with the phase change going in the direction that releases energy. Phase changes are *reversible*. If one kilogram of matter releases L joules of energy in going from phase A to phase B, it will take the same L joules of energy to turn the mass back into phase A. The units of latent heat are energy per unit mass (Joules per kilogram in mks units).

Condensible substances play a central role in the atmospheres of many planets and satellites. On Earth, it is water that condenses, both into liquid water and ice. On Mars, CO_2 condenses into dry ice in clouds and in the form of frost at the surface. On Jupiter and Saturn, not only water but ammonia (NH_3) and a number of other substances condense. The thick clouds of Venus are composed of condensed sulfuric acid. On Titan it is methane, and on Neptune's moon Triton nitrogen itself condenses. Table 2.1 lists the latent heats for the liquid-vapor (evaporation), liquid-solid (fusion) and solid-vapor (sublimation) phase transitions are given for a number of common constituents of planetary atmospheres. Water has an unusually large latent heat; the condensation of 1 kg of water vapor into ice releases nearly five times as much energy as the condensation of 1kg of carbon dioxide gas into dry ice. This is why the relatively small amount of water vapor in Earth's present atmosphere can nonetheless have a great effect on atmospheric structure and dynamics. Ammonia also has an unusually large latent heat, though not so much so as water. In both cases, the anomalous latent heat arises from the considerable energy needed to break hydrogen bonds in the condensed phase.

Like most thermodynamic properties, latent heat varies somewhat with temperature. For example, the latent heat of vaporization of water is $2.5 \cdot 10^6 J/kg$ at 0C, but only $2.26 \cdot 10^6 J/kg$ at 100C. For precise calculations, the variation of latent heat must be taken into account, but nonetheless for many purposes it will be sufficient to assume latent heat to be constant over fairly broad temperature ranges.

The three main phases of interest are solid, liquid and gas (also called vapor), though other phases can be important in exotic circumstances. There is generally a *triple point* in temperature-pressure space where all three phases can co-exist. Above the triple point temperature, the substance undergoes a vapor-liquid phase transition as temperature is decreased or pressure is increased; below the triple point temperature vapor condenses directly into solid, once thermodynamic equilibrium has been attained. For water, the triple point occurs at a temperature of 273.15K and pressure of 6.11mb (see Table 2.1 for other gases). Generally, the triple point temperature can also be taken as an approximation to the "freezing point" – the temperature at which a liquid becomes solid – because the freezing temperature varies only weakly with pressure until very large pressures are reached. Though we will generally take the freezing point to be identical to the triple point in our discussions, the effect of pressure on freezing of liquid can nonetheless be of great importance at the base of glaciers and in the interior of icy planets or moons, and perhaps also in very dense,cold atmospheres.

Typically, the solid phase is more dense than the liquid phase, but water again is exceptional. Water ice floats on liquid water, whereas carbon dioxide ice would sink in an ocean of liquid carbon

dioxide, and methane ice would sink in a methane lake on Titan. This has profound consequences for the climates of planets with a water ocean such as Earth has, since ice formed in winter remains near the surface where it can be more readily melted when summer arrives.

Exercise 2.6.1 Per square meter, how many Joules of energy would be required to evaporate a puddle of Methane on Titan, having a depth of 20m?

Atmospheres can transport energy from one place to another by heating an air parcel by an amount δT, moving the parcel vertically or horizontally, and then cooling it down to its original temperature. This process moves an amount of heat $c_p \delta T$ per unit mass of the parcel. Latent heat provides an alternate way to transport energy, since energy can be used to evaporate liquid into an air parcel until its mixing ratio increases by δr, moving it and then condensing the substance until the mixing ratio returns to its original value. This process transports an amount of heat $L\delta r$ per unit mass of the planet's uncondensible air, and can be much more effective at transporting heat than inducing temperature fluctuations, especially when the latent heat is large. "Ordinary" heat – the kind that feels hot when you touch it, and which is stored in the form of the temperature increase of a substance – is known in atmospheric circles as "sensible" heat.

All gases are condensible at low enough temperatures or high enough pressures. On Earth (in the present climate) CO_2 is not a condensible substance, but on Mars it is. The ability of a gas to condense is characterized by the *saturation vapor pressure*, p_{sat} of that gas, which may be a function of any number of thermodynamic variables. When the partial pressure p_A of gas A is below $p_{sat,A}$, more of the gas can be added, raising the partial pressure, without causing condensation. However, once the partial pressure reaches $p_{sat,A}$, any further addition of A will condense out. The state $p_A = p_{sat,A}$ is referred to as "saturated" with regard to substance A. Each condensed state (e.g. liquid or solid) will have its own distinct saturation vapor pressure. Rather remarkably, for a mixture of perfect gases, the saturation vapor pressure of each component is independent of the presence of the other gases. Water vapor mixed with 1000 mb worth of dry air at a temperature of 300K will condense when it reaches a partial pressure of 38mb; a box of pure water vapor at 300K condenses at precisely the same 38mb. If a substance "A" has partial pressure p_A that is below the saturation vapor pressure, it is said to be "subsaturated," or "unsaturated." The degree of subsaturation is measured by the *saturation ratio* $p_a/p_{sat,A}$, which is often stated as a percent. Applied to water vapor, this ratio is called the *relative humidity*, and one often speaks of the relative humidity of other substances, e.g. "methane relative humidity" instead of saturation ratios. Note that the relative humidity is also equal to the *mixing ratio* of the substance A in a given mixture to the *mixing ratio* the air would have if the substance were saturated. This is different from the ratio of *specific humidity* to *saturation specific humidity*, or the ratio of *molar concentration* to *saturation molar concentration* except when the mixing ratio is small.

It is intuitively plausible that the saturation vapor pressure should increase with increasing temperature, as molecules move faster at higher temperatures, making it harder for them to stick together to form condensate. The temperature dependence of saturation vapor pressure is expressed by a remarkable thermodynamic relation known as the *Clausius-Clapeyron equation*. It is derived from very general thermodynamic principles, via a detailed accounting of the work done in an reversible expansion-contraction cycle crossing the condensation threshold, and requires neither approximation nor detailed knowledge of the nature of the substance condensing. The relation reads

$$\frac{dp_{sat}}{dT} = \frac{1}{T}\frac{L}{\rho_v^{-1} - \rho_c^{-1}} \tag{2.24}$$

where ρ_v is the density of the less condensed phase, ρ_c is the density of the more condensed phase, and L is the latent heat associated with the transformation to the more condensed phase. For vapor

to liquid or solid transitions, $\rho_c \gg \rho_v$, enabling one to ignore the second term in the denominator of Eqn 2.24. Further, upon substituting for density from the perfect gas law, one obtains the simplified form

$$\frac{dp_{sat}}{dT} = \frac{L}{R_A T^2} p_{sat} \tag{2.25}$$

where R_A is the gas constant for the substance which is condensing. If we make the approximation that L is constant, then Eqn 2.24 can be integrated analytically, resulting in

$$p_{sat}(T) = p_{sat}(T_o) e^{-\frac{L}{R_A}(\frac{1}{T} - \frac{1}{T_o})} \tag{2.26}$$

where T_o is some reference temperature. This equation shows that saturation water vapor content is very sensitive to temperature, decaying rapidly to zero as temperature is reduced and increasing rapidly as temperature is increased. The rate at which the change occurs is determined by the characteristic temperature $\frac{L}{R_A}$ appearing in the exponential. For the transition of water vapor to liquid, it has the value 5420K at temperatures near 300K. For CO_2 gas to dry ice, it is $3138K$, and for methane gas to liquid methane it is $1031K$. Equation 2.25 seems to imply that the p_{sat} asymptotes to a constant value when $T \gg L/R_A$. This is a spurious limit, though, since the assumption of constant L invariably breaks down over such large temperature ranges. In fact, L typically approaches zero at some *critical temperature*, where the distinction between the two phases disappears. For water vapor, this *critical point* occurs at a temperature and pressure of 647.1K and 221bars. For carbon dioxide, the critical point occurs for the vapor-liquid transition, at 304.2K and 73.825 bars. Critical points for other atmospheric gases are shown in Table 2.1. At high pressures, the solid/liquid phase boundary does not typically terminate in a critical point, but instead gives way to a bewildering variety of distinct solid phases distinguished primarily by crystal structure.

Exercise 2.6.2 Show that the slope $d \ln p_{sat}/dT$ becomes infinite as $T \to 0$. Show that it decreases monotonically with T provided the latent heat decreases or stays constant as T increases. Show that the curve $p_{sat}(T)$ is infinitely flat near $T = 0$, in the sense that all the derivitives $d^n p_{sat}/dT^n$ vanish there. In Fig. 2.4 why is the curvature of the phase boundary sketched the way it is at low temperature?

Figure 2.4 summarizes the features of a typical phase diagram. Over ranges of a few bars of pressure, the solid-liquid boundary can be considered nearly vertical. In fact the exact form of the Clausius-Clapeyron relation (Eq. 2.24) tells us why the boundary is nearly vertical and how it deviates from verticality. Because the difference in density between solid and liquid is typically quite small while the latent heat of fusion is comparatively large, Eq. 2.24 implies that the slope dp/dT is very large (i.e. nearly vertical. The equation also tells us that in the "normal" case where ice is denser than liquid, the phase boundary tilts to the right, and so the freezing temperature increases with pressure; at fixed pressure, one can cause a cold liquid to freeze by squeezing it. The unusual lightness of water ice relative to the liquid phase implies that instead the phase boundary tilts to the left; one can melt solid ice by squeezing it. Substituting the difference in density between water ice and liquid water, and the latent heat of fusion, into Eq. 2.24, we estimate that $100 bars$ of pressure decreases the freezing point temperature by about $.74K$. This is roughly the pressure caused by about a kilometer of ice on Earth. The effect is small, but can nonetheless be significant at the base of thick glaciers.

Below the triple point temperature, the favored transition is gas/solid,and so the appropriate latent heat to use in the Clausius-Clapeyron relation is the latent heat of sublimation. Above the triple point, the favored transition is gas/liquid, whence one should use the latent heat of

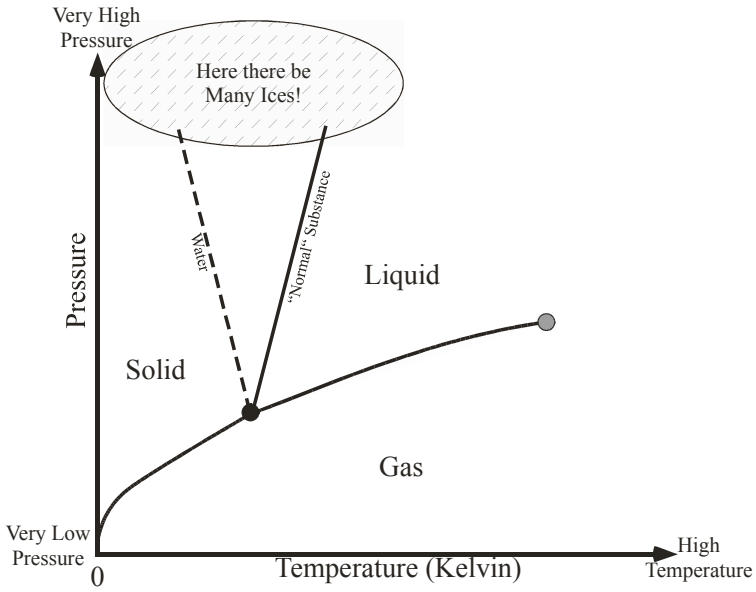

Figure 2.4: The general form of a phase diagram showing the regions of temperature-pressure space where a substance exists in solid, liquid or gaseous forms. The triple point is marked with a black circle while the critical point is marked with a grey circle. The solid-liquid phase boundary for a "normal" substance (whose solid phase is denser than its liquid phase) is shown as a solid curve, whereas the phase boundary for water (ice less dense than liquid) is shown as a dashed curve. The critical point pressure is typically several orders of magnitude above the triple point pressure, while the critical point temperature is generally only a factor of two or three above the triple point temperature. Therefore, the pressure axis on this diagram should be thought of as logarithmic, while the temperature axis should be thought of as linear. This choice of axes also reflects the fact that the pressure must typically be changed by an order of magnitude or more to cause a significant change in the temperature of the solid/liquid phase transition.

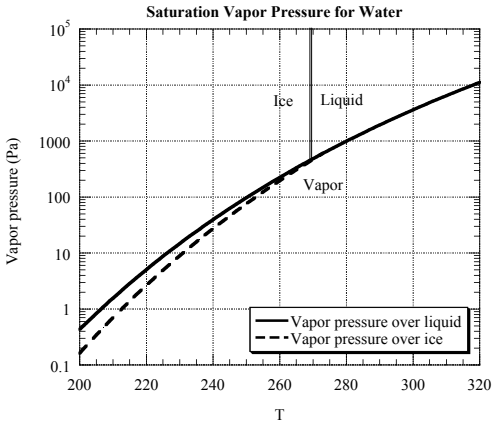

Figure 2.5: Saturation vapor pressure for water, based on the constant-L form of the Clausius-Clapeyron relation. Curves are shown for vapor pressure based on the latent heat of vaporization, and (below freezing) for latent heat of sublimation. The latter is the appropriate curve for subfreezing temperatures.

vaporization. The triple point (T, p) provides a convenient base for use with the simplified Clausius-Clapeyron solution in Eqn. 2.26, or indeed for a numerical integration of the relation with variable L. Results for water vapor are shown in Figure 2.5. These results were computed using the constant L approximation for sublimation and vaporization, but in fact a plot of the empirical results on a logarithmic plot of this type would not be distinguishable from the curves shown. The more exact result does differ from the constant L idealization by a few percent, which can be important in some applications. Be that as it may, the figure reveals the extreme sensitivity of vapor pressure to temperature. The vapor pressure ranges from about .1 Pascals at 200K (the tropical tropause temperature) to 35mb at a typical tropical surface temperature of 300K, rising further to 100mb at 320K. Over this span of temperatures, water ranges from a trace gas to a major constituent; at temperatures much above 320K, it rapidly becomes the dominant constituent of the atmosphere. Note also that the distinction between the ice and liquid phase transitions has a marked effect on the vapor pressure. Because the latent heat of sublimation is larger than the latent heat of vaporization, the vapor pressure over ice is lower than the vapor pressure over liquid would be, at subfreezing temperatures. At 200K, the ratio is nearly a factor of three.

Exercise 2.6.3 Let's consider once more the case of the airliner cruising at an altitude of 300mb, discussed in an earlier Exercise. Suppose that the ambient air at flight level has 100% relative humidity. What is the relative humidity once the air has been brought into the cabin, compressed to 1000mb, and chilled to a room temperature of 290K?

Once the saturation vapor pressure is known, one can compute the molar or mass mixing ratios with respect to the background non-condensible gas, if any, just as for any other pair of gases. The saturation vapor pressure is used in this calculation just like any other partial pressure. For example, the molar mixing ratio is just p_{sat}/p_a, if p_a is the partial pressure of the noncondensible background. Note that, while the saturation vapor pressure is independent of the pressure of the gas with which the condensible substance is mixed, the saturation mixing ratio is not.

Exercise 2.6.4 What is the saturation molar mixing ratio of water vapor in air at the ground in

tropical conditions (1000mb and 300K)? What is the mass mixing ratio? What is the mass-specific humidity? What is the molar mixing ratio (in ppm) of water vapor in air at the tropical tropopause (100mb and 200K)?

2.7 The moist adiabat

When air is lifted, it cools by adiabatic expansion, and if it gets cold enough that one of the components of the atmosphere begins to condense, latent heat is released. This makes the lifted air parcel warmer than the dry adiabat would predict. Less commonly, condensation may occur as a result of subsidence and compression, since the increase of partial pressure of one of the compressed atmospheric components may overwhelm the increase in saturation vapor pressure resulting from adiabatic warming. Whichever direction leads to condensation, if we assume further that the condensation is efficient enough that it keeps the system at saturation, the resulting temperature profile will be referred to as the *moist adiabat*, regardless of whether the condensing substance is water vapor (as on Earth) or something else (CO_2 on Mars or methane on Titan). We now proceed to make this quantitative.

2.7.1 One-component condensible atmospheres

The simplest case to consider is that of a single component atmosphere, which can attain cold enough temperatures to reach saturation and condense. This case is relevant to present Mars, which has an almost pure CO_2 atmosphere that can condense in the cold Winter hemisphere and at upper levels at any time of year. A pure CO_2 atmosphere with a surface pressure on the order of two or three bars is a commonly used model of the atmosphere of Early Mars, though the true atmospheric composition in that instance is largely a matter of speculation. Another important application of a single component condensible atmosphere is the pure steam (water vapor) atmosphere, which occurs when a planet with an ocean gets warm enough that the mass of water which evaporates into the atmosphere dominates the other gases that may be present. This case figures prominenently in the *runaway greenhouse* effect that will be studied in Chapter 4.

For a single component atmosphere, the partial pressure of the condensible substance is in fact the total atmospheric pressure. Therefore, at saturation, the pressure is related to the temperature by the Clausius-Clapeyron relation. To find the saturated moist adiabat, we simply solve for T in terms of p_{sat} in the Clausius-Clapeyron relation, and recall that $p = p_{sat}$ because we are assuming the atmosphere to be saturated everywhere. Using the simplified form of Clausius-Clapeyron given in Eqn 2.26, the saturated moist adiabat is thus

$$T(p) = \frac{T_o}{1 - \frac{RT_o}{L} \ln \frac{p}{p_{sat}(T_o)}} \tag{2.27}$$

where R is the gas constant for the substance making up the atmosphere. This equation is really just an alternate form of Clausius-Clapeyron. It can be thought of as a formula for the "dew-point" or "frost-point" temperature corresponding to pressure p: given a box of gas at fixed pressure p, condensation will occur when the temperature is made lower than the temperature $T(p)$ given by Eq. 2.27.

Without loss of generality, we may suppose that T_o is taken to be the surface temperature, so that $p_{sat}(T_o)$ is the surface pressure p_s. Since the logarithm is negative, the temperature decreases with altitude (recalling that lower pressure corresponds to higher altitude). Further, the factor

multiplying the logarithm is the ratio of the surface temperature to the characteristic temperature L/R. Since the characteristic temperature is large, the prefactor is small, and as a result the temperature of saturated adiabat for a one-component atmosphere varies very little over a great range of pressures. For example, in the case of the CO_2 vapor-ice transition, an atmospheric surface pressure of $7mb$ (similar to that of present Mars) would be in equilibrium with a surface dry-ice glacier at a temperature of $149K$; at $.07mb$ – one one-hundredth of the surface pressure – the temperature on the saturated adiabat would only fall to $122K$.

Exercise 2.7.1 In the above example, what would the temperature aloft have been if there were no condensation and the parcel were lifted along the dry adiabat?

The criterion determining whether condensation occurs on ascent or descent for an arbitrary one-component atmosphere is derived in Problem **??**. Note that for a single-component saturated atmosphere, the supposition that the atmosphere is saturated is sufficient to determine the temperature profile, regardless of the means by which the saturation is maintained. Thus, it does not actually matter to the result whether the saturation is maintained by condensation due to vertical displacement of air parcels, or some other physical mechanism such as radiative cooling of the atmosphere to the point that condensation occurs.

Unless there is a reservoir of condensate at the surface to maintain saturation, it would be rare for an atmosphere to be saturated all the way to the ground. Suppose now that a one-component atmosphere has warm enough surface temperature that the surface pressure is lower than the saturation vapor pressure computed at the surface temperature. In this case, when a parcel is lifted by convection, its temperature will follow the *dry* or *noncondensing* adiabat, until the temperature falls so much that the gas becomes saturated. The level at which this occurs is called the *lifted condensation level*. Above the lifted condensation level, ascent causes condensation and the parcel follows the saturated adiabat. Since the temperature curve along the saturated adiabat falls with altitude so much less steeply than the dry adiabat, it is very easy for the two curves to intersect provided the surface temperature is not exceedingly large. An example for present Summer Martian conditions (specifically, like the warmest sounding in 2.2) is shown in Figure 2.6. A comparison with the Martian profiles in Figure 2.2 indicates that something interesting is going on in the Martian atmosphere. For the warm sounding, whose surface temperature is close to $255K$, the entire atmosphere aloft is considerably warmer than the adiabat, and the temperature nowhere comes close to the condensation threshold, though the very lowest portion of the observed atmosphere, below the 200 Pa level, does follow the dry adiabat quite closely. Clearly, something we haven't taken into account is warming up the atmosphere. A likely candidate for the missing piece is the absorption of solar energy by CO_2 and dust.

Although results like Figure 2.6 show a region of weak temperature dependence aloft which bears a superficial resemblance to the stratosphere seen in Earth soundings (and also at the top of the Venus, Jupiter and Titan soundings), one should not jump to the conclusion that the stratosphere is caused by condensation. This is not generally the case, and there are other explanations for the upper atmospheric temperature structure, which will be taken up in the next few chapters.

2.7.2 Mixtures of condensible with noncondensible gases

As a final step up on the ladder of generality, let's consider a mixture of a condensible substance with a substance that doesn't condense under the range of temperatures encountered in the atmosphere under consideration. This might be a mixture of condensible methane on Titan with

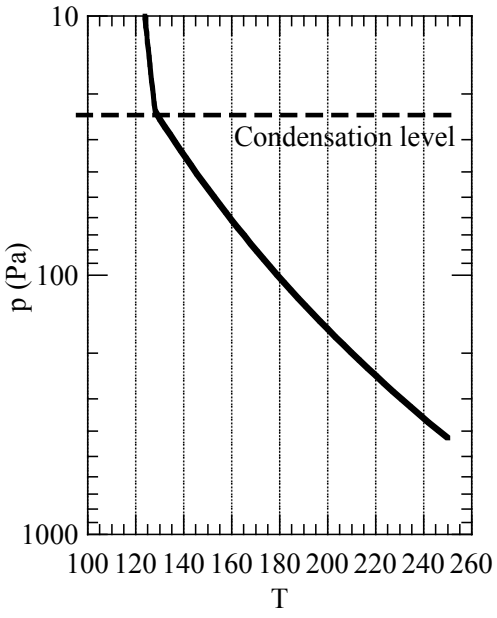

Figure 2.6: The adiabatic profile for a pure CO_2 atmosphere with a surface pressure of 450 Pa (4.5mb) and a surface temperature of 255K. The conditions are similar to those encountered on the warmer portions of present-day Mars.

non-condensible nitrogen, or condensible carbon dioxide on Early Mars with non-condensible nitrogen, or water vapor on Earth with a non-condensible mixture of oxygen and nitrogen. Whatever the substance, we distinguish the properties of the condensible substance with the subscript "c," and those of the non-condensible substance by the subscript "a" (for "air"). We now need to do the energy budget for a parcel of the mixture, assuming that it is initially saturated, and that it is displaced in such a way that condensation (releasing latent heat) must occur in order to keep the parcel from becoming supersaturated. We further introduce the assumption that essentially all of the condensate is immediately removed from the system, so that the heat storage in whatever mass of condensate is left in suspension may be neglected. This is a reasonable approximation for water or ice clouds on Earth, but even in that case the slight effect of the mass of retained condensate on buoyancy can be significant in some circumstances. In other planetary atmospheres the effect of retained condensate could be of greater importance. The temperature profile obtained by assuming condensate is removed from the system is called a *pseudoadiabat*, because the process is not truly reversible. One cannot return to the original saturated state, because the condensate is lost. At the opposite extreme, if all condensate is retained, it can be re-evaporated when the parcel is compressed, allowing for true reversibility.

Let the partial pressure, density, molecular weight, gas constant and specific heat of the noncondensible substance be p_a, ρ_a,M_a, R_a, and c_{pa}, and similarly for the condensible substance. Further, let L be the latent heat of the phase transition between the vapor and condensed phase of the condensible substance, and let $p_{c,sat}(T)$ be the saturation vapor pressure of this substance, as determined by the Clausius-Clapeyron relation. The assumption of saturation amounts to saying that $p_c = p_{c,sat}(T)$; if the parcel weren't at saturation, there would be no condensation and we could simply use the dry adiabat based on a noncondensing mixture of substance "a" and "c."

Now consider a parcel consisting of a mass m_a of noncondensible gas with an initial mass m_c of condensible gas occupying a volume V. Suppose that the temperature is changed by an amount dT, the partial pressure of noncondensible gas is changed by an amount dp_a, and the partial pressure of the condensible gas is changed by an amount dp_c. The mechanical work term for the mixture is $Vd(p_a + p_c)$, but by the definition of density $V = m_a/\rho_a = m_c/\rho_c$. Thus, the total heat budget of the parcel can be written in the form

$$(m_a + m_c)\delta Q = m_a c_{pa} dT - \frac{m_a}{\rho_a} dp_a + m_c c_{pc} dT - \frac{m_c}{\rho_c} dp_c + L dm_c \tag{2.28}$$

where m_c is the mass of condensible substance *in the vapor phase*. Thus, when some vapor is condensed out, dm_c is negative, yielding a negative contribution to δQ when L is positive. Since condensation releases latent heat, this sign may seem counter-intuitive. Recall, however, that condensation puts the system in a *lower energy state*, which is why there is heat available to be "released." If the system is energetically closed ($\delta Q = 0$), then the negative contribution of the latent heat term is offset by an increase in the remaining terms – e.g. an increase in temperature.

Exercise 2.7.2 Show that when there is no condensation ($dm_c = 0$) the adiabat $T(p)$ obtained by setting $\delta Q = 0$ has the same form as the usual dry adiabat, but with R and c_p taking on the appropriate mean values for the mixture.

The heat budget does not contain any term corresponding to heat storage in the condensed phase, since it is assumed that all condensate disappears from the parcel by precipitation. The usual way to change dp_a would be by lifting, causing expansion and reduction of pressure. Now, we divide by $m_c T$, make use of the perfect gas law to substitute for ρ_a and ρ_c, and make use of the fact that $m_c/m_a = (M_c/M_a)(p_c/p_a)$, since m_c/m_a is just the mass mixing ratio, denoted

henceforth by r_c. This yields

$$(1+r_c)\frac{\delta Q}{T} = [c_{pa}\frac{dT}{T} - R_a\frac{dp_a}{p_a}] + [c_{pc}r_c\frac{dT}{T} - r_cR_c\frac{dp_c}{p_c}] + \frac{L}{T}dr_c \tag{2.29}$$

The first two bracketed groups of terms on the right hand side can be recognized as the contribution of the two substances to the noncondensing entropy of the mixture, weighted according to the relative abundance of each species. If there is no condensation, the mixing ratio is conserved as the parcel is displaced to a new pressure, $dr_c = 0$, and the expression reduces to the equivalent of Eqn. 2.9, leading to the dry adiabat for a mixture. At this point, we introduce the saturation assumption, which actually consists of two parts: First, we assume that the air parcel is initially saturated, so that before being displaced, $p_c = p_{c,sat}(T)$ and $r_c = r_{sat} = \epsilon p_{c,sat}(T)/p_a$, where ϵ is the ratio of molecular weights M_c/M_a and $p_{c,sat}(T)$ is determined by the Clausius-Clapeyron relation. Second, we assume that a displacement conserving r_c would cause supersaturation, so that condensation would occur and bring the partial pressure p_c back to the saturation vapor pressure corresponding to the new value of T. Usually, this would occur as a result of ascent and cooling, since cooling strongly decreases the saturation vapor pressure. Typically (though not inevitably), the effect of compressional warming on saturation vapor pressure dominates the effect of increasing partial pressure, so that subsidence of initially saturated air follows the dry adiabat.

Assuming that the displacement causes condensation, we may replace p_c by $p_{c,sat}(T)$ and r_c by r_{sat} everywhere in Eqn. 2.29. Next, we use Clausius-Clapeyron to re-write $dp_{c,sat}$, observing that

$$\frac{dp_{c,sat}}{p_{c,sat}} = d\ln p_{c,sat} = \frac{d\ln p_{c,sat}}{dT}dT \tag{2.30}$$

and

$$dr_{sat} = \epsilon d\frac{p_{c,sat}}{p_a} = \epsilon\frac{p_c}{p_a}d\ln\frac{p_c}{p_a} = r_{sat}\cdot(d\ln p_{c,sat} - d\ln p_a) \tag{2.31}$$

Upon substituting into Equation 2.29 and collecting terms in $d\ln T$ and $d\ln p_a$ we find

$$(1+r_{sat})\frac{\delta Q}{T} = (c_{pa} + (c_{pc} + (\frac{L}{R_cT} - 1)\frac{L}{T})r_{sat})d\ln T - (1 + \frac{L}{R_aT}r_{sat})R_a d\ln p_a \tag{2.32}$$

To obtain the adiabat, we set $\delta Q = 0$, which leads to the following diferential equation defining $\ln T$ as a function of $\ln p_a$:

$$\frac{d\ln T}{d\ln p_a} = \frac{R_a}{c_{pa}}\frac{1 + \frac{L}{R_aT}r_{sat}}{1 + (\frac{c_{pc}}{c_{pa}} + (\frac{L}{R_cT} - 1)\frac{L}{c_{pa}T})r_{sat}} \tag{2.33}$$

Note that this expression reduces to the dry adiabat, as it should, when $r_{sat} \to 0$.

Exercise 2.7.3 What would the the slope $d\ln T/d\ln p_a$ be for a noncondensing mixture of the two gases? (*Hint:* $\ln p = \ln p_a + const.$ in this case). Why doesn't Eq. 2.33 reduce to this value as $L \to 0$? (Hint: Think about the way Clausius-Clapeyron has been used in deriving the moist adiabat, and what it implies for variations of p_c.)

A displaced parcel of atmosphere will follow the moist adiabat if the condensible substance condenses in the course of the displacement, but it will follow the dry adiabat if the displacement causes the condensible substance to become subsaturated. Does condensation occur on ascent (which lowers the total pressure) or descent (which raises the total pressure)? To answer this question, we must compare the moist adiabatic slope $d\ln T/d\ln p$ computed from Eq. 2.33 with

the dry adiabatic slope R/c_p, with R and c_p computed as the appropriate weighted average for the mixture. When the moist adiabatic slope is lower than R/c_p,then lifting a parcel adiabatically creates enough cooling that the parcel becomes supersaturated, and condensation occurs. Conversely, if the moist adiabatic slope is less than R/c_p, condensation occurs on descent instead.

The only problem with applying this criterion is that Eq. 2.33 gives us $d\ln T/d\ln p_a$ whereas we need $d\ln T/d\ln p$. Here, we will restrict attention to the dilute case, where there is little enough condensible substance present that $p \approx p_a$. In this case $d\ln T/d\ln p \approx d\ln T/d\ln p_a$ and $R/c_p \approx R_a/c_{p,a}$. The opposite limit of the condensible-dominated atmosphere is done in Problem **??**, and the general case is done in Problem **??**.

An examination of the properties of gases indicates that c_{pc}/c_{pa} is typically of order unity, whereas $L/(R_cT)$ is typically very large, so long as the temperature is not exceedingly great. If one drops the smaller terms from the denominator of Eq. 2.33, one finds that the temperature gradient along the moist adiabat is weaker than that along the dry adiabat provided $\epsilon L/c_{pa}T > 1$, which is typically the case. In this typical case, condensation occurs on ascent and warms the ascending parcel through the release of latent heat. This behavior can fail when the latent heat is weak or the noncondensible specific heat is very large, whereupon the heat added by condensation has little effect on temperature. It is in this regime that condensation happens on *descent* rather than ascent. Given the thermodynamic properties of common atmospheric constituents, it is not an easy regime to get into. Even the case of H_2O condensation in an H_2 Jovian or Saturnian atmosphere yields condensation on ascent, despite the high specific heat of H_2. In that case, the high specific heat of H_2 is canceled out by the low value of ϵ, which is typical behavior since specific heat tends to scale inversely with molecular weight – substances with low molecular weight have more degrees of freedom per kilogram. The quantity $\epsilon L/c_{pa}T$ decreases with temperature, especially since L approaches zero as the critical point is approached, and this suggests that condensation might occur during descent at very high temperatures. A proper evaluation of the possibility would require taking into account non-ideal gas effects, the temperature dependence of specific heat, the neglected condensate density term in Clausius-Clapeyron, and probably also non-dilute effects.

Everything on the right hand side of Eqn 2.33 is either a thermodynamic constant, or can be computed in terms of $\ln T$ and $\ln p_a$. Therefore, the equation defines a first order ordinary differential equation which can be integrated (usually numerically) to obtain T as a function of p_a. Usually one wants the temperature as a function of total pressure, rather than partial pressure of the noncondensible substance. This is no problem. Once $T(p_a)$ is known, the corresponding total pressure at the same point is obtained by computing $p = p_a + p_{c,sat}(T(p_a))$. To make a plot, or a table, one treats the problem parametrically: computing both T and p as functions of p_a. When the condensible substance is dilute, then $p_{c,sat} << p_a$, and $p \approx p_a$, so Eqn 2.33 gives the desired result directly.

Figure 2.7 shows a family of solutions to Eqn 2.33, for the case of water vapor in Earth air. When the surface temperature is 250K, there is so little moisture in the atmosphere that the profile looks like the dry adiabat right to the ground. As temperature is increased, a region of weak gradients appears near the ground, representing the effect of latent heat on temperature. This layer gets progressively deeper as temperature increases and the moisture content of the atmosphere increases. When the surface temperature is 350K, so much moisture has entered the atmosphere that the surface pressure has actually increased to over 1300mb. Moreover, the moisture-dominated region extends all the way to 10 Pa (.1mb) , and even at 100 Pa (1mb) the atmosphere is 10% water by volume. Thus, for moderate surface temperatures, there is little water high up in the atmosphere. When the surface temperature approaches or exceeds 350K, though, the "cold trap" is lost, and a great deal of water is found aloft, where it is exposed to the destructive ultraviolet light of the sun and the possibility of thermal escape to space. In subsequent chapters, it will be

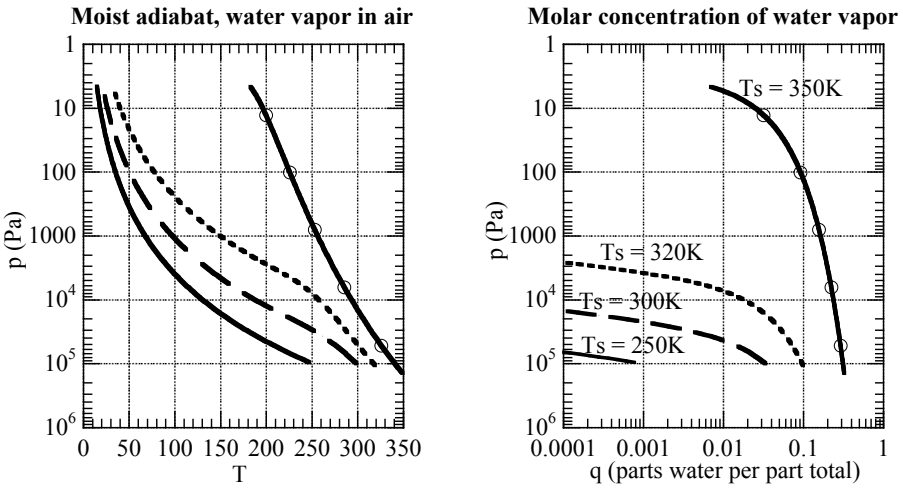

Figure 2.7: The moist adiabat for saturated water vapor mixed with Earth air having a partial pressure of 1 bar at the surface. Results are shown for various values of surface temperature, ranging from 250K to 350K. The left panel shows the temperature profile, while the right shows the profile of molar concentration of water vapor. A concentration value of .1 would mean that one molecule in 10 of the atmosphere is water vapor.

seen that this phenomenon plays a major role in the life cycle of planets, and probably accounts for the present hot, dry state of Venus.

The mixing ratio of the condensible component varies along the moist adiabat, so the reader may wonder whether these variations can lead to compositionally induced buoyancy along the lines discussed in Section 2.4. Suppose a parcel starts on the moist adiabat at pressure p, with temperature $T(p)$, and that the parcel is saturated. For the sake of definiteness, let's suppose also that condensation occurs on ascent, so ascending displacements are saturated and remain on the moist adiabat. Since there is a unique moist adiabat going through the starting point, and since saturation also determines the composition uniquely once temperature and pressure are known, then when the parcel is lifted to pressure p_1 it has the same temperature, composition and density as the ambient air there. Composition doesn't induce buoyancy, because the variations of R/c_p are constrained by the saturation assumption, and have been taken into account in the computation of the moist adiabat. In this sense, the moist adiabat is neutrally stable against condensing ascent. On the other hand, a subsiding air parcel will become subsaturated, and descending along the dry adiabat will become warmer than the surrounding air, causing it to be positively buoyant. In this sense, the moist adiabat is stable against subsidence [4].

Saturated air can go up without expenditure of energy, but it takes energy to push it down.

[4]Compositional effects will alter the buoyancy of subsiding parcels, and in principle could make subsiding parcels negatively buoyant in extreme cases. For typical atmospheric gases the compositional effect is too small to overwhelm the positive buoyancy due to compressional heating, but if one could find a situation where composition rendered subsiding parcels unstable to descent, the resulting convection would be very interesting indeed to study.

However, when an air parcel is initially *subsaturated* and must be lifted some distance before it becomes saturated and follows the moist adiabat, then the full range of issues discussed in Section 2.4 come into play. In that case, the compositional variations in the subsaturated layer can either favor or inhibit the triggering of vertical mixing. On a related note, suppose that the temperature profile $T(p)$ happens to follow the moist adiabat, but that the air is unsaturated; this is in fact the case over most of the Earth's tropics, for dynamical reasons that are beyond the scope of this book. In that case, if a saturated air parcel from the ocean surface is lifted through the dry surroundings, it will follow the moist adiabat but because its composition will be different from the unsaturated surroundings it will in fact be positively buoyant if the condensible has lower molecular weight than the noncondensible background, as is the case for water vapor in Earth air. If the condensible has higher moleculer weight, the parcel will be negatively buoyant and vertical mixing will be choked off. Real atmospheres consist of a mix of saturated and unsaturated regions, and the representation of vertical mixing in these cases pose a considerable challenge. Compositional effects on Earth at its present temperature are slight, but for atmospheres in which the condensible is nondilute, these effects become more and more important. With regard to the modelling of convection, this represents largely unexplored territory. It is probably important for the case of methane on Titan, but would also be important on Earthlike planets having temperatures some tens of degrees or more warmer than the present Earth.

The mass of retained condensate – cloud mass – can also affect the buoyancy of lifted or subsiding parcels, since condensate alters the density of an air parcel. In such cases one also needs to consider cooling due to evaporation of condensate when the air parcel subsides, warms and becomes subsaturated. In quiescent conditions retained condensate mass is expected to be a small fraction of the total mass of an air parcel, since a large condensate loading usually leads to coalescence of cloud droplets and subsequent removal by precipitation. The effects of condensate loading are small, but significant, to convection in the present Earth's atmosphere, but the possibility should be kept in mind that condensate loading may play a more prominent role in planetary atmospheres that have not been as extensively studied as Earth's.

2.7.3 Moist static energy

When the condensible substance is dilute, it is easy to define a *moist static energy* which is a generalization of the dry static energy defined in Eq. 2.23. This is accomplished by multiplying Eq. 2.29 by T to obtain the heat budget, dropping the terms proportional to r_c (which are small in the dilute limit), and making use of $R_a T/p_a = 1/\rho_a \approx 1/\rho$. The hydrostatic relation is used to rewrite the pressure work term in precisely the same way as was done for dry static energy. Then, if we further assume that the temperature range of interest is small enough that L may be regarded as constant, we obtain

$$\delta Q = d(c_{pa}T + gz + Lr_c) \tag{2.34}$$

Note that this relation does not assume that the condensible mixing ratio r_c is saturated. For adiabatic flow in the dilute case, this moist static energy will be conserved following an air parcel whether or not condensation occurs. If energy is added to or taken away from the parcel, then the change in moist static energy is equal to the net amount of energy added or taken away, per unit mass of the parcel.

When the condensate is nondilute, things are a bit more complicated. In this case significant amounts of mass can be lost from the atmosphere in the course of condensation, and in essence the precipitation of condensate can take away significant amounts of energy with it. In order to deal with the heat storage in condensate, one must make use of the specific heat of the condensed

phase, which we'll refer to as $c_{pc\ell}$ (regardless of whether the condensate is liquid or solid); the behavior of this specific heat is inextricably linked to the changes in latent heat with temperature through the thermodynamic relation

$$\frac{dL}{dT} = c_{pc} - c_{pc\ell} \tag{2.35}$$

which is valid when the condensate density is much greater than the vapor density. This relation is essentially a consequence of energy conservation. Given this link, we will take into account the change of latent heat with temperature in carrying out the following analyis.

First let's analyze the energy budget per unit mass of the noncondensible substance. We'll write the pressure work term in Eq. 2.28 in the alternate form $(m_a/\rho_a)dp$. On dividing the equation by m_a and using Eq. 2.35 and the hydrostatic relation we obtain

$$(1 + r_c)\delta Q = d[(c_{pa} + r_c c_{pc\ell})T + (1 + r_c)gz + L(T)r] - (c_{pc\ell}T + gz)dr_c \tag{2.36}$$

The quantity in square brackets is thus identified as the moist static energy per unit mass of noncondensible substance; it will be denoted by the symbol $\mathfrak{M}$. The second term on the right hand side, involving dr_c, represents the sink of moist static energy due to the heat and potential energy carried away by the condensate.

This form of moist static energy can be inconvenient to use, because the $(1 + r_c)$ weighting on the left hand side makes it hard to do the energy budget of a column of air knowing only the net input of energy into the column. The expression also becomes inconvenient when the atmosphere becomes dominated by the condensible substance, leading to very large values of r_c. We can formulate the moist static energy per unit *total* mass of the gas by dividing Eq. 2.36 by $(1 + r_c)$. After carrying out a few basic manipulations, we find

$$\delta Q = d[\frac{\mathfrak{M}}{1 + r_c}] - (c_{pc\ell}T + gz - \frac{\mathfrak{M}}{1 + r_c})d\ln(1 + r_c) \tag{2.37}$$

The term $\mathfrak{M}/(1 + r_c)$ is the desired moist static energy per unit of total gaseous mass, and the second term is the corresponding sink due to precipitation.

Exercise 2.7.4 Rewrite the expression $\mathfrak{M}/(1 + r_c)$ in terms of the mass specific concentration q. What is the form of this expression when $q \approx 1$? What happens to the precipitation sink term in this limit?

An alternate approach to dealing with moist static energy in the nondilute case is to write an energy budget per unit total mass (condensate included) for an air parcel that retains its condensate. To allow for precipitation, one then deals explicitly with the energy loss occuring when some of the retained condensate is removed from the air parcel. This approach is especially useful in situations when the mass of retained condensate can be appreciable. It can be carried out using the moist entropy expressions derived in Emanuel (1994) (see Further Readings). Further modifications to the expression for moist static energy are required if both ice and liquid phases are present in the atmosphere, in order to account for the latent heat of the solid/liquid phase transition.

Because the change in moist static energy of an atmospheric column can be determined from the energy fluxes into the top of the column and the energy fluxes between the bottom of the column and the underlying surface, moist static energy provides a convenient basis for diagnostics of atmospheric energy flows. For the same reason, it provides a convenient basis for

the formulation of simplified vertically-averaged energy balance models of climate. Such models are taken up briefly in Chapter 9. Moist static energy is also important to the formulation of simplified representations of buoyancy-driven vertical mixing an a statically unstable layer of the atmosphere. In such applications, when radiation or some other process creates a region of static instability, it is presumed that mixing will proceed until some layer of the atmosphere is reset to a state of neutral stability. The constraint that the moist static energy following mixing should be the same as that of the initial state (within a suitably chosen layer of atmosphere) plays a crucial role in determining what the temperature profile is following mixing.

2.8 For Further Reading

For a deeper discussion of thermodynamics, see

- Feynman RP, Leighton RB and Sands M 2005: *The Feynman Lectures on Physics, Vol 1.* Addison Wesley.

A physically based derivation of Clausius-Clapeyron is found in Section 45-3, and a good discussion of entropy is found in Section 44-6. A very thorough discussion of moist thermodynamics and of the representation of the effects of moist convection on the Earth's atmosphere can be found in

- Emanuel KA 1994: *Atmospheric Convection.* Oxford University Press.

An example of the computation of the dilute moist adiabat for a non-ideal noncondensible background gas can be found in the Appendix of

- Kasting JF 1991: CO_2 Condensation and the Climate of Early Mars. *Icarus* **94** 1-13.

The extension to the nondilute case would be straightforward were it not for the fact that the saturation vapor pressure can no longer be computed independently of the noncondensible gas state. This subject is under investigation for the N_2/CH_4 system on Titan, and may also be relevant for non-ideal mixed CO_2/H_2O mixtures on hot, wet planets such as Early Venus. The planetary implications of non-ideal non-dilute behavior are still at a very early stage of development.

A very nice analysis of the way the mixing of inhomogeneities increases entropy is given in

- Pauluis O and Held IM: Entropy budget of an atmosphere in radiative-convective equilibrium. Part II: Latent heat transport and moist processes. *J. Atmos. Sci.* **59** 140-149.

This particular analysis applies to the mixing of moist and dry air, and shows that the mixing is a big part of what makes moist convection irreversible.

For thermodynamic properties of gases and associated phase transitions, the NIST Chemistry WebBook and the Air Liquide Gas Encyclopedia, found at

- `http://webbook.nist.gov/`
- `http://encyclopedia.airliquide.com/encyclopedia.asp`

are very useful. Properties for more exotic conditions can often be found in the *Journal of Physical and Chemical Reference Data* and results pertinent to planetary atmospheres are often reported in the journal *Icarus*.

Chapter 3

Elementary models of radiation balance

3.1 Overview

Our objective is to understand the factors governing the climate of a planet. In this chapter we will be concerned with energy balance and planetary temperature. Certainly, there is more to climate than temperature, but equally certainly temperature is a major part of what is meant by "climate," and greatly affects most of the other processes which come under that heading.

From the preceding chapter, we know that the temperature of a chunk of matter provides a measure of its energy content. Suppose that the planet receives energy at a certain rate. If uncompensated by loss, energy will accumulate and the temperature of some part of the planet will increase without bound. Now suppose that the planet loses energy at a rate that increases with temperature. Then, the temperature will increase until the rate of energy loss equals the rate of gain. It is this principle of energy balance that determines a planet's temperature. To quantify the functional dependence of the two rates, one must know the nature of both energy loss and energy gain.

The most familiar source of energy warming a planet is the absorption of light from the planet's star. This is the dominant mechanism for rocky planets like Venus, Earth and Mars. It is also possible for energy to be supplied to the surface by heat transport from the deep interior, fed by radioactive decay, tidal dissipation, or high temperature material left over from the formation of the planet. Heat flux from the interior is a major player in the climates of some gas giant planets, notably Jupiter and Saturn, because fluid motions can easily transport heat from the deep interior to the outer envelope of the planet. The sluggish motion of molten rock, and even more sluggish diffusion of heat through solid rock, prevent internal heating from being a significant part of the energy balance of rocky planets. Early in the history of a planet, when collisions are more common, the kinetic energy brought to the planet in the course of impacts with asteroids and planetesimals can be a significant part of the planet's energy budget.

There are many ways a planet can gain energy, but essentially only one way a planet can lose energy. Since a planet sits in the hard vaccuum of outer space,and its atmosphere is rather tightly bound by gravity, not much energy can be lost through heated matter streaming away from the planet. The only significant energy loss occurs through emission of electromagnetic radiation,

most typically in the infrared spectrum. The quantification of this rate, and the way it is affected by a planet's atmosphere, leads us to the subject of *blackbody radiation*.

3.2 Blackbody radiation

It is a matter of familiar experience that a sufficiently hot body emits light – hence terms like "red hot" or "white hot." Once it is recognized that light is just one form of electromagnetic radiation, it becomes a natural inference that a body with any temperature at all should emit some form of electromagnetic radiation, though not necessarily visible light. Thermodynamics provides the proper tool for addressing this question.

Imagine a gas consisting of two kinds of molecules, labeled A and B. Suppose that the two species interact strongly with each other, so that they come into thermodynamic equilibrium and their statistical properties are characterized by the same temperature T. Now suppose that the molecules A are ordinary matter, but that the "molecules" B are particles of electromagnetic radiation ("photons") or, equivalently, electromagnetic waves. If they interact strongly with the A molecules, whose energy distribution is characterized by their temperature T in accord with classical thermodynamics, the energy distribution of the electromagnetic radiation should also be characterized by the same temperature T. In particular, for any T there should be a unique distribution of energy amongst the various frequencies of the waves. This spectrum can be observed by examining the electromagnetic radiation leaving a body whose temperature is uniform. The radiation in question is known as *blackbody radiation* because of the assumption that radiation interacts strongly with the matter; any radiation impinging on the body will not travel far before it is absorbed, and in this sense the body is called "black" even though, like the Sun, it may be emitting light. Nineteenth century physicists found it natural to seek a theoretical explanation of the observed properties of blackbody radiation by applying well-established thermodynamical principles to electromagnetic radiation as described by Maxwell's classical equations. The attempt to solve this seemingly innocuous problem led to the discovery of quantum theory, and a revolution in the fundamental conception of reality.

Radiation is characterized by direction of propagation and frequency (and also polarization, which will not concern us). For electromagnetic radiation, the frequency ν and wavelength λ are related by the *dispersion relation* $\nu\lambda = c$, where c is a constant with the dimensions of velocity. Because visible light is a familiar form of electromagnetic radiation, c is usually called "the speed of light." The wavenumber, defined by $n = \lambda^{-1} = \nu/c$ is often used in preference to frequency or wavelength. The wavenumber can be viewed as the frequency measured in alternate units, and so we will often refer to wavenumber and frequency interchangeably. Although mks units are preferred throughout this book, we follow spectroscopic convention and make an exception for wavenumber when dealing with infrared radiation, which will usually be measured in cm^{-1} since it yields comfortable and familiar ranges of numbers. Wavelengths themselves will sometimes be measured in μm (microns, or $10^{-6}m$). Figure 3.1 gives the approximate regions of the electromagnetic spectrum corresponding to common names such as "Radio Waves" and so forth.

If a field of radiation consists of a mixture of different frequencies and directions, the mixture is characterized by a *spectrum*, which is a function describing the proportions of each type of radiation making up the blend. A spectrum is a *density* describing the amount of electromagnetic energy contained in a unit volume of the space (3D position, frequency, direction) needed to characterize the radiation, just as the mass density of a three dimensional object describes the distribution of mass in three-dimensional space.

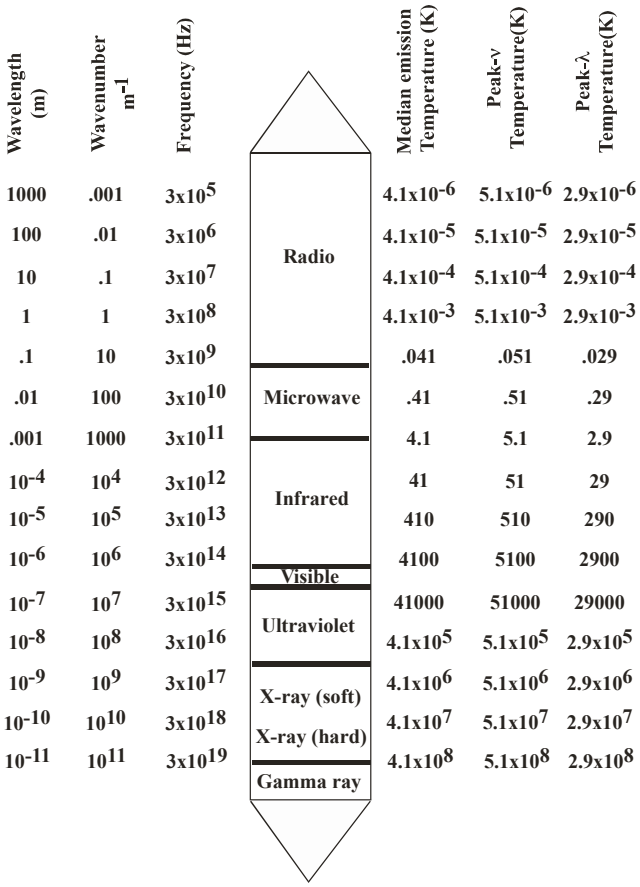

Figure 3.1: The electromagnetic spectrum. The Median Emission Temperature is the temperature of a blackbody for which half of the emitted power is below the given frequency (or equivalently, wavelength or wavenumber). The Peak-ν Temperature is the temperature of a blackbody for which the peak of the Planck density in frequency space is at the stated frequency. The Peak-λ Temperature is the temperature of a blackbody for which the peak of the Planck density in wavelength space is at the stated wavelength.

Before proceeding, we must pause and talk a bit about how the "size" of collections of directions are measured in three dimensions. For collections of directions on the plane, the measure of the "size" of the set of directions between two directions is just the angle between those directions. The angle is typically measured in radians; the measure of the angle in radians is the length of the arc of a unit circle whose opening angle is the angle we are measuring. The set of all angles in two dimensions is then 2π radians for example. A collection of directions in three dimensional space is called a *solid angle*. A solid angle can sweep out an object more complicated than a simple arc, but the "size" or measure of the solid angle can be defined through a generalization of the radian, known as the *steradian*. The measure in steradians of a solid angle made by a collection of rays emanating from a point P is defined as the area of the patch of the unit sphere centered on P which the rays intersect. For example, a set of directions tracing out a hemisphere has measure 2π steradians, while a set of directions tracing out the entire sphere (i.e. all possible directions) has measure 4π steradians. If we choose some specific direction (e.g. the vertical) as a reference direction, then a direction in three dimensional space can be specified in terms of two angles, θ and ϕ, where *theta* is the angle between the reference direction and the direction we are specifying, and ϕ is the angle along a circle centered on the reference direction. These angles define a spherical polar coordinate system with the reference direction as axis; $0 \geq \theta \leq \pi$ and $0 \geq \phi \leq 2\pi$. In terms of the two direction angles, the differential of solid angle Ω is $d\Omega = \sin\theta d\theta \; d\phi = -(d\cos\theta)(d\phi)$. Generally, when writing the expression for $d\Omega$ in the latter form we drop the minus sign and just remember to flip the direction of integration to make the solid angle turn out positive. We recover the area of the unit sphere by integrating $d\Omega$ over $\cos\theta = -1$ to $\cos\theta = 1$ and $\phi = 0$ to $\phi = 2\pi$. A similar integration shows that the set of directions contained within a cone with vertex angle $\Delta\theta$ measured relative to the altitude of the cone has measure $2\pi(1 - \cos\Delta\theta)$ steradians. A narrow cone with $\Delta\theta \ll 1$ has measure $\pi(\Delta\theta)^2$ steradians.

We wish to characterize the energy in the vicinity of a point $\vec{r}$ in three dimensional space, with frequency near ν and direction near that given by a unit vector $\hat{n}$. The energy spectrum $\Sigma(\vec{r}, \nu, \hat{n})$ at this point is defined such that the energy contained in a finite but small sized neighborhood of the point $(\vec{r}, \nu, \hat{n})$ is $\Sigma dV d\nu d\Omega$, where dV is a small volume of space, $d\nu$ is the width of the frequency band we wish to include, and $d\Omega$ measures the range of solid angles we wish to include.

Since electromagnetic waves in a vacuum move with constant speed c, the energy *flux* through a flat patch perpendicular to $\hat{n}$ with area dA is simply $c\Sigma dA d\nu d\Omega$, which defines the flux spectrum $c\Sigma$. In *mks* units, the flux spectrum has units of $(Watts/m^2)/(Hz \cdot steradian)$, where the Hertz (Hz) is the unit of frequency, equal to one cycle per second. The flux spectrum defined in this way is usually called the *spectral irradiance*; integrated over all frequencies, it is called the *irradiance*.

Exercise 3.2.1 The mks unit of energy is the *Joule*, J, which is 1 $Newton \cdot meter/sec$. A Watt (W) is $1J/sec$. A typical resting human in not-too-cold weather requires about $2000 Calories/day$. (A $Calorie$ is the amount of energy needed to increase the temperature of $1Kg$ of pure water by $1K$.) Convert this to a power consumption in W, using the fact that $1 Calorie = 4184J$.

On the average, the flux of Solar energy reaching the Earth's surface is about $240 W/m^2$. Assuming that food plants can convert Solar energy to usable food calories with an efficiency of 1%, what is the maximum population the Earth could support? (The radius of the Earth is about $6371km$)

The bold assumption introduced by Planck is that electromagnetic energy is exchanged only in amounts that are multiples of discrete *quanta*, whose size depends on the frequency of the radiation, in much the same sense that a penny is the quantum of US currency. Specifically, the quantum of energy for electromagnetic radiation having frequency ν is $\Delta E = h\nu$, where h

is now known as *Planck's constant*. It is (so far as currently known) a constant of the universe, which determines the granuarity of reality. h is an exceedingly small number ($6.626 \cdot 10^{-34} Joule - seconds$), so quantization of energy is not directly manifest as discreteness in the energy changes of everyday objects. A 1 watt blue nightlight (wavelength .48 μm, or frequency $6.24 \cdot 10^{14} Hz$ emits $2.4 \cdot 10^{18}$ photons each second, so it is no surprise that the light appears to be a continuous stream. If a bicycle were hooked to an electrical brake that dissipated energy by driving a blue light, emitting photons, the bike would indeed slow down in discontinuous increments, but the velocity increment, assuming the bike and rider to have a mass of $80kg$, would be only $10^{-10} m/s$; if one divides a $1m/s$ decrease of speed into 10^{10} equal parts, the deceleration will appear entirely continuous to the rider. Nonetheless, the aggregate effect of microscopic graininess of energy transitions exert a profound influence on the macroscopic properties of everyday objects. Blackbody radiation is a prime example of this.

Once the quantum assumption was introduced, Planck was able to compute the irradiance (flux spectrum) of blackbody radiation with temperature T using standard thermodynamic methods. The answer is

$$B(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{e^{h\nu/kT} - 1} \tag{3.1}$$

where k is the Boltzmann thermodynamic constant defined in Chapter 2. $B(\nu, T)$ is known as the *Planck function*. Note that the Planck function is independent of the direction of the radiation; this is because blackbody radiation is *isotropic*, i.e. equally intense in all directions. In a typical application of the Planck function, we wish to know the flux of energy exiting the surface of a blackbody through a small nearly flat patch with area dA, over a frequency band of width $d\nu$. Since energy exits through this patch at all angles, we must integrate over all directions. However, energy exiting in a direction which makes an angle θ to the normal to the patch contributes a flux $(BdAd\nu d\Omega)\cos\theta$ through the patch, since the component of flux parallel to the patch carries no energy through it. Further, using the definition of a steradian, $d\Omega = 2\pi d\cos\theta$ for the set of all rays making an angle θ relative to the normal to the patch. Integrating $B\cos\theta d\Omega$ from $\theta = 0$ to $\theta = \pi/2$, and using the fact that B is independent of direction, we then find that the flux through the patch is $\pi BdAd\nu$. This is also the amount of electromagnetic energy in a frequency band of width $d\nu$ that would pass each second through a hoop enclosing area dA (from one chosen side to the other), placed in the interior of an ideal blackbody; an equal amount passes through the hoop in the opposite sense.

The way the angular distribution of the radiation is described by the Planck function can be rather confusing, and requires a certain amount of practice to get used to. The following exercise will test the readers' comprehension of this matter.

Exercise 3.2.2 A radiation detector flies on an airplane a distance H above an infinite flat plain with uniform temperature T. The detector is connected to a watt-meter which reports the total radiant power captured by the detector. The detector is sensitive to rays coming in at angles $\leq \delta\theta$ relative to the direction in which the detector is pointed. The area of the aperture of the detector is δA. The detector is sensitive to frequencies within a small range $\delta\nu$ centered on ν_0.

If the detector is pointed straight down, what is the power received by the detector? What is the size of the "footprint" on the plain to which the detector is sensitive? How much power is emitted by this footprint in the detector's frequency band? Why is this power different from the power received by the detector?

How do your answers change if the detector is pointed at an angle of 45^o relative to the vertical, rather than straight down?

The Planck function depends on frequency only through the dimensionless variable $u = h\nu/(kT)$. Recalling that each degree of freedom has energy $\frac{1}{2}kT$ in the average, we see that u is half the ratio of the quantum of energy at frequency ν to the typical energy in a degree of freedom of the matter with which the electromagnetic energy is in equilibrium. When u is large, the typical energy in a degree of freedom cannot create even a single photon of frequency ν, and such photons can be emitted only by those rare molecules with energy far above the mean. This is the essence of the way quantization affects the blackbody distribution – through inhibition of emission of high-frequency photons. On the other hand, when u is small, the typical energy in a degree of freedom can make many photons of frequency ν, and quantization imposes less of a constraint on emission. The characteristic frequency kT/h defines the crossover between the classical world and the quantum world. Much lower frequencies are little affected by quantization, whereas much higher frequencies are strongly affected. At 300K, the crossover frequency is $6240 GigaHz$, corresponding to a wavenumber of $20814 m^{-1}$, or a wavelength of 48 μm; this is in the far infrared range.

In terms of u, the Planck function can be rewritten

$$B(\nu, T) = \frac{2k^3T^3}{h^2c^2}\frac{u^3}{e^u - 1} \tag{3.2}$$

In the classical limit, $u \ll 1$, and $u^3/(\exp(u)-1) \approx u^2$. Hence, $B \approx 2kT\nu^2/c^2$, which is independent of h. In a classical world, where $h = 0$, this form of the spectrum would be valid for all frequencies, and the emission would increase quadratically with frequency without bound; a body with any nonzero temperature would emit infrared at a greater rate than microwaves, visible light at a greater rate than infrared, ultraviolet at a greater rate than visible, X-rays at a greater rate than ultraviolet, and so forth. Bodies in equilibrium would cool to absolute zero almost instantaneously through emission of a burst of gamma rays, cosmic rays and even higher frequency radiation. This is clearly at odds with observations, not least the existence of the Universe. We are saved from this catastrophe by the fact that h is nonzero, which limits the range of validity of the classical form of B. At frequencies high enough to make $u \gg 1$, then $u^3/(\exp(u) - 1) \approx u^3 exp(-u)$ and the spectrum decays somewhat more slowly than exponentially as frequency is increased. The peak of B occurs at $u \approx 2.821$, implying that the frequency of maximum emission is $\nu \approx (2.821k/h)T \approx 58.78 \cdot 10^9 T$. The peak of the frequency spectrum increases linearly with temperature. This behavior, first deduced empirically long before it was explained by quantum theory, is known as the *Wien Displacement Law*.

Because the emission decays only quadratically on the low frequency side of the peak, but decays exponentially on the high frequency side, bodies emit appreciable energy at frequencies much lower than the peak emission, but very little at frequencies much higher. For example, at one tenth the peak frequency, a body emits at a rate of 4.8% of the maximum value. However, at ten times the peak frequency, the body emits at a rate of only $8.9 \cdot 10^{-9}$ of the peak emission. The microwave emission from a portion of the Earth's atmosphere with temperature $250K$ (having peak emission in the infrared) is readily detectable by satellites, whereas the emission of visible light is not.

Since B is a density, one cannot obtain the corresponding distribution in wavenumber or wavelength space by simply substituting for ν in terms of wavenumber or wavelength in the formula for B. One must also take into account the transformation of $d\nu$. For example, to get the flux density in wavenumber space (call it B_n) we use $B(\nu, T)d\nu = B(n \cdot c, T)d(n \cdot c) = cB(n \cdot c, T)dn$, whence $B_n(n, T) = cB(n \cdot c, T)$. Thus, transforming to wavenumber space changes the amplitude but not the shape of the flux spectrum. The Planck density in wavenumber space is shown for various temperatures in Figure 3.2. Because the transformation of the density from frequency to wavenumber space only changes the labeling of the vertical axis of the graph, one can obtain

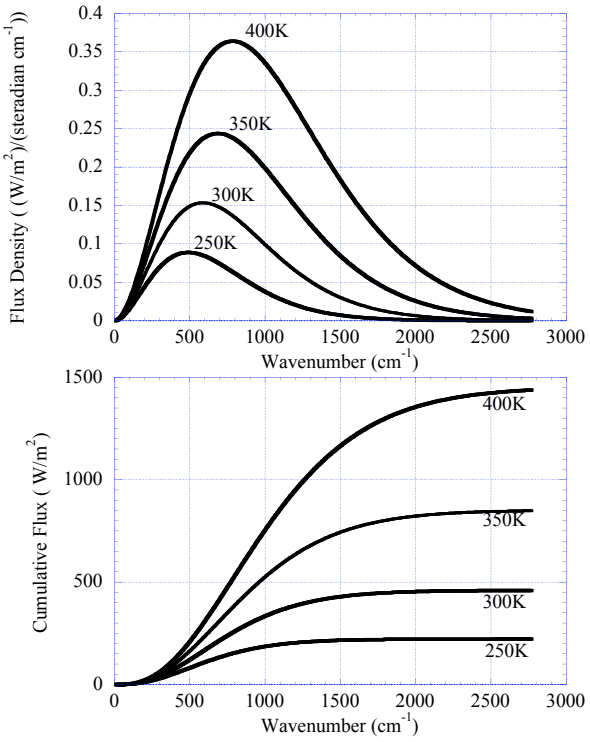

Figure 3.2: The spectrum of blackbody radiation for the various temperatures indicated on the curves. Upper Panel: The Planck density in wavenumber space. Lower Panel: The cumulative emission as a function of wavenumber. Note that the density has been tranformed such that the density times dn is the power per unit solid angle per unit area radiated in a wavenumber interval of width dn.

the wavenumber of maximum emission in terms of the frequency of maximum emission using $n_{max} = \nu_{max}/c$. An important property of the Planck function, readily verified by a simple calculation, is that $dB/dT > 0$ for all wavenumbers. This means that the Planck function for a large temperature is strictly above one for a lower temperature, or equivalently, that increasing temperature increases the emission at each individual wavenumber.

If one transforms to wavelength space, however,

$$B(\nu,T)d\nu = B(c/\lambda,T)d(c/\lambda) = -\frac{c}{\lambda^2}B(c/\lambda,T)d\lambda = \frac{2k^5T^5}{h^4c^3}\frac{u^5}{e^u-1}d\lambda = B_\lambda d\lambda \tag{3.3}$$

where $u = h\nu/kT = hc/\lambda kT$, as before. Transforming to wavelength space changes the shape of the flux spectrum. B_λ has its maximum at $u \approx 4.965$, which is nearly twice as large as the value for the wavenumber or frequency spectrum.

Since the location of the peak of the flux spectrum depends on the coordinate used to measure position within the electromagnetic spectrum, this quantity has no intrinsic physical meaning, apart from being a way to characterize the shape of the curve coming out of some particular kind of measuring apparatus. A more meaningful quantity can be derived from the *cumulative flux spectrum*, value at a given point in the spectrum is the same regardless of whether we use wavenumber, wavelength, $\log\lambda$ or any other coordinate to describe the position within the spectrum. The cumulative flux spectrum is defined as

$$F_{cum}(\nu,T) = \int_0^\nu \pi B(\nu',T)d\nu' = \int_\infty^\lambda \pi B_\lambda(\lambda',T)d\lambda' \tag{3.4}$$

Note that in defining the cumulative emission we have included the factor π which results from integrating over all angles of emission in a hemisphere. $F_{cum}(\nu,T)$ thus gives the power emitted per square meter for all frequencies less than ν, or equivalently, for all wavelengths greater than c/ν. This function is shown for various temperatures in the lower panel of Fig. 3.2, where it is plotted as a function of wavenumber. The value of ν for which $F_{cum}(\nu,T)$reaches half the net emission $F_{cum}(\infty,T)$ provides a natural characterization of the spectrum. We will refer to this characteristic frequency as the *median emission frequency*. The median emission wavelength and wavenumber is defined analogously. Whether one uses frequency, wavelength or some other measure, the median emission is attained at $u \approx 3.503$. For any given coordinate used to describe the spectrum, the (angle-integrated) Planck density in that coordinate is the derivative of the cumulative emission with respect to the coordinate. Hence the peak in the Planck density just gives the point at which the cumulative emission function has its maximum slope. This depends on the coordinate used, unlike the point of median emission. Figure 3.1 shows the the portion of the spectrum in which blackbodies with various temperatures dominantly radiate. For example, a body with a temperature of around $4K$ radiates in the microwave region; this is the famous "Cosmic Microwave Background Radiation" left over from the Big Bang [1]. A body with a temperature of $300K$ radiates in the infrared, one with a temperature of a few thousand degrees radiates in the visible, and one with a temperature of some tens of thousands of degrees would radiate in the ultraviolet.

Next, we evaluate $F_{cum}(\infty,T)$, to obtain the total power F exiting from each unit area of the surface of a blackbody:

$$F = \int_0^\infty \pi B(\nu,T)d\nu = \int_0^\infty \pi B(u,T)\frac{kT}{h}du = [\frac{2\pi k^4}{h^3c^2}\int_0^\infty \frac{u^3}{e^u-1}du]T^4 = \sigma T^4 \tag{3.5}$$

[1] What is remarkable about this observed cosmic radiation is not so much that it is in the microwave region, but that it has a blackbody spectrum, which says much about the interaction of radiation with matter in the early moments of the Universe.

where [2] $\sigma = 2\pi^5 k^4/(15c^2h^3) \approx 5.67 \cdot 10^{-8} Wm^{-2}K^{-4}$. The constant σ is known as the *Stefan-Boltzmann constant*, and the law $F = \sigma T^4$ is the *Stefan-Boltzmann law*. This law was originally deduce from observations, and Boltzmann was able to derive the fourth-power scaling in temperature using classical thermodynamic reasoning. However, classical physics yields an infinite value for the constant σ. The formula for σ clearly reveals the importance of quantum effects in determining this constant, since σ diverges like $1/h^3$ if we try to pass to the classical limit by making h approach zero.

An important property of an ideal blackbody is that the radiation leaving its surface depends only on the temperature of the body. If a blackbody is interposed between an observer and some other object, all properties of the object will be hidden from the observer, who will see only blackbody radiation corresponding to the temperature of the blackbody. This remark allows us to make use of blackbody theory to determine the emission from objects whose temperature varies greatly from place to place, even though blackbody theory applies, strictly speaking, only to extensive bodies with uniform temperature. For example, the temperature of the core of the Earth is about $6000K$, but we need not know this in order to determine the radiation emitted from the Earth's surface; the outermost few millimeters of rock, ice or water at the Earth's surface contain enough matter to act like a blackbody to a very good approximation. Hence, the radiation emitted from the surface depends only on the temperature of this outer skin of the planet. Similarly, the temperature of the core of the Sun is about $16,000,000K$ and even at a distance from the center equal to 90% of the visible radius, the temperature is above $600,000K$. However, the Sun is encased in a layer a few hundred kilometers thick which is sufficiently dense to act like a blackbody, and which has a temperature of about $5780K$. This layer is known as the *photosphere*, because it is the source of most light exiting the Sun. Layers farther out from the center of the Sun can be considerably hotter than the photosphere, but they have a minimal effect on solar radiation because they are so tenuous. In Chapter 4 we will develop more precise methods for dealing with tenuous objects, such as atmospheres, which peter out gradually without having a sharply defined boundary.

An ideal blackbody would be opaque at all wavelengths, but it is a common situation that a material acts as a blackbody only in a limited range of wavelengths. Consider the case of window glass: It is transparent to visible light, but if you could see it in the infrared it would look as opaque as stone. Because it interacts strongly with infrared light, window glass emits blackbody radiation in the infrared range. At temperatures below a few hundred K, there is little blackbody emission at wavelengths shorter than the infrared, so at such temperatures the net power per unit area emitted by a pane of glass with temperature T is very nearly σT^4, even though it doesn't act like a blackbody in the visible range. Liquid water, and water ice, behave similarly. Crystalline table salt, and carbon dioxide ice, are nearly transparent in the infrared as well as in the visible, and in consequence emit radiation at a much lower rate than expected from the blackbody formula. (They would make fine windows for creatures having infrared vision). There is, in fact, a deep and important relation between absorption and emission of radiation, which will be discussed in Section 3.5.

3.3 Radiation balance of planets

As a first step in our study of the temperature of planets, let's consider the following idealized case:

[2] The definite integral $\int_0^\infty (u^3/(e^u - 1))du$ was determined by Euler, as a special case of his study of the behavior of the Riemann zeta function at even integers. It is equal to $6\zeta(4) = \pi^4/15$

- The only source of energy heating the planet is absorption of light from the planet's host star.
- The *albedo*, or proportion of sunlight reflected, is spatially uniform.
- The planet is spherical, and has a distinct solid or liquid surface which radiates like a perfect blackbody.
- The planet's temperature is uniform over its entire surface.
- The planet's atmosphere is perfectly transparent to the electromagnetic energy emitted by the surface.

The uniform-temperature assumption presumes that the planet has an atmosphere or ocean which is so well stirred that it is able to rapidly mix heat from one place to another, smoothing out the effects of geographical fluctuations in the energy balance. The Earth conforms fairly well to this approximation. The equatorial annual mean temperature is only 4% above the global mean temperature of $286K$, while the North polar temperature is only 10% below the mean. The most extreme deviation occurs on the high Antarctic plateau, where the annual mean South polar temperature is 21% below the global mean. The surface temperature of Venus is even more uniform than that of Earth. That of Mars, which in our era, has a thin atmosphere and no ocean, is less uniform. Airless, rocky bodies like the Moon and Mercury do not conform at all well to the uniform temperature approximation.

Light leaving the upper layers of the Sun and most other stars takes the form of blackbody radiation. It is isotropic, and its flux and flux spectrum conform to the blackbody law corresponding to the temperature of the photosphere, from which the light escapes. Once the light leaves the surface of the star, however, it expands through space and does not interact significantly with matter except where it is intercepted by a planet. Therefore, it is no longer blackbody radiation, though it retains the blackbody spectrum. In the typical case of interest, the planet orbits its star at a distance that is much greater than the radius of the star, and itself has a radius that is considerably smaller than the star and is hence yet smaller than the orbital distance. In this circumstance, all the rays of light which intersect the planet are very nearly parallel to the line joining the center of the planet to the center of its star; the sunlight comes in as a nearly *parallel beam*, rather than being isotropic, as would be the case for true blackbody radiation. The parallel-beam approximation is equivalent to saying that, as seen from the planet, the Sun occupies only a small portion of the sky, and as seen from the Sun the planet also occupies only a small portion of the sky. Even for Mercury, with a mean orbital distance of $58,000,000km$, the Sun (whose radius is $695,000km$) occupies an angular width in the sky of only about $2 \cdot 695,000/58,000,000$ radians, or 1.4^o.

The solar flux impinging on the planet is also reduced, as compared to the solar flux leaving the photosphere of the star. The total energy per unit frequency leaving the star is $4\pi r_\circledast^2(\pi B(\nu, T_\circledast))$, where $r_\circledast$ is the radius of the star and $T_\circledast$ is the temperature of its photosphere. At a distance r from the star, the energy has spread uniformly over a sphere whose surface area is $4\pi r^2$; hence at this distance, the energy flux per unit frequency is $\pi B r_\circledast^2/r^2$, and the total flux is $\sigma T_\circledast^4 r_\circledast^2/r^2$. The latter is the flux seen by a planet at orbital distance r, in the form of a beam of parallel rays. It is known as the solar "constant" , and will be denoted by $L_\circledast$, or simply L where there is no risk of confusion with latent heat. The solar (or stellar) "constant" depends on a planet's orbit, but the *luminosity* of the star is an intrinsic property of the star. The stellar luminosity is the net power output of a star, and if the star's emission can be represented as blackbody radiation, the luminosity is given by $\mathcal{L}_\circledast = 4\pi r_\circledast^2 T_\circledast^4$.

We are now equipped to compute the energy balance of the planet, subject to the preceding simplifying assumptions. Let a be the planet's radius. Since the cross-section area of the planet is πa^2 and the solar radiation arrives in the form of a nearly parallel beam with flux $L_\circledast$, the energy per unit time impinging on the planet's surface is $\pi a^2 L_\circledast$; the rate of energy absorption is $(1-\alpha)\pi a^2 L_\circledast$, where α is the albedo. The planet loses energy by radiating from its entire surface, which has area $4\pi a^2$. Hence the rate of energy loss is $4\pi a^2 \sigma T^4$, where T is the temperature of the planet's surface. In equilibrium the rate of energy loss and gain must be equal. After cancelling a few terms, this yields

$$\sigma T^4 = \frac{1}{4}(1-\alpha)L_\circledast \tag{3.6}$$

Note that this is independent of the radius of the planet. The factor $\frac{1}{4}$ comes from the ratio of the planet's cross-sectional area to its surface area, and reflects the fact that the planet intercepts only a disk of the incident solar beam, but radiates over its entire spherical surface. This equation can be readily solved for T. If we substitute for $L_\circledast$ in terms of the photospheric temperature, the result is

$$T = \frac{1}{\sqrt{2}}(1-\alpha)^{1/4}\sqrt{\frac{r_\circledast}{r}}T_\circledast \tag{3.7}$$

Formula 3.7 shows that the blackbody temperature of a planet is much less than that of the photosphere, so long as the orbital distance is large compared to the stellar radius. From the displacement law, it follows that the planet loses energy through emission at a distinctly lower wavenumber than that at which it receives energy from its star. This situation is illustrated in Figure 3.3. For example, the energy received from our Sun has a median wavenumber of about 15000 cm^{-1}, equivalent to a wavelength of about .7 μm. An isothermal planet at Mercury's orbit would radiate to space with a median emission wavenumber of 1100 cm^{-1}, corresponding to a wavelength of 9 μm. An isothermal planet at the orbit of Mars would radiate with a median wavenumber of 550 cm^{-1}, corresponding to a wavelength of 18 μm.

Exercise 3.3.1 A planet with zero albedo is in orbit around an exotic hot star having a photospheric temperature of $100,000K$. The ratio of the planet's orbit to the radius of the star is the same as for Earth (about 215). What is the median emission wavenumber of the star? In what part of the electromagnetic spectrum does this lie? What is the temperature of the planet? In what part of the electromagnetic spectrum does the planet radiate? Do the same if the planet is instead in orbit around a brown dwarf star with a photospheric temperature of $600K$.

The separation between absorption and emission wavenumber will prove very important when we bring a radiatively active atmosphere into the picture, since it allows the atmosphere to have a different effect on incoming vs. outgoing radiation. Since the outgoing radiation has longer wavelength than the incoming radiation, the flux of emitted outgoing radiation is often referred to as *outgoing longwave radiation*, and denoted by OLR. For a non-isothermal planet, the OLR is a function of position (e.g. latitude and longitude on an imaginary sphere tightly enclosing the planet and its atmosphere). We will also use the term to refer to the outgoing flux averaged over the surface of the sphere, even when the planet is not isothermal. As for the other major term in the planet's energy budget, we will refer to the electromagnetic energy received from the planet's star as the *shortwave* or *solar* energy. Our own Sun has its primary output in the visible part of the spectrum, but it also emits significant amounts of energy in the ultraviolet and near-infrared, both of which are shorter in wavelength than the OLR by which planets lose energy to space.

Formula 3.7 is plotted in Figure 3.4 for a hypothetical isothermal planet with zero albedo. Because of the square-root dependence on orbital distance, the temperature varies only weakly with distance, except very near the star. Neglecting albedo and atmospheric effects, Earth would

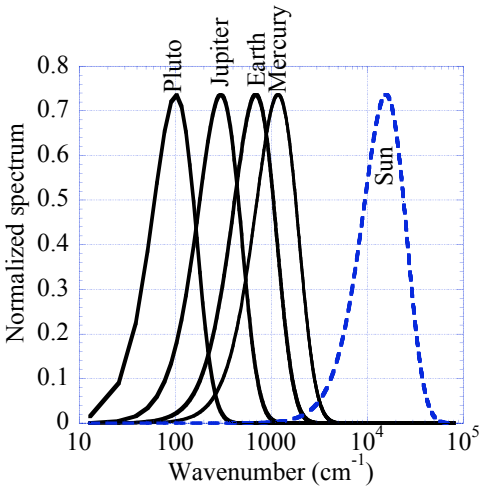

Figure 3.3: The Planck density of radiation emitted by the Sun and selected planets in radiative equilibrium with absorbed solar radiation (based on the observed shortwave albedo of the planets). The Planck densities are transformed to a logarithmic spectral coordinate, and all are normalized to unit total emission.

have a mean surface temperature of about $280K$. Venus would be only 50K warmer than the Earth and Mars only $53K$ colder. At the distant orbit of Jupiter, the blackbody equilibrium temperature falls to $122K$, but even at the vastly more distant orbit of Neptune the temperature is still as high as $50K$. The emission from all of these planets lies in the infrared range, though the colder planets radiate in the deeper (lower wavenumber) infrared. An exception to the strong separation between stellar and planetary temperature is provided by the "roasters" – a recently discovered class of extrasolar giant planets with $\frac{r}{r_\circledcirc}$ as low as 5. Such planets can have equilibrium blackbody temperatures as much as a third that of the photosphere of the parent star. For these planets, the distinction between the behavior of incoming and outgoing radiation is less sharp.

It is instructive to compare the ideal blackbody temperature with observed surface temperature for the three Solar System bodies which have both a distinct surface and a thick enough atmosphere to enforce a roughly uniform surface temperature: Venus, Earth and Saturn's moon Titan. For this comparison, we calculate the blackbody temperature using the observed planetary albedos, instead of assuming a hypothetical zero albedo planet as in Fig. 3.4. Venus is covered by thick, highly reflective clouds, which raise its albedo to .75. The corresponding isothermal blackbody temperature is only $232K$ (as compared to $330K$ in the zero albedo case). This is far less than the observed surface temperature of $740K$. Clearly, the atmosphere of Venus exerts a profound warming effect on the surface. The warming arises from the influence of the atmosphere on the infrared emission of the planet, which we have not yet taken into account. Earth's albedo is on the order of .3, leading to a blackbody temperature of $255K$. The observed mean surface temperature is about 285K. Earth's atmosphere has a considerably weaker warming effect than that of Venus, but it is nonetheless a very important warming, since it brings the planet from subfreezing temperatures where the oceans would almost certainly become ice-covered, to temperatures where liquid water can exist over most of the planet. The albedo of Titan is .21, and using the solar constant

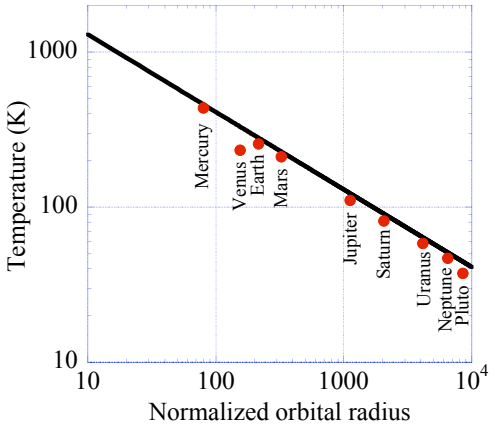

Figure 3.4: The equilibrium blackbody temperature of an isothermal spherical zero-albedo planet, as a function of distance from a Sun having a photospheric temperarature of 5800K. The orbital distance is normalized by the radius of the Sun. Dots show the equilibrium blackbody temperature of the Solar System planets, based on their actual observed albedos.

at Saturn's orbit we find a black body temperature of $85K$. The observed surface temperature is about $95K$, whence we conclude that the infrared effects of Titan's atmosphere moderately warm the surface.

The way energy balance determines surface temperature is illustrated graphically in Figure 3.5. One first determines the way in which the mean infrared emission per unit area depends on the mean surface temperature T_s; for the isothermal blackbody calculation, this curve is simply σT_s^4. The equilibrium temperature is determined by the point at which the OLR curve intersects the curve giving the absorbed solar radiation (a horizontal line in the present calculation). In some sense, the whole subject of climate comes down to an ever-more sophisticated heirarchy of calculations of the curve $OLR(T_s)$; our attention will soon turn to the task of determining how the OLR curve is affected by an atmosphere. With increasing sophistication, we will also allow the solar absorption to vary with T_s, owing to changing clouds, ice cover, vegetation cover, and other characteristics.

We will now consider an idealized thought experiment which illustrates the essence of the way an atmosphere affects OLR. Suppose that the atmosphere has a temperature profile $T(p)$ which decreases with altitude,according to the dry or moist adiabat. Let p_s be the surface pressure, and suppose that the ground is strongly thermally coupled to the atmosphere by turbulent heat exchanges, so that the ground temperature cannot deviate much from that of the immediately overlying air. Thus, $T_s = T(p_s)$. If the atmosphere were transparent to infrared, as is very nearly the case for nitrogen or oxygen, the OLR would be σT_s^4. Now, let's stir an additional gas into the atmosphere, and assume that it is well mixed with uniform mass concentration q. This gas is transparent to solar radiation but interacts strongly enough with infrared that when a sufficient amount is mixed into a parcel of air, it turns that parcel into an ideal blackbody. Such a gas, which is fairly transparent to the incoming shortwave stellar radiation but which interacts

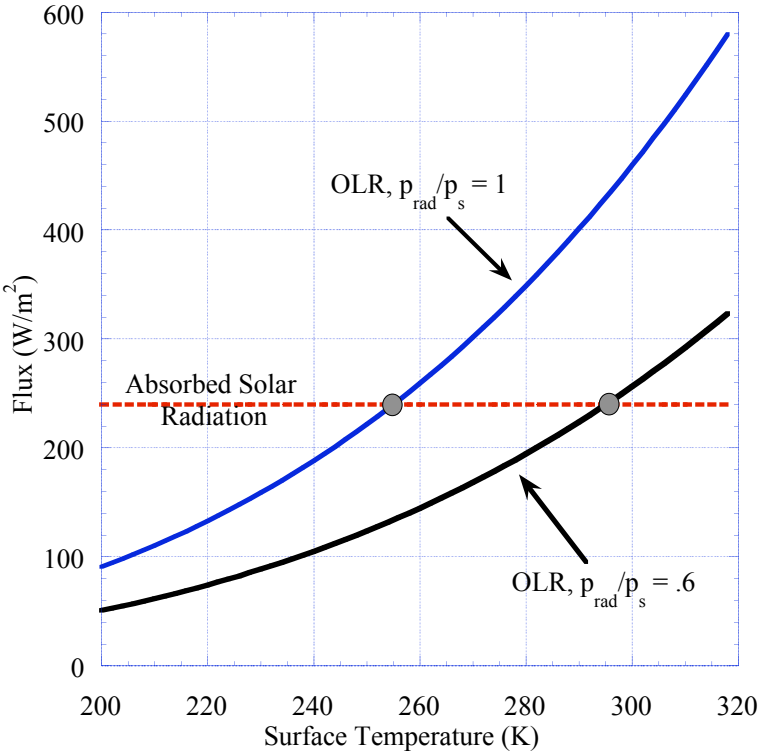

Figure 3.5: Determination of a planet's temperature by balancing absorbed solar energy against emitted longwave radiation. The horizontal line gives the absorbed solar energy per unit surface area, based on an albedo of .3 and a Solar constant of $1370 W/m^2$. The OLR is given as a function of surface temperature. The upper curve assumes the atmosphere has no greenhouse effect ($p_{rad} = p_s$), while the lower OLR curve assumes $p_{rad}/p_s = .6$, a value appropriate to the present Earth.

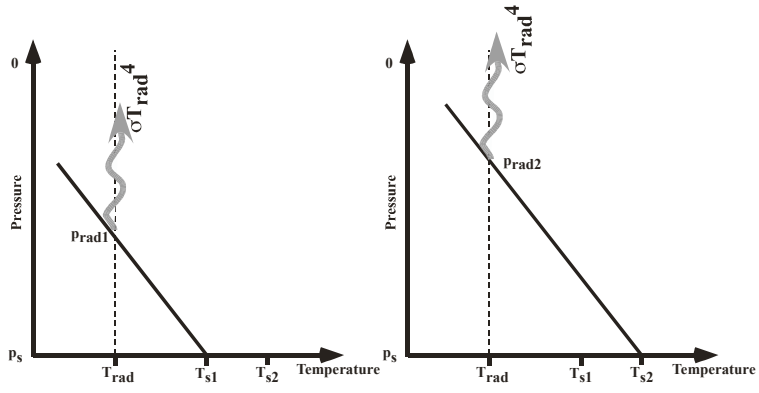

Figure 3.6: Sketch illustrating how the greenhouse effect increases the surface temperature. In equilibrium, the outgoing radiation must remain equal to the absorbed solar radiation, so T_{rad} stays constant. However, as more greenhouse gas is added to the atmosphere, p_{rad} is reduced, so one must extrapolate temperature further along the adiabat to reach the surface.

strongly with the outgoing (generally infrared) emitted radiation is called a *greenhouse gas*, and the corresponding effect on planetary temperature is called the *greenhouse effect*. Carbon dioxide, water vapor and methane are some examples of greenhouse gases, and the molecular properties that make a substance a good greenhouse gas will be discussed in Chapter 4. The mass of greenhouse gas that must be mixed into a column of atmosphere with base of 1 m^2 in order to make that column act begin to act like a blackbody is characterized by the *absorption coefficient* κ, whose units are m^2/kg. Here we'll assume κ to be independent of frequency, temperature and pressure, though for real greenhouse gases, κ depends on all of these. Since the mass of greenhouse gas in a column of thickness Δp in pressure coordinates is $q\Delta p/g$, then the definition of κ implies that the slab acts like a blackbody when $\kappa q \Delta p/g > 1$. When $\kappa q p_s/g < 1$ then the entire mass of the atmosphere is not sufficient to act like a blackbody and the atmosphere is said to be *optically thin*. For optically thin atmospheres, infrared radiation can escape from the surface directly to space, and is only mildly attenuated by atmospheric absorption. When $\kappa q p_s/g \gg 1$, the atmosphere is said to be *optically thick*.

If the atmosphere is optically thick, we can slice the atmosphere up into a stack of slabs with thickness Δp_1 such that $\kappa q \Delta p_1/g = 1$. Each of these slabs radiates like an ideal blackbody with temperature approximately equal to the mean temperature of the slab. Recall, however, that another fundamental property of blackbodies is that they are perfect *absorbers* (though if they are only blackbodies in the infrared, they will only be perfect absorbers in the infrared). Hence *infrared radiation escapes to space only from the topmost slab*. The OLR will be determined by the temperature of this slab alone, and will be insensitive to the temperature of lower portions of the atmosphere. The pressure at the bottom of the topmost slab is Δp_1. We can thus identify Δp_1 as the characteristic pressure level from which radiation escapes to space, which therefore will be called p_{rad} in subsequent discussions. The radiation escaping to space – the OLR– will then be approximately $\sigma T(p_{rad})^4$. Because temperature decreases with altitude on the adiabat the OLR is less than σT_s^4 to the extent that $p_{rad} < p_s$. As shown in Figure 3.5, a greenhouse gas acts like an

insulating blanket, reducing the rate of energy loss to space at any given surface temperature. All other things being equal the equilibrium surface temperature of a planet with a greenhouse gas in its atmosphere must be greater than that of a planet without a greenhouse gas, in order to radiate away energy at a sufficient rate to balance the absorbed solar radiation. The key insight to be taken from this discussion is that *the greenhouse effect only works to the extent that the atmosphere is colder at the radiating level then it is at the ground.*

For real greenhouse gases, the absorption coefficient varies greatly with frequency. Such gases act on the OLR by making the atmosphere very optically thick at some frequencies, less optically thick at others, and perhaps even optically thin at still other frequencies. In portions of the spectrum where the atmosphere is more optically thick, the emission to space originates in higher (and generally colder) parts of the atmosphere. In reality,then, the infrared escaping to space is a blend of radiation emitted from a range of atmospheric levels, with some admixture of radiation from the planet's surface as well. The concept of an *effective radiating level* nonetheless has merit for real greenhouse gases. It does not represent a distinct physical layer of the atmosphere, but rather characterizes the mean depth from which infrared photons escape to space. As more greenhouse gas is added to an atmosphere, more of the lower parts of the atmosphere become opaque to infrared, preventing the escape of infrared radiation from those regions. This increases the altitude of the effective radiating level (i.e. decreases p_{rad}). Some of the implications of a frequency-dependent absorption coefficient are explored in Problem **??**, and the subject will be taken up at great length in Chapter 4.

From an observation of the actual OLR emitted by a planet, one can determine an equivalent blackbody radiating temperature T_{rad} from the expression $\sigma T_{rad}^4 = OLR$. This temperature is the infrared equivalent of the Sun's photospheric temperature; it is a kind of mean temperature of the regions from which infrared photons escape, and p_{rad} represents a mean pressure of these layers. For planets for which absorbed solar radiation is the only significant energy source, T_{rad} is equal to the ideal blackbody temperature given by Eq. 3.7. The arduous task of relating the effective radiating level to specified concentrations of real greenhouse gases is treated in Chapter 4.

Figure 3.7 illustrates the reduction of infrared emission caused by the Earth's atmosphere. At every latitude, the observed OLR is much less than it would be if the planet radiated to space at its observed surface temperature. At the Equator the observed OLR is $238 W/m^2$, corresponding to a radiating temperature of $255K$. This is much less than the observed surface temperature of $298K$, which would radiate at a rate of $446 W/m^2$ if the atmosphere didn't intervene. It is interesting that the gap between observed OLR and the computed surface emission is less in the cold polar regions, and especially small at the Winter pole. This happens partly because, at low temperatures, there is simply less infrared emission for the atmosphere to trap. However, differences in the water content of the atmosphere, and differences in the temperature profile, can also play a role. These effects will be explored in Chapter 4.

Gases are not the only atmospheric constituents which affect OLR. Clouds consist of particles of condensed substance small enough to stay suspended for a long time. They can profoundly influence OLR. Gram for gram, condensed water interacts much more strongly with infrared than does water vapor. In fact, a mere 20 grams of water in the form of liquid droplets of a typical size is sufficient to turn a column of air $500m$ thick by one meter square into a very nearly ideal blackbody. To a much greater extent than for greenhouse gases, a water cloud layer in an otherwise infrared-transparent atmosphere really can be thought of as a discrete radiating layer. The prevalance of clouds in the high, cold regions of the tropical atmosphere accounts for the dip in OLR near the equator, seen in Figure 3.7. Clouds are unlike greenhouse gases, though, since they also strongly reflect the incoming solar radiation. It's the tendency of these two large effects to partly cancel that makes the problem of the influence of clouds on climate so challenging. Not all

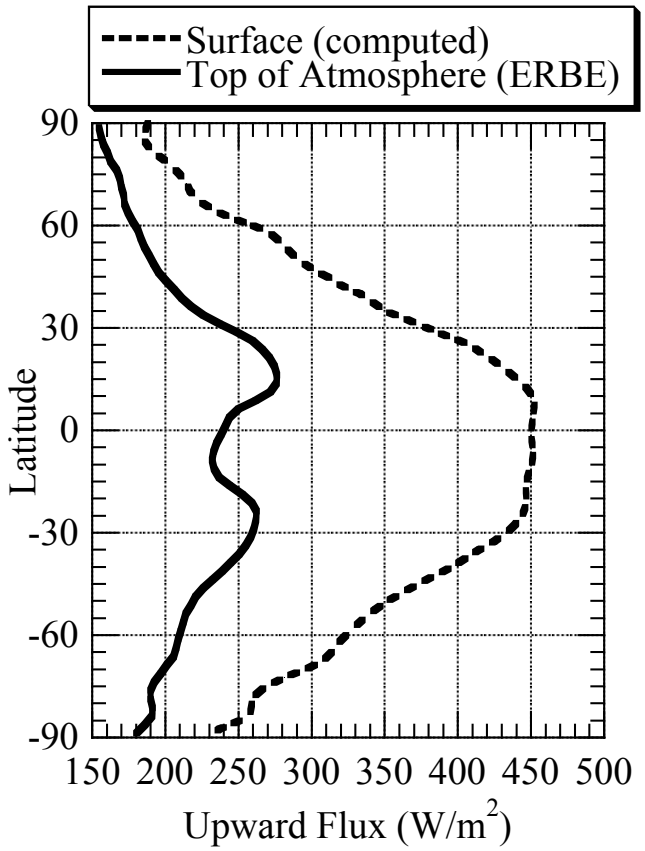

Figure 3.7: The Earth's observed zonal-mean OLR for January, 1986. The observations were taken by satellite instruments during the Earth Radiation Budget Experiment (ERBE), and are averaged along latitude circles. The figure also shows the radiation that would be emitted to space by the surface (σT_s^4) if the atmosphere were transparent to infrared radiation.

condensed substances absorb infrared as well as water does. Liquid methane (imporant on Titan) and CO2 ice (important on present and early Mars) are comparatively poor infrared absorbers. They affect OLR in a fundamentally different way, through reflection instead of absorption and emission. This will be discussed in Chapter 5.

In a nutshell, then, here is how the greenhouse effect works: From the requirement of energy balance, the absorbed solar radiation determines the effective blackbody radiating temperature T_{rad}. This is not the surface temperature; it is instead the temperature encountered at some pressure level in the atmosphere p_{rad}, which characterizes the infrared opacity of the atmosphere, specifically the typical altitude from which infrared photons escape to space. The pressure p_{rad} is determined by the greenhouse gas concentration of the atmosphere. The surface temperature is determined by starting at the fixed temperature T_{rad} and extrapolating from p_{rad} to the surface pressure p_s using the atmosphere's lapse rate, which is approximately governed by the appropriate adiabat. Since temperature decreases with altitude over much of the depth of a typical atmosphere, the surface temperature so obtained is typically greater than T_{rad}, as illustrated in Figure 3.6. Increasing the concentration of a greenhouse gas decreases p_{rad}, and therefore increases the surface temperature because temperature is extrapolated from T_{rad} over a greater pressure range. It is very important to recognize that greenhouse warming relies on the decrease of atmospheric temperature with height, which is generally due to the adiabatic profile established by convection. The greenhouse effect works by allowing a planet to radiate at a temperature colder than the surface, but for this to be possible, there must be some cold air aloft for the greenhouse gas to work with.

For an atmosphere whose temperature profile is given by the dry adiabat, the surface temperature is

$$T_s = (p_s/p_{rad})^{R/c_p} T_{rad}. \tag{3.8}$$

With this formula, the Earth's present surface temperature can be explained by taking $p_{rad}/p_s = .67$, whence $p_{rad} \approx 670mb$. Earth's actual radiating pressure is somewhat lower than this estimate, because the atmosperic temperature decays less strongly with height than the dry adiabat. The high surface temperature of Venus can be accounted for by taking $p_{rad}/p_s = .0095$, assuming that the temperature profile is given by the noncondensing adiabat for a pure CO_2 atmosphere. Given Venus' $93bar$ surface pressure, the radiating level is $880mb$ which, interestingly, is only slightly less than Earth's surface pressure. Earth radiates to space from regions quite close to its surface, whereas Venus radiates only from a thin shell near the top of the atmosphere. Note that from the observed Venusian temperature profile in Fig. 2.2, the radiating temperature ($253K$) is encountered at $p = 250mb$ rather than the higher pressure we estimated. As for the Earth, our estimate of the precise value p_{rad} for Venus is off because the ideal-gas noncondensing adiabat is not a precise model of the actual temperature profile. In the case of Venus, the problem most likely comes from the ideal-gas assumption and neglect of variations in c_p, rather than condensation.

The concept of radiating level and radiating temperature also enables us to make sense of the way energy balance constrains the climates of gas giants like Jupiter and Saturn, which have no distinct surface. The essence of the calculation we have already done for rocky planets is to use the top of atmosphere energy budget to determine the parameters of the adiabat, and then extrapolate temperature to the surface along the adiabat. For a non-condensing adiabat, the atmospheric profile compatible with energy balance is $T(p) = T_{rad}(p/p_{rad})^{R/c_p}$. This remains the appropriate temperature profile for a (noncondensing) convecting outer layer of a gas giant, and the only difference with the previous case is that, for a gas giant, there is no surface to act as a natural lower boundary for the adiabatic region. At some depth, convection will give out and the adiabat must be matched to some other temperature model in order to determine the base of the convecting region, and to determine the temperature of deeper regions. There is no longer

	Observed OLR (W/m^2)	Absorbed Solar Flux (W/m^2)	T_{rad} (actual)	T_{rad} (Solar only)
Jupiter	14.3	12.7	126K	110K
Saturn	4.6	3.8	95K	81K
Uranus	.52	.93	55K	58K
Neptune	.61	.38	57K	47K

Table 3.1: The energy balance of the gas giant planets, with inferred radiating temperature. The solar-only value of T_{rad} is the radiating temperature that would balance the observed absorbed solar energy, in the absence of any internal heat source.

any distinct surface to be warmed by the greenhouse effect, but the greenhouse gas concentration of the atmosphere nonetheless affects $T(p)$ through p_{rad}. For example, adding some additional greenhouse gas to the convecting outer region of Jupiter's atmosphere would decrease p_{rad}, and therefore increase the temperature encountered at, say, the 1 bar pressure level.

The energy balance suffices to uniquely determine the temperature profile because the non-condensing adiabat is a one-parameter family of temperature profiles. The saturated adiabat for a mixture of condensing and noncondensing gases is also a one parameter family, defined by Eq. 2.33, and can therefore be treated similarly. If the appropriate adiabat for the planet had more than one free parameter, additional information beyond the energy budget would be needed to close the problem. On the other hand, a single component condensing atmosphere such as described by Eq. 2.27 yields a temperature profile with no free parameters that can be adjusted so as to satisfy the energy budget. The consequences of this quandary will be taken up as part of our discussion of the runaway greenhouse phenonenon, in Chapter 4.

Using infrared telescopes on Earth and in space, one can directly measure the OLR of the planets in our Solar System. In the case of the gas giants, the radiated energy is substantially in excess of the absorbed solar radiation. Table 3.1 compares the observed OLR to the absorbed solar flux for the gas giants. With the exception of Uranus, the gas giants appear to have a substantial internal energy source, which raises the radiating temperature to values considerably in excess of it would be if the planet were heated by solar absorption alone. Uranus is anomalous, in that it actually appears to be emitting less energy than it receives from the sun. Uncertainties in the observed OLR for Uranus would actually allow the emission to be in balance with solar absorption, but would still appear to preclude any significant internal energy source. This may indicate a profound difference in the internal dynamics of Uranus. On the other hand, the unusually large tilt of Uranus' rotation axis means that Uranus has an unusually strong seasonal variation of solar heating, and it may be that the hemisphere that has been observed so far has not yet had time to come into equilibrium, which would throw off the energy balance estimate.

Because it is the home planet, Earth's radiation budget has been very closely monitored by satellites. Indirect inferences based on the rate of ocean heat uptake indicate that the top of atmosphere radiation budget is currently out of balance, the Earth receiving about $1W/m^2$ more from Solar absorption than it emits to space as infrared [3]. This is opposite from the imbalance that would be caused by an internal heating. It is a direct consequence of the rapid rise of CO_2 and other greenhouse gases, caused by the bustling activities of Earth's human inhabitants. The rapid greenhouse gas increase has cut down the OLR, but because of the time required to warm up the oceans and melt ice, the Earth's temperature has not yet risen enough to restore the energy balance.

[3] At the time of writing, top-of-atmosphere satellite measurements are not sufficiently accurate to permit direct observation of this imbalance

Exercise 3.3.2 A typical well-fed human in a resting state consumes energy in the form of food at a rate of $100W$, essentially all of which is put back into the surroundings in the form of heat. An astronaut is in a spherical escape pod of radius r, far beyond the orbit of Pluto, so that it receives essentially no energy from sunlight. The air in the escape pod is isothermal. The skin of the escape pod is a good conductor of heat, so that the surface temperature of the sphere is identical to the interior temperature. The surface radiates like an ideal blackbody.

Find an expression for the temperature in terms of r, and evaluate it for a few reasonable values. Is it better to have a bigger pod or a smaller pod? In designing such an escape pod, should you include an additional source of heat if you want to keep the astronaut comfortable?

How would your answer change if the pod were cylindrical instead of spherical? If the pod were cubical?

Bodies such as Mercury or the Moon represent the opposite extreme from the uniform-temperature limit. Having no atmosphere or ocean to transport heat, and a rocky surface through which heat is conducted exceedingly slowly, each bit of the planet is, to a good approximation, thermally isolated from the rest. Moreover, the rocky surface takes very little time to reach its equilibrium temperature, so the surface temperature at each point is very nearly in equilibrium with the instantaneous absorbed solar radiation, with very little day-night or seasonal averaging. In this case, averaging the energy budget over the planet's surface gives a poor estimate of the temperature, and it would be more accurate to compute the instantaneous equilibrium temperature for each patch of the planet's surface in isolation. For example, consider a point on the planet where the Sun is directly overhead at some particular instant of time. At that time, the rays of sunlight come in perpendicularly to a small patch of the ground, and the absorbed solar radiation per unit area is simply $(1-\alpha)L_\circledast$; the energy balance determing the ground temperature is then $\sigma T^4 = (1-\alpha)L_\circledast$, without the factor of $\frac{1}{4}$ we had when the energy budget was averaged over the entire surface of an isothermal planet. For Mercury, this yields a temperature of $622K$, based on the mean orbital distance and an albedo of .1. This is similar to the observed maximum temperature on Mercury, which is about $700K$ (somewhat larger than the theoretical calculation because Mercury's highly elliptical orbit brings it considerably closer to the Sun than the mean orbital position). The Moon, which is essentially in the same orbit as Earth and shares its Solar constant, has a predicted maximum temperature of $384K$, which is very close to the observed maximum. In contrast, the maximum surface temperature on Earth stays well short of $384K$, even at the hottest time of day in the hottest places. The atmosphere of Mars in the present epoch is thin enough that this planet behaves more like the no-atmosphere limit than the uniform-temperature limit. Based on a mean albedo of .25,the local maximum temperature should be $297K$, which is quite close to the observed maximum temperature.

More generally speaking, when doing energy balance calculations the temperature we have in mind is the temperature averaged over an appropriate portion of the planet and over an appropriate time interval, where what is "appropriate" depends on the response time and the efficiency of the heat transporting mechanisms of the planet under considerations. Correspondingly, the appropriate incident solar flux to use is the incident solar flux per unit of radiating surface, averaged consistently with temperature. We will denote this mean solar flux by the symbol S. For an isothermal planet $S = \frac{1}{4}L_\circledast$, while at the opposite extreme $S = L_\circledast$ for the instantaneous response at the subsolar point. In other circumstances it might be appropriate to average along a latitude circle, or over a hemisphere. A more complete treatment of geographical, seasonal and diurnal temperature variations will be given in Chapter 7.

Exercise 3.3.3 Consider a planet which is tide-locked to its Sun, so that it always shows the same

Surface type	**Albedo**
Clean new H_2O snow	.85
Bare Sea ice	.5
Clean H_2O glacier ice	.6
Deep Water	.1
Sahara Desert sand	.35
Martian sand	.15
Basalt (any planet)	.07
Granite	.3
Limestone	.36
Grassland	.2
Deciduous forest	.14
Conifer forest	.09
Tundra	.2

Table 3.2: Typical values of albedo for various surface types. These are only representative values. Albedo can vary considerably as a function of detailed conditions. For example, the ocean albedo depends on the angle of the solar radiation striking the surface (the value given in the table is for near-normal incidence), and the albedo of bare sea ice depends on the density of air bubbles.

face to the Sun as it proceeds in its orbit (just as the Moon always shows the same face to the Earth). Estimate the mean temperature of the day side of the planet, assuming the illuminated face to be isothermal, but assuming that no heat leaks to the night side.

3.4 Ice-albedo feedback

Albedo is not a static quantity determined once and for all time when a planet forms. In large measure, albedo is determined by processes in the atmosphere and at the surface which are highly sensitive to the state of the climate. Clouds consist of suspended tiny particles of the liquid or solid phase of some atmospheric constituent; such particles are very effective reflectors of visible and ultraviolet light, almost regardless of what they are made of. Clouds almost entirely control the albedos of Venus, Titan and all the gas giant planets, and also play a major role in Earth's albedo. In addition, the nature of a planet's surface can evolve over time, and many of the surface characteristics are strongly affected by the climate. Table 3.2 gives the albedo of some common surface types encountered on Earth. The proportions of the Earth covered by sea-ice, snow, glaciers, desert sands or vegetation of various types are determined by temperature and precipitation patterns. As climate changes, the surface characteristics change too, and the resulting albedo changes feed back on the state of the climate. It is not a "chicken and egg" question of whether climate causes albedo or albedo causes climate; rather it is a matter of finding a consistent state compatible with the physics of the way climate affects albedo and the way albedo affects climate. In this sense, albedo changes lead to a form of climate *feedback*. We will encounter many other kinds of feedback loops in the climate system.

Among all the albedo feedbacks, that associated with the cover of the surface by highly reflective snow or ice plays a distinguished role in thinking about the evolution of the Earth's climate. Let's consider how albedo might vary with temperature for a planet entirely covered by a water ocean – a reasonable approximation to Earth, which is $\frac{2}{3}$ ocean. We will characterize the climate by the global mean surface temperature T_s, but suppose that, like Earth, the temperature

is somewhat colder than T_s at the poles and somewhat warmer than T_s at the Equator. When T_s is very large, say greater than some threshold temperature T_o, the temperature is above freezing everywhere and there is no ice. In this temperature range, the planetary albedo reduces to the relatively low value (call it α_o) characteristic of sea water. At the other extreme, when T_s is very, very low, the whole planet is below freezing, the ocean will become ice-covered everywhere, and the albedo reduces to that of sea ice, which we shall call α_i. We suppose that this occurs for $T_s < T_i$, where T_i is the threshold temperature for a globally frozen ocean. In general T_i must be rather lower than the freezing temperature of the ocean, since when the mean temperature $T_s = T_{freeze}$ the equatorial portions of the planet will still be above freezing. Between T_i and T_o it is reasonable to interpolate the albedo by assuming the ice cover to decrease smoothly and monotonically from 100% to zero. The phenomena we will emphasize are not particularly sensitive to the detailed form of the interpolation, but the quadratic interpolation

$$\alpha(T) = \begin{cases} \alpha_i & \text{for } T \le T_i, \\ \alpha_o + (\alpha_i - \alpha_o)\frac{(T-T_o)^2}{(T_i-T_o)^2} & \text{for } T_i < T < T_o \\ \alpha_o & \text{for } T \ge T_o \end{cases} \tag{3.9}$$

qualitatively reproduces the shape of the albedo curve which is found in detailed calculations. In particular, the slope of albedo vs temperature is large when the temperature is low and the planet is nearly ice-covered, because there is more area near the Equator, where ice melts first. Conversely, the slope reduces to zero as the temperature threshold for an ice-free planet is approached, because there is little area near the poles where the last ice survives; moreover, the poles receive relatively little sunlight in the course of the year, so the albedo there contributes less to the global mean than does the albedo at lower latitudes. Note that this description assumes an Earthlike planet, which on average is warmest near the Equator. As will be discussed in Chapter 7, other orbital configurations could lead to the poles being warmer, and this would call for a different shape of albedo curve.

Ice albedo feedback of a similar sort could arise on a planet with land, through snow accumulation and glacier formation on the continents. The albedo could have a similar temperature dependence, in that glaciers are unlikely to survive where temperatures are very much above freezing, but can accumulate readily near places that are below freezing – *provided there is enough precipitation*. It is the latter requirement that makes land-based snow/ice albedo feedback much more complicated than the oceanic case. Precipitation is determined by complex atmospheric circulation patterns that are not solely determined by local temperature. A region with no precipitation will not form glaciers no matter how cold it is made. The present state of Mars provides a good example: its small polar glaciers do not advance to the Equator, even though the daily average equatorial temperature is well below freezing. Still, for a planet like Earth with a widespread ocean to act as a source for precipitation, it may be reasonable to assume that most continental areas will eventually become ice covered if they are located at sufficiently cold latitudes. In fairness, we should point out that even the formation of sea ice is considerably more complex than we have made it out to be, particularly since it is affected by the mixing of deep unfrozen water with surface waters which are trying to freeze.

Earth is the only known planet that has an evident ice/snow albedo feedback, but it is reasonable to inquire as to whether a planet without Earth's water-dominated climate could behave analogously. Snow is always "white" more or less regardless of the substance it is made of, since its reflectivity is due to the refractive index discontinuity between snow crystals and the ambient gas or vacuum. Therefore, a snow-albedo feedback could operate with substances other than water (e.g. nitrogen or methane). Titan presents an exotic possibility, in that its surface is bathed in a rain of tarry hydrocarbon sludge, raising the speculative possibility of "dark glacier" albedo

feedbacks. Sea ice forming on Earth's ocean gets its high albedo from trapped air bubbles, which act like snowflakes in reverse. The same could happen for ices of other substances, but sea-ice albedo feedback is likely to require a water ocean. The reason is that water, alone among likely planetary materials, floats when it freezes. Ice forming on, say, a carbon dioxide or methane ocean would sink as soon as it formed, preventing it from having much effect on surface albedo.

Returning attention to an Earthlike waterworld, we write down the energy budget

$$(1 - \alpha(T_s))\frac{L_\circledast}{4} = OLR(T_s) \tag{3.10}$$

This determines T_s as before, with the important difference that the Solar absorption on the left hand side is now a function of T_s instead of being a constant. Analogously to Fig. 3.5, the equilibrium surface temperature can be found by plotting the absorbed Solar radiation and the OLR vs. T_s on the same graph. This is done in Fig. 3.8, for four different choices of $L_\circledast$. In this plot, we have taken $OLR = \sigma T^4$, which assumes no greenhouse effect [4]. In contrast with the fixed-albedo case, the ice-albedo feedback allows the climate system to have *multiple equilibria*: there can be more than one climate compatible with a given Solar constant, and additional information is required to determine which state the planet actually settles into. The nature of the equilibria depends on $L_\circledast$. When $L_\circledast$ is sufficiently small (as in the case $L_\circledast = 1516W/m^2$ in Fig. 3.8) there is only one solution, which is a very cold globally ice-covered Snowball state, marked Sn_1 on the graph. Note that the Solar constant that produces a unique Snowball state exceeds the present Solar constant at Earth's orbit. Thus, were it not for the greenhouse effect, Earth would be in such a state, and would have been for its entire history. When *L_odot* is sufficiently large (as in the case $L_\circledast = 2865W/m^2$ in Fig. 3.8) there is again a unique solution, which is a very hot globally ice-free state, marked H on the graph. However, for a wide range of intermediate $L_\circledast$, there are three solutions: a Snowball state (Sn_2), a partially ice covered state with a relatively large ice sheet (e.g. A), and a warmer state (e.g. B) which may have a small ice sheet or be ice free, depending on the precise value of $L_\circledast$. In the intermediate range of Solar constant, the warmest state is suggestive of the present or Pleistocene climate when there is a small ice-cap, and suggestive of Cretaceous-type hothouse climates when it is ice-free. In either case, the frigid Snowball state is available as an alternate possibility.

As the parameter $L_\circledast$ is increased smoothly from low values, the temperature of the the Snowball state increases smoothly but at some point an additional solution discontinuously comes into being at a temperature far from the previous equilibrium, and splits into a pair as $L_\circledast$ is further increased. As $L_\circledast$ is increased further, at some point, the intermediate temperature state merges with the snowball state, and disappears. This sort of behavior, in which the behavior of a system changes discontinuously as some control parameter is continuously varied, is an example of a *bifurcation*.

Finding the equilibria tells only part of the story. A system placed exactly at an equilbrium point will stay there forever, but what if it is made a little warmer than the equilibrium? Will it heat up yet more, perhaps aided by melting of ice, and ultimately wander far from the equilibrium? Or will it cool down and move back toward the equilibrium? Similar questions apply if the state is made initially slightly cooler than an equilibrium. This leads us to the question of *stability*. In order to address stability, we must first write down an equation describing the time evolution of the system. To this end, we suppose that the mean energy storage per unit area of the planet's surface can be written as a function of the mean temperature; let's call this function $E(T_s)$. Changes in the energy storage could represent the energy required to heat up or cool down a layer of water of

[4] Of course, this is an unrealistic assumption, since a waterworld would inevitably have at least water vapor – a good greenhouse gas – in its atmosphere

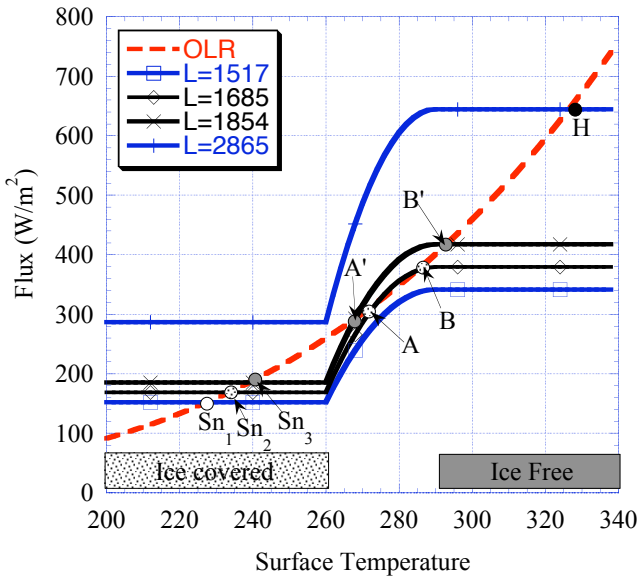

Figure 3.8: Graphical determination of the possible equilibrium states of a planet whose albedo depends on temperature in accordance with Eq. 3.9. The OLR is computed assuming the atmosphere has no greenhouse effect, and the albedo parameters are $\alpha_o = .1, \alpha_i = .6, T_i = 260K$ and $T_o = 290K$. The Solar constant for the various solar absorption curves is indicated in the legend.

some characteristic depth, and could also include the energy needed to melt ice, or released by the freezing of sea water. For our purposes, all we need to know is that E is a monotonically increasing function of T_s. The energy balance for a time-varying system can then be written

$$\frac{dE(T_s)}{dt} = \frac{dE}{dT_s}\frac{dT_s}{dt} = G(T_s) \tag{3.11}$$

where $G = \frac{1}{4}(1-\alpha(T_s))L_\circledast - OLR(T_s)$. We can define the generalized heat capacity $\mu(T) = dE/dT$, which is positive by assumption. Thus,

$$\frac{dT_s}{dt} = \frac{G(T_s)}{\mu(T_s)} \tag{3.12}$$

By definition, $G = 0$ at an equilibrium point T_{eq}. Suppose that the slope of G is well-defined near T_{eq} – in formal mathematical language, we say that G is continuously differentiable at T_{eq}, meaning that the derivative of G exists and is a continous function for T_s in some neighborhood of T_{eq}. Then, if $dG/dT_s < 0$ at T_s, it will also be negative for some finite distance to the right and left of T_s. This is the case for points a and c in the net flux curve sketched in Fig. 3.9. If the temperature is made a little warmer than T_{eq} in this case, $G(T_s)$ and hence $\frac{dT_s}{dt}$ will become negative and the solution will move back toward the equilibrium. If the temperature is made a little colder than T_{eq}, $G(T_s)$ and hence $\frac{dT_s}{dt}$ will become positive, and the solution will again move back toward the equilibrium. In contrast, if $dG/dT_s > 0$ near the equilibrium, as for point c in the sketch, a temperature placed near the equilibrium moves away from it, rather than towards it. Such equilibria are *unstable*. If the slope happens to be exactly zero at an equilibrium, one must look to higher derivatives to determine stability. These are "rare" cases, which will be encountered only for very special settings of the parameters. If the d^2G/dT^2 is non zero at the equilibrium, the curve takes the form of a parabola tangent to the axis at the equilibrium. If the parabola opens upwards, then the equilibrium is stable to displacements to the left of the equilibrium, but unstable to displacements to the right. If the parabola opens downwards, the equilibrium is unstable to displacements to the left but stable to displacements to the right. Similar reasoning applies to the case in which the first non-vanishing derivative is higher order, but such cases are hardly ever encountered.

Exercise 3.4.1 Draw a sketch illustrating the behavior near marginal equilibria with $d^2G/dT^2 > 0$ and $d^2G/dT^2 < 0$. Do the same for equilibria with $d^2G/dT^2 = 0$, having $d^3G/dT^3 > 0$ and $d^3G/dT^3 < 0$

It is rare that one can completely characterize the behavior of a nonlinear system, but one dimensional problems of the sort we are dealing with are exceptional. In the situation depicted in Fig. 3.9, G is positive and dT/dt is positive throughout the interval between b and c. Hence, a temperature placed anywhere in this interval will eventually approach the solution c arbitrarily closely – it will be *attracted* to that stable solution. Similarly, if T is initially between a and b, the solution will be attracted to the stable equilibrium a. The unstable equilibrium b forms the boundary between the *basins of attraction* of a and c. No matter where we start the system within the interval between a and c (and somewhat beyond, depending on the shape of the curve further out), it will wind up approaching one of the two stable equilibrium states. In mathematical terms, we are able to characterize the *global behavior* of this system, as opposed to just the *local behavior* near equilibria.

At an equilibrium point, the curve of solar absorption crosses the OLR curve, and the stability criterion is equivalent to stating that the equilibrium is stable if the slope of the solar

curve is less than that of the OLR curve where the two curves intersect. Using this criterion, we see that the intermediate-temperature large ice-sheet states, labeled A and A' in Fig. 3.8, are unstable. If the temperature is made a little bit warmer then the equilibrium the climate will continue to warm until it settles into the warm state (B or B') which has a small or nonexistent ice sheet. If the temperature is made a little bit colder than the equilibrium, the system will collapse into the snowball state (Sn_2 or Sn_3). The unstable state thus defines the boundary separating the basin of attraction of the warm state from that of the snowball state.

Moreover, if the net flux $G(T)$ is continous and has a continuous derivative (i.e. if the curve has no "kinks" in it), then the sequence of consecutive equilibria always alternates between stable and unstable states. For the purpose of this theorem, the rare marginal states with $dG/dT = 0$ should be considered "wildcards" that can substitute for either a stable or unstable state. The basic geometrical idea leading to this property is more or less evident from Figure 3.9, but a more formalized argument runs as follows: Let T_a and T_b be equilibria, so that $G(T_a) = G(T_b) = 0$. Suppose that the first of these is stable, so $dG/dT < 0$ at T_a, and also that the two solutions are consecutive, so that $G(T)$ does not vanish for any T between T_a and T_b. Now if $dG/dT < 0$ at T_b, then it follows that $G > 0$ just to the left of T_b. The slope near T_a similarly implies that $G < 0$ just to the right of T_a. Since G is continuous, it would follow that $G(T) = 0$ somewhere between T_a and T_b. This would contradict our assumption that the two solutions are consecutive. In consequence, $dG/dT \geq 0$ at T_b. Thus, the state T_b is either stable or marginally stable, which proves our result. The proof goes through similarly if T_a is unstable. Note that we didn't actually need to make use of the condition that dG/dT be continuous everywhere: it's enough that it be continuous near the equilibria, so we can actually tolerate a few kinks in the curve.

A consequence of this result is that, if the shape of $G(T)$ is controlled continously by some parameter like $L_\circledast$, then new solutions are born in the form of a single marginal state which, upon further change of $L_\circledast$ splits into a stable/unstable or unstable/stable pair. The first member of the pair will be unstable if there is a pre-existing stable solution immediately on the cold side of the new one, as is the case for the Snowball states Sn in Fig. 3.8. The first member will be stable if there is a pre-existing unstable state on cold side, or a pre-existing stable state on the warm side (e.g. the state H in Fig. 3.8). What we have just encountered is a very small taste of the very large and powerful subject of *bifurcation theory*.

3.4.1 Faint Young Sun, Snowball Earth and Hysteresis

We we now have enough basic theoretical equipment to take a first quantitative look at the Faint Young Sun problem. To allow for the greenhouse effect of the Earth's atmosphere, we take $p_{rad} = 670mb$, which gives the correct surface temperature with the observed current albedo $\alpha = .3$. How much colder does the Earth get if we ratchet the Solar constant down to $960W/m^2$, as it was 4.7 billion years ago when the Earth was new? As a first estimate, we can compute the new temperature from Eq. 3.8 holding p_{rad} and the albedo fixed at their present values. This yields $261K$. This is substantially colder than the present Earth. The fixed albedo assumption is unrealistic,however, since the albedo would increase for a colder and more ice-covered Earth, leading to a substantially colder temperature than we have estimated. In addition, the strength of the atmospheric greenhouse effect could have been different for the Early Earth, owing to changes in the composition of the atmosphere.

An attempt at incorporating the ice-albedo feedback can be made by using the energy balance Eq. 3.10 with the albedo parameterization given by Eq. 3.9. For this calculation, we choose constants in the albedo formula that give a somewhat more realistic Earthlike climate than

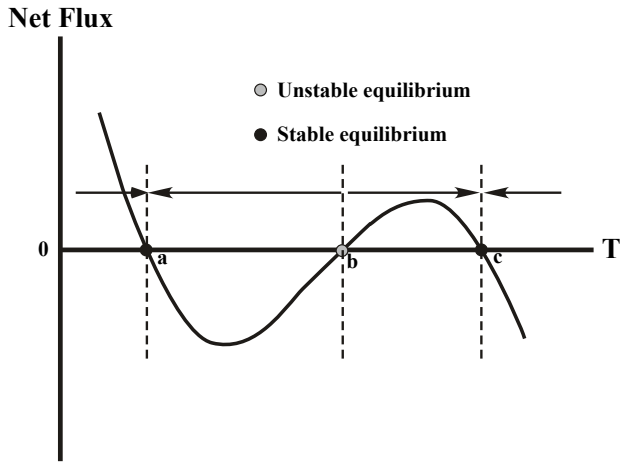

Figure 3.9: Sketch illustrating stable vs. unstable equilibrium temperatures.

those used in Figure 3.8. Specifically, we set $\alpha_o = .28$ to allow for the albedo of clouds and land, and $T_o = 295$ to allow a slightly bigger polar ice sheet. The position of the equilibria can be determined by drawing a graph like Fig. 3.8, or by applying a root-finding algorithm like Newton's method to Eq. 3.10. The resulting equilibria are shown as a function of $L_\circledast$ in Figure 3.10, with p_{rad} held fixed at $670mb$. Some techniques for generating diagrams of this type are developed in Problem **??**. For the modern Solar constant, and $p_{rad} = 670mb$, the system has a stable equilibrium at $T_s = 286K$, close to the observed modern surface temperature, and is partially ice covered. However, the system has a second stable equilibrium, which is a globally ice-covered Snowball state having $T_s = 249K$. Even today, the Earth would stay in a Snowball state if it were somehow put there. The two stable equilibria are separated by an unstable equilibrium at $T_s = 270K$, which defines the boundary between the set of initial conditions that go to the "modern" type state, and the set that go to a Snowball state. The attractor boundary for the modern open-ocean state is comfortably far from the present temperature, so it would not be easy to succumb to a Snowball.

Now we turn down the Solar constant, and re-do the calculation. For $L_\circledast = 960W/m^2$, there is only a single equilibrium point if we keep $p_{rad} = 670mb$. This is a stable Snowball state with $T_s = 228K$. Thus, if the Early Earth had the same atmospheric composition as today,leading to a greenhouse effect no stronger than the present one, the Earth would have inevitably been in a Snowball state. The open ocean state only comes into being when $L_\circledast$ is increased to $1330W/m^2$, which was not attained until the relatively recent past. This contradicts the abundant geological evidence for prevalent open water throughout several billion years of Earth's history. Even worse, if the Earth were initially in a stable snowball state four billion years ago, it would stay in that state until $L_\circledast$ increases to $1640W/m^2$, at which point the stable snowball state would disappear and the Earth would deglaciate. Since this far exceeds the present Solar constant, the Earth would be globally glaciated today. This even more obviously contradicts the data.

The currently favored resolution to the paradox of the Faint Young Sun is the supposition that the atmospheric composition of the early Earth must have resulted in a stronger greenhouse effect than the modern atmosphere produces. The prime candidate gases for mediating this change are CO_2 and CH_4. The radiative basis of the idea will be elaborated further in Chapter 4, and some ideas about why the atmosphere might have adjusted over time so as to maintain an equable

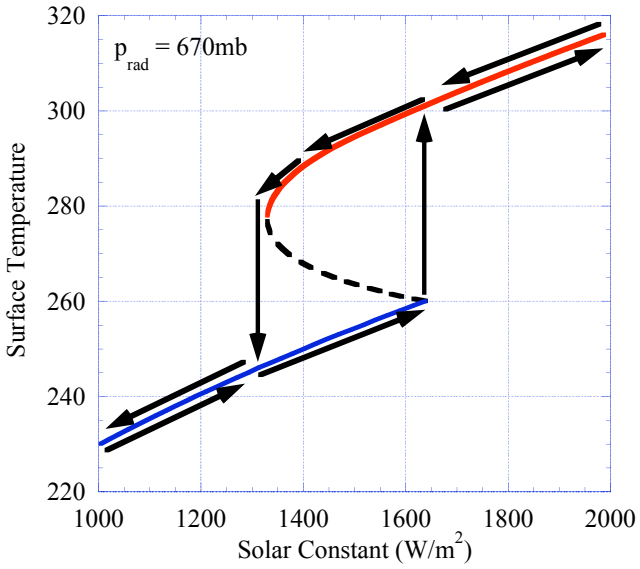

Figure 3.10: Hysteresis diagram obtained by varying $L_\circledast$ with p_{rad}/p_s fixed at .67. Arrows indicate path followed by the system as $L_\circledast$ is first increased, then decreased. The unstable solution branch is indicated by a dashed curve.

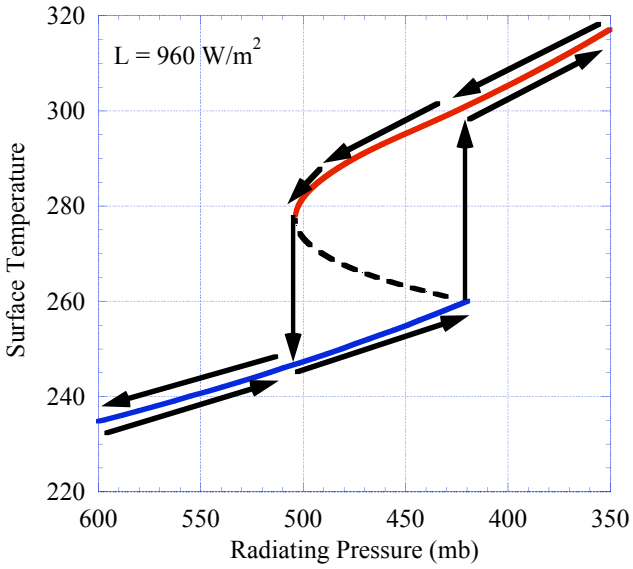

Figure 3.11: As in Fig. 3.10, but varying p_{rad} with $L_\circledast = 960 W/m^2$.

climate despite the brightening Sun are introduced in Chapter 8. Fig. 3.11 shows how the equilibria depend on p_{rad}, with $L_\circledast$ fixed at $960 W/m^2$. Whichever greenhouse gas is the Earth's savior, if it is present in sufficient quantities to reduce p_{rad} to $500mb$ or less, then a warm state with an open ocean exists (the upper branch in Fig. 3.11). However, for $420mb < prad < 500mb$ a stable snowball state also exists, meaning that the climate that is actually selected depends on earlier history. If the planet had already fallen into a Snowball state for some reason, the early Earth would stay in a Snowball unless the greenhouse gases build up sufficiently to reduce p_{rad} below $420mb$ at some point.

Figures 3.10 and 3.11 illustrate an important phenomenon known as *hysteresis*: the state in which a system finds itself depends not just on the value of some parameter of the system, but the history of variation of that parameter. This is possible only for systems that have multiple stable states. For example, in 3.10 suppose we start with $L_\circledast = 1000 W/m^2$, where the system is inevitably in a Snowball state with $T = 230K$. Let's now gradually increase $L_\circledast$. When $L_\circledast$ reaches $1500 W/m^2$ the system is still in a Snowball state, having $T = 254K$, since we have been following a stable solution branch the whole way. However, when $L_\circledast$ reaches $1640 W/m^2$, the Snowball solution disappears, and the system makes a sudden transition from a Snowball state with $T = 260K$ to the only available stable solution, which is an ice-free state having $T = 301K$. As $L_\circledast$ increases further to $2000 W/m^2$, we follow the warm, ice-free state and the temperature rises to $316K$. Now suppose we begin to gradually dim the Sun, perhaps by making the Solar system pass through a galactic dust cloud. Now, we follow the upper, stable branch as $L_\circledast$ decreases, so that when we find ourselves once more at $L_\circledast = 1500 W/m^2$ the temperature is $294K$ and the

system is in a warm, ice-free state rather than in the Snowball state we enjoyed the last time we were there. As $L_\circledast$ is decreased further, the warm branch disappears at $L_\circledast = 1330W/m^2$ and the system drops suddenly from a temperature of $277K$ into a Snowball state with a temperature of $246K$, whereafter the Snowball branch is again followed as $L_\circledast$ is reduced further. The trajectory of the system as $L_\circledast$ is increased then decreased back to its original value takes the form of an open loop, depicted in Fig. 3.10.

The thought experiment of varying $L_\circledast$ in a hysteresis loop is rather fanciful, but many atmospheric processes could act to either increase or decrease the greenhouse effect over time. For the very young Earth, with $L_\circledast = 960W/m^2$, the planet falls into a Snowball when p_{rad} exceeds $500mb$, and thereafter would not deglaciate until p_{rad} is reduced to $420mb$ or less (see Fig. 3.11). The boundaries of the hysteresis loop, which are the critical thresholds for entering and leaving the Snowball, depend on the solar constant. For the modern solar constant, the hysteresis loop operates between $p_{rad} = 690mb$ and $p_{rad} = 570mb$. It takes less greenhouse effect to keep out of the Snowball now than it did when the Sun was fainter, but the threshold for initiating a Snowball in modern conditions is disconcertingly close to the value of p_{rad} which reproduces the present climate.

The fact that the freeze-thaw cycle can exhibit hysteresis as atmospheric composition changes is at the heart of the Snowball Earth phenomenon. An initially warm state can fall into a globally glaciated Snowball if the atmospheric composition changes in such a way as to sufficiently weaken the greenhouse effect. Once the threshold is reached, the planet can fall into a Snowball relatively quickly – in a matter of a thousand years or less – since sea ice can form quickly. However, to deglaciate the Snowball, the greenhouse effect must be increased far beyond the threshold value at which the planet originally entered the Snowball state. Atmospheric composition must change drastically in order to achieve such a great increase, and this typically takes many millions of years. When deglaciation finally occurs, it leaves the atmosphere in a hyper-warm state, which only gradually returns to normal as the atmospheric composition evolves in such a way as to reduce the greenhouse effect. As discussed in Chapter 1, there are two periods in Earth's past when geological evidence suggests that one or more Snowball freeze-thaw cycles may have occurred. The first is in the Paleoproterozoic, around 2 billion years ago. At this time, $L_\circledast \approx 1170W/m^2$, and the thresholds for initiating and deglaciating a Snowball are $p_{rad} = 600mb$ and $p_{rad} = 500mb$ in our simple model. For the Neoproterozoic, about 700 million years ago, $L_\circledast \approx 1290W/m^2$ and the thresholds are at $p_{rad} = 650mb$ and $p_{rad} = 540mb$.

The boundaries of the hysteresis loop shift as the Solar constant increases, but there is nothing obvious in the numbers to suggest why a Snowball state should have occurred in the Paleoproterozoic and Neoproterozoic but not at other times. Hysteresis associated with ice-albedo feedback has been a feature of the Earth's climate system throughout the entire history of the planet. Hysteresis will remain a possibility until the Solar constant increases sufficiently to render the Snowball state impossible even in the absence of any greenhouse effect (i.e. with $p_{rad} = 1000mb$). Could a Snowball episode happen again in the future, or is that peril safely behind us? These issues require an understanding of the processes governing the evolution of Earth's atmosphere, a subject that will be taken up in Chapter 8.

Exercise 3.4.2 Assuming an ice albedo of .6, how high does $L_\circledast$ have to become to eliminate the possibility of a snowball state? Will this happen within the next five billion years? What if you assume there is enough greenhouse gas in the atmosphere to make $p_{rad}/p_s = .5$?

Note: The evolution of the Solar constant over time is approximately $L_\circledast(t) = L_{\circledast p} \cdot (.7 + (t/22.975) + (t/14.563)^2)$, where t is the age of the Sun in billions of years ($t = 4.6$ being the current age) and $L_{\circledast p}$ is the present Solar constant. This fit is reasonably good for the first 10 billion years

of Solar evolution.

The "cold start" problem is a habitability crisis that applies to waterworlds in general. If a planet falls into a Snowball state early in its history, it could take billions of years to get out if one needs to wait for the Sun to brighten. The time to get out of a Snowball could be shortened if greenhouse gases build up in the atmosphere, reducing p_{rad}. How much greenhouse gas must build up to deglaciate a snowball? How long would that take? What could cause greenhouse gases to accumulate on a Snowball planet? These important questions will be taken up in subsequent chapters.

Another general lesson to be drawn from the preceding discussion is that the state with a stable, small icecap is very fragile, in the sense that the planetary conditions must be tuned rather precisely for the state to exist at all. For example, with the present Solar constant, the stable small icecap solution first appears when p_{rad} falls below $690mb$. However, the icecap shrinks to zero as p_{rad} is reduced somewhat more, to $615mb$. Hence, a moderate strengthening in the greenhouse effect would, according to the simple energy balance model, eliminate the polar ice entirely and throw the Earth into an ice-free Cretaceous hothouse state. The transition to an ice-free state of this sort is continous in the parameter being varied; unlike the collapse into a snowball state or the recovery from a snowball, it does not result from a bifurcation. In light of its fragility, it is a little surprising that the Earth's present small-icecap state has persisted for the past two million years, and that similar states have occurred at several other times in the past half billion years. Does the simple energy-balance model exaggerate the fragility of the stable small-icecap state? Does some additional feedback process adjust the greenhouse effect so as to favor such a state while resisting the peril of the Snowball? These are largely unresolved questions. Attacks on the first question require comprehensive dynamical models of the general circulation, which we will not encounter in the present volume. We will take up, though not resolve, the second question in Chapter 8. It is worth noting that small-icecap states like those of the past two million years appear to be relatively uncommon in the most recent half billion years of Earth's history, for which data is good enough to render a judgement about ice cover. The typical state appears to be more like the warm relatively ice-free states of the Cretaceous, and perhaps this reflects the fragility of the small-icecap state.

The simple models used above are too crude to produce very precise hysteresis boundaries. Among the many important effects left out of the story are water vapor radiative feedbacks, cloud feedbacks, the factors governing albedo of sea ice, ocean heat transports and variations in atmospheric heat transport. The phenomena uncovered in this exposition are general, however and can be revisited across a heirarchy of models. Indeed, the re-examination of this subject provides an unending source of amusement and enlightenment to climate scientists.

3.4.2 Climate sensitivity, radiative forcing and feedback

The simple model we have been studying affords us the opportunity to introduce the concepts of *radiative forcing, sensitivity coefficient* and *feedback factor*. These diagnostics can be applied across the whole spectrum of climate models, from the simplest to the most comprehensive.

Suppose that the mean surface temperature depends on some parameter Λ, and we wish to know how sensitive T is to changes in that parameter. For example, this parameter might be the Solar constant, or the radiating pressure. It could be some other parameter controlling the strength of the greenhouse effect, such as CO_2 concentration. Near a given Λ, the sensitivity is characterized by $dT/d\Lambda$.

Let G be the net top-of-atmosphere flux, such as used in Eq. 3.11. To allow for the fact that the terms making up the net flux depend on the parameter Λ, we write $G = G(T, \Lambda)$. If we take the derivative of the the energy balance requirement $G = 0$ with respect to Λ, we find

$$0 = \frac{\partial G}{\partial T}\frac{dT}{d\Lambda} + \frac{\partial G}{\partial \Lambda} \tag{3.13}$$

so that

$$\frac{dT}{d\Lambda} = -\frac{\frac{\partial G}{\partial \Lambda}}{\frac{\partial G}{\partial T}} \tag{3.14}$$

The numerator in this expression is a measure of the *radiative forcing* associated with changes in Λ. Specifically, changing Λ by an amount $\delta\Lambda$ will perturb the top-of-atmosphere radiative budget by $\frac{\partial G}{\partial \Lambda}\delta\Lambda$, requiring that the temperature change so as to bring the energy budget back into balance. For example, if Λ is the Solar constant L, then $\frac{\partial G}{\partial \Lambda} = \frac{1}{4}(1-\alpha)$. If Λ is the radiating pressure p_{rad}, then $\frac{\partial G}{\partial \Lambda} = -\frac{\partial OLR}{\partial p_{rad}}$. Since OLR goes down as p_{rad} is reduced, a reduction in p_{rad} yields a positive radiative forcing. This is a warming influence.

Radiative forcing is often quoted in terms of the change in flux caused by a standard change in the parameter, in place of the slope $\frac{\partial G}{\partial \Lambda}$ itself. For example, the radiative forcing due to CO_2 is typically described by the change in flux caused by doubling CO_2 from its pre-industrial value, with temperature and everything else is held fixed. This is practically the same thing as $\frac{\partial G}{\partial \Lambda}$ if we take $\Lambda = \log_2 pCO_2$, where pCO_2 is the partial pressure of CO_2. Similarly, the climate sensitivity is often described in terms of the temperature change caused by the standard forcing change, rather than the slope $\frac{dT}{d\Lambda}$. For example, the notation ΔT_{2x} would refer to the amount by which temperature changes when CO_2 is doubled.

The denominator of Eq. 3.14 determines how much the equilibrium temperature changes in response to a given radiative forcing. For any given magnitude of the forcing, the response will be greater if the denominator is smaller. Thus, the denominator measures the *climate sensitivity*. An analysis of ice-albedo feedback illustrates how a feedback process affects the climate sensitivity. If we assume that albedo is a function of temperature, as in Eq. 3.9, then

$$\frac{\partial G}{\partial T} = -\frac{1}{4}L\frac{\partial \alpha}{\partial T} - \frac{\partial OLR}{\partial T} \tag{3.15}$$

With this expression, Eq. 3.14 can be rewritten

$$\frac{dT}{d\Lambda} = -\frac{1}{1+\Phi}\left[\frac{\frac{\partial G}{\partial \Lambda}}{\frac{\partial OLR}{\partial T}}\right] \tag{3.16}$$

where

$$\Phi = \frac{1}{4}L\frac{\frac{\partial \alpha}{\partial T}}{\frac{\partial OLR}{\partial T}} \tag{3.17}$$

In writing this equation we primarily have ice-albedo feedback in mind, but the equation is valid for arbitrary $\alpha(T)$ so it could as well describe a variety of other processes. The factor in square brackets in Eqn. 3.16 is the sensitivity the system would have if the response were unmodified by the change of albedo with temperature. The first factor determines how the sensitivity is increased or decreased by the feedback of temperature on albedo. If $-1 < \Phi < 0$ then the feedback increases the sensitivity – the same radiative forcing produces a bigger temperature change than it would in the absence of the feedback. When $\Phi = -\frac{1}{2}$, for example, the response to the forcing is twice what it would have been in the absence of the feedback. The sensitivity becomes infinite as $\Phi \to -1$,

and for $-2 < \Phi < -1$ the feedback is so strong that it actually reverses the sign of the response as well as increasing its magnitude. On the other hand, if $\Phi > 0$, the feedback reduces the sensitivity. In this case it is a *stabilizing feedback*. The larger Φ gets, the more the response is reduced. For example, when $\Phi = 1$ the response is half what it would have been in the absence of feedback. Note that the feedback term is the same regardless of whether the radiative forcing is due to changing L, p_{rad} or anything else.

As an example, let's compute the feedback parameter Φ for the albedo-temperature relation given by Eq. 3.9, under the conditions shown in Fig. 3.10. Consider in particular the upper solution branch, which represents a stable partially ice-covered climate like that of the present Earth. At the point $L = 1400W/m^2, T = 288K$ on this branch, we find $\Phi = -.333$. Thus, at this point the ice-albedo feedback increases the sensitivity of the climate by a factor of about 1.5. At the bifurcation point $L \approx 1330W/m^2, T \approx 277K$, $\Phi \to -1$ and the sensitivity becomes infinite. This divergence merely reflects the fact that the temperature curve is vertical at the bifurcation point. Near such points, the temperature change is no longer linear in radiative forcing. It can easily be shown that the temperature varies as the square root of radiative forcing near a bifurcation point, as suggested by the plot.

The ice-albedo feedback increases the climate sensitivity, but other feedbacks could be stabilizing. In fact Eq. 3.17 is valid whatever the form of $\alpha(T)$, and shows that the albedo feedback becomes a stabilizing influence if albedo increases with temperature. This could conceivably happen as a result of vegetation feedback, or perhaps dissipation of low clouds. The somewhat fanciful Daisyworld example in the Workbook section at the end of this chapter provides an example of such a stabilizing feedback.

The definition of the feedback parameter can be generalized as follows. Suppose that the energy balance function G depends not only on the control parameter Λ, but also on some other parameter R which varies systematically with temperature. In the previous example, $R(T)$ is the temperature-dependent albedo. We write $G = G(T, R(T), \Lambda)$. Following the same line of reasoning as we did for the analysis of ice-albedo feedback, we find

$$\Phi = \frac{\partial G}{\partial R}\frac{\partial R}{\partial T}\frac{\partial G}{\partial T} \tag{3.18}$$

For example, if R represents the concentration of water vapor on Earth, or of methane on Titan, and if R varies as a function of temperature, then the feedback would influence G through the OLR. Writing $OLR = OLR(T, R(T), \Lambda)$, then the feedback parameter is

$$\Phi = \frac{\frac{\partial OLR}{\partial R}\frac{\partial R}{\partial T}}{\frac{\partial OLR}{\partial T}} \tag{3.19}$$

assuming the albedo to be independent of temperature in this case. Now, since OLR increases with T and OLR decreases with R, the feedback will be destabilizing ($\Phi < 0$) if R increases with T. (One might expect R to increase with T because Clausius-Clapeyron implies that the saturation vapor pressure increases sharply with T, making it harder to remove water vapor by condensation, all other things being equal). Note that in this case the water vapor feedback does not lead to a runaway, with more water leading to higher temperatures leading to more water in a never-ending cycle; the system still attains an equilibrium, though the sensitivity of the equilibrium temperature to changes in a control parameter is increased.

3.5 Partially absorbing atmospheres

The assumption underpinning the blackbody radiation formula is that radiation interacts so strongly with matter that it achieves thermodynamic equilibrium at the same temperature as the matter. It stands to reason, then, that if a box of gas contains too few molecules to offer much opportunity to intercept a photon, the emission will deviate from the blackbody law. Weak interaction with radiation can also arise from aspects of the structure of a material which inhibit interaction, such as the crystal structure of table salt or carbon dioxide ice. In either event, the deviation of emission from the Planck distribution is characterized by the *emissivity*. Suppose that $I(\nu, \hat{n})$ is the observed flux of radiation at frequency ν emerging from a body in the direction $\hat{n}$. Then the emissivity $e(\nu, \hat{n})$ is defined by the expression

$$I(\nu, \hat{n}) = e(\nu, \hat{n})B(\nu, T) \tag{3.20}$$

where T is the temperature of the collection of matter we are observing. Note that in assigning a temperature T to the body, we are assuming that the matter itself is in a state of thermodynamic equilibrium. The emissivity may also be a function of temperature and pressure. We can also define a mean emissivity over frequencies, and all rays emerging from a body. The mean emissivity is

$$\bar{e} = \frac{\int_{\nu,\Omega} e(\nu, \hat{n})B(\nu, T)\cos\theta d\nu d\Omega}{\sigma T^4} \tag{3.21}$$

where θ is the angle of the ray to the normal to the body's surface and the angular integration is taken over the hemisphere of rays leaving the surface of the body. With this definition, the net flux emerging from any patch of the body's surface is $F = \bar{e}\sigma T^4$. Even if e does not depend explicitly on temperature, $\bar{e}$ will be temperature dependent if e is frequency dependent, since the relative weighting of different frequencies, determined by $B(\nu, T)$ changes with temperature.

A blackbody has unit emissivity at all frequencies and directions. A blackbody also has unit absorptivity, which is just a restatement of the condition that blackbodies interact strongly with the radiation field. For a non-black body, we can define the absorptivity $a(\nu, \hat{n})$ by shining light at a given frequency and direction at the body and measuring how much is reflected and how much comes out the other side. Specifically, suppose that we shine a beam of electromagnetic energy with direction $\hat{n}$, frequency ν and flux F_{inc} at the test object. Then we measure the *additional* energy flux coming out of the object once this beam is turned on. This outgoing flux may come out in many different directions, because of scattering of the incident beam; in exotic cases, even the frequency could differ from the incident radiation. Let T and R be the transmitted and reflected energy flux, integrated over all angles and frequencies. Then, the absorptivity is defined by taking the ratio of the flux of energy left behind in the body to the incident flux. Thus,

$$a(\nu, \hat{n}) = \frac{F_{inc} - (T + R)}{F_{inc}} \tag{3.22}$$

The Planck function is unambigously the natural choice of a weighting function for defining the mean emissivity $\bar{e}$ for an object with temperature T. There is no such unique choice for defining the mean absorptivity over all frequencies and directions. The appropriate weighting function is determined by the frequency and directional spectrum of the incident radiation which requires a detailed knowledge of its source. If the incident radiation is a blackbody with temperature T_{source} then $\bar{a}$ should be defined with a formula like Eq. 3.20, using $B(\nu, T_{source})$ as the weighting function. Note that the weighting function is defined by the temperature of the *source* rather than by the temperature of the the object doing the absorbing. As was the case for mean emissivity, the temperature dependence of the weighting function implies that $\bar{a}$ will vary with T_{source} even if $a = a(\nu)$ and is not explicitly dependent on temperature.

Absorptivity and emissivity might appear to be independent characteristics of an object, but observations and theoretical arguments reveal an intimate relation between the two. This relation, expressed by *Kirchhoff's Law of Radiation* is a profound property of the interaction of radiation with matter that lies at the heart of all radiative transfer theory. Kirchoff's Law states that the emissivity of a substance at any given frequency equals the absorptivity measured at the same frequency. It was first inferred experimentally. The hard-working spectroscopists of the late nineteenth centuries employed their new techniques to measure the emission spectrum $I(\nu, \hat{n}, T)$ and absorptivity $a(\nu, \hat{n}, T)$ of a wide variety of objects at various temperatures. Kirchhoff found that, with the exception of a few phosphorescent materials whose emission was not linked to temperature, all the experimental data collapsed onto a single universal curve, independent of the material, once the observed emission was normalized by the observed absorptivity. In other words, virtually all materials fit the relation $I(\nu, \hat{n}, T)/a(\nu, \hat{n}, T) = f(\nu, T)$ with the same function f. If we take the limit of a perfect absorber – a perfectly "black" body – then $a = 1$ and we find that f is in fact what we have been calling the *Planck* function $B(\nu, T)$. In fact, it was this extrapolation to a perfect absorber that originally led to the formulation of the notion of blackbody radiation. Since $f = B$ and $I = eB$, we recover the statement of Kirchoff's law in the form $e/a = 1$.

The thought experiment sketched in Fig. 3.12 allows us to deduce Kirchhoff's law for the mean absorptivity and emissivity from the requirements of the Second Law of Thermodynamics. We consider two infinite slabs of a blackbody material with temperature T_o, separated by a gap. Into the gap, we introduce a slab of partially transparent material with mean absorptivity $\bar{a}(T_1)$ and mean emissivity $\bar{e}(T_1)$, where T_1 is the temperature of the test material. Note that this system is energetically closed. We next require that the radiative transfer between the blackbody material and the test object cause the system to evolve toward an isothermal state. In other words we are *postulating* that radiative heat transfers satisfy the Second Law. A *necessary* condition for radiative transfer to force the system to evolve towards an isothermal state is that the isothermal state $T_o = T_1$ be an equilibrium state of the system; if it weren't an initially isothermal state would spontaneously generate temperature inhomogeneities. Energy balance requires that $2\bar{a}(T_o)\sigma T_o^4 = 2\bar{e}(T_1)\sigma T_1^4$. Kirchhoff's law then follows immediately by setting $T_o = T_1$ in the energy balance, which then implies $\bar{a}(T_o) = \bar{e}(T_o)$. Note that the mean absorptivity in this statement is defined using the Planck function at the common temperature of the two materials as the weighting function.

A modification of the preceding argument allows us to show that in fact the emissivity and absorptivity should be equal at each individual frequency, and not just in the mean. To simplify the argument, we will assume that e and a are independent of direction. The thought experiment we employ is similar to that used to justify Kirchhoff's Law in the mean, except that this time we interpose frequency-selective mirrors between the test object and the blackbody material, as shown in Fig. 3.13. The mirrors allow the test object to exchange radiant energy with the blackbody only in a narrow frequency band $\Delta\nu$ around a specified frequency ν. The energy budget for the test object now reads $2e(\nu)B(\nu, T_1)\Delta\nu = 2a(\nu)B(\nu, T_o)\Delta\nu$. Setting $T_1 = T_o$ so that the isothermal state is an equilibrium, we find that $e(\nu) = a(\nu)$.

The preceding argument, presented in the form originally given by Kirchhoff, is the justification commonly given for Kirchhoff's Law. It is ultimately unsatisfying, as it applies equilibrium thermodynamic reasoning to a system in which the radiation field is manifestly out of equilibrium with matter; in the frequency-dependent form, it invokes the existence of mirrors with hypothetical material properties; worse, it takes as its starting point that radiative heat transfer will act like other heat transfers to equalize temperature, whereas we really ought to be able to demonstrate such a property from first principles of the interaction of radiation with molecules. The great mathematician David Hilbert, was among many who recognized these difficulties; in 1912

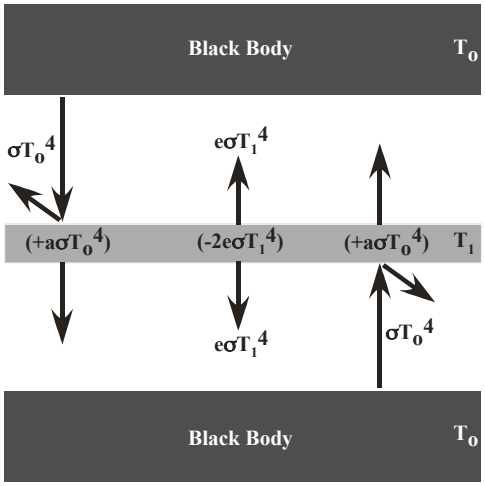

Figure 3.12: Sketch illustrating thought experiment for demonstating Kirchoff's Law in the mean over all wavenumbers. In the annotations on the sketch, $a = \bar{a}(T_o)$ and $e = \bar{e}(T_1)$.

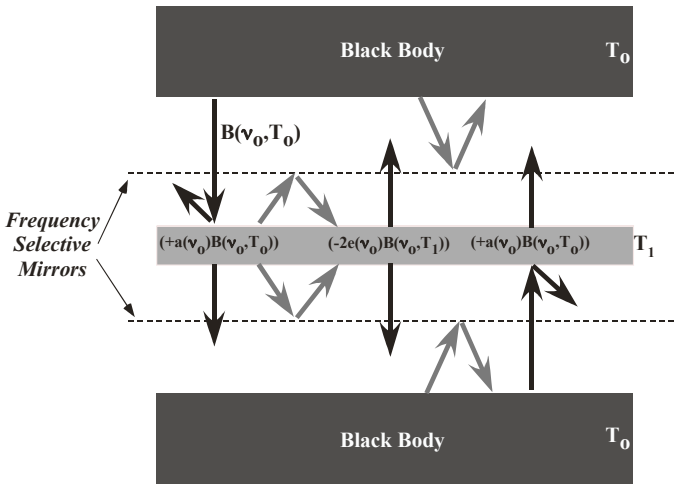

Figure 3.13: Sketch illustrating thought experiment for demonstrating Kirchoff's Law for a narrow band of radiation near frequency ν_0. The thin dashed lines represent ideal frequency-selective mirrors, which pass frequencies close to ν_0, but reflect all others without loss.

he presented a formal justification that eliminated the involvement of hypothetical ideal selective mirrors. The physical content of Hilbert's proof is that one doesn't need an ideal mirror, if one requires that a sufficient variety of materials with different absorbing and emitting properties will all come into an isothermal state at equilibrium. Hilbert's deriviation nonetheless relied on an assumption that radiation would come into equilibrium with matter at each individual wavelength considered separately. While Kirchoff did the trick with mirrors, Hilbert, in essence, did the trick with axioms instead, leaving the microscopic justification of Kirchoff's Law equally obscure. It is in fact quite difficult to provide a precise and concise statement of the circumstances in which a material will comply with Kirchoff's Law. Violations are quite commonplace in nature and in engineered materials, since it is quite possible for a material to store absorbed electromagnetic energy and emit it later, perhaps at a quite different frequency. A few examples that come to mind are phosphorescent ("glow in the dark") materials, fluorescence (e.g. paints that glow when exposed to ultraviolet, or "black" light), frequency doubling materials (used in making green laser pointers), and lasers themselves. In Nature, such phenomena involve insignificant amounts of energy, and are of no known importance in determining the energy balance of planets. We will content ourselves here with the statement that all known liquid and solid planetary materials, as well as the gases making up atmospheres, conform very well to Kirchoff's Law, except perhaps in the most tenuous outer reaches of atmospheres where the gas itself is not in thermodynamic equilibrium.

When applying Kirchhoff's law in the mean, careful attention must be paid to the weighting function used to define the mean absorptivity. For example, based on the incident Solar spectrum, the Earth has a mean albedo of about .3, and hence a mean absorptivity of .7. Does this imply that the mean emissivity of the Earth must be .7 as well? In fact, no such implication can be drawn, because Kirchhoff's Law only requires that the mean emissivity and absorptivity are the same when averaged over identical frequency weighting functions. Most of the Earth's thermal emission is in the infrared, not the visible. Kirchhoff's law indeed requires that the *visible* wavelength emissivity is .7, but the net thermal emission of the Earth in this band is tiny compared to the infrared, and contributes almost nothing to the Earth's net emission. Specifically, the Planck function implies that, at $255K$ the emission in visible wavelengths is smaller than the emission in infrared wavelengths by a factor of about 10^{-19}. Thus, if the infrared emission from some region were $100W/m^2$, the visible emission would be only $10^{-17}W/m^2$. Using $\Delta E = h\nu$ to estimate the energy of a photon of visible light, we find that this amounts to an emission of only 50 visible light photons each second, from each square meter of radiating surface. This tiny outgoing *thermal* emission of visible light should not be confused with the much larger outgoing flux of *reflected* solar radiation.

It is a corollary of Kirchhoff's law that $e \leq 1$. If the emissivity were greater than unity, then by Kirchhoff's Law, the absorptivity would also have to be greater than unity. In consequence, the amount of energy absorbed by the body per unit time would be greater than the amount delivered to it by the incident radiation. By conservation of energy, that would imply the existence of an internal energy source. However, any internal energy source would ultimately be exhausted, violating the assumption that the system is in a state of equilibrium which can be maintained indefinitely.

3.6 Optically thin atmospheres: The skin temperature

Since the density of an atmosphere always approaches zero with height, in accordance with the hydrostatic law, one can always define an outer layer of the atmosphere that has so few molecules in it that it will have low infrared emissivity. We will call this the *skin layer*. What is the

temperature of this layer? Suppose for the moment that it is transparent to solar radiation, and that atmospheric motions do not transport any heat into the layer; thus, it is heated only by infrared upwelling from below. Because the emissivity of the skin layer is assumed small, little of the upwelling infrared will be absorbed, and so the upwelling infrared is very nearly the same as the OLR. The energy balance is between absorption and emission of infrared. Since the skin layer radiates from both its top and bottom, the energy balance reads

$$2e_{ir}\sigma T_{skin}^4 = e_{ir}OLR. \tag{3.23}$$

Hence,

$$T_{skin} = \frac{1}{2^{\frac{1}{4}}}(\frac{OLR}{\sigma})^{\frac{1}{4}} = \frac{1}{2^{\frac{1}{4}}}T_{rad} \tag{3.24}$$

where T_{rad} is defined as before. Thus, the skin temperature is colder than the blackbody radiating temperature by a factor of $2^{-\frac{1}{4}}$. The skin temperature is the natural temperature the outer regions of an atmosphere would have in the absence of *in situ* heating by solar absorption or other means. Note that the skin layer does not need any interior heat transfer mechanism to keep it isothermal, since the argument we have applied to determine T_{skin} applies equally well to any sublayer of the skin layer.

A layer that has low emissivity, and hence low absorptivity, in some given wavelength band is referred to as being *optically thin* in this band. A layer could well be optically thick in the infrared, but optically thin in the visible, which is in fact the case for strong greenhouse gases.

Now let's suppose that the entire atmosphere is optically thin, right down to the ground, and compute the pure radiative equilibrium in this system in the absence of heat transfer by convection. We'll also assume that the atmosphere is completely transparent to the incident Solar radiation. Let S be the incident Solar flux per unit surface area, appropriate to the problem under consideration (e.g. $\frac{1}{4}L_\circledast$ for the global mean or $L_\circledast$ for temperature at the subsolar point on a planet like modern Mars). Since the atmosphere has low emissivity, the heating of the ground by absorption of downwelling infrared emission coming from the atmosphere can be neglected to lowest order. Since the ground is heated only by absorbed Solar radiation, its temperature is determined by $\sigma T_s^4 = (1-\alpha)S$, just as if there were no atmosphere at all. In other words, $p_{rad} = p_s$ because the atmosphere is optically thin, so that the atmosphere does not affect the surface temperature no matter what its temperature structure turns out to be. Next we determine the atmospheric temperature. The whole atmosphere has small *but nonzero* emissivity so that the skin layer in this case extends right to the ground. The atmosphere is then isothermal, and its temperature T_a is just the skin temperature $2^{-1/4}T_s$.

The surface is thus considerably warmer than the air with which it is in immediate contact. There would be nothing unstable about this situation if radiative transfer were truly the only heat transfer mechanism coupling the atmosphere to the surface. In reality, the air molecules in contact with the surface will acquire the temperature of the surface by heat conduction, and turbulent air currents will carry the warmed air away from the surface, forming a heated, buoyant layer of air. This will trigger convection, mixing a deep layer of the atmosphere within which the temperature profile will follow the adiabat. The layer will grow in depth until the temperature at the top of the mixed layer matches the skin temperature, eliminating the instability. This situation is depicted in Figure 3.14. The isothermal, stably stratified region above the mixed region is the stratosphere in this atmosphere, and the lower, adiabatic region is the troposphere; the boundary between the two is the tropopause. We have just formulated a theory of tropopause height for optically thin atmospheres. To make it quantitative, we need only require that the adiabat starting at the surface temperature match to the skin temperature at the tropopause. Let p_s be the surface

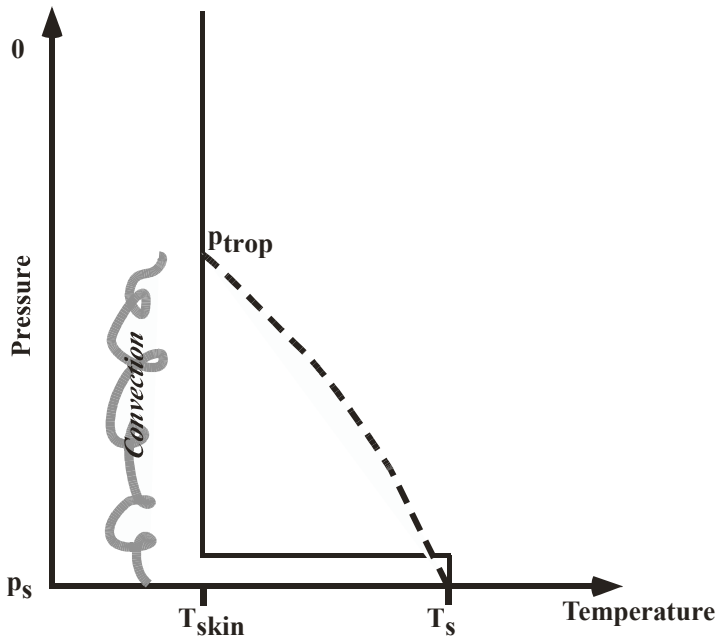

Figure 3.14: The unstable pure radiative equilibrium for an optically thin atmosphere (solid line) and the result of adjustment to the adiabat by convection (dashed line). The adjustment of the temperature profile leaves the surface temperature unchanged in this case, because the atmosphere is optically thin and has essentially no effect on the OLR.

pressure and p_{trop} be the tropopause pressure. For the dry adiabat, the requirement is then $T_s(p_{trop}/p_s)^{R/c_p} = T_{skin}$. Since $T_s = 2^{1/4}T_{skin}$, the result is

$$\frac{p_{trop}}{p_s} = 2^{-\frac{c_p}{4R}} \tag{3.25}$$

Note that the tropopause pressure is affected by R/c_p, but is independent of the insolation S.

The stratosphere in the preceding calculation differs from the observed stratosphere of Earth in that it is isothermal rather than warming with altitude. The factor we have left out is that real stratospheres often contain constituents that absorb solar radiation. To rectify this shortcoming, let's consider the effect of solar absorption on the temperature of the skin layer. Let e_{ir} be the infrared emissivity, which is still assumed small, and a_{sw} be the shortwave (mostly visible) absorptivity, which will also be assumed small. Note that Kirchhoff's Law does not require $e_{ir} = a_{sw}$, as the emissivity and absorptivity are at different wavelengths. The solar absorption of incident radiation is $a_{sw}S$. We'll assume that the portion of the solar spectrum which is absorbed by the atmosphere is absorbed so strongly that it is completely absorbed before reaching the ground. This is in fact the typical situation for solar near-infrared and ultraviolet. In this case, one need not take into account absorption of the upwelling solar radiation reflected from the surface.

Exercise 3.6.1 Show that if the atmosphere absorbs uniformly throughout the solar spectrum,

then the total absorption in the skin layer is $(1 + (1 - a_{sw})\alpha_g)a_{sw}S$, where α_g is the solar albedo of the ground. Show that the *planetary albedo* – i.e. the albedo observed at the top of the atmosphere – is $(1 - a_{sw})^2\alpha_g$.

The energy balance for the skin layer now reads

$$e_{ir}\sigma T^4 = e_{ir}OLR + a_{sw}S \tag{3.26}$$

Hence,

$$T = T_{skin}(1 + \frac{a_{sw}}{e_{ir}}\frac{S}{OLR})^{\frac{1}{4}} \tag{3.27}$$

where T_{skin} is the skin temperature in the absence of Solar absorption. The formula shows that Solar absorption always increases the temperature of the skin layer. The temperature increases as the ratio of shortwave absorption to infrared emissivity is made larger. So long as the temperature remains less than the Solar blackbody temperature, the system does not violate the Second Law of Thermodynamics, since the radiative transfer is still acting to close the gap between the cold atmospheric temperature and the hot Solar temperature. As the atmospheric temperature approaches that of the Sun, however, it would no longer be appropriate to use the infrared emissivity, since the atmosphere would then be radiating in the shortwave range. Kirchoff's Law would come into play, requiring $a/e = 1$. This would prevent the atmospheric temperature from approaching the photospheric temperature.

If the shortwave absorptivity is small, the skin layer can be divided into any number of sublayers, and the argument applies to determine the temperature of each one individually. This is so because the small absorptivity of the upper layers do not take much away from the Solar beam feeding absorption in the lower layers. We can then infer that the temperature of an absorbing stratosphere will increase with height if the absorption increases with height, making a_{sw}/e_{ir} increase with height.

Armed with our new understanding of the optically thin outer portions of planetary atmospheres, let's take another look at a few soundings. The skin temperature, defined in Eq 3.24, provides a point of reference. It is shown for selected planets in Table 3.3. Except for the Martian case, these values were computed from the global mean OLR, either observed directly (for Jupiter) or inferred from the absorbed Solar radiation. In the case of present Mars, the fast thermal response of the atmosphere and surface makes the global mean irrelevant. Hence, assuming the atmosphere to be optically thin, we compute the skin temperature based on the upwelling infrared from a typical daytime summer surface temperature corresponding to the Martian soundings of Figure 2.2. The tropical Earth atmosphere sounding shown in Fig. 2.1 shows that the temperature increases sharply with height above the tropopause. This suggests that solar absorption is important in the Earth's stratosphere. For Earth, the requisite solar absorption is provided by ozone, which strongly absorbs Solar ultraviolet. This is the famous "ozone layer," which shields life on the surface from the sterilizing effects of deadly Solar ultraviolet rays. However, it is striking and puzzling that virtually the entire stratosphere is substantially colder than the skin temperature based on the global mean radiation budget. The minimum temperature in the sounding is $188K$, which is fully $26K$ below the skin temperature. If anything, one might have expected the tropical temperatures to exceed the global mean skin temperatures, because the local tropospheric temperatures are warmer than the global mean. A reasonable conjecture about what is going on is that high, thick tropical clouds reduce the local OLR, thus reducing the skin temperature. However, the measured tropical OLR In Fig. 3.7 shows that at best clouds reduce the tropical OLR to $240W/m^2$, which yields the same $214K$ skin temperature computed from the global mean budget. Apart from possible effects of dynamical heat transports, the only way the temperature can fall below the skin temperature is

	Skin temperature
Venus	213.
Earth	214.
Mars (255K sfc)	214.
Jupiter	106.
Titan	72.

Table 3.3: Computed skin temperatures of selected planets.

if the infrared emissivity becomes greater than the infrared absorptivity. This is possible, without violating Kirchoff's law, if the spectrum of upwelling infrared is significantly different from the spectrum of infrared emitted by the skin layer. We will explore this possibility in the next chapter.

Referring to Fig. 2.2 we see that the temperature of the Martian upper atmosphere declines steadily with height, unlike Earth; this is consistent with Mars' CO_2 atmosphere, which has only relatively weak absorption in the Solar near infrared spectrum. The Martian upper atmosphere presents the same quandary as Earth's though, in that the temperatures fall well below the skin temperature estimates. Just above the top of the Venusian troposphere, there is an isothermal layer with temperature $232K$, just slightly higher than the computed skin temperature. However, at higher altitudes, the temperature falls well below the skin temperature, as for Mars.

Between $500mb$ and $100mb$, just above Titan's troposphere, Titan has an isothermal layer with temperature of $75K$, which is very close to the skin temperature. Above $100mb$, the atmosphere warms markedly with height, reaching $160K$ at $10mb$. The solar absorption in Titan's stratosphere is provided mostly by organic haze clouds. Jupiter, like Titan, has an isothermal layer just above the troposphere, whose temperature is very close to the skin temperature. Jupiter's atmosphere also shows warming with height; its upper atmosphere becomes nearly isothermal at $150K$, which is $44K$ warmer than the skin temperature. This indicates the presence of solar absorbers in Jupiter's atmosphere as well, though the solar absorption is evidently more uniformly spread over height on Jupiter than it is on Earth or Titan.

We have been using the term "stratosphere" rather loosely, without having attempted a precise definition. It is commonly said, drawing on experience with Earth's atmosphere, that a stratosphere is an atmosphereic layer within which temperature increases with height. This would be an overly restrictive and Earth-centric definition. The dynamically important thing about a stratosphere is that it is much more stably stratified than the troposphere, i.e. that its temperature goes down less steeply than the adiabat appropriate to the planet under consideration. The stable stratification of a layer indicates that convection and other dynamical stirring mechanisms are ineffective or absent in that layer, since otherwise the potential temperature would become well mixed and the temperature profile would become adiabatic. An isothermal layer is stably stratified, because its potential temperature increases with height; even a layer like that of Mars' upper atmosphere, whose temperature decreases gently with height, can be stably stratified. We have shown that an optically thin stratosphere is isothermal in the absence of solar absorption. Indeed, this is often taken as a back-of-the envelope model of stratospheres in general, in simple calculations. In the next chapter, we will determine the temperature profile of stratospheres that are not optically thin.

In a region that is well mixed in the vertical, for example by convection, temperature will decrease with height. Dynamically speaking, such a mixed layer constitutes the troposphere. By contrast the stratosphere may be defined as the layer above this, within which vertical mixing plays a much reduced role. Note, however, that the temperature minimum in a profile need not

be coincident with the maximum height reached by convection; as will be discussed in Chapter 4, radiative effects can cause the temperature to continue decreasing with height above the top of the convectively mixed layer. Yet a further complication is that, in midlatitudes, large scale winds associated with storms are probably more important than convection in carrying out the stirring which establishes the tropopause.

We conclude this chapter with a few comparisons of observed tropopause heights with the predictions of the optically thin limit. We'll leave Venus out of this comparison, since its atmosphere is about as far from the optically thin limit as one could get. On Mars, using the dry adiabat for CO_2 and a $5mb$ surface pressure puts the tropopause at $2.4mb$, which is consistent with the top of the region of steep temperature decline seen in the daytime Martian sounding in Fig. 2.2. For Titan, we use the dry adiabat for N_2 and predict that the tropopause should be at $816mb$,which is again consistent with the sounding. If we use the methane/nitrogen moist adiabat instead of the nitrogen dry adiabat, we put the tropopause distinctly higher, at about $440mb$. Because the moist adiabatic temperature decreases less rapidly with height than the dry adiabat, one must go to greater elevations to hit the skin temperature (as in Fig. 3.14). The tropopause height based on the saturated moist adiabat is distinctly higher than seems compatible with the sounding, from which we infer that the low levels of Titan must be undersaturated with respect to Methane. Using $R/c_p = \frac{2}{7}$ for Earth air and $1000mb$ for the surface pressure, we find that the Earth's tropopause would be at $545mb$ in the optically thin, dry limit. This is somewhat higher in pressure (lower in altitude) that the actual midlatitude tropopause, and very much higher in pressure than the tropical tropopause. Earth's real atmosphere is not optically thin, and the lapse rate is less steep than the dry adiabat owing to the effects of moisture. The effects of optical thickness will be treated in detail in Chapter 4, but we can already estimate the effect of using the moist adiabat. Using the computation of the water-vapor/air moist adiabat described in Chapter 2, the tropopause rises to $157mb$, based on a typical tropical surface temperature of $300K$ and the skin temperature estimated in Table 3.3. This is much closer to the observed tropopause (defined as the temperature minimum in the sounding), with the remaining mismatch being accounted for by the fact that the minimum temperature is appreciably colder than the skin temperature.

3.7 For Further Reading

For more information on electromagnetic waves and electromagnetic radiation, I recommend:

- Jackson JD 1998: *Classical Electrodynamics*. Wiley
- Feynman RP, Leighton RB and Sands M 2005: *The Feynman Lectures on Physics, Vol 2*. Addison Wesley.

An engaging and accessible intellectual history of the quantum theory can be found in

- Pais A 1991: *Niels Bohr's Times*. Oxford University Press

For a derivation of the Planck distribution, see Chapter 1 of

- Rybicki GB and Lightman AP 2004: /it Radiative Processes in Astrophysics. Wiley-VCH.

The reader seeking a comprehensive introduction to nonrelativistic quantum theory will find it in Volume 1 of

- Messiah A 1999: *Quantum Mechanics*. Dover

In the Dover edition, this book is a bargain, and repays a lifetime of close study.

Chapter 4

Radiative transfer in temperature-stratified atmospheres

4.1 Overview

Our objective in this chapter is to treat the computation of a planet's energy loss by infrared emission in sufficient detail that the energy loss can be quantitatively linked to the actual concentration of specific greenhouse gases in the atmosphere. Unlike the simple model of the greenhouse effect described in the preceding chapter, the infrared radiation in a real atmosphere does not all come from a single level; rather, a bit of emission is contributed from each level (each having its own temperature), and a bit of this is absorbed at each intervening level of the atmosphere. The radiation comes out in all directions, and the rate of emission and absorption is strongly dependent on frequency. Dealing with all these complexities may seem daunting, but in fact it can all be boiled down to a conceptually simple set of equations which suffice for a vast range of problems in planetary climate.

It was shown in Chapter 3 that there is almost invariably an order of magnitude separation in wavelengths between the shortwave spectrum at which a planet receives stellar radiation and the longwave (generally infrared) spectrum at which energy is radiated to space. This is true throughout the Solar system, for cold bodies like Titan and hot bodies like Venus, as well as for bodies like Earth that are habitable for creatures like ourselves. The separation calls for distinct sets of approximations in dealing with the two kinds of radiation. Infrared is both absorbed and emitted by an atmosphere, at typical planetary temperatures. However, the long infrared wavelengths are not appreciably scattered by molecules or water clouds, so scattering can be neglected in many circumstances. One of the particular challenges of infrared radiative transfer is the intricate dependence of absorption and emission on wavelength. The character of this dependence is linked to the quantum transitions in molecules whose energy corresponds to infrared photons; it requires an infrared-specific description.

In contrast, planets do not emit significant amounts of radiation in the shortwave spectrum, though shortwave scattering by molecules and clouds is invariably significant; absorption of shortwave radiation arises from quite different molecular processes than those involved in infrared

absorption, and its wavelength-dependence has a correspondingly different character. Moreover, solar radiation generally reaches the planet in the form of a nearly parallel beam, whereas infrared from thermal emission by the planet and it's atmosphere is more nearly isotropic. The approximations pertinent to shortwave radiation will be taken up in Chapter 5, where we will also consider the effects of scattering on thermal infrared.

We'll begin with a general formulation of the equations of plane-parallel radiative transfer without scattering, in Section 4.2. Though we will be able to derive certain general properties of the solutions of these equations, the equations are not very useful in themselves because of the problem of wavelength dependence. To gain further insight, a detailed examination of an idealized model with wavelength-independent infrared emissivity will be presented in Section 4.3. A characterization of the wavelength dependence of the absorption of real gases, and methods for dealing with that dependence, will be given in Sections 4.4 and 4.5.

4.2 Basic Formulation of Plane Parallel Radiative Transfer

We will suppose that the properties of the radiation field and the properties of the medium through which it travels are functions of a single coordinate, which we will take to be the pressure in a hydrostatically balanced atmosphere. (Recall that in such an atmosphere there is a one to one correspondence between pressure and altitude). This is the *plane-parallel* assumption. Although the properties of planetary atmospheres vary geographically with horizontal position within the spherical shell making up the atmosphere, in most cases it suffices to divide up the sphere into patches of atmosphere which are much larger in the horizontal than they are deep, and over which the properties can be considered horizontally uniform. In this case, vertical radiative transfer is much more important than horizontal transfer, and the atmosphere can be divided up into a large number of columns that act independently, insofar as radiative transfer is concerned.

In this section, we will develop an approximate form of the equations of plane parallel radiative transfer. The errors introduced in this approximation are small enough that the resulting equations are sufficiently accurate to form basis of the infrared radiative transfer component of virtually all large scale climate models. These equations will certainly be good enough for addressing the broad-brush climate questions that are our principal concern.

4.2.1 Optical thickness and the Schwarzschild equations

Although the radiation field varies in space only as a function of pressure, p, its intensity depends also on direction. Let $I(p, \hat{n}, \nu)$ be the flux density of electromagnetic radiation propagating in direction $\hat{n}$, measured at point p. This density is just like the Planck function $B(\nu, T)$, except that we allow it to depend on direction and position. The technical term for this flux density is *spectral irradiance*. Now we suppose that the radiation propagates through a thin layer of atmosphere of thickness δp as measured by pressure. The absorption of energy at frequency ν is proportional to the number of molecules of absorber encountered; assuming the mixing ratio of the absorber to be constant within the layer for small δp, the number of molecules encountered will be proportional to δp, in accord with the hydrostatic law. By Kirchoff's law, the absorptivity and emissivity of the layer are the same; we'll call the value $\delta\tau_\nu$, and keep in mind that in general it will be a function of ν. Let θ be the angle between the direction of propagation $\hat{n}$ and the vertical, as shown in Figure 4.1. Now, let $\Delta\tau_\nu^*$ be the emissivity (and absorptivity) of the layer for radiation propagating in the direction $\theta = 0$. We may define the proportionality between emissivity and pressure through the

relation $\delta\tau_\nu^* = -\kappa\Delta p/g$ where g is the acceleration of gravity and κ is an absorption coefficient. It has units of area per unit mass, and can be thought of as an absorption cross-section per unit mass – in essence, the area taken out of the incident beam by the absorbers contained in a unit mass of atmosphere. In general κ is a function of frequency, pressure, temperature and the mixing ratios of the various greenhouse gases in the atmosphere. Passing to the limit of small δp, we can define an *optical thickness* coordinate through the differential equation

$$\frac{d\tau_\nu^*}{dp} = -\frac{1}{g}\kappa \tag{4.1}$$

Since pressure decreases with altitude, τ_ν^* increases with altitude. Radiation propagating at an angle θ relative to the vertical acts just like vertically propagating radiation, except that the thickness of each layer through which the beam propagates, and hence the number of absorbing molecules encountered, is increased by a factor of $1/\cos\theta$. Hence, the optical thickness for radiation propagating with angle θ is simply $\tau_\nu = \tau_\nu^*/\cos\theta$. The equations of radiative transfer can be simplified by using either τ_ν or τ_ν^* in place of pressure as the vertical coordinate.

The specific absorption cross section κ depends on the number of molecules of each greenhouse gas encountered by the beam and the absorption properties characteristic to each kind of greenhouse gas molecule. Letting q_i be the mass-specific concentration of greenhouse gas i, we may write

$$\kappa(\nu, p, T) = \sum_{i=0}^{n} \kappa_i(\nu, p, T) q_i(p) \tag{4.2}$$

The specific concentrations q_i depend on p because we are using pressure as the vertical coordinate, and the concentration of the gas may vary with height;a well-mixed greenhouse gas would have constant q_i. The dependence of the coefficients κ_i on p and T arises from certain aspects of the physics of molecular absorption, to be discussed in Section 4.4.

Eq. 4.1 defines the optical thickness $\tau_\nu^*(p_1, p_2)$ for the layer between pressures p_1 and p_2. Unless κ is constant, it is not proportional to $|p_1 - p_2|$, but it is a general consequence of the definition that $\tau_\nu^*(p_1, p_2) = \tau_\nu^*(p_1, p') + \tau_\nu^*(p', p_2)$ if p' is between p_1 and p_2. Consider an atmosphere with a single greenhouse gas having concentration $q(p)$. Then, if κ_G is independent of p the optical thickness can be expressed as $\tau_\nu^*(p_1, p_2) = \kappa_G \ell$ where ℓ is the *path*, defined by

$$\ell(p_1, p_2) = -\int_{p_1}^{p_2} q(p)\frac{dp}{g} \tag{4.3}$$

The boundaries of the layer are generally chosen so as to make the path and optical thickness positive. The path is the mass of greenhouse gas in the layer, per square meter of the planet's surface, and in *mks* units has units of kg/m^2. If the greenhouse gas is well mixed then $\ell = q \cdot (p_1 - p_2)/g$. Now, it often happens that κ_G increases linearly with pressure – a phenomenon known as *pressure broadening* (alternatively *collisional broadening* for reasons that will eventually become clear. If we write $\kappa_G(p) = \kappa_G(p_o) \cdot (p/p_o)$, then we can define an *equivalent path*

$$\ell_e = -\int_{p_1}^{p_2} q(p)\frac{p}{p_o}\frac{dp}{g} \tag{4.4}$$

such that $\tau_\nu^*(p_1, p_2) = \kappa_G(p_o)\ell_e$ much as before. The equivalent path still has units of mass per unit area, but because of the pressure weighting will differ from the actual path. For example, if the greenhouse gas is well-mixed then

$$\ell_e = \frac{1}{g} q \frac{1}{2p_o}(p_1^2 - p_2^2) = q\frac{p_1 - p_2}{g}\frac{p_1 + p_2}{2p_o} \tag{4.5}$$

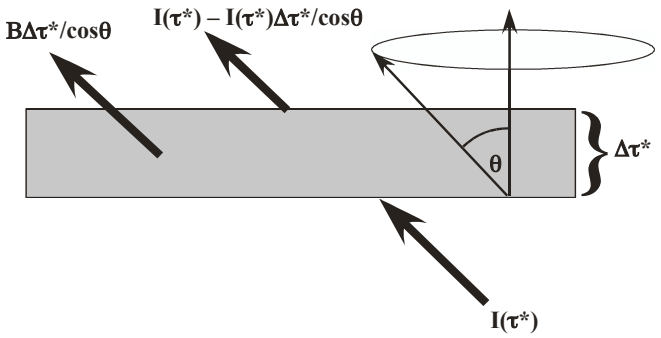

Figure 4.1: Sketch of the radiative energy balance for a slab of atmosphere illuminated by incident radiation from below.

The equivalent path is thus the actual path weighted by the ratio of the mean pressure to the reference pressure.

Consider now the situation illustrated in Fig. 4.1, in which radiation at a given frequency and angle is incident on slab of atmosphere from below. In general, part of the incident radiation is scattered into other directions. However, for infrared and longer wave radiation interacting with gases, such scattering is negligible; scattering is also negligible for infrared interacting with condensed cloud particles made of substances such as water, which are strong absorbers. Here, we shall neglect scattering, though it will be brought back into the picture in Chapter 5. The radiation at the same angle which comes out the top of the slab is then the incident flux minus the small amount absorbed in the slab, plus the small amount emitted. Thus

$$I(\tau_\nu^* + \delta\tau_\nu^*, \hat{n}, \nu) = (1 - \frac{\delta\tau_\nu^*}{\cos\theta})I(\tau_\nu^*, \hat{n}, \nu) + B(\nu, T(\tau_\nu^*))\frac{\delta\tau_\nu^*}{\cos\theta} \tag{4.6}$$

or, passing to the limit of small $\delta\tau_\nu^*$,

$$\frac{d}{d\tau_\nu^*}I(\tau_\nu^*, \hat{n}, \nu) = -\frac{1}{\cos\theta}[I(\tau_\nu^*, \hat{n}, \nu) - B(\nu, T(\tau_\nu^*))] \tag{4.7}$$

For a precise solution, one needs to solve this equation separately for each θ and then integrate over angles to get the net upward and downward fluxes. The angular distribution of radiation changes with distance from the source, since radiation propagating near the direction $\theta = 0$ or $\theta = \pi$ decays more gradually than radiation with θ nearer to $\pi/2$. Hence, radiation that starts out isotropic at the source (as is the blackbody emission) tends to become more forward-peaked as it propagates. For some specialized problems, it is indeed necessary to solve for the angular distribution explicitly in this fashion, which is rather computationally demanding. Fortunately, the isotropy of the blackbody source term tends to keep longwave radiation isotropic enough to allow one to make do with a much more economical approximate set of equations.

We can derive an equation for the net upward flux per unit frequency, I_+, by multiplying Eq. 4.7 by $\cos\theta$ and integrating over all solid angles in the upward-pointing hemisphere. Integrating over the downward hemisphere yields the net downward flux I_-. However, because of the factor $1/\cos\theta$ on the right hand side of Eq. 4.7, the hemispherically averaged intensity appearing on the right hand side is not I_+. Instead, it is $\int I(\tau^*, \hat{n}, \nu)d\Omega$, or equivalently $\int_0^{\pi/2} 2\pi I(\tau^*, \theta, \nu)\sin\theta d\theta$. One cannot proceed further without some assumption about the angular distribution. If we assume that the distribution remains approximately isotropic, by virtue of the isotropic source B, then

$I(\tau^*, \theta, \nu)$ is independent of θ, and hence the problematic integral becomes $2\pi I \int_0^{\pi/2} \sin\theta d\theta$ which is equal to $2I_+$ under the assumption of isotropy. This result yields a closed equation for I_+. It states that, if the radiation field remains approximately isotropic, the decay rate is the same as for unidirectional radiation propagating with an angle $\bar{\theta}$ such that $\cos\bar{\theta} = \frac{1}{2}$, i.e. $\bar{\theta} = 60^o$. From now on we will deal only with this approximate angle-averaged form of the equations, and use $\tau_\nu = \tau^*/\cos\bar{\theta}$ as our vertical coordinate. The choice $\cos\bar{\theta} = \frac{1}{2}$ is by no means a unique consequence of the assumption of isotropy. The fact is that an isotropic distribution is not an exact solution of Eq. 4.7 except in a few very special limits, so that the choice we make is between different errors of roughly the same magnitude. If we had calculated $\cos\bar{\theta}$ by multiplying Eq. 4.7 by $(cos\theta)^2$ instead of $\cos\theta$ before averaging over angles, we would have concluded $\cos\bar{\theta} = \frac{2}{3}$ and this would be an equally valid choice within the limitations of the isotropic approximation. Sometimes, a judicious choice of $\cos\bar{\theta}$ is used to maximize the fit to an angle-resolved calculation in some regime of particular interest. For the most part we will simply use $\cos\bar{\theta} = \frac{1}{2}$ in our calculations unless there is a compelling reason to adopt a different value.

In terms of τ_ν, the equations for the upward and downward flux are

$$\begin{aligned} \frac{d}{d\tau_\nu} I_+ &= -I_+ + \pi B(\nu, T(\tau_\nu)) \\ \frac{d}{d\tau_\nu} I_- &= I_- - \pi B(\nu, T(\tau_\nu)) \end{aligned} \tag{4.8}$$

These are known as the *two-stream equations*, and will serve as the basis for all subsequent discussion of radiative transfer in this book, save that we will incorporate the neglected scattering term in Chapter 5.

Because of the neglect of scattering, the equations for I_+ and I_- are uncoupled, and each consists of a linear, inhomogeneous first order differential equation. The solution can be obtained by substituting $I_+ = A(\tau_\nu)\exp(-\tau_\nu)$, and similarly for I_-, which reduces the problem to evaluation of a definite integral for A. The result is

$$\begin{aligned} I_+(\tau_\nu, \nu) &= I_+(0)e^{-\tau_\nu} + \int_0^{\tau_\nu} \pi B(\nu, T(\tau'_\nu))e^{-(\tau_\nu - \tau'_\nu)} d\tau'_\nu \\ I_-(\tau_\nu, \nu) &= I_-(\tau_\infty)e^{-(\tau_\infty - \tau_\nu)} + \int_{\tau_\nu}^{\tau_\infty} \pi B(\nu, T(\tau'_\nu))e^{-(\tau'_\nu - \tau_\nu)} d\tau'_\nu \end{aligned} \tag{4.9}$$

where τ'_ν is a dummy variable and τ_∞ is the optical thickness of the entire atmosphere, i.e. $\tau^*_\nu(p_s, 0)/\cos\bar{\theta}$. Note that τ_∞ depends on ν in general, though we have suppressed the subscript for the sake of readability. The top of the atmosphere ($p = 0$) is at τ_∞. The physical content of these equations is simple: $I_+(\tau_\nu, \nu)$ consists of two parts. The first is the portion of the emission from the ground which is transmitted by the atmosphere (the first term in the expression for I_+). The second is the radiation emitted by the atmosphere itself, which appears as an exponentially-weighted average (the second term in the expression for I_+) of the emission from all layers below τ_ν, with more distant layers given progressively smaller weights. Similarly, $I_-(\tau_\nu, \nu)$ is an exponentially-weighted average of the emission from all layers above τ_ν, plus the transmission of incident downward flux. The atmospheric emission to space will be most sensitive to temperatures near the top of the atmosphere. This emission will dominate the OLR when the atmosphere is fairly opaque to the radiation emitted from the ground, whereas the transmitted ground emission will dominate when the atmosphere is fairly transparent. The downward radiation into the ground will be most sensitive to temperatures nearest the ground.

In the long run, it will save us some confusion if we introduce special notation for temperatures and fluxes at the boundaries; this will prove especially important when there is occasion to

switch back and forth between pressure and optical thickness as a vertical coordinate. The temperature at the top of the atmosphere ($p = 0$ or $\tau = \tau_\infty$) will be denoted by T_∞,and the temperature of the air at the bottom of the atmosphere ($p = p_s$ or $\tau = 0$) will be called T_{sa}. For planets with a solid or liquid surface this is the temperature of the gas in immediate contact with the surface. For such planets, one must distinguish the temperature of the air from the temperature of the surface (the "ground") itself, which will be called T_g. The outgoing and incoming fluxes at the top of the atmosphere will be called $I_{+,\infty}(\nu)$ and $I_{-,\infty}(\nu)$ respectively, while the upward and downward fluxes at the bottom of the atmosphere will be called $I_{+,s}(\nu)$ and $I_{-,s}(\nu)$

For planets with a liquid or solid surface, we require that $I_{+,s}(\nu)$ be equal to the upward flux emitted by the ground, which is $e(\nu)B(\nu, T_g)$, where $e(\nu)$ is the emissivity of the ground. Continuity of the fluxes is required because, the air being in immediate contact with the ground, there is no medium between the two which could absorb or emit radiation, nor is there any space where radiation "in transit" could temporarily reside. We generally assume that there is no infrared radiation incident on the top of the atmosphere, so that the upper boundary condition is $I_{-,\infty} = 0$. The incident solar radiation does contain some near-infrared, but this is usually treated separately as part of the shortwave radiation calculation (see Chapter 5). For planets orbiting stars with cool photospheres, such as red giant stars, it might make sense to allow $I_{-,\infty}$ to be nonzero and treat the incoming infrared simultaneously with the internally generated thermal infrared. Since the radiative transfer equations are linear in the intensities, it is a matter of taste whether to treat the incoming stellar infrared in this way, or as part of the calculation dealing with the shorter wave part of the incoming stellar spectrum.

For gas or ice giant planets, which have no distinct solid or liquid surface, we do not usually try to model the whole thermal structure of the planet all the way to its center. It typically suffices to specify the temperature and convective heat flux from the interior at some level which is sufficiently deep that the density has increased to the point that the fluid making up the atmosphere can be essentially considered a blackbody. Once the optically thick regime is reached, one doesn't need to know the temperature deeper down in order to do the radiation calculation, any more than one needs to know the temperature of the Earth's core to do radiation on Earth.

The weighting function appearing in the integrands in Eq. 4.9 is the *transmission function*. Written as a function of pressure, it is

$$\mathfrak{T}_\nu(p_1, p_2) = e^{-|\tau_\nu(p_1) - \tau_\nu(p_2)|} \tag{4.10}$$

$\mathfrak{T}_\nu(p_1, p_2)$ is the proportion of incident energy flux at frequency ν which is transmitted through a layer of atmosphere extending from p_1 to p_2; whatever is not transmitted is absorbed in the layer. Note that $\mathfrak{T}_\nu(p, p')d\tau'_\nu = d\mathfrak{T}_\nu$ (with p held constant), if $p < p'$, and $\mathfrak{T}_\nu(p, p')d\tau'_\nu = -d\mathfrak{T}_\nu$ if $p > p'$. Using this result Eq. 4.9 can be re-written

$$\begin{aligned} I_+(\tau_\nu, \nu) &= I_{+,s}(\nu)\mathfrak{T}_\nu(p, p_s) - \int_{p'=p}^{p_s} \pi B(\nu, T(p'))d\mathfrak{T}_\nu(p, p') \\ I_-(\tau_\nu, \nu) &= I_{-,\infty}(\nu)\mathfrak{T}_\nu(0, p) + \int_{p'=0}^{p} \pi B(\nu, T(p'))d\mathfrak{T}_\nu(p, p') \end{aligned} \tag{4.11}$$

In the integrals above, the differential of $\mathfrak{T}_\nu$ is meant to be taken with p held fixed. Integration by parts then yields the following alternate form of the solution to the two-stream equations:

$$\begin{aligned} I_+(p, \nu) &= \pi B(\nu, T(p)) + (I_{+,s}(\nu) - \pi B(\nu, T_{sa}))\mathfrak{T}_\nu(p, p_s) + \int_p^{p_s} \pi\mathfrak{T}_\nu(p, p')dB(\nu, T(p')) \\ I_-(p, \nu) &= \pi B(\nu, T(p)) + (I_{-,\infty}(\nu) - \pi B(\nu, T_\infty))\mathfrak{T}_\nu(0, p) - \int_0^{p} \pi\mathfrak{T}_\nu(p, p')dB(\nu, T(p')) \end{aligned} \tag{4.12}$$

Neither of these forms of the solution is particularly convenient for analytic work, but either one can be used to good advantage when carrying out approximate integrations via the trapezoidal rule (see Section 4.4.6). For analytical work, and some kinds of numerical integration, it helps to rewrite the integrand using $dB = (dB/dT)(dT/dp')dp'$. The result is

$$\begin{aligned} I_+(p,\nu) &= \pi B(\nu, T(p)) + (I_{+,s}(\nu) - \pi B(\nu, T_{sa}))\mathfrak{T}_\nu(p, p_s) + \int_p^{p_s} \pi \mathfrak{T}_\nu(p, p') \frac{dB}{dT}|_{T(p')} \frac{dT}{dp'} dp' \\ I_-(p,\nu) &= \pi B(\nu, T(p)) + (I_{-,\infty}(\nu) - \pi B(\nu, T_\infty))\mathfrak{T}_\nu(0, p) - \int_0^{p} \pi \mathfrak{T}_\nu(p, p') \frac{dB}{dT}|_{T(p')} \frac{dT}{dp'} dp' \end{aligned} \tag{4.13}$$

A considerable advantage of any of the forms in Eq. 4.11, 4.12 or 4.13 is that the integration variable p' is no longer dependent on frequency. This will prove particularly useful when we come to consider real gases, for which the optical thickness has an intricate dependence on frequency. The first two terms in the expression for the fluxes in either Eq. 4.12 or 4.13 give the exact result for an isothermal atmosphere; in each case, the first of the two terms represents the contribution of the local blackbody radiation, whereas the second accounts for the modifying effect of the boundaries. The boundary terms vanish at points far from the boundary, where $\mathfrak{T}$ is small. Note that the boundary term for I_+ vanishes identically if the upward flux at the boundary has the form of blackbody radiation with temperature equal to the surface air temperature. For a planet with a solid or liquid surface, this would be the case if the ground temperature equals the surface air temperature and the ground has unit emissivity.

The main reason for dealing with radiative transfer in the atmosphere is that one needs to know the amount of energy deposited in or withdrawn from a layer of atmosphere by radiation. This is the radiative heating rate (with negative heating representing a cooling). It is obtained by taking the derivative of the net flux, which gives the difference between the energy entering and leaving a thin layer. The heating rate per unit optical thickness, per unit frequency, is thus

$$\mathfrak{H}_\nu = -\frac{d}{d\tau_\nu}(I_+(\tau_\nu, \nu) - I_-(\tau_\nu, \nu)) \tag{4.14}$$

This must be integrated over all frequencies to yield the net heating rate. For making inferences about climate, one ordinarily requires the heating rate per unit mass rather than the heating rate per unit optical depth. This is easily obtained using the definition of optical depth, specifically,

$$H_\nu = g\frac{d}{dp}(I_+ - I_-) = g\frac{d\tau_\nu}{dp}\frac{d}{d\tau_\nu}(I_+ - I_-) = \frac{\kappa}{\cos\bar{\theta}}\mathfrak{H}_\nu \tag{4.15}$$

When integrated over frequency this heating rate has units W/kg. One can convert into a temperature tendency K/s by dividing this value by the specific heat c_p.

4.2.2 Some special solutions of the Two-Stream equations

Beer's law

Suppose that the atmosphere is too cold to radiate significantly at the frequency under consideration. In that case, $B(\nu, T) \approx 0$ and the internal source vanishes. This would be the case, for example, if ν is in the visible light range and the temperature of the atmosphere is Earthlike. In this case, the solutions are simply $I_+ = I_+(0)\exp(-\tau_\nu)$ and $I_- = I_-(\tau_\infty)exp(\tau_\nu - \tau_\infty)$. The exponential attenuation of radiation is known as *Beer's Law*. Here we've neglected scattering, but in Chapter 5 we'll see that a form of Beer's law still applies even if scattering is taken into account.

Infinite isothermal medium

Consider next an unbounded isothermal medium. In this case, it is readily verified that $I_+ = I_- = \pi B(\nu, T)$ is an exact solution to 4.9. The right hand sides of the equations vanish, but the derivatives on the left hand sides vanish also, because T is independent of τ_ν. Hence, in an unbounded isothermal medium, the radiation field reduces to uniform blackbody radiation.

Since the fluxes are independent of τ_ν, the radiative heating rate vanishes, from which we recover the fact that blackbody radiation is in equilibrium with an extended body of isothermal matter.

Exercise 4.2.1 Derive this result from Eq. 4.9; from Eq. 4.13.

Finite-thickness isothermal slab

Now let's consider an isothermal layer of finite thickness, embedded in an atmosphere which is completely transparent to radiation at frequency ν. We suppose further that there is no radiation at this frequency incident on the layer from either above or below. We are free to define $\tau_\nu = 0$ at the center of the layer, so that $\tau_\nu = \frac{1}{2}\tau_\infty$ at the top of the layer and $\tau_\nu = -\frac{1}{2}\tau_\infty$ at the bottom. The boundary conditions corresponding to no incident flux are $I_- = 0$ at the top of the layer and $I_+ = 0$ at the bottom of the layer. The solution $I_+ = I_- = \pi B$ is still a *particular* solution within the layer, since the layer is isothermal, but it does not satisfy the boundary conditions. A homogeneous solution must be added to each flux in order to satisfy the boundary conditions. The homogeneous solutions are just exponentials, and so we easily find that the full solution within the layer is

$$\begin{aligned} I_+(\tau_\nu, \nu) &= [1 - \exp(-(\tau + \frac{\tau_\infty}{2}))]\pi B \\ I_-(\tau_\nu, \nu) &= [1 - \exp(+(\tau - \frac{\tau_\infty}{2}))]\pi B \end{aligned} \tag{4.16}$$

Exercise 4.2.2 Derive this result from Eq. 4.13.

The radiation emitted out the top of the layer is $I_+(\tau_\infty, \nu)$, or $(1 - \exp(-\tau_\infty))\pi B$, which reduces to the blackbody value πB when the layer is optically thick for the frequency in question, i.e. $\tau_\infty(\nu) \gg 1$. The same applies for the emission out of the bottom of the layer, *mutatis mutandum*. Note that in the optically thick limit, $I_+ = I_- = \pi B$ through most of the layer, and the inward-directed intensities only fall to zero in the two relatively thin skin layers near the top and bottom of the slab.

In the opposite extreme, when the slab is optically thin, both τ and τ_∞ are small. Using the first order Taylor expansion of the exponentials, we find that the emission out the top of the layer is $\tau_\infty \pi B$, and similarly for the bottom of the layer. Hence, τ_∞ in this case is just the bulk emissivity of the layer. This is consistent with the way we constructed the Schwarzschild equations, which can be viewed as a matter of stacking a great number of individually optically thin slabs upon each other.

Substituting into Eq. 4.14, the heating rate for this solution is

$$\mathfrak{H}_\nu = -[\exp(-\tau) + \exp(\tau)] \exp(-\frac{\tau_\infty}{2})\pi B \tag{4.17}$$

In the optically thick case, the heating rate is nearly zero in the interior of the slab, but there is strong radiative cooling within about a unit optical depth of each surface. In this case the radiation drains heat out of a thin skin layer near each surface, causing intense cooling there. In the optically thin limit, the cooling is distributed uniformly throughout the slab.

It turns out that condensed water is a much better infrared absorber than the same mass of water vapor. Hence, an isolated absorbing layer such as we have just considered can be thought of as a very idealized model of a cloud. The following slight extension makes the connection with low lying stratus clouds, such as commonly found over the oceans, more apparent.

Exercise 4.2.3 Instead of being suspended in an infinite transparent medium, suppose that the cloud is in contact with the ground, and that the ground has the same temperature as the cloud. We still assume that the air above the cloud is transparent to radiation at the frequency under consideration. Compute the upward and downward fluxes, and the radiative heating rate, in this case.

This exercise shows that convection in boundary layer stratus clouds can be driven by cooling at the top, rather than heating from below. This is rather important, since the reflection of sunlight by the cloud makes it hard to warm up the surface. Entrainment of dry air due to top-driven convection is one of the main mechanisms for dissipating such clouds.

Optically thick limit

We now depart from the assumption of constant temperature. While allowing T to vary in the vertical, we assume the atmosphere to be *optically thick* at frequency ν. This means that a small change in pressure p amounts to a large change in the optical thickness coordinate τ_ν. Referring, to Eq. 4.1, we see that the assumption of optical thickness is equivalent to the assumption that $\kappa\delta p/g \gg 1$, where δp is the typical amount by which one has to change the pressure in order for the temperature to change by an amount comparable to its mean value. For most atmospheres, it suffices to take δp to be the depth of the whole atmosphere, namely p_s, so that the optical thickness assumption becomes $\tau_\infty = \kappa p_s/g \gg 1$ Since κ depends on frequency, an atmosphere may be optically thick near one frequency, but optically thin near another.

The approximate form of the fluxes in the optically thick limit can be most easily derived from the integral expression in the form given in Eq. 4.9. Consider first the expression for I_+. Away from the immediate vicinity of the bottom boundary, the boundary term proportional to $\mathfrak{T}_\nu(p, p_s)$ is exponentially small and can be dropped. To simplify the integral, we note that $\mathfrak{T}_\nu(p, p')$ is very small unless p' is close to p. Therefore, as long as the temperature gradient is a continuous function of p', it varies little over the range of p' for which the integrand contributes significantly to the integral. Hence dT/dp' can be replaced by its value at p, which can then be taken outside the integral. Likewise, dB/dT can be evaluated at $T(p)$, so that this term can also be taken outside the integral. Finally, if one is not too close to the bottom boundary,

$$\int_p^{p_s} \mathfrak{T}_\nu(p', p)dp' \approx \int_p^{\infty} \mathfrak{T}_\nu(p', p)dp' = \frac{g\cos\bar{\theta}}{\kappa}\int_\tau^{\infty} \mathfrak{T}_\nu(\tau', \tau)d\tau' = \frac{g\cos\bar{\theta}}{\kappa}, \tag{4.18}$$

whence the upward flux in the optically thick limit becomes

$$I_+(\nu, p) = \pi B(\nu, T(p)) + \pi\frac{g\cos\bar{\theta}}{\kappa}\frac{dB}{dT}|_{T(p)}\frac{dT}{dp} \tag{4.19}$$

Near the bottom boundary, the neglected boundary term would have to be added to this expression. In addition, Eq. 4.18 would need to be corrected to allow for the fact that there is not room for $\int \mathfrak{T}$ to integrate out to its asymptotic value for an infinitely thick layer.

Using identical reasoning, the downward flux becomes

$$I_-(\nu,p) = \pi B(\nu,T(p)) - \pi \frac{g\cos\bar{\theta}}{\kappa}\frac{dB}{dT}|_{T(p)}\frac{dT}{dp} \tag{4.20}$$

so long as one is not too near the top of the atmosphere. Near the top of the atmosphere, the neglected boundary term becomes significant.

In both expressions the second term, proportional to the temperature gradient, becomes progressively smaller as κ is made larger and the atmosphere becomes more optically thick. To lowest order, then, the upward and downward fluxes are both equal to the blackbody radiation flux at the local temperature. In this sense, the optically thick limit looks "locally isothermal." The term proportional to the temperature gradient represents a small correction to the locally isothermal behavior. In the expression for I_+, for example, if $dT/dp > 0$ the correction term makes the upward flux somewhat greater than the local blackbody value. This makes sense, because a small portion of the upwelling radiation comes from lower layers where the temperature is warmer than the local temperature. Note that the correction term depends on ν through the frequency dependence of κ, as well as through the frequency dependence of dB/dT.

The radiation exiting the top of the atmosphere ($I_{+,\infty}$) is of particular interest, because it determines the rate at which the planet loses energy. In the optically thick approximation, we find that as long as dT/dp is finite at $p = 0$, $I_{+,\infty}$ becomes close to $\pi B(\nu, T_\infty)$ as the atmosphere is made more optically thick. Hence, at frequencies where the atmosphere is optically thick, the planet radiates to space like a blackbody with temperature equal to that of the upper regions of the atmosphere – the regions "closest" to outer space.

Similarly, the downward radiation ($I_{-,s}(\nu)$) from the atmosphere into the ground – sometimes called the *back radiation* – is of interest because it characterizes the radiative effect of the atmosphere on the surface energy budget. In the optically thick limit, $I_{-,s}(\nu) = \pi B(\nu, T_{sa})$ to lowest order, so that the atmosphere radiates to the ground like a blackbody with temperature equal to the low level air temperature. If $dT/dp > 0$ at the ground, as is typically the case, then the correction term slightly reduces the downward radiation, because some of the radiation into the ground comes from higher altitudes where the air is colder. Suppose now that the surface temperature T_g is equal to the air temperature T_{sa}, and that the surface has unit emissivity at the frequency under consideration. In that case, the net radiative heating of the ground is

$$I_{-,s}(\nu) - B(\nu,T_g) = I_{-,s}(\nu) - B(\nu,T_{sa}) = -\pi \frac{g\cos\bar{\theta}}{\kappa}\frac{dB}{dT}|_{T_{sa}}\frac{dT}{dp}|p_s \tag{4.21}$$

at frequencies where the solar flux is negligible. This is negative when $dT/dp > 0$, representing a radiative cooling of the ground. The radiative cooling vanishes in the limit of large κ. In the optically thick limit, then, the surface cannot get rid of heat by radiation unless the ground temperature becomes larger than the low level air temperature. Remember, though, that the radiative heating of the ground is but one term in the surface energy budget coupling the surface to the atmosphere. Turbulent fluxes of moisture and heat also exchange energy between the surface and the atmosphere, and these become dominant when the radiative term is weak.

In the optically thick limit, the net flux is

$$I_+ - I_- = 2\pi \frac{g \cos\bar{\theta}}{\kappa} \frac{dB}{dT} \frac{dT}{dp} \tag{4.22}$$

whence the radiative heating rate is

$$H_\nu = \frac{d}{dp}[D(\nu, p)\frac{dT}{dp}] \tag{4.23}$$

where

$$D(\nu, p) = 2\pi \frac{g^2 \cos\bar{\theta}}{\kappa} \frac{dB}{dT}|_{T(p)} \tag{4.24}$$

Hence, in the optically thick limit, the heating and cooling caused by radiative transfer acts just like a thermal diffusion in pressure coordinates, with the diffusivity given by $D(\nu, p)$. Since $dB/dT > 0$, the radiative diffusivity is always positive. It becomes weak as κ becomes large. Note that the diffusive approximation to the heating is is only valid when one is not too close to the top and bottom of the atmosphere. Near the boundaries, the neglected boundary terms contribute an additional heating which is exponentially trapped near the top and bottom of the atmosphere. The effect of the boundary terms is explored in Problem **??**

Consider an atmosphere which is transparent to solar radiation, and within which heat is redistributed only by infrared radiative transfer. Eq 4.15 then requires that the net upward flux $I_+ - I_-$ must be independent of altitude when integrated over all wavenumbers. This constant flux is nonzero, since the infrared flux through the system is set by the rate at which infrared escapes from the top of the atmosphere – namely, the OLR. Integrating Eq. 4.22 over the infrared yields an expression for dT/dp in terms of the OLR and the frequency-integrated diffusivity; because both OLR and diffusivity are positive, it follows that $dT/dp > 0$ for an optically thick atmosphere in pure infrared radiative equilibrium – that is, the temperature decreases with altitude. The more optically thick the atmosphere becomes, the smaller is D, and hence the stronger is the temperature variation in equilibrium. Pure radiative equilibrium will be discussed in detail in Sections 4.3.4 and 4.7, and the optically thick limit is explored in Problem **??**.

Optically thin limit

The *optically thin* limit is defined by $\tau_\infty \ll 1$. Since $\tau_\nu \leq \tau_\infty$ and $\tau'_\nu \leq \tau$ in Eq. 4.9, all the exponentials in the expression for the fluxes are close to unity. Moreover, the integral is carried out over the small interval $[0, \tau_\nu]$, and hence is already of order τ_∞ or less. It is thus a small correction to the first term, and we may set the exponentials in the integrand to unity and still have an expression that is accurate to order τ_∞. The boundary terms are not integrated, though, so we must retain the first two terms in the Taylor series expansion of the exponential to achieve the same accuracy. With these approximations, the fluxes become

$$\begin{aligned} I_+(\tau_\nu) &= (1 - \tau_\nu) I_{+,s} + \int_0^{\tau_\nu} \pi B(\nu, T(\tau'_\nu)) d\tau'_\nu \\ I_-(\tau_\nu) &= (1 - (\tau_\infty - \tau_\nu)) I_{-,\infty} + \int_{\tau_\nu}^{\tau_\infty} \pi B(\nu, T(\tau'_\nu)) d\tau'_\nu \end{aligned} \tag{4.25}$$

In this case, the upward flux is the sum of the upward flux from the boundary (diminished by the slight atmospheric absorption on the way up) with the sum of the unmodified blackbody emission from all the layers below the point in question. The downward flux is interpreted similarly.

In order to discuss the radiation escaping the top of the atmosphere and the back-radiation into the ground, we introduce the mean emission temperature $\bar{T}_\nu$, defined by solving the relation

$$B(\bar{T}_\nu, \nu) = \frac{1}{\tau_\infty} \int_0^{\tau_\infty} B(\nu, T(\tau'_\nu)) d\tau'_\nu \tag{4.26}$$

With this definition, the boundary fluxes are

$$\begin{aligned} I_{+,\infty} &= (1-\tau_\infty) I_+(0) + \tau_\infty \pi B(\bar{T}_\nu, \nu) \\ I_{-,s} &= (1-\tau_\infty) I_{-,\infty}) + \tau_\infty \pi B(\bar{T}_\nu, \nu) \end{aligned} \tag{4.27}$$

According to this expression, an optically thin atmosphere acts precisely like an isothermal slab with temperature $\bar{T}_\nu$ and (small) emissivity τ_∞. It is only in the optically thin limit that the radiative effect of the atmosphere mimics that of an isothermal slab.

Substituting the approximate form of the fluxes into the expression for radiative heating rate, we find

$$H_\nu = \frac{\kappa}{\cos\bar{\theta}} \cdot [(I_{+,s} + I_{-,\infty}) - 2\pi B(\nu, T(p))] \tag{4.28}$$

This is small, because κ is small in the optically thin limit. The first pair of terms are always positive, and represent heating due to the proportion of incident fluxes which are absorbed in the atmosphere. The second term is always negative, and represents cooling by blackbody emission of the layer of air at pressure p. In contrast to the general case or the optically thick case, the cooling term is purely a function of the local temperature; radiation emitted by each layer escapes directly to space or to the ground, without being significantly captured and re-emitted back by any other layer.

Typical greenhouse gases are optically thin in some spectral regions and optically thick in others. We have seen that the infrared heating rate becomes small in both limits. From this result, we deduce the following general principle: *The infrared heating rate of an atmosphere is dominated by the spectral regions where the optical thickness is order unity.* If an atmosphere is optically thick throughout the spectrum, the heating is dominated by the least thick regions; if it is optically thin throughout, it is dominated by the least thin regions.

4.3 The Grey Gas Model

We will see in Section 4.4 that for most atmospheric gases κ, and hence the optical thickness, has an intricate dependence on wavenumber. This considerably complicates the solution of the radiative transfer equations, since the fluxes must be solved for individually on a very dense grid of wavenumbers, and then the results integrated to yield the net atmospheric heating, which is the quantity of primary interest. The development of shortcuts that can improve on a brute-force integration is an involved business, which in some respects is as much art as science, and leads to equations whose behavior can be difficult to fathom. The radiative transfer equations become much simpler if the optical thickness is independent of wavenumber. This is known as the *grey gas approximation*. For grey gases, the Schwartzschild equations can be integrated over wavenumber, yielding a single differential equation for the net upward and downward flux. More specifically, we shall assume only that the optical thickness is independent of wavenumber within the infrared spectrum, and that the temperature of the planet and its atmosphere is such that essentially all the emission of radiation lies in the infrared. Instead of integrating over all wavenumbers, we integrate only over the infrared range, thus obtaining a set of equations for the net infrared flux. Because

of the assumption regarding the emission spectrum, the integrals of the Planck function $\pi B(\nu, T)$ over the infrared range can be well approximated by σT^4.

With the exception of clouds of strongly absorbing condensed substances like water, the grey gas model yields a poor representation of radiative transfer in real atmospheres, for which the absorption is typically strongly dependent on wavenumber. Nonetheless, a thorough understanding of the grey gas model provides the starting point for any deeper inquiry into atmospheric radiation. Here, we can find many of the fundamental phenomena laid bare, because one can get much farther before resorting to detailed numerical computations. Further, grey gas radiation has proved valuable as a placeholder radiation scheme in theoretical studies involving the coupling of radiation to fluid dynamics, when one wants to focus on dynamical phenomena without the complexity and computational expense of real gas radiative transfer. Sometimes, a simple scheme which is easy to understand is better than an accurate scheme which defies comprehension.

The grey gas versions of the two-stream Schwarzschild equations are obtained by making τ_ν independent of frequency and integrating the resulting equations over all frequencies. The result is

$$\begin{aligned} \frac{d}{d\tau} I_+ &= -I_+ + \sigma T(\tau)^4 \\ \frac{d}{d\tau} I_- &= I_- - \sigma T(\tau)^4 \end{aligned} \tag{4.29}$$

Grey gas versions of the solutions given in the previous section can similarly be obtained by integrating the relations over all frequencies, taking into account that τ is now independent of ν. The expressions have precisely the same form as before, except that I_+ and I_- now represent total flux integrated over all longwave frequencies, and every occurence of πB is replaced by σT^4. To avoid unnecessary proliferation of notation, when the context allows little possibility of confusion we will use the same symbols I_+ and I_- to represent the longwave-integrated flux as we used earlier to represent the frequency-dependent flux spectrum. When we need to emphasize that a flux is a frequency dependent spectrum, we will include the dependence explicitly (as in "$I_+(\nu)$" or "$I_+(\nu, p)$"; when we need to emphasize that a flux represents the longwave-integrated net flux, we will use an overbar (as in $\bar{I}_+$).

4.3.1 OLR and back-radiation for an optically thin grey atmosphere

The OLR and surface back-radiation for an optically thin grey atmosphere are obtained by integrating Eq. 4.27 over all frequencies. The result is

$$\begin{aligned} I_{+,\infty} &= (1 - \tau_\infty) I_+(0) + \tau_\infty \sigma \bar{T}^4 \\ I_{-,s} &= (1 - \tau_\infty) I_{-,\infty} + \tau_\infty \sigma \bar{T}^4 \end{aligned} \tag{4.30}$$

where the mean atmospheric emission temperature is given by

$$\bar{T}^4 = \frac{1}{\tau_\infty} \int_0^{\tau_\infty} T^4 d\tau \tag{4.31}$$

The first term in the expression for $I_{+,\infty}$ is the proportion of upward radiation from the ground which escapes without absorption by the intervening atmosphere, while the second is the emission to space added by the atmosphere. In the expression for $I_{-,s}$ the first term is the proportion of incoming infrared flux which reaches the surface without absorption, while the second is the

downward emission from the atmosphere. Note that for an optically thin atmosphere the atmospheric emission to space is identical to the atmospheric emission to the ground; in this regard the atmosphere radiates like an isothermal slab with temperature $\bar{T}$. According to Eq. 4.27, a nongrey atmosphere behaves similarly, if it is optically thin *for all frequencies*.

4.3.2 Radiative properties of an all-troposphere dry atmosphere

Let's consider an atmosphere for which the convection is so deep that it establishes a dry adiabat throughout the depth of the atmosphere. Thus, $T(p) = T_s(p/p_s)^{R/c_p}$ all the way to $p = 0$. We wish to compute the OLR for this atmosphere, which is done by substituting this $T(p)$ into the grey-gas form of Eq. 4.9 and evaluating the integral for I_+ at $\tau = \tau_\infty$, i.e. the top of the atmosphere. Since the temperature is expressed as a function of pressure, it is necessary to substitute for pressure in terms of optical thickness in order to carry out the integral. We'll suppose that κ is a constant, so that $\tau_\infty - \tau = \kappa p/g = \tau_\infty p/p_s$. Using this to eliminate pressure from $T(p)$, the integral for OLR becomes

$$\begin{aligned} OLR &= I_{+,s}e^{-\tau_\infty} + \int_0^{\tau_\infty} \sigma T_s^4 (\frac{\tau_\infty - \tau'}{\tau_\infty})^{4R/c_p} e^{-(\tau_\infty - \tau')} d\tau' \\ &= I_{+,s}e^{-\tau_\infty} + \int_0^{\tau_\infty} \sigma T_s^4 (\frac{\tau_1}{\tau_\infty})^{4R/c_p} e^{-\tau_1} d\tau_1 \\ &= I_{+,s}e^{-\tau_\infty} + \sigma T_s^4 \tau_\infty^{-4R/c_p} \int_0^{\tau_\infty} \tau_1^{4R/c_p} e^{-\tau_1} d\tau_1 \end{aligned} \tag{4.32}$$

The second line is derived by introducing a new dummy variable $\tau_1 = \tau_\infty - \tau'$. This is the the optical depth measured relative to the top of the atmosphere, and the re-expressed integral is computed by integrating from the top down, rather than from the ground up. The first term on the right hand side of Eq. 4.32 represents the proportion of the upward surface radiation which survives absorption by the atmosphere and reaches space. The second term is the net emission from the atmosphere. In the optically thin limit, the integral becomes small and the exponential in the first term approaches unity; thus, the OLR approaches the emission from the ground, $I_{+,s}$. As the atmosphere is made more optically thick, the boundary term becomes exponentially small, and the integral becomes more and more dominated by the emission from the upper reaches of the atmosphere. However, to obtain the optically thick limit, we cannot use the grey gas form of Eq 4.19, since dT/dp becomes infinite at $p = 0$ when the dry adiabat extends all the way to the top of the atmosphere.

In the optically thick limit, $\tau_\infty \gg 1$, the first term becomes exponentially small and the upper limit of the integral can be replaced by ∞, yielding the expression

$$OLR = \sigma T_s^4 \tau_\infty^{-4R/c_p} \Gamma(1 + \frac{4R}{c_p}) \tag{4.33}$$

where Γ is the Gamma function, defined by $\Gamma(s) \equiv \int_0^\infty \zeta^{s-1} exp(-\zeta) d\zeta$. Using integration by parts, $\Gamma(s) = (s-1)\Gamma(s-1)$, while $\Gamma(1) = 1$, so $\Gamma(n) = (n-1)!$. For Earth air, $4R/c_p = \frac{8}{7}$ so $\Gamma(1+4R/c_p)$ will be close to $\Gamma(2)$, which is unity; in fact it is approximately 1.06. For any of the gases commonly found in planetary atmospheres, $\Gamma(1 + 4R/c_p)$ will be an order unity constant. As the atmosphere is made more optically thick, the OLR goes down algebraically like τ_∞^{-4R/c_p}, becoming much less than the value σT_s^4 prevailing for a transparent atmosphere. The OLR approaches zero as τ_∞ is made large because the temperature vanishes at the top of the atmosphere, and as the atmosphere

is made more optically thick, the OLR is progressively more dominated by the emission from the cold upper reaches of the atmosphere.

The calculation can be related to the conceptual greenhouse effect model introduced in the previous chapter by computing the effective radiating pressure p_{rad}. Recall that $\sigma T_{rad}^4 = OLR$, so

$$\sigma T_s^4 (\frac{p_{rad}}{p_s})^{4R/c_p} = OLR = \sigma T_s^4 \tau_\infty^{-4R/c_p} \Gamma(1 + \frac{4R}{c_p}) \tag{4.34}$$

whence $(p_{rad}/p_s) = \tau_\infty^{-1}(\Gamma(1 + 4R/c_p))^{c_p/4R}$. This formula implies that the radiation to space comes essentially from the top unit optical depth of the atmosphere. If an atmosphere has optical depth $\tau_\infty = 100$, then it is only the layer between roughly the top of the atmosphere ($\tau = 100$) and $\tau = 99$ which dominates the OLR. For the all-troposphere model, the maximum temperature of the top unit optical depth approaches zero as the atmosphere is made more optically thick, because this entire layer corresponds to pressures approaching zero ever more closely as κ is made larger.

If S is the absorbed solar radiation per unit area of the planet's surface, then the surface temperature in balance with S is obtained by setting the OLR equal to S. Solving for the surface temperature, we find that in the optically thick all-troposphere limit, the surface temperature is

$$T_s = (S/\sigma)^{\frac{1}{4}} \Gamma(1 + \frac{4R}{c_p})^{-\frac{1}{4}} \cdot \tau_\infty^{R/c_p} \tag{4.35}$$

The first term is the temperature the planet would have in the absence of any atmosphere. As τ_∞ increases, the surface becomes warmer without bound. This constitutes our simplest quantitative model of the greenhouse effect for a temperature-stratified atmosphere. Note that the greenhouse warming depends on the lapse rate. For an isothermal atmosphere ($R/c_p = 0$) there is no greenhouse warming. For fixed optical depth, the greenhouse warming becomes larger as the R/c_p, and hence the lapse rate, becomes large. For Venus, the absorbed solar radiation is approximately $163 W/m^2$, owing to the high albedo of the planet. For a pure CO_2 atmosphere $R/c_p \approx .2304$, for which $\Gamma \approx .969$. Then, the $737K$ surface temperature of Venus can be accounted for if $\tau_\infty = 156$, which is a very optically thick atmosphere. This is essentially the calculation used by Carl Sagan to infer that the dense CO_2 atmosphere of Venus could give it a high enough surface temperature to account for the then-mysterious anomalously high microwave radiation emitted by the planet (microwaves being directly emitted to space by the hot surface without significant absorption by the atmosphere).

Exercise 4.3.1 This exercise illustrates the fact that if the Earth's atmosphere acted like a grey gas, then a doubling of CO_2 would make us toast! Using Eq. 4.35, find the τ_∞ that yields a surface temperature of $285K$ for the Earth's absorbed solar radiation (about $270 W/m^2$ allowing a crude correction for net cloud effects). Now suppose we double the greenhouse gas content of the atmosphere. If the Earth's greenhouse gases were grey gases, this would imply doubling the value of τ_∞ from the value you just obtained. What would the resulting temperature be? Note that this rather alarming temperature doesn't even fully take into account the amplifying effect of water vapor feedback.

An examination of the radiative heating rate profile for the all-troposphere case provides much insight into the processes which determine where the troposphere leaves off and where a stratosphere will form. We'll assume that $I_{-,\infty} = 0$ and that the turbulent heat transfer at the ground is efficient enough that $T_{sa} = T_g$. Consider first the optically thin limit, for which the grey

gas version of Eq. 4.28 is

$$H_\nu = \frac{\kappa}{\cos\bar{\theta}} \cdot [\sigma T_g^4 - 2\sigma T(p)^4] \tag{4.36}$$

assuming the stated boundary conditions. Since the radiative heating rate is nonzero, the temperature profile will not be in a steady state unless some other source of heating and cooling is provided to cancel the radiative heating. According to Eq. 4.36, the atmosphere is cooling at low altitudes, where $T > T_g/2^{\frac{1}{4}}$, i.e. where the local temperature is greater than the skin temperature. The cooling will make the atmosphere's potential temperature lower than the ground temperature, which allows the air in contact with the ground to be positively buoyant. The resulting convection brings heat to the radiatively cooled layer, allowing a steady state to be maintained if the convection is vigorous enough. However, in the upper atmosphere, where $T < T_g/2^{\frac{1}{4}}$ the atmosphere is being *heated* by upwelling infrared radiation, and there is no obvious way that convection could provide the cooling needed to make this region a steady state. Instead, the atmosphere in this region is expected to warm until a stratosphere in pure radiative equilibrium forms. Indeed, the tropopause as estimated by the boundary between the region of net heating and net cooling is located at the point where $T(p)$ equals the skin temperature; this is precisely the same result as we obtained in the steady state model of the tropopause for an optically thin atmosphere, as discussed in Section 3.6.

In the optically thick limit it is easiest to infer the infrared heating profile from an examination of the expression for net infrared flux, which becomes

$$I_+ - I_- = 2\frac{g\cos\bar{\theta}}{\kappa}(4\sigma T^3)\frac{dT}{dp} = 8\sigma T_g^4 \frac{R}{c_p}\frac{g^2\cos\bar{\theta}}{\kappa p_s}(\frac{p}{p_s})^{4R/c_p - 1} \tag{4.37}$$

in the all-troposphere grey-gas case. Recall that this expression breaks down in thin layers within roughly a unit optical depth of the bottom and top boundaries. The formula shows that whether the bulk of an optically thick atmosphere is heating or cooling depends on the lapse rate. The formula is valid even if κ depends on pressure and temperature. For constant κ, if $4R/c_p > 1$ the optically thick net flux decreases with height, and most of the atmosphere is *heated* by infrared radiation, and hence we expect a deep stratosphere and shallow troposphere. If $4R/c_p < 1$, corresponding to a weaker temperature lapse rate, most of the atmosphere instead experiences infrared radiative cooling, so we expect a deep troposphere. For constant κ, most gases would produce a deep stratosphere in the optically thick limit. This conclusion is greatly altered by pressure broadening. If $\kappa(p) = (p/p_s)\kappa(p_s)$, then the appearance of the pressure factor in the denominator alters the flux profile. Specifically, we now only require that $4R/c_p < 2$ in order for the flux to increase with height, yielding radiative cooling throughout the depth where the approximation is valid. Most atmospheric gases satisfy this criterion, and therefore most gases ought to yield a deep troposphere in the optically thick limit. Real gases are typically optically thick at some wavenumbers but optically thin at others, and we shall see in Section 4.8 how the competition plays out

Figure 4.2 shows numerically computed profiles of net infrared flux ($I_+ - I_-$) without pressure broadening, for a range of optical thicknesses, with $R/c_p = \frac{2}{7}$. In this case, $4R/c_p > 1$, and we expect deep heating in the optically thick limit. For $\tau_\infty = 50$ the profile does follow the optically thick approximate form over most of the atmosphere, and exhibits a decrease in flux with height, implying deep heating. There is a thin layer of cooling near the ground, where the optically thick formula breaks down. When $\tau_\infty = 10$, the flux only conforms to the optically thick limit near the center of the atmosphere; there is a region of infrared cooling that extends from the ground nearly to 70% of the surface pressure. The case $\tau_\infty = 1$ looks quite like the optically thin limit, with the lower half of the atmosphere cooling and the upper half heating. Numerical computations incorporating pressure broadening confirm the predictions of the optically thick

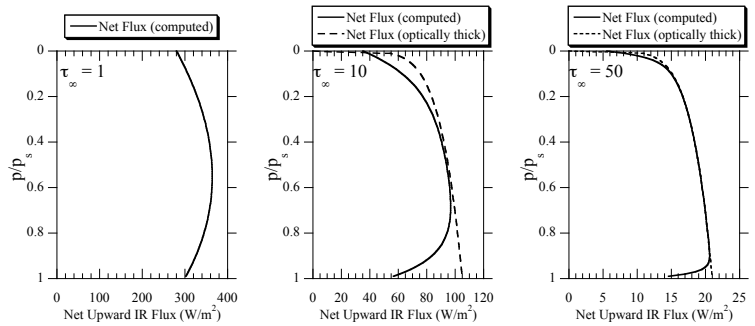

Figure 4.2: Net infrared flux $(I_+ - I_-)$ for the all-troposphere grey-gas model, for $\tau_\infty = 1, 10$ and 50. In the latter two cases, the dashed line gives the result of the optically thick approximation. The surface temperature is fixed at $300K$, and the temperature profile is the dry adiabat with $R/c_p = \frac{2}{7}$.

formula. Specifically, the boundary between cooling and heating is unchanged for optically thin atmospheres, but rises to $p/p_s = .24$ for $\tau_\infty = 10$ and to $p/p_s = .11$ for $\tau_\infty = 50$. The profiles are not shown here, but are explored in Problem **??**.

The troposphere is defined as the layer stirred by convection, and since hot air rises, buoyancy driven convection transports heat upward where it is balanced by radiative cooling. Therefore, at least the upper region of a troposphere invariably experiences radiative cooling. In the calculation discussed above, the layer with cooling, fated to become the troposphere, occurs in the lower portion of the atmosphere. In Figure 4.2 one notices that the radiative cooling decreases as the atmosphere is made more optically thick, suggesting that tropospheric convection becomes more sluggish in an optically thick atmosphere, there being less radiative cooling to be offset by convective heating. However, one should note also that this sequence of calculations is done with fixed surface temperature, and that the OLR decreases as optical thickness is made larger. Hence, in the optically thick cases, it takes less absorbed solar radiation to maintain the surface temperature of the planet. There is less flux of energy through the system, and correspondingly less convection. Mars, at a more distant orbit than Earth, receives less solar energy; if Mars were given an atmosphere with enough greenhouse effect to warm it up to Earthlike temperatures, one would expect the radiative cooling in its troposphere to be less than Earth's, and one would expect the convection to be more sluggish.

The presence of a stratosphere causes the OLR to exceed the values implied by the all-troposphere calculation, since the upper portions of an atmosphere with a stratosphere are warmer than the all-troposphere model would predict. If the stratosphere is optically thin, it has a minor effect on the OLR; in essence, the all-troposphere OLR formula provides a good estimate if the effective radiating level is below the tropopause. If the stratosphere becomes optically thick, then the OLR is in fact determined by the stratospheric structure. Problem **??** explores some aspects of the effect of an optically thick stratosphere on OLR. Puzzling out the effect of the stratosphere on OLR is rather tricky, because the tropopause height itself depends on the optical thickness of the atmosphere. An optically thin atmosphere obviously can't have an optically thick stratosphere, but an optically thick atmosphere can nevertheless have an optically thin stratosphere if the tropopause height increases rapidly enough with τ_∞. The grey gas radiative cooling profiles discussed above suggest that the stratosphere becomes optically thick when $4R/c_p > 1$. In contrast, for $4R/c_p < 1$ the radiatively cooled layer extends toward the top of the atmosphere in the optically thick limit,

and hence the stratosphere could remain optically thin.

4.3.3 A first look at the runaway greenhouse

We have seen in Chapter 2 that the mass of an atmosphere in equilibrium with a reservoir of condensed substance (e.g. a water ocean) is not fixed. It increases with temperature in accordance with the dictates of the Clausius-Clapeyron relation. If the condensible substance is a greenhouse gas, then the optical thickness τ_∞ increases with temperature. This tends to reduce the OLR, offsetting or even reversing the tendency of rising temperature to increase the OLR. What are the implications of this for the dependence of OLR on surface temperature, and for planetary energy balance? The resulting phenomena are most commonly thought about in connection with the effects of a water ocean on evolution of a planet's climate, but the concept generalizes to any condensible greenhouse gas in equilibrium with a large condensed reservoir. We'll take a first look at this problem here, in the context of the grey-gas model.

In the general case, we'd like to consider an atmosphere in which the condensible greenhouse gas is mixed with a noncondensible background of fixed mass (which may also have a greenhouse effect of its own). This is the case for water vapor in the Earth's atmosphere, for methane on Titan, and probably also for water vapor in the early atmosphere of Venus. It could also have been the case for mixed nitrogen-CO_2 atmospheres on Early Mars, with CO_2 playing the role of the condensible component. We will eventually take up such atmospheres, but the difficulty in computing the moist adiabat for a two-component atmosphere introduces some distractions which get in the way of grasping the key phenomena. Hence, we'll start with the simpler case in which the atmosphere consists of a pure condensible component in equilibrium with a reservoir (an "ocean," or perhaps a glacier). In this case, the saturated moist adiabat is given by the simple analytic formula Eq. 2.27, obtained by solving the simplified form of the Clausius-Clapeyron relation for temperature in terms of pressure. We've already seen in Chapter 2 that a mixed atmosphere is dominated by the condensible component at large temperatures, so if we are primarily interested in the large-temperature behavior, the use of the one-component condensible atmosphere is not at all a bad approximation.

We write $T(p) = T_o/(1 - \frac{RT_o}{L}\ln\frac{p}{p_o})$, where (p_o, T_o) are a *fixed* reference temperature and pressure on the saturation curve, such as the triple point temperature and pressure. If the surface pressure is p_s, then the surface temperature is $T_s = T(p_s)$. Hence, specifying surface pressure is equivalent to specifying surface temperature in this problem. To keep the algebra simple, we'll assume a constant specific absorption κ. Then $\tau_\infty = \kappa p_s/g$, which increases as T_s is made larger. Further, for constant specific absorption, $p/p_o = (\tau_\infty - \tau')/\tau_o$ where $\tau_o = \kappa p_o/g$. Now, the choice of the reference temperature and pressure (p_o, T_o) is perfectly arbitrary, and we'll get the same answer now matter what choice we make (within the accuracy of the approximate form of Clausius-Clapeyron we are using). Hence, we are free to set $p_o = g/\kappa$ so that $\tau_o = 1$. T_o then implicitly depends on κ, and becomes larger as κ gets smaller. T_o is the temperature at the level of the atmosphere where the optical depth measured relative to the top of the atmosphere is unity.

Substituting the one-component $T(p)$ into the integral giving the solution to the Schwarzschild equation, and substituting for pressure in terms of optical thickness, we find

$$\begin{aligned} OLR &= I_+(0)e^{-\tau_\infty} + \int_0^{\tau_\infty} \sigma \frac{T_o^4}{(1 - \frac{RT_o}{L}\ln\frac{p}{p_o})^4} e^{-(\tau_\infty - \tau')} d\tau' \\ &= I_+(0)e^{-\tau_\infty} + \sigma T_o^4 \int_0^{\tau_\infty} \frac{1}{(1 - \frac{RT_o}{L}\ln\tau_1)^4} e^{-\tau_1} d\tau_1 \end{aligned} \tag{4.38}$$

where we have in the second line defined a new dummy variable $\tau_1 = \tau_\infty - \tau'$ as before. The surface temperature enters the expression for OLR only through τ_∞, which is proportional to surface pressure. In the optically thin limit, the integral on the right hand side of the expression is small (because τ_∞ is small). This happens at low surface temperatures, because p_s is small when the surface temperature is small. The OLR then reduces to the first term, which is approximately $I_+(0)$,i.e. the unmodified upward radiation from the surface. In the optically thick limit, which occurs for high surface temperatures, the term proportional to $I_+(0)$ is negligible, and the second term dominates. This term consists of the flux σT_o^4 multiplied by a non-dimensional integral. Recall that T_o is a constant dependent on the thermodynamic and infrared optical properties of the gas making up the atmosphere; it does not change with surface temperature. Because of the decaying exponential in the integrand, the integral is dominated by the contribution from the vicinity of $\tau_1 = 0$, and will therefore become independent of τ_∞ for large τ_∞ [1]. In the optically thick (high temperature) limit, then, the integral is a function of RT_o/L alone. From this we conclude that the OLR becomes independent of surface temperature in the limit of large surface temperature (and hence large τ_∞). This limiting OLR is known as the *Kombayashi-Ingersoll limit*. It was originally studied in connection with the long-term history of water on Venus, using a somewhat different argument than we have presented here. We shall use the term to refer to a limiting OLR arising from the evaporation of any volatile greenhouse gas reservoir, whether computed using a grey gas model or a more realistic radiation model.

It is readily verified that the integral multiplying σT_o^4 approaches unity as RT_o/L approaches zero. In fact, for typical atmospheric gases L/R is a very large temperature, on the order of several thousand Kelvins. Hence, unless the specific absorption is exceedingly small, RT_o/L tends to be small, typically on the order of .1 or less. For $RT_o/L = .1$, the integral has the value of .905. Thus, the limiting OLR is essentially σT_o^4. Recalling that T_o is the temperature of the moist adiabat at one optical depth unit down from the top of the atmosphere, we see that the limiting OLR behaves very nearly as if all the longwave radiation were emitted from a layer one optical depth unit from the top of the atmosphere.

Figure 4.3 shows some results from a numerical evaluation of the integral in Eq. 4.38. For small surface temperatures, there is little atmosphere, and the OLR increases like σT_s^4. As the surface temperature is made larger, the atmosphere becomes thicker and the OLR eventually asymptotes to a limiting value, as predicted. In accordance with the argument given above, the limiting OLR should be slightly less than the blackbody flux corresponding to the temperature T_o found one optical depth down from the top of the atmosphere. This temperature depends on g/κ, which is the pressure one optical depth down from the top. For $g/\kappa = 100Pa$, solving the simplified Clausius-Clapeyron relation for T at $100Pa$ yields $T_0 = 250.3K$, whence $\sigma T_o^4 = 222.6W/m^2$; for $g/\kappa = 1000Pa$, $T_o = 280.1K$ and $\sigma T_o^4 = 349.0W/m^2$. These values are consistent with the numerical results shown in the graph.

Note that for a given atmospheric composition (which determines κ) the Kombayashi-Ingersoll limit depends on the acceleration of gravity. A planet with stronger surface gravity will have a higher Kombayashi-Ingersoll limit than one with weaker gravity. An explicit formula for the dependence on g/κ is obtained by substituting for T_o using the formula for the single-component saturated adiabat, Eq. 2.27. Thus, the limiting OLR can be written in the form

$$OLR_\infty = A(L/R)\sigma T_o^4 = A(L/R)\frac{\sigma (L/R)^4}{\ln(\kappa p^*/g)^4} \tag{4.39}$$

where $A(L/R)$ is the order unity constant discussed previously and $p^* = p_{ref}\exp(L/RT_{ref})$. In

[1] Technically, the integral diverges at extremely large τ_∞, because the denominator of the integrand can vanish. This is an artifact of assuming a constant latent heat and has no physical significance.

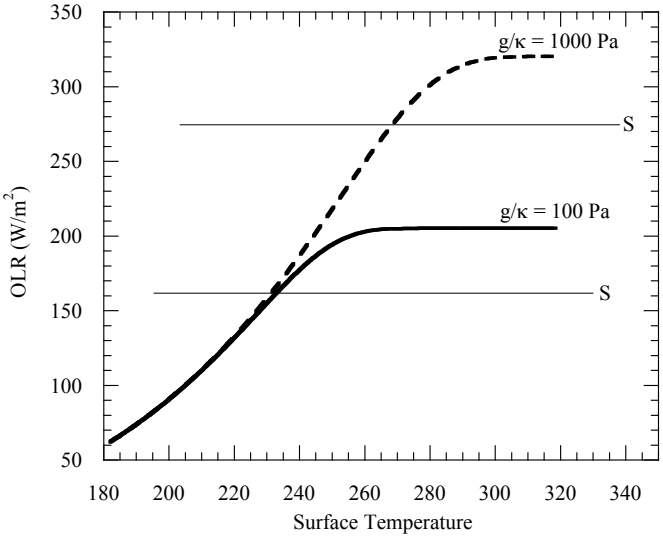

Figure 4.3: OLR vs surface temperature for a one-component grey gas condensible atmosphere in equilibrium with a reservoir. Calculations were done for thermodynamic parameters L and R corresponding to water vapor. Results are shown for two different values of g/κ, where κ is the specific cross section of the gas and g is the acceleration of gravity.

the formula for p^*, (p_{ref}, T_{ref}) is any point on the saturation vapor pressure curve, for example the triple point temperature and pressure. p^* is an enormous pressure ($2.3 \cdot 10^{11} Pa$ for water vapor), so Eq. 4.39 predicts that the Kombayashi-Ingersoll limit increases as surface gravity increases, since increasing g makes the logarithm in the denominator smaller. The apparent singularity when $\kappa p^*/g = 1$ is spurious, as the approximations we have made break down long before that value is reached.

We are now prepared to describe the runaway greenhouse phenomenon. Let S be the absorbed solar radiation per unit surface area of the planet, and let the limiting OLR computed above be OLR_{max}. If $S < OLR_{max}$ the planet will come to equilibrium in the usual way, warming up until it loses energy by infrared radiation at the same rate as it receives energy from its star. But what happens if $S > OLR_{max}$? In this case, as long as there is still an ocean or other condensed reservoir to feed mass into the atmosphere, the planet cannot get rid of all the solar energy it receives no matter how much it warms up; hence the planet continues to warm until the surface temperature becomes so large that the entire ocean has evaporated into the atmosphere. The temperature at this point depends on the mass and composition of the volatile reservoir. For example, the Earth's oceans contain enough mass to raise the surface pressure to about $100 bars$ if dumped into the atmosphere in the form of water vapor. The ocean has been exhausted when the saturation vapor pressure reaches this value. Using the simplified exponential form of the Clausius-Clapeyron relation to extrapolate the vapor pressure from the sea level boiling point (1 bar at $373.15K$), we estimate that this vapor pressure is attained at a surface temperature of about $550K$. This estimate is inaccurate, because the latent heat of vaporization varies appreciably over the range of temperatures involved. A more exact value based on measurements of properties of steam is $584K$, but the grim implications for survival or emergence of life as we know it are largely the same.

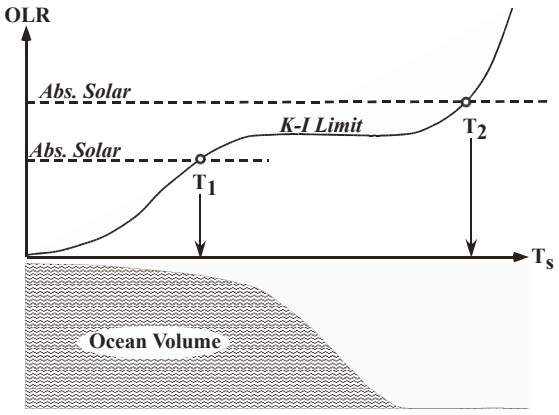

Figure 4.4: Schematic picture of the termination of a runaway greenhouse upon depletion of the volatile reservoir.

At temperatures larger than that at which the ocean is depleted, the mass of the atmosphere becomes fixed and no longer increases with temperature. The greenhouse gas content of the atmosphere – which in the present case is the entire atmosphere – no longer increases with temperature. As a result, the OLR is once more free to increase as the surface becomes warmer, and the planet will warm up until it reaches an equilibrium at a temperature warmer than that at which the ocean is depleted. The additional warming required depends on the gap between the Kombayashi-Ingersoll limit and the absorbed solar radiation. This situation is depicted schematically in Figure 4.4. Once the ocean is gone, the lower atmosphere is unsaturated and air can be lifted some distance before condensation occurs. The resulting atmospheric profile is on the dry adiabat in the lower atmosphere, transitioning to the moist adiabat at the altitude where condensation starts. The situation is identical to that depicted for CO_2 in Fig. 2.6. Rain will still form in the condensing layer. Much of it will evaporate in the lower noncondensing layer; some of it may reach the ground, but the resulting puddles would tend to rapidly evaporate back into the highly undersaturated lower atmosphere. As surface temperature is made larger, the altitude where condensation sets in moves higher, until at very large temperatures the atmosphere behaves like a noncondensing dry system (albeit one where the entire atmosphere may consist of water vapor).

The runaway greenhouse phenomenon may explain how Venus wound up with such a radically different climate from Earth, despite having started out in a rather similar state. The standard story goes something like this: Venus started with an ocean, and with most of it's CO_2 bound up in rocks as is the case for Earth. However, it was just enough closer to the Sun to trigger a runaway greenhouse. Once the entire ocean had evaporated into the atmosphere, there was so much water vapor in the upper atmosphere that it could be broken apart by energetic solar ultraviolet rays, whereafter the light hydrogen could escape to space. The highly reactive oxygen left behind would react to form minerals at the surface. Once there was no more liquid water in play, the reactions that bind up carbon dioxide in rocks could no longer take place (as will be explained in Chapter 8), so all the planet's CO_2 outgassed from volcanism and stayed in the atmosphere, leading to the hot, dry super-dense atmosphere of modern Venus.

Assuming habitability to require a reservoir of liquid water, the Kombayashi-Ingersoll limit for water determines the inner orbital limit for habitability, since if the Solar constant exceeds the

limiting flux a runaway will ensue and any initial ocean will not persist. It also determines how long it takes before the planet's Sun gets bright enough to trigger a runaway, and thus sets the lifetime of a water-dependent biosphere (Earth's included). Accurate calculations of the Kombayashi-Ingersoll limit are therefore of critical importance to understanding the limits of habitability both in time and orbital position. The grey gas model is not good enough to determine the value of p_o appropriate to a given gas, and so cannot be used for accurate evaluations of the runaway greenhouse threshold. We can at least say that, all other things being equal, a planet with larger surface gravity will be less susceptible to the runaway greenhouse. This is so because $p_o = g/\kappa$, whence larger g implies larger p_o, which implies in turn larger $T(p_o)$ and hence a larger limiting OLR. This observation may be relevant to the class of extrasolar planets known as "Large Earths."

We will revisit the runaway greenhouse using more realistic radiation physics in Section 4.6. Some effects of the effects of the stratosphere and of clouds will be brought into the picture in Chapter 5.

The runaway greenhouse phenomenon is usually thought of in conjunction with water vapor, but the concept applies equally well to any situation where there is a volatile reservoir of greenhouse gas, whether it be in solid or liquid form. For example, one could have a runaway greenhouse in association with the sublimation of a large CO_2 ice cap, or in association with the evaporation of a Methane or Ammonia ocean. In fact, the Kombayashi-Ingersoll limit determines whether a planet would develop a reservoir of condensed substance at its surface (a glacier or ocean), given sustained outgassing of that substance in the absence of any chemical sink. As the gas builds up in the atmosphere, the pressure increases and it would eventually tend to condense at the surface, preventig any further gaseous accumulation. However, the greenhouse effect of the gas warms the surface,which increases the saturation vapor pressure. The Kombayashi-Ingersoll limit tells us which effect wins out as surface pressure increases. Earth is below the threshold for water, so we have a water ocean. Venus is above the threshold for water and CO_2, so both accumulate as gases in the atmosphere (apart from possible escape to space). When we revisit the problem with real-gas radiation, we will be able to say whether CO_2 would form a condensed reservoir on Earth or Mars, or CH_4 on Titan, given sustained outgassing in the absence of a chemical sink.

4.3.4 Pure radiative equilibrium for a grey gas atmosphere

For the temperature profiles discussed in Sections 4.3.2 and 4.3.3, the net infrared radiative heating computed from Eq. 4.14 is nonzero at virtually all altitudes; generally the imbalance acts to cool the lower atmosphere and warm the upper atmosphere. In using such solutions to compute OLR and back-radiation, we are presuming that convective heat fluxes will balance the cooling and keep the troposphere in a steady state. The upper atmosphere will continue to heat, and ultimately reach equilibrium creating a stratosphere, but in the all-troposphere idealization we presume that the stratosphere is optically thin enough that it doesn't much affect the OLR.

Now, we'll investigate solutions for which, in contrast, the net radiative heating vanishes individually at each altitude. Such solutions are in *pure radiative equilibrium*, as apposed to *radiative-convective equilibrium*. First we'll consider the case in which the only radiative heating is supplied by infrared; later we'll bring heating by atmospheric solar absorption into the picture. Pure radiative equilibrium is the opposite extreme from the all-troposphere idealization, and tells us much about the nature of the stratosphere, and the factors governing the tropopause height.

Assuming the atmosphere to be transparent to solar radiation, pure radiative equilibrium requires that the frequency-integrated longwave radiative heating $\mathfrak{H}$ vanish for all τ. From the grey gas version of Eq. 4.14, we then conclude that $I_+ - I_-$ is independent of τ. Applying the upper

boundary condition, we find that this constant is $I_+(\tau_\infty)$, which is the OLR. Now, by taking the difference between the equations for I_+ and I_- we find

$$0 = \frac{d}{d\tau}(I_+ - I_-) = -(I_+ + I_-) + 2\sigma T^4 \tag{4.40}$$

which gives us the temperature in terms of $(I_+ + I_-)$. Next, taking the sum of the equations for I_+ and I_- yields

$$\frac{d}{d\tau}(I_+ + I_-) = -(I_+ - I_-) \tag{4.41}$$

This is easily solved by noting that $-(I_+ - I_-) = const = -OLR$. In consequence,

$$2\sigma T^4 = (I_+ + I_-) = (1 + \tau_\infty - \tau)OLR \tag{4.42}$$

where we have again used the boundary condition at τ_∞. This expression gives us the pure radiative equilibrium temperature profile $T(\tau)$. In pure radiative equilibrium, the temperature always approaches the skin temperature at the top of the atmosphere, where $\tau = \tau_\infty$. This recovers the result obtained in the previous chapter, in Section 3.6. When the atmosphere is optically thin, $\tau_\infty - \tau$ is small throughout the atmosphere, and the entire atmosphere becomes isothermal with temperature equal to the skin temperature. When the atmosphere is not optically thin, the temperature decreases gently with height, approaching the skin temperature as the top of the atmosphere is approached.

Eq. 4.42 also gives us the upward and downward fluxes, since we now know both $I_+ - I_-$ and $I_+ + I_-$ at each τ. In particular, the downward flux into the ground is

$$I_-(0) = \frac{1}{2}((I_+ + I_-) - (I_+ - I_-)) = \frac{1}{2}((1 + \tau_\infty)OLR - OLR) = \frac{1}{2}\tau_\infty OLR \tag{4.43}$$

For an optically thin atmosphere, the longwave radiation returned to the ground by the atmosphere is only a small fraction of that emitted to space. As the atmosphere becomes optically thick, the radiation returned to the ground becomes much greater than that emitted to space, because the radiative equilibrium temperature near the ground becomes large and the optical thickness implies that the radiation into the ground is determined primarily by the low level temperature. If we assume the planet to be in radiative equilibrium with the absorbed solar radiation $(1-\alpha)S$, where α is the albedo of the ground, then $OLR = (1-\alpha)S$ and the radiative energy budget of the ground is

$$\sigma T_s^4 = (1 - \alpha)S + I_-(0) = (1 - \alpha)S \cdot (1 + \frac{1}{2}\tau_\infty) \tag{4.44}$$

where T_s is the surface temperature. This, together with the temperature profile determined by Eq. 4.42, determines what the thermal state of the system would be in the absence of heat transport mechanisms other than radiation. For an optically thin atmosphere, the surface temperature is only slightly greater than the no-atmosphere value. As τ_∞ becomes large, the surface temperature increases without bound. Note that, while this formula yields a greenhouse warming of the surface, the relation between surface temperature and τ_∞ is different from that given by the all-troposphere radiative convective calculation in Eq. 4.35, because the pure radiative equilibrium temperature profile is different from the adiabat which would be established by convection.

Let's now compare the surface temperature with the temperature of the air in immediate contact with the surface. From Eq. 4.42 we find that the low level air temperature is determined by $\sigma T(0)^4 = (1 - \alpha)S \cdot (\frac{1}{2} + \frac{1}{2}\tau_\infty)$ Taking the ratio,

$$\frac{T(0)}{T_s} = (\frac{\frac{1}{2} + \frac{1}{2}\tau_\infty}{1 + \frac{1}{2}\tau_\infty})^{1/4} \tag{4.45}$$

Thus, the surface is always warmer than the overlying air in immediate contact with it. In the previous chapter, we saw that this was the case for pure radiative equilibrium in an optically thin atmosphere, but now we have generalized it ot arbitrary optical thickness. In the optically thin limit, the formula reduces to our earlier result, $T(0) = 2^{-1/4}T_s$. In the optically thick limit, $T_s - T(0) = \frac{1}{4}T_s/\tau_\infty$, whence the temperature jump (relative to surface temperature) falls to zero as the atmosphere is made more optically thick. As we already discussed in Section 3.6, cold air immediately above a warmer surface constitutes a very unstable situation. Under the action of diffusive or turbulent heat transfer between the surface and the nearby air, a layer of air near the surface will heat up to the temperature of the surface, whereafter it will be warmer than the air above it. Being buoyant, it will rise and lead to convection, which will stir up some depth of the atmosphere and establish an adiabat – creating a troposphere.

In pure radiative equilibrium, the surface heating inevitably gives rise to convection. However, it is also possible that the temperature profile in the interior of the atmosphere may become unstable to convection, even without the benefit of a surface. This is a particularly important possibility to consider for gas giant planets, which have no distinct surface to absorb solar radiation and stimulate convection. To determine stability, we must compute the lapse rate dT/dp in radiative equilibrium, and see if it is steeper than that of the adiabat (moist or dry) appropriate to the atmosphere. Taking the derivative of 4.42 with respect to optical thickness, we find

$$8\sigma T^3 \frac{dT}{d\tau} = -OLR \tag{4.46}$$

whence, using $d/dp = (d\tau/dp)(d/d\tau)$, we find

$$\frac{d\ln T}{d\ln p} = -\frac{1}{4}\frac{1}{(1+\tau_\infty-\tau)}p\frac{d\tau}{dp} \tag{4.47}$$

Stability is determined by comparing the slope of the adiabat to the radiative-equilibrium slope we have just computed. For the dry adiabat, the atmosphere is stable where

$$\frac{R}{c_p} \geq -\frac{1}{4}p\frac{d\tau}{dp}\frac{1}{(1+\tau_\infty-\tau)} \tag{4.48}$$

The factor p appearing on the right hand side of this equation guarantees that the upper portion of the atmosphere will always be stable, unless $d\tau/dp$ blows up like $1/p$ or faster as $p \to 0$. Moreover, optically thin atmospheres are always stable throughout the depth of the atmosphere. This is so because the denominator is close to unity in the optically thin limit, while $-pd\tau/dp = \kappa p/(g\cos\bar{theta}) < \tau_\infty \ll 1$. Since optically thin atmospheres are nearly isothermal in pure radiative equilibrium, it is hardly surprising that they are statically stable.

In the case of constant absorption coefficient κ, we have $pd\tau/dp = -\kappa p/g\cos\bar{\theta}$, which is just $\tau - \tau_\infty$. Thus, the stability condition becomes

$$\frac{R}{c_p} \geq \frac{1}{4}\frac{\tau_\infty-\tau}{(1+\tau_\infty-\tau)} \tag{4.49}$$

The right hand side has its maximum at the ground $\tau = 0$, and the maximum value is $\frac{1}{4}\tau_\infty/(1+\tau_\infty)$. The more optically thick the atmosphere is, the more unstable it is near the ground. For large optical thickness, the stability criterion becomes the remarkably simple statement $4R/c_p \geq 1$. Dry Earth air, with $R/c_p = 2/7$, just misses being unstable by this criterion, and pure noncondensing water vapor is almost precisely on the boundary. Pure noncondensing CO_2, NH_3 and CH_4 just barely satisfy the condition for instability near the ground when the atmosphere is optically thick.

Typical atmospheres, however, will be more unstable than the constant κ calculation suggests. As will be explained in Section 4.4 ,*collisional broadening* typically causes the absorption coefficient to increase linearly with pressure, out to pressures of several bars. With collisional broadening, $\kappa(p) = \kappa(p_s) \cdot (p/p_s)$ for a combination of well-mixed greenhouse gases with pressure-independent concentrations. In this case $-pd\tau/dp = 2\tau_\infty$ at the surface, so that the dry adiabatic stability condition in the optically thick limit becomes $2R/c_p \geq 1$. The extra factor of 2 compared to the case without pressure broadening destabilizes all well-mixed atmospheres, provided they are sufficiently optically thick near the ground. The maximum value of the stability parameter occurs for mono-atomic gases like Helium, which has $2R/c_p = .8$ and easily meets the criterion for instability.

Other processes can destabilize the atmosphere as well. For example, on the moist adiabat the slope $-d\ln T/d\ln p$ is always less than the dry adiabatic slope, which deepens the layer within which the radiative-equilibrium atmosphere is unstable. In addition, a sharp decrease of optical thickness with height tends to destabilize the atmosphere, particularly if it occurs in a place where the atmosphere is optically thick. This happens whenever there is a layer where the concentration of absorbers decreases strongly with height. Because of Clausius-Clapeyron, this situation often happens in the lower portions of atmospheres in equilibrium with a reservoir of condensible greenhouse gas (water vapor on Earth or Methane on Titan, for example). In this case, the condensible substance has a destabilizing effect through its influence on the radiative equilibrium, as well as through its effect on the adiabat. Condensed water is a good infrared absorber, so radiative equilibrium drives cloud tops to be unstable. In contrast, the atmosphere is stabilized where the infrared absorber concentration increases strongly with height; this situation is less typical, but can happen at the bottom of a water cloud.

4.3.5 Effect of atmospheric solar absorption on pure radiative equilibrium

Now we will examine how the absorption of solar radiation within an atmosphere affects the temperature structure of the atmosphere in radiative equilibrium. The prime application of this calculation is to understand the thermal structure of stratospheres. Under what circumstances does the temperature of a stratosphere increase with height? The effect of solar absorption on gas giant planets like Jupiter is even more crucial. There being no distinct surface to absorb sunlight, *all* solar driving of the atmosphere for gas giants comes from deposition of solar energy within the atmosphere. In this case, the profile of absorption determines in large measure where, if anywhere, the radiative equilibrium atmosphere is unstable to convection, and therefore where a troposphere will tend to form. The answer determines whether convection on gas giants is driven in part by solar heating as opposed to ascent of buoyant plumes carrying heat from deep in the interior of the planet.

In the Earth's stratosphere, solar absorption is largely due to the absorption of ultraviolet by ozone. On Earth as well as other planets having appreciable water in their atmospheres, absorption of solar near-infrared by water vapor and water clouds is important. CO_2 also has significant near-infrared absorption, which is relatively unimportant at present-day CO_2 concentrations on Earth, but becomes significant on the Early Earth when CO_2 concentrations were much higher; solar near-infrared absorption by CO_2 is of course important in the CO_2 dominated atmospheres of Mars (present and past) and Venus. Solar absorption by dust is important to the Martian thermal structure throughout the depth of the atmosphere. On Titan, it is solar absorption by organic haze clouds that control the thermal structure of the upper atmosphere. Solar absorption is also crucial to the understanding of the influence of greenhouse gases like CH_4 and SO_2, which strongly absorb

sunlight in addition to being radiatively active in the thermal infrared. Strong solar absorption also would occur in the high-altitude dust and soot cloud that would be lofted in the wake of a global thermonuclear war or asteroid impact (the "Nuclear Winter" problem).

Since the Schwartzschild equations in this chapter are used to describe the infrared flux alone, the addition of solar heating does not change these equations. The heating due to solar absorption only alters the condition for local equilibrium, which now involves the deposition of solar as well as infrared flux. We write the solar heating rate per unit optical depth in the form $Q_\circledast = dF_\circledast/d\tau$, where $F_\circledast$ is the net downward solar flux as a function of infrared optical depth. At the top of the atmosphere, $F_\circledast = (1-\alpha)S$, where α is the planetary albedo – that is, the albedo measured at the top of the atmosphere. Since atmospheres at typical planetary temperatures do not emit significantly in the solar spectrum, there is no internal source of solar flux and therefore $F_\circledast$ must decrease monotonically going from the top of the atmosphere to the ground.

The net radiative heating at a given position is now the sum of the infrared and solar term, i.e.

$$-\frac{d}{d\tau}(I_+ - I_-) + \frac{d}{d\tau}F_\circledast = 0 \tag{4.50}$$

Integrating this equation and requiring that the top of atmosphere energy budget be in balance with the local absorbed solar radiation, we find

$$(I_+ - I_-) - F_\circledast = 0 \tag{4.51}$$

At the top of the atmosphere, this reduces to $OLR - (1-\alpha)S = 0$, which is the requirement for top of atmosphere energy balance. Because the solar absorption does not change the infrared Schwartzschild equations, Eq. 4.41 is unchanged from the case of pure radiative equiibrium without solar absorption. Substituting Eq 4.51 and integrating, we obtain

$$I_+ + I_- = \int_\tau^{\tau_\infty} F_\circledast(\tau')d\tau' + (1-\alpha)S \tag{4.52}$$

In writing this expression we have made use of the boundary condition $I_+ - I_- = OLR = (1-\alpha)S$ at the top of the atmosphere. The heat balance equation 4.40 needs to be slightly modified, since the infrared cooling now balances the solar heating, instead of being set to zero. Thus,

$$\frac{d}{d\tau}F_\circledast = \frac{d}{d\tau}(I_+ - I_-) = -(I_+ + I_-) + 2\sigma T^4 \tag{4.53}$$

from which we infer

$$2\sigma T^4 = \frac{d}{d\tau}F_\circledast + \int_\tau^{\tau_\infty} F_\circledast(\tau')d\tau' + (1-\alpha)S \tag{4.54}$$

This gives the vertical profile of temperature in terms of the vertical profile of the solar flux; the previous case (without solar absorption) can be recovered by setting $F_\circledast = const = (1-\alpha)S$. At the top of the atmosphere, the integral in Eq. 4.53 vanishes, and the temperature becomes identical to the temperature of a skin layer heated by solar absorption, derived in Chapter 3 (Eq. 3.27).

Taking the derivative with respect to τ yields

$$\frac{d}{d\tau}2\sigma T^4 = \frac{d^2}{d\tau^2}F_\circledast - F_\circledast \tag{4.55}$$

This equation provides a simple criterion determining when the solar absorption causes the temperature to increase with height. When there is no absorption, $F_\circledast$ is a constant and since it

is positive the temperature decreases with height. The quantity $\sqrt{|F_{\circledast}^{-1} d^2 F_{\circledast}/d\tau^2|}$ is local solar flux decay rate, expressed in units of infrared optical depth. Where the local solar decay rate is less than unity – meaning solar flux is attenuated at a lower rate than infrared – the radiative equilibrium temperature decreases with height. Where the local solar decay rate is greater than unity, the temperature increases with height. Note that it is the solar extinction rate *relative to the infrared extinction rate* that counts. One can make the temperature increase with height either by increasing the solar opacity or decreasing the infrared opacity (generally by decreasing the greenhouse gas concentration). The profile of solar absorption is also sensitive to the vertical distribution of solar absorbers. Where solar absorbers increases sharply with height, as is the case for ozone on Earth or organic haze on Titan, the stratospheric temperature increases with height.

By way of illustration, let's suppose that the net downward solar flux decays exponentially as it penetrates the atmosphere. Specifically, let $F_{\circledast} = (1-\alpha)S\exp(-(\tau_\infty - \tau)/\tau_S)$, where τ_S is a constant. τ_S is the decay rate of solar radiation, measured in infrared optical depth units. When τ_S is large, solar absorption is weak compared to infrared absorption, and one must go a great distance before the solar beam is appreciably attenuated. Conversely, when τ_S is small, solar absorption is strong and the solar beam decays to zero over a distance so short that infrared is hardly attenuated at all. With the assumed form of the solar flux, the temperature profile is given by

$$2\frac{\sigma T^4}{(1-\alpha)S} = 1 + \tau_S + (\frac{1}{\tau_S} - \tau_S)e^{-(\tau_\infty - \tau)/\tau_S} \tag{4.56}$$

If $\tau_S > 1$ the temperature decreases with height, and if $\tau_S < 1$ the temperature increases with height. Defining the skin temperature as $T_{skin} \equiv (\frac{1}{2\sigma}(1-\alpha)S))^{1/4}$ the temperature at the top of the atmosphere is $(1+1/\tau_S)T_{skin}$, which reduces to the skin temperature when τ_S is large and becomes much greater than the skin temperature when τ_S is small. If the atmosphere is deep enough that essentially all solar radiation is absorbed before reaching the ground, then the exponential term vanishes in the deep atmosphere and the deep atmosphere becomes isothermal with temperature $(1+\tau_S)T_{skin}$. Thus, when τ_S is small, all the solar radiation is absorbed within a thin layer near the top of the atmosphere. The temperature increases rapidly with height in this layer, but the bulk of the atmosphere below is approximately isothermal at the skin temperature. The strong solar absorption causes the deep atmosphere, and the ground (if there is one) , to be colder than it would have been in the absence of an atmosphere. This *anti-greenhouse effect* arises because the deep atmosphere is heated only by downwelling infrared emitted by the solar-absorbing layer. This downward radiation equals the upward radiation loss to space, which must equal $(1-\alpha)S$ to satisfy the top of atmosphere balance. The deep atmosphere falls to the skin temperature because it is being illuminated by this flux from one side, but is radiating from both its top and bottom. This limit is relevant to the *nuclear winter* phenomenon, in which energetic explosions and fires loft a global or hemispheric solar-absorbing soot and dust cloud to high altitudes. The same situation would occur in the aftermath of a large bolide (asteroid or comet) impact. In either case, the atmosphere below the soot layer would become as frigid as the depth of winter, but moreover the relaxation to a uniform temperature state would shut off the convection which in large measure drives the hydrological cycle.

It can happen that the atmosphere is deep enough to absorb all solar radiation before it reaches the ground, even if the rate of solar absorption is weak and $\tau_S >> 1$. This would happen if the atmosphere is so optically thick in the infrared that $\tau_/infty/\tau_S >> 1$ despite τ_S being large. In this case, the deep atmosphere is still isothermal, but it becomes much hotter than the skin temperature – indeed it becomes hotter without bound as τ_S is made larger. In this case, it is the top of the atmosphere which equilibrates at the relatively cold skin temperature, while the deep atmosphere exhibits a strong greenhouse effect. Because the deep atmosphere is isothermal, it is

stable and will not generate a troposphere. This is a possible model for the internal state of a gas or ice giant with little internal heat flux, whose atmosphere is optically thick in the infrared but for which there is only weak solar absorption in the deep atmosphere.

The solution given in Eq. 4.56 shows that one can account for the temperature increase in the Earth's stratosphere if the upper atmosphere strongly absorbs solar radiation. As a model of the Earth's atmosphere, however, it has the shortcoming that if one makes the stratospheric absorption strong enough to yield a temperature increase with height, essentially all the solar beam is depleted in the stratosphere, leaving an isothermal lower atmosphere that won't convect and create a troposphere. What happens in Earth's real stratosphere is that ozone is a good absorber only in the ultraviolet part of the Solar spectrum. Once the ultraviolet is depleted, the remaining flux making its way into the lower atmosphere is only weakly absorbed by the atmosphere. In terms of Eq. 4.55, the solar decay rate $\sqrt{|F_{\circledast}^{-1} d^2 F_{\circledast}/d\tau^2|}$ is large in the upper atmosphere, where there is still plenty of ultraviolet to absorb; in consequence, the temperature increases with height there. In the lower atmosphere the solar decay rate becomes small, so that the radiative equilibrium temperature decreases with height. There is also plenty of solar radiation left to heat the ground, destabilize the atmosphere, and create a troposphere.

The solution in Eq. 4.56 also explains why human-caused increases in CO_2 over the past century have led to tropospheric warming but stratospheric cooling, as illustrated in Figure 1.17. Increasing the greenhouse gas concentration is equivalent to increasing τ_∞. If one plots temperature as a function of pressure for a sequence of increasing τ_∞, the phenomenon is immediately apparent in cases where the upper level solar absorption is sufficiently strong. The behavior is explored in Problems **??**. Without solar absorption, increasing τ_∞ warms the atmosphere at every level, though the amount of warming decreases with height as the temperature asymptotes to the skin temperature. With solar absorption, however, the increased infrared cooling of the upper atmosphere offsets more and more of the warming due to solar absorption, leading to a cooling there. In the real atmosphere convection modifies the temperature profile in the lower atmosphere. Further, one must take into account real gas infrared and solar absorption in order to quantitatively account for the observed temperature trends. Nonetheless, the grey-gas pure radiative equilibrium calculation captures the essence of the mechanism.

The general lesson to take away from this discussion is that solar absorption near the top of the atmosphere stabilizes the atmosphere, reduces the greenhouse effect, and cools the lower portion of the atmosphere and also the ground. This is important in limiting the effectiveness of greenhouse gases like CH_4 and SO_2, which significantly absorb solar radiation when their concentration becomes very high. It is also the way high altitude solar-absorbing haze clouds on Titan and perhaps Early Earth act to cool the troposphere. The soot and dust clouds lofted by an asteroid impact act similarly. In contrast, solar absorption concentrated near the ground has an effect which is not much different from simply reducing the albedo of the ground itself.

4.4 Real gas radiation: Basic principles

4.4.1 Overview: OLR through thick and thin

It would be exceedingly bad news for planetary habitability if real greenhouse gases were grey gases (see Exercise 4.3.1). Greenhouse gas concentrations would have to be tuned exceedingly accurately to maintain a planet in a habitable temperature range, and there would be little margin for error. Thus, it is of central importance that, for real gases, OLR varies much more gradually

with greenhouse gas concentration than it would for an idealized grey gas [2]. This is another area in which the quantum nature of the Universe directly intervenes in macroscopic phenomena governing planetary climate.

Infrared Radiative transfer is a very deep and complex subject, and mastery of the material in this section will still not leave the reader prepared to write state-of-the-art radiation codes. Nor will we cover the myriad engineering tricks large and small which are needed to make a radiation code fast enough to embed in a general circulation model, where it will need to be invoked a dozen times per model day at each of several thousand grid points. We do aspire to provide enough of the basic physics to allow the student to understand why OLR is less sensitive to the concentration of a typical real gas than to a grey gas, and to help the student develop some intuition about the full possible range of behaviors of greenhouse gases on Earth and other planets, now and in the distant past or future. Such an understanding should extend even to greenhouse gases that are not at present commonly considered in the context of climate, or implemented in standard "off the shelf" radiation models. What would you do, for example, if you found yourself wondering whether SO_2 or H_2S significantly affected the climate of Early Earth or Mars? The grey gas model does not provide an adequate first attack on such problems. We thus aspire to provide enough of the basic algorithmic equipment to allow the student to build simplified radiative models from scratch, that get the OLR and infrared heating profiles roughly correct.

Even though we will have recourse to a "professionally" written radiation code in Section 4.5, we'd like to at least draw back the curtain a little bit, so that the reader will not be left with the all-too-common notion that radiation routines are black boxes, the internal workings of which can only be understood by the high priesthood of radiative transfer. Hopefully, this will also open the door to entice more people into innovative work on the subject.

Since the main point is to understand how the wavenumber dependence of absorptivity affects the sensitivity of OLR to greenhouse gas concentration, we'll begin with a discussion of the spectrum of outgoing longwave radiation in an idealized case. Let's consider a planet whose surface radiates like an ideal blackbody in the infrared, having an atmosphere whose air temperature at the surface T_{sa} is equal to the ground temperature T_g. The temperature $T(p)$ is monotonically decreasing with height in the troposphere, and is patched continuously to an isothermal stratosphere having temperature T_{strat}. The atmosphere consists mostly of infrared-transparent N_2 and O_2 with a surface pressure of $10^5 Pa$, like Earth. Unlike Earth, the only greenhouse gas is a mythical substance (call it Oobleck), which is a bit like CO_2, but much simpler to think about. It has the same molecular weight as CO_2, but it's absorption coefficient $\kappa_{Oob}(\nu)$ has an absorption band centered on wavenumber $\nu_o = 600 cm^{-1}$. Within $100 cm^{-1}$ of ν_o, κ_{Oob} has the constant value κ_o. Outside of this limited range of wavenumbers, Oobleck is transparent to infrared, i.e. $\kappa_{Oob} = 0$. To make life even simpler for the atmospheric physicists of this planet, κ_{Oob} is independent of both temperature and pressure. Like real CO_2, the specific concentration of Oobleck (q_{Oob}) is constant throughout the depth of the atmosphere.

What does the spectrum of OLR look like for this planet? The answer is shown in the left panel of Figure 4.5. In this figure, we have assumed that the Oobleck molecule has an absorptivity of $1 m^2/kg$. Then, with a molar concentration of $300 ppmv$ (like CO_2 in the 1960's), the specific concentration is $4.6 \cdot 10^{-4}$ and the optical thickness $\kappa_o q_{Oob} p_s / g \cos\bar{\theta}$ is 9.4 within the absorption band. Since the atmosphere is optically thick in this wavenumber region, infrared radiation in this part of the spectrum exits the atmosphere with the temperature of the stratosphere. This is exactly what we see in the graph. Outside the absorption band, the atmosphere is transparent,

[2] Lest there be any misunderstanding, we must emphasize at this point that "less sensitive" does not mean "insensitive." If CO_2 were a grey gas, then doubling it's concentration, as we are poised to do within the century, would be unquestionably lethal. Because CO_2 is not in fact a grey gas, the results may be merely catastrophic.

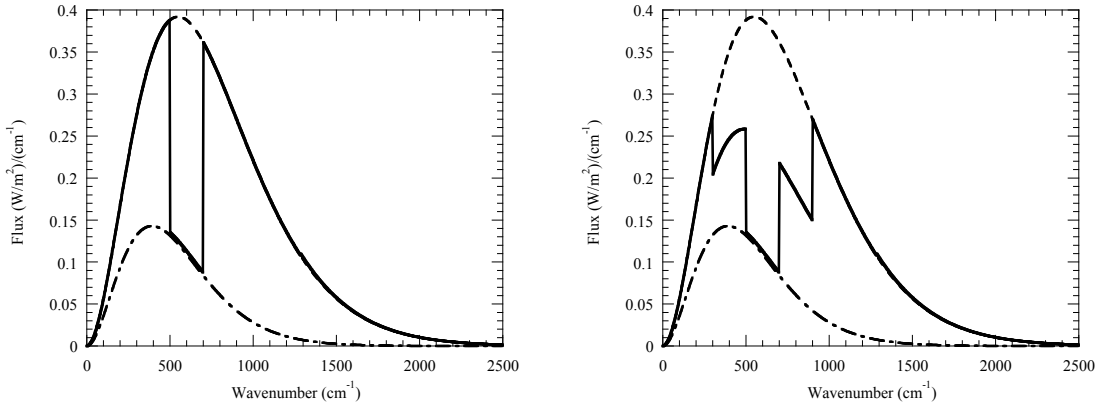

Figure 4.5: The OLR spectrum for a hypothetical gas which has a piecewise constant absorption coefficient. The dashed-dotted lower curve is the blackbody spectrum corresponding to the stratospheric temperature T_{strat} while the dashed upper curve is the blackbody spectrum corresponding to the surface temperature T_s. The calculation was carried out for $T_s = 280K$ and $T_{strat} = 200K$, and with a greenhouse gas concentration sufficient to make the optical thickness ≈ 10 in the central absorption band. Left panel: The gas has an absorption coefficient of $1m^2/kg$ within a single absorption band extending from 500 to 700 cm^{-1}. Right Panel: The gas has additional weak absorption bands from 300 to 500 cm^{-1} and 700 to 900 cm^{-1}, within which the absorption coefficient is $.125m^2/kg$.

and hence infrared leaves the top of the atmosphere at the much higher temperature of the ground. The overall appearance of the OLR spectrum is that the greenhouse gas has "dug a ditch" in the spectrum of OLR, or perhaps "taken a bite" out of it. The ditch in the spectrum reduces the total OLR of the planet, but not so much so as if the absorption were strong throughout the spectrum, as would be the case for a grey gas. This is the typical way that real greenhouse gases work: they make the atmosphere optically thick in a limited part of the spectrum, while leaving it fairly transparent elsewhere. The strength of the greenhouse effect is not so much a matter of how deep the ditch, but how wide.

Oobleck is a very contrived substance, but the above exercise gives us a fair idea of what to look for when interpreting real observations of the spectrum of OLR. Figure 4.6, giving the OLR spectrum of Mars observed at two times of day by the TES instrument on Global Surveyor, is a case in point. Mars has an essentially pure CO_2 atmosphere complicated only be optically thin ice clouds and dust clouds (which can be very thin between major dust storms). The planet thus provides perhaps the purest illustration of the CO_2 radiative effect available in the Solar system. In Figure 4.6 a CO_2 "ditch" centered on about $650cm^{-1}$ is evident both in the afternoon and sunset spectra. At the trough of this ditch, the radiation exits the atmosphere with a radiating temperature of about $170K$ both in the afternoon and sunset cases. This temperature is similar to the coldest temperature encountered in the upper atmosphere of Mars in the Summer (see Fig 2.2), and is compatible with the strong decrease of temperature with height seen in the soundings. Away from CO_2 ditch, the atmosphere appears transparent, and the emission resembles the blackbody emission from a land surface having temperature $265K$ in the afternoon case and $212K$ in the sunset case. These numbers are compatible with the observed range of ground temperature on Mars, cross-checked by near-surface data from landers.

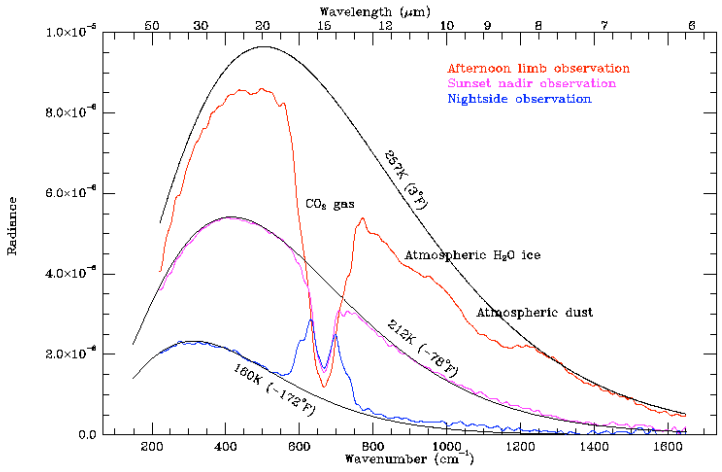

Figure 4.6: Some representative OLR spectra for Mars, observed by the Thermal Emission Spectrometer on Mars Global Surveyor at various times of day.

In a situation like that shown in the left panel of Figure 4.5, the OLR is as low as it is going to get, provided the stratospheric temperature is held fixed. Increasing the greenhouse gas concentration q_G cannot lower the OLR further since, in the spectral region where the gas is radiatively active the atmosphere is already radiating at the coldest available temperature [3]. Suppose, however, that instead of the gas being transparent outside the central absorption band, there is a set of weaker absorption bands waiting in the wings on either side of the primary band – a gas we may call "Two-Band Oobleck." In this case, illustrated in the right panel of Figure 4.5, the effect of the weaker bands on OLR is not yet saturated, and increases in q_G will cause the OLR to go down until these bands, too, are saturated. But what if there are yet-weaker absorption bands waiting a bit farther out? Then further increases of q_G will yield additional decreases in the OLR. One can imagine making the process continuous by making the width of the bands smaller, and the jump in absorption coefficients between adjacent bands smaller. Real greenhouse gases act very much like this, as they almost invariably have absorption whose overall strength decays strongly with distance in wavenumber from a central peak. The rate at which the absorption decays with distance determines the rate at which the OLR decreases as the greenhouse gas concentration is made larger.

From Eq. 4.11, 4.12 or 4.13, if we know the transmission function, we can carry out the integral needed to obtain the radiative fluxes. As we shall see shortly, in most cases the dependence of κ on wavenumber is so intricate that solving the problem by doing a brute-force integral over wavenumber is prohibitive if one aims to do the calculation enough times to gain some insight from modelling a climate (even in a single dimension). In any event, doing the calculation with enough spectral resolution to directly resolve all the wiggles in $\kappa(\nu)$ provides much more information about

[3] This example is somewhat contrived, since increasing the concentration of a greenhouse gas generally cools the stratosphere. However, it serves to illustrate the way additional weak absorption bands influence the OLR.The additional OLR reduction from cooling of the stratosphere as q_G increases is a secondary effect. Since the temperature there is already so low, it wouldn't throw off the result very much to simply replace the OLR at the depths of the ditch to zero.

spectral variability than is needed in most cases. What we really want is to understand something about the properties of the transmission function averaged over a finite sized spectral region of width Δ, centered on a given frequency ν. Specifically, let's choose Δ to be small enough that the Planck function B and its derivative dB/dT are both approximately constant over the spectral interval of width Δ. In that case, when the solution for the flux given in Eq. 4.12 or its alternate forms is averaged over Δ, B can be treated as nearly independent of ν and taken outside the average. In consequence, the resulting band-averaged equations have precisely the same form as the original ones, save that the fluxes are replaced by average fluxes like

$$\bar{I}_+(\nu, p) = \frac{1}{\Delta}\int_{\nu-\Delta/2}^{\nu+\Delta/2} I_+(\nu', p)d\nu' \tag{4.57}$$

and the transmission function is replaced by

$$\bar{\mathfrak{T}}_\nu(p, p') = \frac{1}{\Delta}\int_{\nu-\Delta/2}^{\nu+\Delta/2} \mathfrak{T}_{\nu'}(p, p')d\nu' \tag{4.58}$$

We need to learn how to derive properties of $\bar{\mathfrak{T}}_\nu(p, p')$. The essential challenge is that the nonlinear exponential function stands between the statistics of κ_ν and the statistics of $\mathfrak{T}_\nu$.

The transmission function satisfies the *multiplicative property*, that

$$\mathfrak{T}_\nu(p_1, p_2) = \mathfrak{T}_\nu(p_1, p')\mathfrak{T}_\nu(p', p_2) \tag{4.59}$$

if p' is between p_1 and p_2. The multiplicative property means that the transmission along a path through the atmosphere can be obtained by taking the product of the transmissions along any number of constitutent parts of the path. The band-average transmission loses this valuable property, because for two general functions f and g, $\int f(\nu)g(\nu)d\nu \neq (\int f(\nu)d\nu)(\int g(\nu)d\nu)$. The equality holds only if the two functions are uncorrelated, which is not generally the case for the transmission in two successive parts of a path. In the first part of the path, the strongly absorbed frequencies are used up first, and are no longer available for absorption in the second part of the path. The system has memory, and one can think of the light as becoming "tired," or depleted more and more in the easily absorbed frequencies the longer it travels, with the result that the absorption in the latter parts of the path are weaker than they would be if fresh light were being absorbed.

4.4.2 The absorption spectrum of real gases

We will now take a close look at the absorption properties of CO_2, in order to introduce some general ideas about the nature of the absorption of infrared radiation by molecules in a gas. Continuing to use CO_2 as an example, these ideas will be developed in Sections 4.4.3,4.4.4 and 4.4.6 into a computationally efficient means of calculating infrared fluxes in a real-gas atmosphere. A survey of the spectral characteristics of selected other greenhouse gases will be given in Sections 4.4.7 and 4.4.8.

Figure 4.7 shows the absorption coefficient of CO_2 as a function of wavenumber, for pure CO_2 gas at a pressure of $1bar$ and a temperature of $296K$. In some spectral regions, e.g. 1700-1800 cm^{-1}, CO_2 at this temperature and pressure is essentially transparent. This is a *window region* through which infrared can easily escape to space if no other greenhouse gas intervenes. For a $285K$ blackbody, $60W/m^2$ can be lost through this window. There are two major bands in which absorption occurs. For Earthlike temperatures, the lower wavenumber band, from about

450 to 1100 cm^{-1} is by far the most important. At $285K$ the blackbody emission in this band is $218W/m^2$ out of a total of $374W/m^2$, so the absorption in this band is well tuned to intercept terrestrial infrared and to thus reduce OLR. The blackbody emission in the higher wavenumber band, from 1800 to 2500 cm^{-1}, is only $6W/m^2$. This band has a minor effect on OLR for Earth, but it can become important for much hotter planets like Venus, and even for Earth is important for the absorption of solar near-infrared. Within either band, the absorption coefficient varies by more than eight orders of magnitude.

The absorption does not vary randomly. It is arrange around six peaks (three in each major band), with the overall envelope of the absorption declining approximately exponentially with distance from the peak. However, there is a great deal of fine-scale variation within the overall envelope. Zooming in on a typical region in the inset to Figure 4.7 we see that the absorption can vary by an order of magnitude over a wavenumber range of only a few tenths of a cm^{-1}. Most significantly, the absorption peaks sharply at a discrete set of frequencies.

Why does the absorption peak at preferred frequencies? In essence, molecules are like little radio receivers, tuned to listen to light only at certain specific frequencies. Since energy is conserved, the absorption or emission of a photon must be accompanied by a change in the internal energy state of the molecule. It is a consequence of quantum mechanics that the internal energy of a molecule can only take on values drawn from a finite set of possible energy states, the distribution of which is determined by the structure of the molecule. If there are N state, there are $N(N-1)/2$ possible transitions, and each one leads to a possible absorption/emission line as illustrated in Figure 4.8. Transitions between different energy states of a molecule's electron configuration do not significantly contribute to infrared absorption in most planetary situations. The energy states involved in infrared absorption and emission are connected with displacement of the nuclei in the molecule, and take the form of vibrations or rotations. Every molecule has an equilibrium configuration, in which each nucleus is placed so that the sum of the electromagnetic forces from the other nuclei and from the electron cloud sum up to zero. A displacement of the nuclear positions will result in a restoring force that brings the system back toward equilibrium, leading to vibrations. The nucleii can be thought of as being connected with quantum-mechanical springs (one between each pair of nucleii) of different spring constants, and the vibrations can be thought of as arising from a set of coupled quantum-mechanical oscillators. Rigid molecules, held together by rigid rods rather than springs, would have rotational states but not vibrational states. The fact that molecules are not rigid causes the rotational states to couple to the vibrational states, through the coriolis and centrifugal forces.

Noble gases (He, Ar, etc.) are monatomic, have only electron transitions, and are not active in the infrared. A diatomic molecule (Fig. 4.9) has a set of energy levels associated with the oscillation caused by pulling the nuclei apart and allowing them to spring back; it also has a set of energy levels associated with rotation about either of the axes perpendicular to the line joining the nuclei. Centrifugal force couples the stretching to the rotation. Triatomic molecules (Fig. 4.10) have an even richer set of vibrations and rotations, especially if their equilibrium state is bent rather than linear (Fig. 4.11). Polyatomic molecules like CH_4, NH_3, SF_6 and the chlorofluorcarbons (e.g. CFC-12, which is CCl_2F_2) have yet more complex modes of vibration and rotation. As the set of energy states becomes richer and more complex, the set of differences between states fills in more and more of the spectrum, making the molecule a better infrared absorber.

For a molecule to be a good infrared absorber and emitter, it is not enough that it have transitions whose energy corresponds to the infrared spectrum. In order for a photon to be absorbed or emitted, the associated molecular motions must also couple strongly to the electromagnetic field. Although the quantum nature of radiation is crucial for many purposes, when it comes to the interaction of infrared or longer wavelength radiation with molecules, one can productively

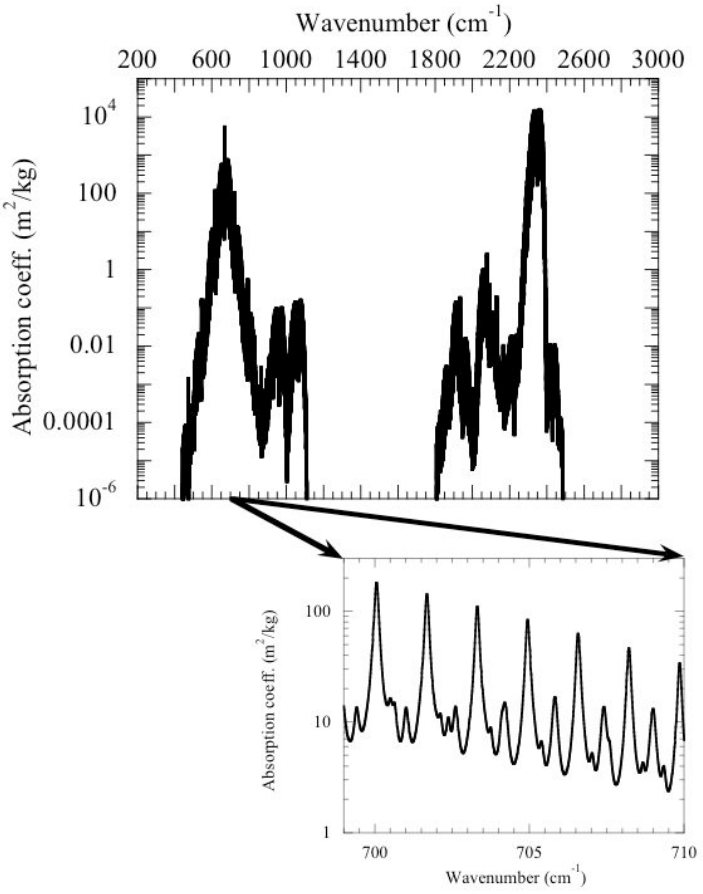

Figure 4.7: The absorption coefficient vs. wavenumber for pure CO_2 at a temperature of $293K$ and pressure of $10^5 Pa$. This graph is not the result of a measurement by a single instrument, but is synthesized from absorption data from a large number of laboratory measurements of spectral features, supplemented by theoretical calculations. The inset shows the detailed wavenumber dependence in a selected spectral region.

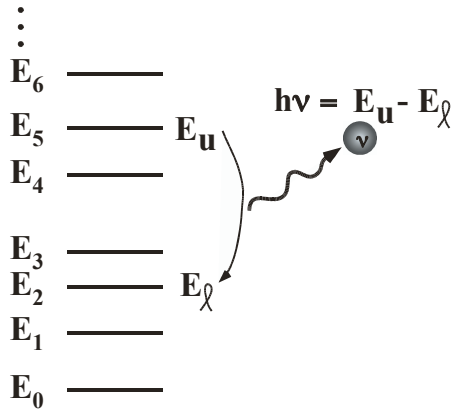

Figure 4.8: Schematic of emission of a photon by transition from a higher energy state to a lower energy state

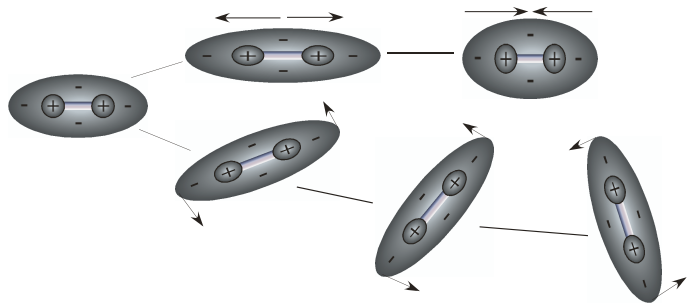

Figure 4.9: Vibration and rotation modes of a diatomic molecule made of a pair of identical atmos, with associated charge distributions

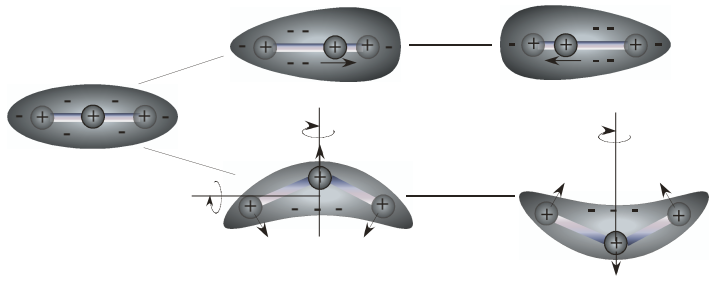

Figure 4.10: Vibration and rotation modes of a linear symmetric triatomic molecule (like CO_2), with associated charge distributions

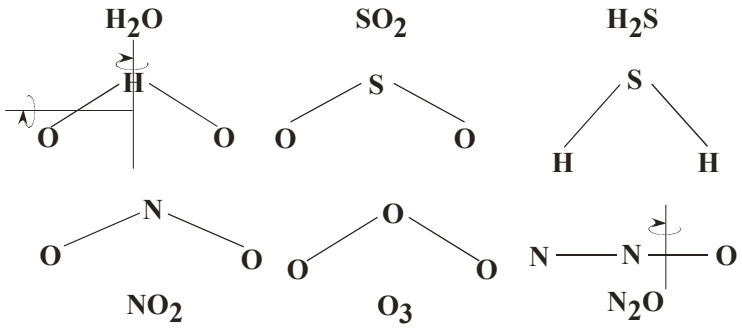

Figure 4.11: Some polar triatomic molecules. Two different modes of rotation are indicated for the H_2O molecule. There is a third mode of rotation about an axis perpendicular to the page.

think of the interaction in semiclassical terms. The reason is that the wavelength of infrared is on the order of 10 μm, which is two to three orders of magnitude larger than the size of the molecules we will be considering. Thus, one can think of the infrared light as providing a large scale fluctuating electric and magnetic field which alters the environment in which the molecule finds itself, and exerts a force on the constituent parts of the molecule. This force displaces the nuclei and electron cloud, and excites vibration or rotation. Conversely, a vibrating or rotating molecule creates a moving charge distribution, which classically radiates an electromagnetic wave. While one must fully take into account quantum effects in describing molecular motion, one need not for our purposes confront the much harder problem of quantizing the electromagnetic field as well (the problem of "quantum field theory"). The only way in which we make use of the quantum nature of the electromagnetic field is in converting the energy difference $E_u - E_\ell$ into a frequency of light, via $\Delta E = h\nu$.

The strongest interaction is between an electromagnetic field and a particle with a net charge. A charged particle will experience a net force when subjected to an electric field, which will cause the particle to accelerate. However, ions are extremely rare throughout most of a typical planetary atmosphere. The molecules involved in determining a planet's energy balance are almost invariably electrically neutral. The next best thing to having a net charge is to have a disproportionate part of the molecule's negatively charged electron cloud bunched up on one side of the molecule, while a compensating excess of positive charged nuclei are at the other side. This creates a *dipole moment*, which experiences a net torque when placed in an electric field, causing the dipole axis to try to align with the field. Interactions associated with higher order moments than the dipole lead to absorption many orders of magnitude weaker than the dipole absorption, and can be ignored for most planetary climate purposes.

Many common atmospheric molecules have no dipole moment in their unperturbed equilibrium state. Such *nonpolar* molecules can nonetheless couple strongly to the electromagnetic field. They do so because vibration and rotation can lead to a dipole moment through distortion of the equilibrium positions of the electron cloud and the nucleii. As illustrated in Figure 4.9, diatomic molecules made of two identical atoms, do not acquire a dipole moment under the action of either rotation or stretching. Symmetric diatomic molecules, such as N_2, O_2 and H_2 in fact have plenty of rotational and vibrational transitions that are in the infrared range. Because the associated molecular distortions have no dipole moment, however, these gases are essentially transparent to infrared unless they are strongly perturbed by frequent collisions. This is why the most common gases in Earth's atmosphere – N_2 and O_2 – do not contribute to Earth's greenhouse effect.

However, it is important to recognize that situations in which diatomic molecules become good greenhouse gases are in fact quite common in planetary atmospheres. When there are frequent collisions, such as happen in the high density atmospheres of Titan and on all the giant planets, diatomic molecules acquire enough of a dipole moment during the time collisions are taking place that the electromagnetic field can indeed interact with their transitions quite strongly. This makes N_2 and H_2 the most important greenhouse gases on Titan, and H_2 a very important greenhouse gas on all the gas giant planets. In terms of volume of atmosphere affected, Hydrogen is by far the most important greenhouse gas in the entire Solar System. Collision-induced absorption of this type forms a *continuum* in which the absorption is a very smooth function of wavenumber, without any significant line structure. Polyatomic molecules can also have significant continua, existing alongside the line spectra. Continuum absorption will be discussed in Section 4.4.8.

CO_2 is a linear molecule with the two oxygens symmetrically disposed about the central carbon, as illustrated in Figure 4.10. A uniform stretch of such a molecule does not create a dipole moment, but a vibrational mode which displaces the central atom from one side to the other does. In addition, bending modes of CO_2 have a fluctuating dipole moment, which can in turn be further influenced by rotation. Both these modes are illustrated schematically in Figure 4.10. Modes of this sort make CO_2 a very good greenhouse gas - the more so because the typical energies of the transitions involved happen to correspond to frequencies near the peak of the Planck function for Earthlike temperatures.

Some molecules – called *polar* have a dipole moment even in their undisturbed state. Most common diatomic gases made of two different elements – notably HF and HCl – are polar, and their vibrational and rotational modes cause fluctuations in the dipole which make them quite good infrared absorbers. They are not commonly thought of as greenhouse gases, because they are highly chemically reactive and do not appear in radiatively significant quantities in any known planetary atmosphere. However, one must keep an open mind about such things. Most triatomic atmospheric gases (H_2O, SO_2,O_3, NO_2 and H_2S, among others) are polar. CO_2, a symmetric linear molecule with the carbon at the center, is a notable exception. Ammonia (NH_3) is also polar, having its three hydrogens sticking out on one side like legs of a tripod attached to the nitrogen atom at the other side. Polar molecules couple strongly to the electromagnetic field, and their asymmetry also gives them a rich set of coupled rotation and vibration modes with many opportunities for transitions corresponding to the infrared spectrum. The spectrum is enriched because rotation about the axis with the largest moment of inertia (shown as the vertical axis for the water molecule in Figure 4.11) causes the wing molecules to fling outwards, changing the bond angle and the dipole moment. The molecule can also rotate about an axis perpendicular to the plane of the Figure, leading to distinct set of energy levels. Further, energy can be stored in rotations about the axis with minimum moment of inertia (shown as horizonal in the Figure). For a linear molecule like CO_2, rotation about the corresponding axis has essentially no energy.

Let's return now to the matter of how the OLR varies as a function of the greenhouse gas content of an atmosphere. Essentially, we revisit the discussion of Figure 4.5, but this time in the context of the actual absorption spectrum of CO_2 instead of the hypothetical gas discussed earlier. We use the same temperature profile as in Figure 4.5, but the OLR is computed using the fully resolved absorption coefficient as a function of wavenumber, reproduced in the lower panel of the Figure 4.12. The horizontal lines in this panel indicate the spectral regions that are optically thick for CO_2 paths of $\frac{1}{10}$, 1, and 1000 kg/m^2. The upper panel shows the corresponding OLR for the same three paths. The OLR was computed at the full spectral resolution of the lower panel, but was smoothed over bands of width $10cm^{-1}$ to make the pattern easier to see. The smoothing is done in such a way that the integral of the smoothed OLR curve over wavenumber yields the same net value as the integral over the original unsmoothed curve. This calculation of OLR is

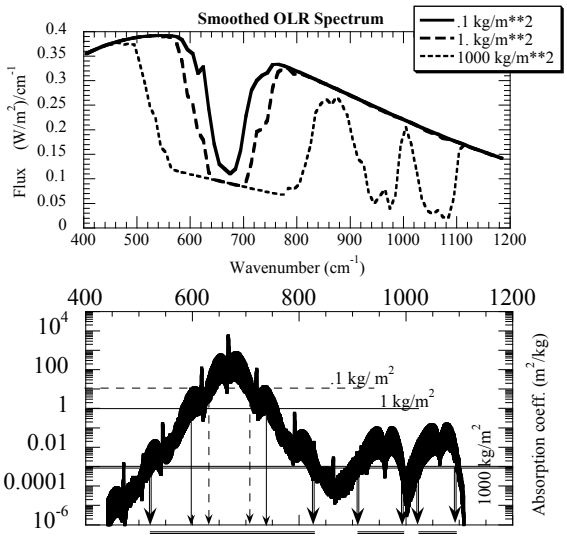

Figure 4.12: Lower panel: The absorption coefficient for CO_2 at 1 bar and $300K$, in the wavenumber range of interest for Earthlike and Marslike planets. The horizontal lines show the wavenumber range within which the optical thickness exceeds unity for CO_2 paths of $\frac{1}{10}$, 1 and 1000 kg/m^2. Upper panel: The corresponding OLR for the three path values, computed for the same temperature profiles as in Figure 4.5. The OLR has been averaged over bands of width 10 cm^{-1}.

still not completely realistic, since, to keep things simple, it is carried out as if the absorption coefficient were uniform throughout the depth of the atmosphere. In reality, κ_{CO2} varies with both temperature and pressure, though we'll see eventually that to a good approximation this variation can be handled through the introduction of an *equivalent path*, which is generally somewhat more than half the actual full-atmosphere path, if the reference pressure at which absorption coefficients are stated is taken to be the surface pressure.

Figure 4.12 explains why the OLR reduction is approximately logarithmic in greenhouse gas concentration for CO_2 and similar greenhouse gases. The key thing to note is that the absorption coefficient in the principal band centered on $675cm^{-1}$ decays exponentially with distance from the center. Hence, as the CO_2 path is increased by a factor of 10, from $\frac{1}{10}$ to 1 kg/m^2, the width of the ditch within which the radiating temperature is reduced to cold stratospheric values increases only like the logarithm of the ratio of paths. This is true for paths as small as $.01kg/m^2$ and as large as $100kg/m^2$. However, when the path gets as large as $1000kg/m^2$, the weak absorption bands on the shoulder, near 950 and 1050 cm^{-1} start to become important, and enhance the optical thickness beyond what one would expect on the basis of the central absorption peak. 1000 kg/m^2 corresponds to a partial pressure of CO_2 of about $100mb$ for Earth's gravity, or equivalently a molar mixing ratio of about 10 % for Earth's current surface pressure. This is far in excess of any CO_2 concentration on Earth likely to have been attained in the past 300 million years, but is well within the range of what has been contemplated at the end of a Neoproterozoic Snowball episode, or earlier during the Faint Young Sun period. Many greenhouse gases also have a central absorption peak with exponential skirts, and these will also exhibit a nearly logarithmic dependence of OLR on the concentration of the corresponding greenhouse gas.

With the notable exception of the collision-induced continuum discussed in Section 4.4.8, the absorption spectrum of a gas is built by summing up the contributions of the thousands of spectral lines from each of the radiatively active constituents of the gas. To proceed further, then, we must look more deeply into the nature of the lines and how they are affected by pressure and temperature.

4.4.3 I walk the line

An individual spectral line is described by a line *position* (i.e. the wavenumber at the center), a line *shape*, a line *strength* (or *intensity*), and a line *width*. The line shape is described by a nondimensional function of nondimensional argument, $f(x)$, normalized so that the total area under the curve is unity. The contribution of a single spectral line to the absorption coefficient for substance G can then be written

$$\kappa_G(\nu, p, T) = \frac{S}{\gamma} f(\frac{\nu - \nu_c}{\gamma}) \tag{4.60}$$

where ν_c is the frequency of the center of the line, S is the line intensity and γ is the line width. Note that $\int \kappa_G d\nu = S$. As a line is made broader, the area remains fixed, so that the absorption in the wings increases at the expense of decreased absorption near the center.

The pressure and temperature dependence of κ_G enters almost entirely through the pressure and temperature dependence of S and γ. The line center ν_c can be regarded as independent of pressure and temperature for the purposes of computation of planetary radiation balance. At very low pressures (below $1000Pa$), one may also need to make the line shape dependent on pressure.

Every line has an intrinsic width determined by the charactistic time for spontaneous decay of the higher energy state (analogous to a radioactive half-life). This width is far too narrow to be of interest in planetary climate problems. In addition, the lines of a molecule in motion will experience *Doppler broadening*, associated with the fact that a molecule moving towards a light source will see the frequency shifted to higher values, and conversely for a molecule moving away. For molecules in thermodynamic equilibrium, the velocities have a Gaussian distribution, and so the line shape becomes $f(x) = exp(-x^2)/\sqrt{\pi}$. The width is $\gamma = \gamma(T) = \nu_c \frac{v}{c}$, where $v = \sqrt{2RT}$, R being the gas constant for the molecule in question. v is a velocity, which is essentially the typical speed of a molecule at temperature T. For CO_2 at $250K$, the Doppler line width for a line with center $600cm^{-1}$ is only about $.0006cm^{-1}$.

The type of line broadening of primary interest in planetary climate problems is *collisional broadening*, alternatively called *pressure broadening*. Collisional broadening arises because the kinetic energy of a molecule is not quantized, and therefore if a molecule has experienced a collision sufficiently recently, energy can be borrowed from the kinetic energy in order to make up the difference between the photon's energy and the energy needed to jump one full quantum level. The theory of this process is exceedingly complex, and in many regards incomplete. There is a simple semi-classical theory that predicts that collision-broadened lines should have the *Lorentz line shape* $f(x) = 1./(\pi \cdot (1 + x^2))$, and this shape seems to be supported by observations, at least within a hundred widths or so of the line center. For the Lorentz shape, absorption decays rather slowly with distance from the center; 10 half-widths γ from the center, the Lorentz absorption has decayed to only $\frac{1}{101}$ of its peak value, whereas the Gaussian doppler-broadened line has decayed to less that 10^{-43} of its peak. There are both theoretical and observational reasons to believe that the very far tails of collision broadened lines die off faster than predicted by the Lorentz shape. A full discussion of this somewhat unsettled topic is beyond the level of sophistication which we aspire

to here, but the shape of far-tails has some important consequences for the continuum absorption, which will be taken up briefly in Section 4.4.8.

In the simplest theories leading to the Lorentz line shape, the width of a collision-broadened line is proportional to the mean collision frequency, i.e. the reciprocal of the time between collisions. The Lorentz shape is valid in the limit of infinitesimal duration of collisions; it is the finite time colliding molecules spend in proximity to each other that leads to deviations from the Lorentz shape in the far tails, but there is at present no general theory for the far-tail shape. For many common planetary gases the line width is on the order of a tenth of a cm^{-1} when the pressure is 1 bar and the temperature is around $300K$. For fixed temperature, the collision frequency is directly proportional to pressure, and laboratory experiment shows that the implied proportionality of line width to pressure is essentially exact. Holding pressure fixed, the density goes down in inverse proportion to temperature while the mean molecular velocity goes up like the square root of temperature. This should lead to a collision frequency and line width that scales like $1/\sqrt{T}$. Various effects connected with the way the collision energy affects the partial excitation of the molecule lead to the measured temperature exponent differing somewhat from its ideal value of $\frac{1}{2}$. Putting both effects together, if the width is known at a standard state (p_o, T_o), then it can be extrapolated to other states using

$$\gamma(p,T) = \gamma(p_o,T_o)\frac{p}{p_o}(\frac{T_o}{T})^n \tag{4.61}$$

where n is a line-dependent exponent derived from quantum mechanical calculations and laboratory measurements. It is tabulated along with standard-state line widths in spectral line databases. One must typically go to very low pressures before Doppler broadening starts to become important. For example, for a collision-broadened line with width $.1cm^{-1}$ at 1 bar, the width doesn't drop to values comparable to the Doppler width until the pressure falls to $6mb$ – comparable to the middle stratosphere of Earth or the surface pressure of Mars. Even then, the collision broadening dominates the absorption when one is not too close to the line center, because the Lorentz shape tails fall off so much more gradually than the Gaussian.

Another complication is that the collision-broadened line width depends on the molecules doing the colliding. Broadening by collision between molecules of like type is called *self-broadening*, while that due to dissimilar molecules is called *foreign broadening*. The simple Lorentz theory would suggest a proportionality to collision rate, which is a simple function of the ratio of molecular weights. Not only does the actual ratio of self to foreign broadened width deviate from what would be expected by this ratio, but the ratio actually varies considerably from one absorption line to another. For CO_2, for example, some self-broadened lines have essentially the same width as for the air-broadend case, whereas others can have widths nearly half again as large. For water vapor, the disparity is even more marked. Evidently, some kinds of collisions are better at partially exciting energy levels than others. There is no good theory at present that enables one to anticipate such effects. Standard spectral databases tabulate the self-broadened and air-broadened widths at standard temperature and pressure, but if one were interested in broadening of water vapor by collisions with CO_2 (important for Early Mars) or broadening of NH_3 by collisions with H_2 (on Jupiter or Saturn) , one would have to either find specialized laboratory experiments or extrapolate based on molecular weights and hope for the best.

The line intensities are independent of pressure, but they do increase with temperature. For temperatures of interest in most planetary atmospheres, the temperature dependence of the line intensity is well described by

$$S(T) = S(T_o)(\frac{T}{T_o})^n \exp(-\frac{h\nu_\ell}{k}(\frac{1}{T} - \frac{1}{T_o})) \tag{4.62}$$

where n is the line-width exponent defined above and $h\nu_\ell$ is the energy of the lower energy state in the transition that gives rise to the line. This energy is tabulated in standard spectroscopic databases, and is usually stated as the frequency ν_ℓ. Determination of the lower state energy is a formidable task, since it means that one must assign an observed spectral line to a specific transition. When such an assignment cannot be made, one cannot determine the temperature dependence of the strength of the corresponding line.

Now let's compute the average transmission function associated with a single collision-broadened spectral line in a band of wavenumbers of width Δ. We'll assume that the line is narrow compared to Δ, so that the absorption coefficient can be regarded as essentially zero at the edges of the band. Without loss of generality, we can then situate the line at the center of the band. The mean transmission function is

$$\bar{\mathfrak{T}}(p_1,p_2) = \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2}[\exp(-\frac{1}{g\pi}\int_{p_1}^{p_2}\frac{S(T)\gamma q}{\nu'^2+\gamma^2}dp)]d\nu' \tag{4.63}$$

where $\nu' = \nu - \nu_c$. The argument of the exponential is just the optical thickness of the layer between p_1 and p_2, and to keep the notation simple we will assume the integral to be taken in the sense that makes it positive. The double integral and the nonlinearity of the exponential make this a hard beast to work with, but there are two limits in which the result becomes simple. When the layer of atmosphere between p_1 and p_2 is optically thin even at the center of the line, where absorption is strongest, the line is said to be in the *weak line regime*. All lines are in this regime in the limit $p_2 \to p_1$, though if the line is very narrow or the intensity is very large, the atmospheric layer might have to be made exceedingly small before the weak line limit is approached. For weak lines the exponential can be approximated as $\exp(-\delta\tau) \approx 1. - \delta\tau$, whence

$$\begin{aligned}\bar{\mathfrak{T}}(p_1,p_2) &\approx 1 - \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2}\frac{1}{g\pi\cos\bar{\theta}}\int_{p_1}^{p_2}\frac{S(T)\gamma q}{\nu'^2+\gamma^2}dpd\nu' \\ &= 1 - \frac{1}{\Delta}\frac{1}{g\cos\bar{\theta}}\int_{p_1}^{p_2}S(T)qdp \\ &= 1 - \frac{1}{\Delta}S(T_o)\ell_w\end{aligned} \tag{4.64}$$

where T_o is a constant reference temperature and the *weighted path* for strong lines is

$$\ell_w(p_1,p_2) \equiv \frac{1}{g\cos\bar{\theta}}\int_{p_1}^{p_2}\frac{S(T(p))}{S(T_o)}q(p)dp \tag{4.65}$$

Note that for weak lines, the averaged transmission is independent of the line width. From the expression for $\bar{\mathfrak{T}}$ we can define the *equivalent width* of the line, $W \equiv S(T_o)\ell_w$. To understand the meaning of the equivalent width, imagine that absorption takes *all* of the energy out of the incident beam within a range of wavenumbers of width W, leaving the rest of the spectrum. undisturbed. The equivalent width W is defined such that the amount of energy thus removed is equal the amount removed by the actual absorption, which takes just a little bit of energy out of each wavenumber throughout the spectrum.

When the layer of atmosphere between p_1 and p_2 is optically thick at the line center, the transmission is reduced to nearly zero there. This defines the *strong line* limit. For strong lines, there is essentially no transmission near the line center; all the transmission occurs out on the wings of the lines. Since essentially nothing gets through near the line centers anyway, there is little loss of accuracy in replacing the line shape by it's far-tail form, $\pi^{-1}S\gamma/\nu'^2$. With this approximation

to the line shape, the band-averaged transmission may be written:

$$\begin{aligned}\bar{\mathfrak{T}}(p_1,p_2) &\approx \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2}[\exp(-\frac{1}{\nu'^2}\frac{1}{g\pi\cos\bar{\theta}}\int_{p_1}^{p_2}S(T)\gamma q dp)]d\nu' \\ &= \frac{1}{\Delta}\int_{-\Delta/2}^{\Delta/2}\exp(-\frac{X}{\nu'^2})d\nu' \\ &= \frac{1}{2\zeta_m}\int_{-\zeta_m}^{\zeta_m}\exp(-\frac{1}{\zeta^2})d\zeta\end{aligned} \tag{4.66}$$

where $X \equiv \sqrt{S(T_o)\gamma(p_o,T_o)\ell_s}/\pi$, and the weighted path for strong lines is

$$\ell_s \equiv \frac{1}{g\cos\bar{\theta}}\int_{p_1}^{p_2}\frac{S(T(p))}{S(T_o)}(\frac{T_o}{T})^n\frac{p}{p_o}q(p)dp \tag{4.67}$$

The third line in the expression for $\bar{\mathfrak{T}}$ comes from introducing the rescaled dummy variable $\zeta \equiv \nu'/\sqrt{X}$; the limit of integration then becomes $\zeta_m = \Delta/(2\sqrt{X})$ Unless the path is enormous, ζ_m will be very large, because the averaging interval Δ is invariably taken to be much larger than the typical line width (otherwise there would be little point in averaging). For $\zeta_m >> 1$, the integral in the last line can be evaluated analytically, and is

$$\int_0^{\zeta_m}\exp(-\frac{1}{\zeta^2})d\zeta \approx \zeta_m - \sqrt{\pi} \tag{4.68}$$

(see Problem **??**). Therefore, substituting for X, the expression for $\bar{\mathfrak{T}}$ in the strong limit becomes

$$\bar{\mathfrak{T}}(p_1,p_2) \approx 1 - \frac{1}{\Delta}2\sqrt{S(T_o)\gamma(p_o)\ell_s} \tag{4.69}$$

For strong lines the equivalent width is $W \equiv 2\sqrt{S(T_o)\gamma(p_o)\ell_s}$. In this case, the width of the chunk taken out of the spectrum increases like the square root of the path because the absorption coefficient decreases like $1/\nu'^2$ with distance from the line center, implying that the width of the spectral region within which the atmosphere is optically thick scales like the square root of the path. Unlike weak lines, strong lines really do take almost all of the energy out of a limited segment of the spectrum. The multiplicative property for transmission is equivalent to an additive property for equivalent width. The nonlinearity of the square root linking path to equivalent width in the strong line case thus means that the band-averaged transmission has lost the multiplicative property. As in our earlier general discussion of this property, the loss stems from the progressive depletion of energy in parts of the spectrum near the line center.

The pressure-weighting of the strong-line path reflects the fact that, away from the line centers, the atmosphere becomes more optically thick as pressure is increased and the absorption is spread over a greater distance around each line. Note that if we choose as the reference pressure p_o any pressure that remains between p_1 and p_2, then $\ell_s \to \ell_w$ as $p_1 \to p_2$. In this case, one can use the strong line path ℓ_s regardless of the pressure range, since the strong line path reduces to the correct weak line path for thin layers where weak line approximation becomes valid. A common choice for the reference pressure is the average $(p_1+p_2)/2$ but one could just as well choose one of the endpoints of the interval instead. In the case of a well-mixed greenhouse gas (constant q_G) for a nearly isothermal layer, the equivalent path becomes $1/\cos\bar{\theta}$ times $\frac{1}{2g}q_G(p_2^2-p_1^2)/p_o$, which reduces to the actual mass path $q_G(p_2-p_1)/g$ if p_o is taken to be the average. In this case, one gets the correct transmission by using the conventional mass path with absorption coefficients computed for the average pressure of the layer. This is known as the *Curtiss-Godson* approximation.

In solving radiative transfer problems related to planetary climate, one typically takes the bandwidth Δ large enough that the band contains a great many lines. For example, there are about 600 CO_2 lines in the band between 600 and 625 cm^{-1}. In the weak line limit the transmission is linear in the absorption coefficient, so one can simply sum the equivalent widths of all the lines in the band to obtain the total equivalent width $W = \sum W_i$. For strong lines, the situation is a bit more complicated, because of the nonlinearity of the exponential function. For the same reasons one loses the multiplicative property of transmission upon band averaging, one generally loses the additive property of equivalent widths. There is one important case in which additivity of equivalent widths is retained, however. If the lines are *non-overlapping*, in the sense that they are far apart compared to the width over which each one causes significant absorption, then the absorption from each line behaves almost as if the line were acting in isolation. In this case, each line essentially takes a distinct chunk out of the spectrum, and the equivalent widths can be summed up to yield the net transmission.

The additivity of strong-line equivalent widths breaks down at large paths. Since each W_i increases like the square root of the path, eventually the sum exceeds Δ, leading to the absurdity of a negative transmission. What is going wrong is that, as the equivalent widths become large, the absorption regions associated with each line start to overlap. One is trying to take away the same chunk of the spectrum more than once. This doesn't work for spectra any more than it works for ten hungry people trying to eat an eighth of a pizza each. One approach which has met with considerable success is to assume that the lines are randomly placed, so that the transmission functions due to each line are uncorrelated. This is *Goody's Random Overlap Approximation*. For uncorrelated transmission functions, the band-averaged transmissions can be multiplied, yielding

$$\begin{aligned} \bar{\mathfrak{T}} &\approx (1 - \frac{W_1}{\Delta})(1 - \frac{W_2}{\Delta})...(1 - \frac{W_N}{\Delta}) \\ &= \exp \sum \ln(1 - \frac{W_i}{\Delta}) \\ &\approx \exp(-\frac{1}{\Delta} \sum W_i) \end{aligned} \tag{4.70}$$

The last step follows from the assumption that each individual equivalent width is small compared to Δ. Note that when the sum of the equivalent widths is small compared to Δ, this expression reduces to the previous expression given for individual or non-overlapping lines.

The line parameters laid out above – position, width,strength and temperature scaling – lie at the heart of most real-gas radiative transfer calculations. There being thousands of spectral lines for dozens of substances of interest in planetary climate, teasing out the data one needs from the original literature is a daunting task. Fortunately, a small but dedicated group of spectroscopists [4] have taken on the task of validating, cross-checking and assembling the available line data into a convenient database known as `HITRAN`. It is suitable for most planetary calculations, though it must sometimes be supplemented with information on absorption that is not associated with spectral lines (the *continuum* absorption), and with additional data on weak lines which are important in the extremely hot, dense atmosphere of Venus. Instructions for obtaining the `HITRAN` database, along with sources for additional spectral data of use on Venus, Titan and Jupiter, are given in the references section at the end of this chapter, and the software supplement to this book provides a simple set of routines for reading and performing calculations with the `HITRAN` database.

[4] May they live forever!

4.4.4 Behavior of the band-averaged transmission function

Although the absorption spectrum has very complex behavior, the band-averaged transmission function averages out most of the complexity. The definition of the transmission guarantees that it decays monotonically as $|p_1 - p_2|$ increases and the path increases, but in addition the decay is invariably found to be smooth, proceeding without erratic jumps, kinks or other complex behavior. This smoothness is what makes computationally economical radiative transfer solutions possible, and the various schemes for carrying out the calculation of fluxes amount to different ways of exploiting the smoothness of the band-averaged transmission function.

By way of example, the band-averaged transmission function for CO_2 is shown for three different bands in Fig. 4.13. The calculation of $\bar{\mathfrak{T}}_\nu(p_1, p_2)$ was carried out using a straightforward – and very time consuming – integration of the transmission over frequency; at each frequency in the integrand, one must do an integral of $\kappa_{CO2}(\nu, p)$ over pressure, and each of those κ must be evaluated as a sum over the contributions of up to several hundred lines. Temperature was held constant at $296K$ and a constant mass-specific concentration of .0005 ($330ppmv$) of CO_2 mixed with air was assumed. The pressure p_1 was held fixed at $100mb$, while p_2 was varied from $100mb$ to $1000mb$. This plot thus gives an indication of the upward flux transmitted from each layer of the atmosphere, as seen looking down from the Earth's tropical tropopause. The results are plotted as a function of the pressure-weighted strong-line path, which for constant q and T is $q \cdot (p_2^2 - p_1^2)/(2gp_o \cos \bar{theta})$, where the reference pressure p_o is taken to be $10^5 Pa$. Plotting the results this way makes it easier to compare them with theoretical expectations, and also makes it easier to generalize the results to transmission between different pairs of pressure levels, which will have different amounts of pressure broadening. The rationale for using the strong-line path is that the lines are narrow enough that almost all parts of the spectrum are far from the line centers in comparison to the width, and in such cases the collision-broadened absorption coefficient increases linearly with pressure almost everywhere. This behavior is incorrect near the line centers, but the error in the transmission introduced by this shortcoming is minimal, since the absorption is so strong there the contribution to the transmission is essentially zero anyway. This reasoning – based directly on what we have learned from the strong-line limit – is at the basis of most representations of pressure-broadening effects in radiative calculations. Here, we only are using it as a graphical device, since the transmission itself is computed without approximation. Note that the strong line path becomes proportional to the (weak-line) mass path $q \cdot (p_2 - p_1)/(g \cos \bar{\theta})$ when $p_2 \to p_1$, with proportionality constant p_1/p_o. In the present calculation, when p_2 is at its limit of $1000mb$, the path is about $5kg/m^2$, which is about half the unweighted mass path over the layer. This reflects the fact that the lower pressure over most of the layer weakens the absorption relative to the reference value at $p = p_o$.

Apart from noticing that the transmission function is indeed smooth, we immediately remark that the transmission first declines sharply, as portions of the spectrum with the highest absorption coefficient are absorbed. At larger paths, the spectrum becomes progressively more depleted in easily-absorbed wavenumbers, and the decay becomes slower. For the two strongly absorbing bands in the left panel, the transmission curve becomes nearly vertical at small paths, as suggested by the square-root behavior of the strong line limit. There is guaranteed to be a weak-line region at sufficiently small paths, where the slope becomes finite, but in these bands the region is so tiny it is invisible. In fact, the strong line transmission function in Eq. 4.69 fits the calculated transmission in the 575-600 cm^{-1} band almost exactly throughout the range of paths displayed, when used with the random-overlap modification in Eq. 4.70. For the more strongly absorbing 600-625 cm^{-1} band the fit is very good out to paths of 1.5 kg/m^2, but thereafter the actual transmission decays considerably more rapidly than the strong-line form. This mismatch occurs because the derivation

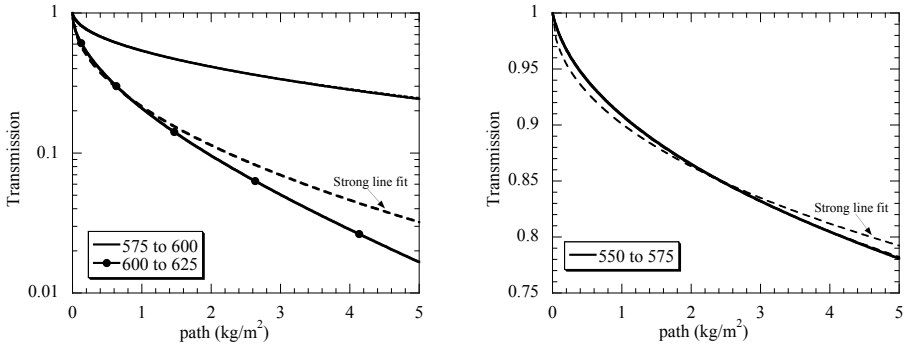

Figure 4.13: The band averaged transmission as a function of path, for the three different bands, as indicated. In each case, the transmission is computed between a fixed pressure $p_1 = 100mb$ and a higher pressure p_2 ranging from $100mb$ to $1000mb$. Calculations were carried out assuming the temperature to be constant at $296K$, with a constant CO_2 specific concentration of $q = .0005$, and assuming a mean propagation angle $\cos\bar{theta} = \frac{1}{2}$. Results are plotted as a function of the pressure-weighted path for strong lines, $q \cdot (p_2^2 - p_1^2)/(2gp_o \cos\bar{theta})$, where $p_o = 1000mb$. In the left panel, the best fit to the strong-line transmission function is shown as a dashed curve; the fit is essentially exact for the $575 - 600cm^{-1}$ band, so the fitted curve isn't visible. For the weaker absorption band in the right panel, fits are shown both for the strong line and the Malkmus transmission function, but the Malkmus fit is essentially exact and can't be distinguished.

of the strong-line transmission function assumes that the absorption coefficients within the band approach zero arbitrarily closely: as more and more radiation is absorbed, there is always some region where the absorption coefficient is arbitrarily close to zero, which leads to ever-slower decay. In reality, overlap between the skirts of the lines leads to finite-depth valleys between the peaks (see the inset of Fig. 4.7), and the absorption is bounded below by a finite positive value. The decay of the transmission at large paths is determined by the local minima in the valleys, and will tend toward exponential decay, rather than the slower decay predicted by the strong line approximation.

For the weakly absorbing band shown in the right panel of Fig. 4.13, a hint of weak-line behavior can be seen at small values of the path, with the result that the behavior diverges noticeably from the best strong-line fit. The representation of the transmission can be improved by adopting a two-parameter fit tailored to give the right answer in both the weak and strong limits. The *Malkmus model* is a handy and widely-used example of this approach. It is defined by

$$\sum W_i = 2\frac{R^2}{S}\frac{p_1}{p_o}\left(\sqrt{1 + \frac{S^2}{R^2}\left(\frac{p_o}{p_1}\right)^2 \ell_s} - 1\right) \tag{4.71}$$

where R and S are the parameters of the fit [5]. The parameters can be identified with characteristics of the absorption spectrum in the band by looking at the weak line (small ℓ_s) and strong line (large ℓ_s) limits. For small ℓ_s, the sum of the equivalent widths is $S \cdot (p_o/p_1)\ell_s = S\ell_w$, so by comparing with Eq. 4.64 we identify S as the sum of the line intensities. For large ℓ, the sum is $2\sqrt{R^2\ell}$, whence on comparison with Eq. 4.69 we identify R^2 as the sum of $\gamma_i(p_o)S_i$ for all the lines in the band.

[5] The factor p_1/p_o deals with the difference between the strong line and weak line paths, and is necessary so that the limits work out properly for small and large path. There is some flexibility in defining this factor. It is common to use $\frac{1}{2}(p_1 + p_2)/p_o$ to make things look more symmetric in $p_1 and p_2$. This slightly changes the way the function interpolates between the weak and strong limits, without changing the endpoint behavior

The parameters R and S can thus be determined directly from the database of line intensities and widths, though in some circumstances it can be advantageous to do a direct fit to the results of a line-by-line calculation like that in 4.13 instead. One uses the Malkmus equivalent-width formula with the random-overlap transformation given in Eq. 4.70, so as to retain validity at large paths. With the Malkmus model, the transmission function in the weakly absorbing 550-575 cm^{-1} band can be fit almost exactly. Since the Malkmus model reduces to the strong line form at large paths, it fits the transmission functions in the left panel of Fig. 4.13 at least as well as the strong line curve did. However, it does nothing to improve the fit of the strongly absorbing case at large paths, since that mismatch arises from a failure of the strong-line assumption itself.

The Malkmus model is a good basic tool to have in one's radiation modelling toolkit. It works especially well for CO_2, and does quite well for a range of other gases as well. There are other fits which have been optimized to the characteristics of different greenhouse gases (e.g. *Fels-Goody* for water vapor), and fits with additional parameters. Most of the curve-fit families have troubles getting the decay of the transmission right when very large paths are involved, though if the trouble only appears after the transmission has decayed to tiny values, the errors are inconsequential.

Empirical fits to the transmission function are a time-honored and effective means of dealing with infrared radiative transfer. This approach has a number of limitations, however. We have already seen some inadequacies in the Malkmus model when the path gets large; patching up these problems leads to fits with more parameters, and finding fits that are well-tailored to the characteristics of some new greenhouse gas one wants to investigate can be quite involved. It also complicates the implementation of the algorithm to have to use different classes of fits for different gases, and maybe even according to the band being considered. A more systematic and general approach is called for. The one we shall pursue now, known as *exponential sums*, has the additional advantage that it can be easily generalized to allow for the effects of scattering, which is not possible with band-averaged fits like the Malkmus model. As a gentle introduction to the subject, let's consider the behavior of the integral

$$\bar{\mathfrak{T}}(\ell) = \frac{1}{\Delta}\int_{\nu_o-\Delta/2}^{\nu_o+\Delta/2} e^{-\kappa_G(\nu)\ell} d\nu \tag{4.72}$$

where κ_G is the absorption coefficient for a greenhouse gas G and ℓ is a mass path. This would in fact be the exact expression for the band-averaged transmission for a simplified greenhouse gas whose absorption coefficient is independent of pressure and temperature. In this case, the path ℓ between pressure p_1 and p_2 is simply the unweighted mass path $|\int qdp|/(g\cos\bar{\theta})$, which reduces to $q|p_1 - p_2|/(g\cos\bar{\theta})$ if the concentration q is constant.

The problem we are faced with is the evaluation of the integral of a function $f(x)$ which is very rapidly varying as a function of x. The ordinary way to approximate the integral is as a Riemann-Stieltjes sum, dividing the interval up into N sub-intervals $[x_j, x_{j+1}]$ and summing the areas of the rectangles, i.e.

$$\int_0^1 f(x)dx \approx \sum_1^N f(\frac{x_j + x_{j+1}}{2})(x_{j+1} - x_j) \tag{4.73}$$

The problem with this approach is that a great many rectangles are needed to represent the complex area under the curve $f(x)$. Instead, we may define the function $H(a)$, which is the sum of the lengths of the intervals for which $f(x) \leq a$, as illustrated in Fig. 4.14. Now, the integral can be approximated instead by the sum

$$\int_0^1 f(x)dx = \int_{f=f_{min}}^{f_{max}} f dH(f) \approx \sum_1^M \frac{f_j + f_{j+1}}{2}(H(f_{j+1}) - H(f_j)) \tag{4.74}$$

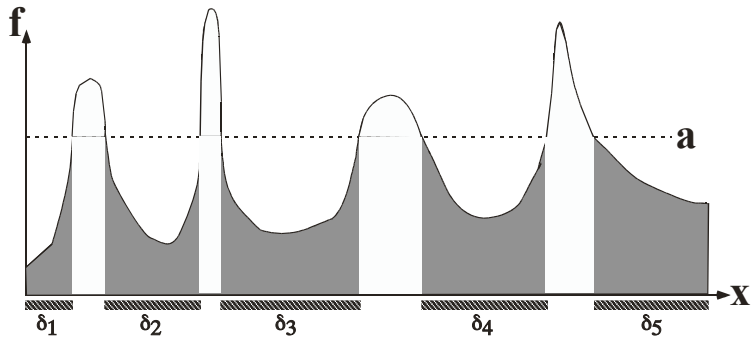

Figure 4.14: Evaluation of the area under a curve by Lebesgue integration

where we have divided the range of the function f (i.e. $[f_1, f_2]$) into M partitions. This representation can be very advantageous if $H(f)$ is a much more smoothly varying function than $f(x)$. To mathematicians, this form of the approximation of an integral by a sum is the first step in the magnificent apparatus of *Lebesgue integration*, leading onwards to what is known as measure theory, which forms the basis of rigorous real analysis.

The idea is to apply the Lebesgue integration technique to the transmission function defined in Eq 4.72, with the absorption coefficient κ_G playing the role of f and the frequency ν playing the role of x. Thus, if $H(a)$ is the sum of the lengths of the frequency intervals in the band for which $\kappa_G \leq a$, then $H(a) = 0$ when a is less than the minimum of κ_G and $H(a)$ approaches the bandwidth Δ when a approaches the maximum of κ_G. The transmission function can then be written approximately as

$$\bar{\mathfrak{T}}(\ell) = \int_{\kappa_{min}}^{\kappa_{max}} e^{-\kappa_G \ell} dH(\kappa_G) \approx \sum_{1}^{M} e^{-(\kappa_{j+1}+\kappa_j)\ell/2}(H(\kappa_{j+1}) - H(\kappa_j)) \tag{4.75}$$

This is the exponential sum formula. It can be regarded as an M term fit to the transmission function, much as the Malkmus model is a two-parameter fit. The Lebesgue integration technique amounts to a simple reshuffling of the terms in the integrand: we collect together all wavenumbers with approximately the same κ, compute the transmission for that value, and then weight the result according to "how many" such wavenumbers there are.

Because the absorption coefficient varies over such an enormous range, it is more convenient to work with $H(\ln \kappa)$ rather than $H(\kappa)$. A typical result for CO_2 is shown in Fig. 4.15, computed for two bands at a pressure of $100mb$. The function is quite smooth, and can be reasonably well characterized by ten points or less. In contrast, given that the typical line width at $100mb$ is only $.01cm^{-1}$, evaluation of the transmission integral in the Riemann form, Eq. 4.73, would require at least 25000 points in a band of width $25cm^{-1}$. Thus, the exponential sum approach is vastly more economical of computer time than a direct line-by-line integration would be.

The decay of the transmission with path length described by Eq. 4.75 is exactly analogous to the decay in time of the concentration of a mixture of radioactive substances with different half-lives. The short-lived things go first, leading to rapid initial decay of concentration; as time goes on, the mixture is increasingly dominated by the long-lived substances, and the decay rate is correspondingly slower. The way the transmission function converges as additional terms are included in the exponential sum formula is illustrated in Figure 4.16. Specifically, we divide the range of absorption coeffients into 20 bins equally spaced in $\log \kappa_G$, and then truncate H so as

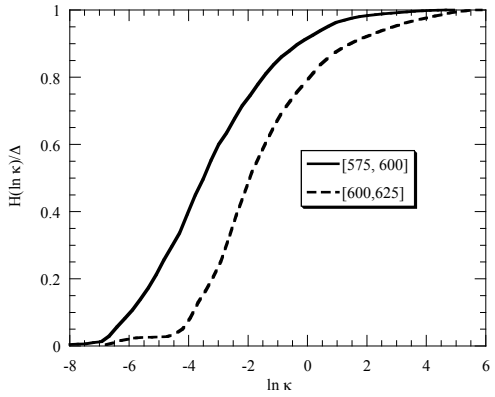

Figure 4.15: Cumulative probability function of the natural log of the absorption coefficient for CO_2. Results are given for the 600-625 cm^{-1} and 575-600 cm^{-1} bands, and were computed at a pressure of $100mb$ and a temperature of $296K$.

to keep only the N largest absorption coefficients, with N ranging from 1 (retaining only the strongest absorption) to 20 (retaining all absorption coefficients including the weakest). When only the strongest absorptions are included, the steep decay of transmission for small paths is correctly represented, but the transmission function decays too strongly at large paths. As more of the weaker absorption terms are included, the weaker decay of the transmission is well represented out to larger and larger paths. The ability to represent the decay rate of transmission over a very large range of paths is one of the two advantages of exponential sums over the Malkmus approach, the other advantage being the ability to incorporate scattering effects. An analytical example exploring related features of the exponential integral in Eq. 4.75 is explore in Problem **??**.

If it weren't for the dependence of absorption coefficient on pressure and temperature, the exponential sum representation would be exact in the limit of sufficiently many terms. The computational economy of exponential sums comes at a cost, however, which is scrambling the information about which absorption coefficient corresponds to which frequency. This is not a problem if one is dealing with a layer with essentially uniform pressure and temperature, but it becomes a cause for concern in the typical atmospheric case where one is computing the transmission over a layer spanning considerable variations in temperature and pressure. The problem is that changing pressure or temperature changes the *shape* of the distribution $H(\kappa)$, and there is no rigorously correct way to deal with this within the exponential sum framework. In the discussion of line shapes, for example, we learned that increasing p reduces the peak absorption, but increases absorption between peaks. In terms of the probability distribution of κ, this means that the largest and smallest values of κ become less prevalent at the same time that the intermediate values become more prevalent. At very large values of pressure where the lines become extremely broad, κ becomes a smooth function of frequency within each band and the probability distribution becomes concentrated on a single mean value of κ. The effect of temperature on the shape of $H(\kappa)$ can be even more complex, since the temperature-dependence coefficients of line strength can differ greatly even for neighboring lines.

All these problems notwithstanding, experience has shown that one can obtain a reasonably accurate approximation to the band-averaged transmission function by assuming that all the

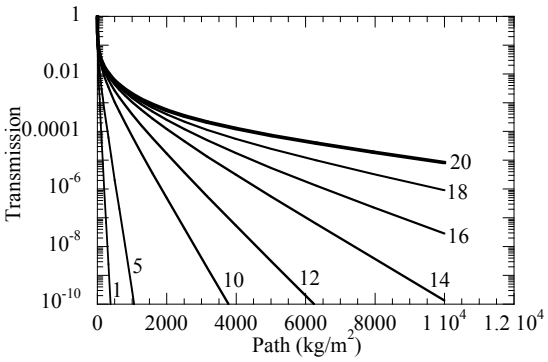

Figure 4.16: Convergence of exponential sum representation of the band averaged transmission function as the number of terms is increased. The calculation is for CO_2 in a wavenumber band near the peak absorption.

absorption coefficients within a band have separable scaling of the form

$$\kappa_G(p,T) = \kappa_G(p_o,T_o)F(p/p_o, T/T_o) \tag{4.76}$$

Given scaling of this sort, one can compute the transmission for a path through an inhomogeneous atmosphere using Eq 4.75, by defining a suitable equivalent path. For example, if the specific concentration of the greenhouse gas is q_G, the absorption scaling is linear in pressure, and the temperature dependence scales according to a function $S(T/T_o)$, then the equivalent path for the layer between p_1 and p_2 is

$$\ell = \frac{1}{g}\int_{p_1}^{p_2} q_G(p)\frac{p}{p_o}S(\frac{T(p)}{T_o})dp \tag{4.77}$$

The temperature scaling function S would be computed separately for each band. The use of linear scaling in pressure is justified by the fact that most of the spectrum is far from line centers, so absorption scales like the strong-line approximation as long as the pressure is not so large that line widths become comparable to the width of the band under consideration. The use of a single temperature-scaling function is harder to justify theoretically, but seems to be supported by numerical experiment.

When q_G is not small, there is one additional complication to take into account when defining the equivalent path. Namely, the absorption coefficient for self-broadened collisions is generally different from that for foreign-broadened collisions, so one must scale the absorption coefficient according to the proportion of self vs. foreign collisions. Like the temperature scaling factor, the ratio of self to foreign broadening can vary considerably even amongst nearby lines. Typically, though, one simply chooses a representative ratio of self to foreign broadening which is assumed constant within each band over which the distribution H is computed. Let's call this ratio a_{self}. The molar concentration of the greenhouse gas is $(\bar{M}/M_G)q_G$, where $\bar{M}$ is the mean molecular weight of the mixture; hence the proportion of collisions which are self-collisions is $(\bar{M}/M_G)q_G$ while the proportion of foreign collisions is $1-(\bar{M}/M_G)q_G$. Then, if H is computed using the foreign-broadened absorption at the standard pressure, the appropriate equivalent path to use in computing the transmission is

$$\ell = \frac{1}{g}\int_{p_1}^{p_2}(1+\frac{\bar{M}}{M_G}q_G(p)(a_{self}-1))q_G(p)\frac{p}{p_o}S(\frac{T(p)}{T_o})dp \tag{4.78}$$

For a pure one-component atmosphere, $(\bar{M}/M_G)q_G = 1$ and one simply uses the self-broadened absorption coefficients in preparing the distribution $H(\kappa)$, rather than going through the intermediary of defining a_{self} for each band.

Because the scaling of absorption coefficient with pressure and temperature is only approximate, it is important to compute H for a reference pressure and temperature that is characteristic of the general range of interest for the atmosphere under consideration, so as to minimize the amount of scaling needed. Typically, one might use a half or a tenth of the surface pressure as the reference pressure, and a mid-tropospheric temperature as the reference temperature; if one were primarily interested in stratospheric phenomena, or if one were computing OLR on a planet like Venus where most of the OLR comes from only the uppermost part of the atmosphere, pressures and temperatures characteristic of a higher part of the atmosphere would be more appropriate.

Modern professionally-written radiative transfer codes attempt to get around the inaccuracies of temperature and pressure scaling by using an extension of the exponential-sums method known as *correlated-k*. The basic idea behind this method is to explicitly compute a database of absorption distribution functions H covering the range of pressure and temperature values encountered in the atmosphere, rather than generating them from rescaling of a single distribution function. The mathematical justification of the way these distribution functions are used to compute the transmission is poor, but the method reduces to exponential sums when scaling is valid. It is thus guaranteed to be no worse than exponential sums, and comparisons with detailed line-by-line calculations indicate that it commonly performs well, though it is hard to say when the method should work and when it shouldn't. The exponential sums approach suffices to give the reader an understanding of the basic principles of real-gas radiative effects, so we will use that method as the basis of most of our further discussion of the subject. The reader wishing to learn how to make use of the correlated-k method is directed to the reference given in the Further Readings section of this chapter.

4.4.5 Dealing with multiple greenhouse gases

We now know how to efficiently compute the band-averaged transmission function for a single greenhouse gas acting alone. It is commonly the case, however, that two or more greenhouse gases are simultaneously present – CO_2, CH_4 and water vapor in Earth's case, for example. How do we compute the averaged transmission function is this situation? The issues are closely related to thosed discussed in Section 4.4.1 in connection with the loss of the multiplicative property in band-averaged transmission functions. Similar reasoning shows that the average transmission function for two greenhouse gases acting together generally differs from the product of the averaged transmission functions of each of the individual gases taken alone. Fortunately, however, the special circumstances under which the multiplicative property holds for multiple gases are expected to be fairly common.

By way of illustration, let's consider an idealized greenhouse gas A with transmission function $\mathfrak{T}_A(\nu)$ in a wavenumber band of width Δ, with the property that $\mathfrak{T}_A = a$ on a set of wavenumber subintervals with length adding up to $r \cdot \Delta$, but with $\mathfrak{T}_A = 1$ elsewhere; naturally, we require $a < 1$ and $r \leq 1$. Consider a second greenhouse gas B with $\mathfrak{T}_B = b$ on a set of wavenumber subintervals with length adding up to $s \cdot \Delta$, and with $\mathfrak{T}_B = 1$ elsewhere. The band-averaged transmission for gas A in isolation is $r \cdot a + (1 - r)$, and for B in isolation is $s \cdot b + (1 - s)$. What is the transmission when both gases act in combination?

The answer depends on how the regions of absorption of gas A – spectral regions where $\mathfrak{T}_A < 1$ – line up with those of gas B. We can distinguish three limiting cases. First, the regions of

absorption of the two gases may be perfectly correlated, in which case $r = s$, $\mathfrak{T}_A(\nu) = a$ precisely where $\mathfrak{T}_B(\nu) = b$, and

$$\begin{aligned}\bar{\mathfrak{T}}(\nu) &= \frac{1}{\Delta}\int_{\nu-\Delta/2}^{\nu+\Delta/2} \mathfrak{T}_A(\nu')\mathfrak{T}_B(\nu')d\nu' \\ &= rab + (1-r) \\ &= \bar{\mathfrak{T}}_A\bar{\mathfrak{T}}_B - r(1-r)(a+b-ab-1)\end{aligned} \tag{4.79}$$

Further, $a + b - ab = a + (1 - a)b$, and so this expression has a value lying between a and b, both of which are less then unity. Hence $(a + b - ab - 1) < 0$ and $r(1 - r)(a + b - ab - 1) \leq 0$, given that $0 \leq r \leq 1$. We conclude that the mean transmission function for the two gases acting in concert is always greater than or equal to the product of the individual transmission functions, the equality applying only when $r = 0$ or $r = 1$, i.e. when there is no absorption or completely uniform absorption. For example, if we take $r = \frac{1}{2}$ and $a = b = frac110$, the mean transmission function for the two gases acting together is $\frac{101}{200}$, or just a bit over a half, whereas the individual transmission functions are $\frac{11}{20}$ each, multiplying out to $\frac{121}{400}$ or just over a quarter. When the absorption regions of the two gases coincide, the gases acting together transmit considerably more radiation than one would infer by allowing each gas to act independently in sequence. This happens because one of the gases uses up some of the frequencies that the other gas would like to absorb.

Exercise 4.4.1 Derive the final equality in Eq. 4.79. Sketch graphs of transmission functions vs. frequency for the two gases for *two different* cases illustrating perfectly correlated absorption regions. Evaluate the mismatch between $\bar{\mathfrak{T}}$ and $\bar{\mathfrak{T}}_A\bar{\mathfrak{T}}_B$ for $a = b = r = \frac{1}{2}$. Allowing a, b and r to vary over all possible values, what is the greatest possible mismatch?

At the other extreme, the absorption regions of the two gases may be completely *disjoint*, so that $\mathfrak{T}_A = 1$ wherever $\mathfrak{T}_B = b < 1$ and $\mathfrak{T}_B = 1$ wherever $\mathfrak{T}_A = a < 1$. For this to be possible, we require $r + s \leq 1$. In the disjoint case,

$$\begin{aligned}\bar{\mathfrak{T}}(\nu) &= \frac{1}{\Delta}\int_{\nu-\Delta/2}^{\nu+\Delta/2} \mathfrak{T}_A(\nu')\mathfrak{T}_B(\nu')d\nu' \\ &= ra + sb + (1-(r+s)) \\ &= \bar{\mathfrak{T}}_A\bar{\mathfrak{T}}_B - rs(1-a)(1-b)\end{aligned} \tag{4.80}$$

In the disjoint case, then, the transmission for the two gases acting together is always *less than* the compounded transmission of the two gases acting independently.

Exercise 4.4.2 Derive the final equality in Eq. 4.80. Sketch some transmission functions illustrating the disjoint case. Put in a few numerical values for a, b, r and s to show the size of the mismatch between $\bar{\mathfrak{T}}$ and $\bar{\mathfrak{T}}_A\bar{\mathfrak{T}}_B$. What is the greatest possible mismatch in the disjoint case?

As the final limiting case, suppose that the absorption of the two gases is *uncorrelated*, so that at any given frequency the probability that $\mathfrak{T}_A = a$ is r regardless of the value of $\mathfrak{T}_B$ there, and the probability that $\mathfrak{T}_B = b$ is s regardless of the value of $\mathfrak{T}_A$ there. This situation is also known as the *random overlap* case. In this case

$$\begin{aligned}\bar{\mathfrak{T}}(\nu) &= \frac{1}{\Delta}\int_{\nu-\Delta/2}^{\nu+\Delta/2} \mathfrak{T}_A(\nu')\mathfrak{T}_B(\nu')d\nu' \\ &= r(1-s)a + s(1-r)b + rsab + (1-r)(1-s) \\ &= \bar{\mathfrak{T}}_A\bar{\mathfrak{T}}_B\end{aligned} \tag{4.81}$$

The reasoning behind the second line is that $r(1-s)$ is the probability that only the first gas is absorbing, $s(1-r)$ is the probability that only the first gas is absorbing, rs is the probability that both gases are absorbing, and $(1-r)(1-s)$ is the probability that neither gas is absorbing. Multiplying out the terms in the product of $\mathfrak{T}_A$ and $\mathfrak{T}_B$ we find that in the random overlap case, the mean transmission of the two gases acting together is precisely the same as the compounded transmission of the two gases acting independently.

The properties illustrated by the three cases just discussed can be generalized to an arbitrary set of transmission functions. Let $\mathfrak{T}_A$ and $\mathfrak{T}_B$ be any two transmission functions, and define the fluctuation

$$\mathfrak{T}'_A = \mathfrak{T}_A - \bar{\mathfrak{T}}_A \tag{4.82}$$

and similarly for $\mathfrak{T}_B$. Then,

$$\overline{\mathfrak{T}_A \mathfrak{T}_B} = \bar{\mathfrak{T}}_A \bar{\mathfrak{T}}_B + \overline{\mathfrak{T}'_A \mathfrak{T}'_B} \tag{4.83}$$

From this we conclude that the transmission of the two gases acting in concert is greater than the product of the individual transmissions if the two transmissions are positively correlated, less than the product of the individual transmissions of the two transmissions are negatively correlated, and equal to the product if the two transmissions are uncorrelated.

In fact, by exercising just a little more mathematical sophistication, it is possible to go further and put an upper bound on the amount by which the mean transmission function for joint action by the two gases deviates from the product of the individual transmission functions. The key is to use a handy and powerful relation known as the *Schwarz Inequality*, which states that for any two functions $f(x)$ and $g(x)$, $(\overline{fg})^2 \leq \overline{f}^2 \overline{g}^2$, where an overbar indicates an average over x. The equality applies only when $f(x)$ is proportional to $g(x)$. Applying the Schwarz Inequality to Eq. 4.83, we find that the deviation satisfies the inequality

$$|\overline{\mathfrak{T}'_A \mathfrak{T}'_B}| \leq \sqrt{\overline{(\mathfrak{T}'_A)^2}} \sqrt{\overline{(\mathfrak{T}'_B)^2}} \tag{4.84}$$

In other words, in the worst case the deviation can become as large as the product of the standard deviations of the two individual transmission functions. Since $0 \leq \mathfrak{T}'_{A,B} \leq 1$, the maximum standard deviation is $\frac{1}{2}$, occuring when each transmission function is zero for half of the frequencies in the band and unity for the other half; the error in random overlap in this case is $+\frac{1}{4}$ if the transmissions are perfectly correlated and $-\frac{1}{4}$ if the transmissions are perfectly anticorrelated. These errors should be compared to the random-overlap value for the limiting case, which is $\frac{1}{4}$. The effects of non-random overlap are potentially severe.

Fortunately, the situation is rarely as bad as the worst-case suggests. The positions of spectral lines are a sensitive function of molecular structure, so it is a reasonable guess that the absorption spectra of dissimilar molecules should be fairly uncorrelated. Thus, in most circumstances one can get a reasonable approximation to the joint transmission function by computing the individual transmission functions for each gas and taking the product of the individual transmission functions. It is fairly easy, if computationally expensive, to test the accuracy of the random-overlap assumption in a given band by computing the correlations of the full frequency-dependent transmission functions in the band. However, finding a general characterization of the correction to random-overlap, and the way the correction depends on the concentrations of the individual gases, is an intricate art which we will not pursue here.

4.4.6 A homebrew radiation model

We have now laid out all the ingredients that go into a real gas radiation model, and are ready to begin assembling them. The ingredients are:

- A means of computing the band-averaged transmission over a specified wavenumber range
- The band-averaged integral (Eq. 4.11,4.12, or 4.13) giving the band-averaged solution to the Schwartzschild equation in terms of the preceding transmission functions

and the recipe is:

- Divide the spectrum into bands of a suitable width
- Prepare in advance: Malkmus coefficients or exponential sum coefficients $H(\ln \kappa)$ for each band, for each greenhouse gas present in significant quantities in the atmosphere
- Program up a function to compute the band averaged transmission in each band, using the coefficients prepared in the previous step.
- If there are multiple greenhouse gases, do the preceding for each individual greenhouse gas and combine the resulting transmission functions, allowing suitably for the nature of the overlap between absorption bands of the competing gases (for advanced chefs only!)
- Use the resulting transmission in a numerical approximation to the integral in Eq. 4.11,4.12 or 4.13 in each band to get the band-averaged fluxes.
- Sum up the fluxes in each band to get the total flux
- Serve up the fluxes to the rest of the climate model and enjoy

In the typical climate simulation application, one is given a list of values of temperature and greenhouse gas concentrations tabulated on a finite grid of pressure levels p_j for $j = 0, ...N$, and one must compute the fluxes based on this information. Either of Eq. 4.11 or 4.12 provides a suitable basis for numerical evaluation when one is working from atmospheric data tabulated on a grid. In writing down the approximate expressions for the flux, we will adopt the convention that $j = 0$ at the top of the atmosphere and that $j = N$ represents the ground. We shall use the superscript (k) to refer to quantities averaged or integrated over the band k, centered on frequency $\nu^{(k)}$ and having width $\Delta^{(k)}$. Let's define the gridded quantities:

$$\begin{aligned} B_j &\equiv B(\nu^{(k)}, \frac{1}{2}(T_j + T_{j+1}))\Delta^{(k)} \\ \bar{\mathfrak{T}}_{ij}^{(k)} &\equiv \bar{\mathfrak{T}}^{(k)}(p_i, p_j) \\ e_{ij} &\equiv \bar{\mathfrak{T}}_{ij}^{(k)} - \bar{\mathfrak{T}}_{i(j+1)}^{(k)} \end{aligned} \tag{4.85}$$

The trapezoidal-rule approximation to the the expression for upward flux in band (k), based on Eq. 4.11, is then simply

$$I_+^{(k)}(p_i) = I_{+,s}^{(k)} \bar{\mathfrak{T}}_{iN}^{(k)} + \sum_{j=i}^{N} B_j e_{ij} \tag{4.86}$$

The expression for the downward flux follows a similar form. B_j is the blackbody emission from layer j, and the flux at a given level is a weighted sum of the emissions from each layer below (for upward flux) or above (for downward flux) layer i. The weighting coefficient e_{ij} characterizes the joint effects of the emissivity at layer j and the absorptivity by all layers between i and j.

Exercise 4.4.3 Write down the analogous trapezoidal-rule approximation to I_-

Exercise 4.4.4 Write down analogous trapezoidal-rule approximations to I_+ and I_- based on the form of the solution given in Eq. 4.12. What would be the advantages of using this form of solution?

To implement Eq. 4.86 and its variants as a computer algorithm, one generally writes a function which computes the transmission between levels p_i and p_j. The rest of the algorithm is independent of the form this function takes, and so one can easily switch from one representation to another (e.g. Malkmus to exponential sums) by simply switching functions. One can equally easily use different representations for different bands. The transmission function requires as arguments the transmission parameters for the band under consideration (e.g. R and S parameters for Malkmus, or the H distribution for exponential sums), as well as enough information to compute the equivalent path. For a well-mixed greenhouse gas, if we are ignoring temperature scaling effects the equivalent path is simply $q_G \frac{1}{2}|(p_1^2 - p_2^2)|/(p_o g \cos\bar{\theta})$, and one can simply make the concentration q_G and the pressures p_1 and p_2 arguments of the transmission function. For an inhomogeneous path, arising when q_G varies with height or one needs to take into account temperature scaling which also varies in height, the path is determined by an integral. In this case, it is inefficient to recompute the path from scratch each time. Since the equivalent path can be computed incrementally using $\ell(p_1, p_2 + \delta p) = \ell(p_1, p_2) + \ell(p_2, p_2 + \delta p)$, it is better to use the equivalent path as an argument to the transmission function, and compute the path from layer i to each layer j iteratively in the same loop in Eq. 4.86 where the weighted emission is computed.

In the preceding algorithm, we have used exponential sums to represent the transmission function appearing in the integral form of the solution to the Schwartzschild Equations. However, because the equations are linear in the fluxes, and because the exponential sum method is a weighted sum of calculations for a number of different absorption coefficients, exactly the same results can be obtained by organizing the calculation in a quite different way. Namely, instead of working from the integral form of the solution, we can work directly with a set of independent Schwartschild equations (Eq. 4.8) – one for each κ going into the exponential sum for a given band; as usual, the band would be chosen narrow enough that $B(\nu, T)$ could be assumed independent of frequency within the band, so we wouldn't need to know anything about which set of wavenumbers each κ correponded to. With a 10-term exponential sum, for example, we would solve the Schwartschild equation for each of the 10 values of κ, then form a weighted sum of the 10 resulting fluxes. This alternate formulation is not available with band-averaged transmission function models such as the Malkmus model. The weighted differential-equation approach offers a number of advantages over the transmission-function approach. A single solution gives the fluxes at all levels, making optimal use of calculations for preceding levels. This is useful when computing heating rate profiles. Moreover, it is easy to use a high-order integrator to obtain high accuracy with fewer levels. These two advantages make the method computationally attractive, but there is a further advantage that is even more compelling for our purposes: it is straightforward to extend this method to incorporate scattering, whereas it is essentially impossible to do so with band-averaged transmission approaches. We will carry out this program in Chapter 5. One could well ask why the weighted differential-equation approach hasn't completely taken over the business of radiation modelling. There is some evidence that, when scattering is unimportant, pressure and temperature scaling can be done more accurately in band-averaged transmission models, but in large measure the transmission models are a holdover from an earlier day when many radiation calculations were done on paper, and when slow computers required highly tuned special approximations in order to speed up the calculation. It does seem that the exponential sum (and its close cousin the *correlated-k*) approach are gradually taking over. We have nonetheless chosen to

introduce the transmission function approach first, because it corresponds better to what is going on in most existing radiation models, and because the form of the solution gives considerable direct insight into the factors governing the fluxes at a given level.

4.4.7 Spectroscopic properties of selected greenhouse gases

Now we will provide a survey of the infrared absorption properties of a few common greenhouse gases, with particular emphasis on the "big 3" that determine much of climate evolution of Earth, Mars and Venus from the distant past through the distant future: CO_2, H_2O and CH_4. In each case, the spectral results shown are for the dominant isotopic form of the gas, e.g. $^{12}C^{16}O_2$ for the case of carbon dioxide. Other isotopic forms (*isotopologues*) can have significantly different spectra, particularly when the substitution of a heavier or lighter isotope changes the symmetry of the molecule, as in HDO or $^{12}C^{16}O^{18}O$. The reader should keep these isotopic effects in mind, since the asymmetric molecules can have have strong absorption in parts of the spectrum where the symmetric molecule has essentially no absorption lines at all. In such cases, the asymmetric isotopologues can be radiatively important even if they are present only in small quantities. When such is the case, one needs to know the isotopic composition of an atmosphere before one can accurately compute the climate. This is a considerable challenge for atmospheres that cannot be directly observed.

This section will deal only with that part of absorption that can be identified as being caused by nearby spectral lines associated with energy transitions of the molecule in question; the *continuum absorption* which is not so directly associated with spectral lines will be discussed separately. Although CH_4 is a greenhouse gas on Titan, most of its contribution there comes from continuum absorption rather than its line spectrum.

Although we focus on some of the more conventional examples, the reader is encouraged to keep an open mind with regard to what might be a greenhouse gas. At present, NO_2 and SO_2 do not exist in sufficient quantities on any known planet to be important as a greenhouse gas, but with different atmospheric chemistries occuring in the past or on as-yet undiscovered planets, the situation could well be different. For that matter, things like SiO_2 that we consider rocks on Earth could be gases and clouds on "roasters" – extrasolar gas giants in near orbits – and there one ought to give some thoughts to their effect on thermal infrared.

When interpreting the absorption spectra to be presented below, it is useful to keep the Planck function in mind. Absorption is not very important where there is little flux to absorb, so the relevant part of the absorption curve varies with the temperature of the planet under consideration. For Titan at $100K$, $\frac{3}{4}$ of the emission is at wavenumbers below 350 cm^{-1}. For Earth at $280K$, $\frac{3}{4}$ of the emission occurs below 1000 cm^{-1}. Mars is slightly colder, and the threshold wavenumber is therefore a bit less. For Venus at $737K$, $\frac{3}{4}$ of the emission is below 2550 cm^{-1}, but moreover the flux beyond this wavenumber amounts to over 4000 W/m^2. This near-infrared flux is vastly in excess of the mere $170W/m^2$ of absorbed solar radiation which maintains the Venusian climate. For Venus one needs to consider the absorption out to higher wavenumbers than for Earth or Mars, and given the large fluxes involved and the huge mass of CO_2 present, exquisite attention to detail and to the effect of weak lines that can normally be ignored is required. For the aforementioned roaster planets, the "thermal" emission extends practically out to the visible range.

An ideal greenhouse gas absorbs well in the thermal infrared part of the spectrum but is transparent to incoming solar radiation. The following survey concentrates on thermal infrared – the part of the spectrum that is involved in OLR– but it should be remembered that some of the gases discussed, such as Methane, absorb quite strongly in the solar near-infrared spectrum.

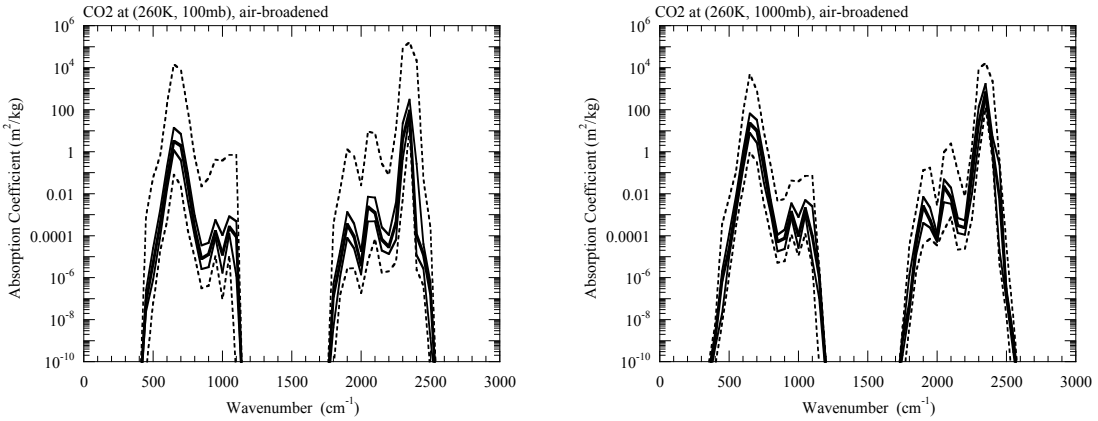

Figure 4.17: The minimum, 25^{th} percentile, median, 75^{th} percentile, and maximum absorption coefficients in bands of width 50 cm^{-1}, computed for CO_2 in air from spectral line data in the `HITRAN` database. The left panel shows results at an air pressure of $100mb$, whereas the right panel gives results at $1bar$. Both are calculated at a temperature of $260K$.

This can compromise their effectiveness as greenhouse gases when they reach concentrations high enough that the solar absorption becomes significant. Even the weaker solar absorption from CO_2 and H_2O can affect the thermal structure of atmospheres in significant ways. The implications of real-gas solar absorption will be taken up in more detail in Chapters 5.

Carbon Dioxide

We have already introduced a fair amount of information about the absorption properties of CO_2 but here we will present the data in a more systematic and general way, which will make it easier to compare CO_2 with other greenhouse gases. In a line graph of a wildly varying function like the absorption, such as shown in Figure 4.7, only only sees the maximum and minimum defining the envelope of the absorption. There is no useful information about the relative probabilities of the values in between. To get around these problems, we divide up the spectrum into slices of a fixed width (50 cm^{-1} in the results presented throughout this section), and then compute the minimum, maximum, median and the 25^{th} and 50^{th} percentiles of the log of the absorption in each band. By plotting these statistics vs. wavenumber, it is possible to present a fairly complete picture of the probability distribution of the absorption. To make the absorption data easier to interpret, we exponentiate the median , quartiles, min and max, and plot the resulting values on a logarithmic vertical axis. Plots of this sort for CO_2 at a temperature of $260K$ are shown in Figure 4.17. The left panel is computed at a pressure of $100mb$ while the right panel is computed at $1000mb$. Both results are for air-broadened lines. As for the other absorption spectra we shall discuss, these results are for the dominant isotopic form of CO_2, i.e. $^{12}C^{16}O_2$. Other isotopic forms can have substantially different properties, particularly the asymmetric forms involving one heavy oxygen and one ligher oxygen.

For Earthlike and Marslike temperatures, only the lower frequency absorption region, from about $400 - 1100cm^{-1}$ affects the OLR, since there is little emission in the range of the higher frequency band. For Earth, the higher frequency band can contribute to Solar absorption, since the Solar radiation reaching the Earth in that part of the spectrum amounts to $7.2W/m^2$ (or about

$3W/m^2$ at the orbit of Mars). However, most of the absorption of Solar near-infrared, whether by water vapor or CO_2, occurs at higher frequencies. Venus, however, has significant surface emission in the range of the higher frequency group, which therefore has a significant effect on that planet's OLR.

The absorption spectrum depends on temperature and pressure, and on whether the lines are broadened by collision with air or by collisions with molecules of the same type (e.g. CO_2 with CO_2 in the present case); rather than present page upon page of graphs, the strategy is to show the results for a single standard temperature and pressure, and make use of appropriate temperature and pressure weighted equivalent paths to apply the standard absorption to a wider variety of atmospheric circumstances. We will adopt $T = 260K$, $p = 100mb$ and air-broadening as our standard conditions when discussing other greenhouse gases in this section. The comparison of the two panels of Figure 4.17 illustrates the nature of the pressure scaling. Increasing the pressure by a factor of 10 shifts the minimum, median and quartiles upward by a similar factor. The main exception to the scaling is the maximum absorption in each band, which is reduced and becomes more tightly clustered to the median. This is the expected behavior, since increased pressure reduces the absorption near the centers of absorption lines. Since this part of the spectrum is rare, and absorbs essentially everything hitting it anyway, the errors in pressure scaling near the line centers are of little consequence. A bit of numerical experimentation indicates that the linear pressure scaling works quite well for pressures below 1 or 2 bars. At higher pressures (certainly by 10 bar) the lines overlap to such an extent that the whole probability distribution collapses on the median, in which case the absorption can be described as a smooth function of the wavenumber, and loses its statistical character. The other gases to be discussed have similar behavior under pressure, and so we will not repeat the description separately for each gas.

To understand a planet's radiation balance, the main thing we need to understand is where in the spectrum the atmosphere is optically thick, and where it is optically thin. Pressure exerts by far the strongest modifying influence on the standard-state absorption coefficients given in Figure 4.17. Hence, we will first discuss the effect of CO_2 on the optical thickness of several planetary atmospheres in terms of the $260K$ air-broadened coefficients using pressure-adjusted paths. Later we will make a few remarks about how self-broadening and temperature affect the results.

The strong line pressure-weighted equivalent path for a well mixed greenhouse gas with specific concentration q is $\ell = \frac{1}{2}q(p_1^2 - p_2^2)/(p_o g \cos\bar{\theta})$. If we take the reference pressure $p_o = 100mb$, the equivalent paths can be multipled by the absorption coefficients in the left panel of Figure 4.17 to obtain optical thickness. Some typical values of the equivalent path are given in Table 4.1. Although Mars at present has much more CO_2 per square meter in its atmosphere than modern Earth, the equivalent paths are quite similar in the two cases because the total pressure on Mars is so much lower. For Earth and Mars temperatures only the lower frequency absorption group is of interest, and within this group it is only the dominant spike in absorption near $675cm^{-1}$ that contributes significantly to the absorption. Both atmospheres have optical thicknesses exceeding unity within the wavenumber band from roughly $610 - 750cm^{-1}$ and are optically thin outside this band. The weak absorption shoulder occurring to the high-wavenumber side of $870cm^{-1}$ has little effect for modern Earth or Mars. For any given wavenumber range, the 10% of the atmosphere nearest the ground (i.e. the lowest $100mb$ on Earth) is one fifth as optically thick as the atmosphere as a whole, meaning that a substantial part of the back-radiation of infrared to the surface attibutable to CO_2 comes from the lowest part of the atmosphere. In contrast, because of low pressure aloft, the uppermost 10% of the atmosphere, which may loosely be thought of as the stratosphere, is comparatively optically thin. If CO_2 were the only factor in play, the greenhouse effect of the present Earth and Mars atmospheres would be quite similar. The two planets are rendered qualitatively different primarily by the greater role of water vapor in the

	Mod. Earth	Early Earth	Mod. Mars	Early Mars	Venus
Whole atm	47.	15000.	30.	10^6	10^9
Bottom 10%	9.	2900.	5.	$2 \cdot 10^5$	$2 \cdot 10^8$
Top 10%	.5	160.	.3	10^4	10^7
Top 1%	.005	1.5	.003	100.	10^5

Table 4.1: Pressure-weighted equivalent paths for various planetary situations, in units of kg/m^2. The weighted path is based on a $100mb$ reference pressure, so these paths are intended to be used with the absorption coefficients in the left-hand panel of Figure 4.17 . A mean slant path $\cos\bar{\theta} = \frac{1}{2}$ is assumed. The Modern Earth case is based on $300ppmv$ CO_2 in a $1bar$ atmosphere, while the Early Earth case assumes 10% (molar) CO_2 in a $1bar$ atmosphere. Modern Mars is based on a $10mb$ pure CO_2 atmosphere, while Early Mars assumes a $2bar$ pure CO_2 atmosphere. The Venus case consists of a $90bar$ pure CO_2 atmosphere.

warmer atmosphere of modern Earth, and by differences in vertical temperature structure between the two atmospheres (arising largely from differences in solar absorption).

On Early Earth – either during the high CO_2 phase after a long-lived Snowball, or during the Faint Young Sun period – the CO_2 may have been 10% or more of the atmosphere by molar concentration. In this case, the equivalent paths are vastly greater than at present. The Earth atmosphere in this case is optically thick from about $520 - 830cm^{-1}$, and the higher wavenumber shoulder just barely begins to be significant. It becomes rapidly more so as the CO_2 molar concentration is increased beyond 10% . This is good to keep in mind when doing radiation calculations at very high CO_2, since many approximate radiation codes designed for the modern Earth only incorporate the effect of the principal absorption feature centered on $675cm^{-1}$. Even with such high CO_2 concentrations, the atmosphere is optically thick in only a limited portion of the spectrum, so that the net greenhouse effect will be modest unless other greenhouse gases come into play. This remark is particularly germane to Snowball Earth, where the cold temperatures allow little water vapor in the atmosphere. Without help from water vapor, CO_2 has only limited power to warm up a Snowball Earth to the point of deglaciation.

On a hypothetical Early Mars with a $2bar$ pure CO_2 atmosphere, the equivalent path is orders of magnitude greater than the Early Earth case. This renders the atmosphere nearly opaque within the wavenumber range $500 - 1100cm^{-1}$. However a great deal of OLR can still escape through the CO_2 window regions, so continuum absorption, water vapor and other greenhouse gases will play a key role in deciding whether the gaseous greenhouse effect can explain a warm,wet Early Mars climate.

Since the equivalent path assumes absorption increases linearly with pressure, the equivalent paths given for the bottom 10% and for the entirety of the Venus atmosphere yield overestimates of the true optical depths, given that the increase of absorption with pressure weakens substantially above $10bars$. Even assuming that the equivalent paths are overestimated by a factor of 10, the implied optical depths for the lowest $9bars$ of the Venus atmosphere, and for the entire Venus atmosphere, remain so huge that these layers can be considered essentially completely opaque to infrared outside the CO_2 window regions. In fact even the top $1bar$ (about 1%) of the Venus atmosphere has an optical thickness greater than unity throughought the $500 - 1100cm^{-1}$ and $1800 - 2500cm^{-1}$ spectral regions. Within these regions, essentially all the OLR comes from the relatively cold top $1bar$ of the atmosphere.

The window region, where CO_2 has no absorption lines, presents a challenge to the explanation of the surface temperature of Venus, particularly since the peak of the Planck function

for the observed Venusian surface temperature lies in this region. In fact, if the $1200 - 1700 cm^{-1}$ band were really completely transparent to infrared, then emission through this region alone would reduce the Venus surface temperature to a mere $355K$ (assuming a globally averaged absorbed solar radiation of $163 W/m^2$). Some energy also leaks out through the low frequency window region, which would reduce the temperature even further. We must plug the hole in the spectrum if we are to explain the high surface temperature of Venus. The high altitude sulfuric acid clouds of Venus play some role, by reflecting infrared back to the surface; there may also be some influence of the less common isotopes of CO_2, since asymmetric versions of the molecule (e.g. made with one ^{16}O and one ^{18}O) can have lines in the window region. The principal factor at play is the *continuum absorption* to be discussed in Section 4.4.8. Figure 4.17 was computed taking into account only the relatively nearby contributions of each absorption line (within roughly 1000 line widths), whereas collisions can allow some absorption to occur much farther from the line centers. This far-tail absorption spills over into the window region, particularly at high pressure. It cannot be reliably described in terms of the Lorentz line shape, and therefore requires separate consideration.

The difficulties posed by Venus do not end with the window region¿ For planets having Earthlike temperatures, the thermal radiation beyond $2300 cm^{-1}$ is insignificant, so we needn't be too concerned about the absorption properties there. However, for a planet with a surface temperature of $700K$, the ground emission on the shortwave side of $2300 cm^{-1}$ is 3853 W/m^2, compared to under $200 W/m^2$ of absorbed solar radiation for Venus. In these circumstances it simply won't do to assume the atmosphere transparent at high wavenumbers The `HITRAN` spectral database is a monumental accomplishment, but it is still rather Earth-centric, and lacks the weak lines needed to deal with high temperature atmospheres; there may also be continuum absorption in the shortwave region, especially at high pressures. In order to deal with the high wavenumber part of the Venus problem, one needs to employ specialized high-temperature CO_2 databases, which are less verified and more in a state of flux than `HITRAN`. Among specialized databases in use for Venus, the `HITEMP` database (described in the supplementary reading at the end of this Chapter) is particularly convenient, because it at least uses the same data format as `HITRAN`. It will be left to the reader to explore the use of this database. Suffice it to say that there appears to be sufficient absorption in the high wavenumber region to raise the radiating level to altitudes where the temperature is low enough that one may not need to consider shortwave emission in computing OLR.

The preceding discussion was based on air-broadened absorption at $260K$, whereas self-broadened data would be more appropriate to the pure-CO_2 atmospheres and in all cases one must think about whether the increase of line strength with temperature substantially alters the picture presented. A re-calculation of Figure 4.17 for self-broadened pure CO_2 indicates that the self-broadened absorption is generally about 30% stronger than the air-broadened case,though there are a few bands in which the enhancement is as little as 13%. This is quantitatively significant, but the enhancement factor is too small to alter the general picture presented above. The temperature dependence can have a more consequential effect. In the strong line approximation, valid away from line centers, the temperature affects the absorption only in the form of the product of line strength and line width, $S(T)\gamma(p,T)$, which yields a temperature dependence of the form $\kappa_{CO2} \sim \exp(-T^*/T)$ for some coefficient T^*. The coefficient differs from line to line, but we can still attempt to fit this form to the computed temperature dependence of the median absorption within each $50 cm^{-1}$ band. The result is shown in Figure 4.18. This still only gives an incomplete picture of the effect of temperature on absorption, since the other quartiles may have different scaling coefficients. Within an exponential-sum framework, however, one can do little else than pick a base temperature most appropriate to the planet under consideration, compute the probabilility distribution for that case,and then assume that each absorption coefficient in a band scales with the same function of temperature. This procedure in any event gives an estimate of the magnitude

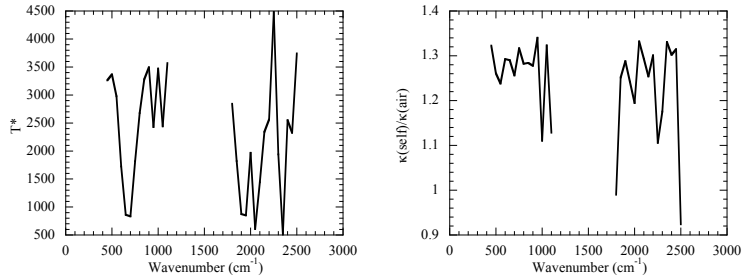

Figure 4.18: Left panel: Temperature dependence coefficient for air-broadened CO_2 at $100mb$. The coefficient is computed for the median absorption in bands of width $50cm^{-1}$, as shown in Figure 4.17. The median absorption within each band increases with temperature in proportion to $\exp(-T^*/T)$. Right panel: The ratio of self-broadened to air-broadened median absorption coefficients.

of the temperature effect. From the Figure, it is seen that the temperature dependence varies greatly with wavenumber. It is low near the principal absorption spike at $675cm^{-1}$, with T^* as low as $1000K$. This value increases the absorption by a factor of 1.7 going from $260K$ to $300K$, and decreases absorption by a factor of 2 going from $260K$ to $220K$. Where $T^* = 3000K$, the corresponding ratios are 4.7 and .14 . Therefore CO_2 is significantly more optically thick than our previous estimates in the warmer-lower reaches of the present Earth atmosphere, and significantly less optically thick in the cold tropopause regions. For the $737K$ surface temperature of Venus, the absorption is enhanced by a factor of 12 when $T^* = 1000K$ but by a substantial factor of 1750 when $T^* = 3000K$. This just makes an already enormously optically thick part of the atmosphere even thicker. Outside the window regions, most of the atmosphere of Venus is in the optically thick limit where very slow radiative diffusion transfers heat; infrared radiative cooling in the deeper parts of the Venus atmosphere is determined almost exclusively by what is going on in the window regions.

Water Vapor

Figure 4.19 shows the standard-state absorption spectrum for water vapor. Unlike CO_2, the H_2O molecule, which has more complex geometry, has lines throughout the spectrum, so there is no completely transparent window region. Water vapor nonetheless has two window regions where the absorption is very weak; it will turn out that continuum absorption from far tails excluded from the computation in the figure substantially increases the absorption in these window regions (Section 4.4.8). The peak absorption coefficient for water vapor has a similar magnitude to that for CO_2, but water vapor absorbs well over a far broader portion of the spectrum than CO_2. In particular, the H_2O absorption has a peak within both the $1000cm^{-1}$ and longwave window regions of CO_2. This critically affects the greenhouse warming on Earth and on Early Mars, but it plays little role on present-day Venus, which has little water vapor in its atmosphere. It should not be concluded that water vapor overwhelms the greenhouse effect of CO_2, however. It would be more precise to say that the water vapor greenhouse effect complements that of CO_2. CO_2 absorbs strongly near the peak of the Planck function for Earthlike temperatures, but the water vapor absorption is nearly two orders of magnitude weaker there. Further, for Earthlike planets, water vapor condenses and therefore disappears in colder regions of the planet; it is only the long-lived CO_2 greenhouse effect that can persist in cold parts of the atmosphere.

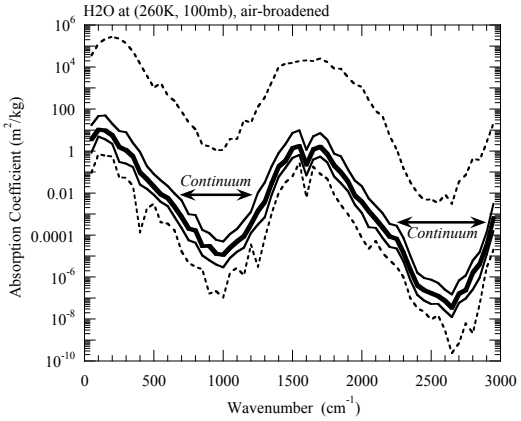

Figure 4.19: As in Figure 4.17, but for H_2O. The continuum regions marked on the figure are regions where the measured absorption substantially exceeds that computed from the spectral lines.

On a planet without a substantial condensed water reservoir, water vapor could be a well-mixed noncondensing greenhouse gas much like CO_2 on modern or early Earth. In most known cases of interest, though, atmospheric water vapor is in equilibrium with a reservoir –an ocean or glacier – which fills the atmosphere to the point that the atmospheric water vapor content is limited by the saturation vapor pressure. The prime cases of interest are water vapor feedback on Earth and on Early Mars, and the runaway greenhouse on Early Venus. The runaway greenhouse is also relevant to the ultimate far-future fate of the Earth, and the evolution of hypothetical water-rich extrasolar planets. The status of water vapor as a greenhouse gas whose concentration is limited,via condensation, by temperature does not derive from any special properties of water. One tends to focus on a role of this sort for water simply because planets that are "habitable" to the only example of life with which we are presently familiar seem to require a planet with liquid water and an operating temperature range in which the saturation vapor pressure is high enough for water vapor to be present in sufficient quantities to be active as a greenhouse gas. On present and past Mars, as well on a hypothetical early Snowball Earth, CO_2 can be limited by condensation, and on Titan today CH_4 condenses, while NH_3 and other gases have condensation layers on the gas giant planets.

First, let's think about the effect of water on OLR, supposing that the atmosphere is saturated at each altitude. The water vapor greenhouse effect is determined by a competition between two factors. Water vapor causes the greatest optical thickness near the ground, where both pressure-broadening and saturation vapor pressure are highest. However, a strong greenhouse effect requires optical thickness at higher altitudes, where the temperature is substantially colder than the surface, in order to reduce the radiating temperature of the planet. For this reason, water vapor in the mid to upper troposphere is more important than water vapor near the ground. Consider a typical Earth tropical case, on the moist adiabat with $300K$ temperature near the ground. If the low level air is saturated, then the equivalent path of the lowest $100mb$ of the atmosphere is about $400kg/m^2$. This is sufficient to make the lower atmosphere optically thick except within the window regions, so that the atmosphere would radiate to the ground at the near-surface air temperature except within the windows. Most of the infrared cooling of the ground occurs in the windows, but we'll see eventually that at temperatures of $300K$ and above, the continuum substantially closes off the window region as well. This near-surface opacity doesn't much affect

the OLR, however. Near $400mb$, the temperature is about $260K$,and in saturation the equivalent path between $400mb$ and $500mb$ is only $25kg/m^2$. This relatively small amount of water vapor is still sufficient to make the layer optically thick between $1350cm^{-1}$ and $1900cm^{-1}$, as well as on the low frequency side of $450cm^{-1}$. This substantially reduces the OLR. At the tropopause level, in contrast, the temperature is about $200K$, the pressure is $100mb$, and the equivalent path from $100mb$ to $200mb$ is only $.01kg/m^2$. At such low concentrations, water vapor is optically thin practically throughout the spectrum. At lower surface temperatures, the dominant greenhouse effect of water vapor comes from correspondingly lower altitudes. With a $273K$ surface temperature, the $400mb$ temperature on the dry air adiabat is only $210K$, and the water vapor opacity is inconsequential there. One has to go down to about $600 - 700mb$, where the path is about $2kg/m^2$, to get a significant water vapor greenhouse effect. On a low CO_2 Snowball Earth soon after freeze-up, where the tropical surface temperature is under $250K$, the water vapor greenhouse effect is essentially negligible. The water vapor greenhouse effect only starts to play a role when the planet has warmed to the point that the tropical temperatures approach the melting point.

As was the case for CO_2, the absorption coefficient for water vapor decays roughly exponentially with distance in wavenumber space from each peak of absorption. This has similar consequences for OLR as were discussed previously in connection with CO_2. Because of the exponential envelope of absorption, doubling or halving the water vapor content of a layer of the atmosphere has approximately the same effect on the optical thickness of that layer regardless of whether the base amount being doubled or halved is very large (say 200 kg/m^2) or very small (say 2 kg/m^2). There may not be much water to work with in the Earth's mid troposphere, but nonetheless halving or doubling the amount would have a significant effect on OLR. This remark is particularly significant because there are dynamical effects which in fact keep the Earth's mid-troposphere substantially undersaturated. Although there are regions with relative humidity as low as 10%, it would still significantly increase the OLR if the relative humidity were reduced still more to 5%, and conversely it would significantly decrease the OLR if the relative humidity were increased to 20%. Because of the spectral position of the absorption peaks relative to the shape of the Planck curve, the effect of water vapor concentration on OLR is not as precisely logarithmic as is the case for CO_2. Nonetheless, it is fair to say that the change in water vapor content *relative to the amount initially present* gives a more true idea of the radiative impact of the change than would the change in the absolute number of kilograms of water present in a layer.

With regard to water vapor, then, it is clear that subsaturation is important. However, the determination of the degree of subsaturation involves intrinsically fluid dynamical processes, and we will not have much to say about this important issue in this book. Similar considerations would apply to any radiatively active condensible substance in a planetary atmosphere.

Figure 4.20 shows the temperature dependence and the ratio of self to air-broadened absorption for water vapor. The general range of temperature sensitivity is much the same as it was for CO_2. However, whereas self-broadened absorption for CO_2 is only a few tens of percents stronger than air-broadened absorption, the self-broadened H_2O absorption is fully five to seven times stronger than the air-broadened case. This is extremely important to the runaway greenhouse, which involves portions of the atmosphere which consist largely of water vapor. Moreover, when a species has a molar concentration in air of, say, 10% or less one wouldn't ordinarily have to worry much about self-broadening, since collisions with air are so much more common than self-collisions. However, because of the great amplification of self-broadened absorption for water vapor, the self-broadening in fact starts to become dominant even at molar concentrations of around 10%. At the Earth's surface, this concentration is achieved at a temperature of $320K$. With a dry air partial pressure of $100mb$, this concentration would be achieved at temperatures near $280K$; this situation is relevant to a hot planet on the verge of a runaway greenhouse. The strong

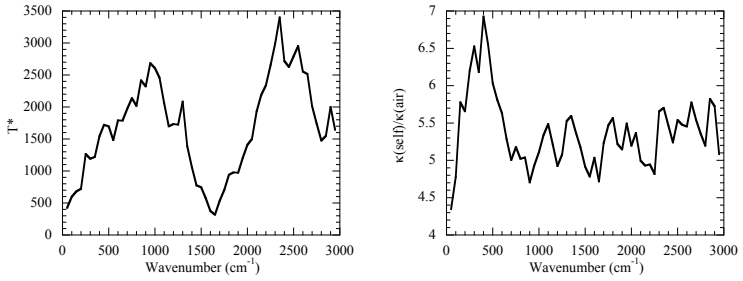

Figure 4.20: Left panel: Temperature dependence coefficient for air-broadened H_2O at $100mb$, as defined in Fig. 4.18. Right panel: Ratio of self-broadened to air-broadened median absorption for H_2O at $100mb$ and $260K$. As usual, medians are computed in bands of width $50cm^{-1}$.

enhancement of self-broadened absorption relative to air-broadened absorption is far in excess of what would be anticipated from simple effects associated with the different molecular weights of the colliders. The sensitivity to the nature of the colliding molecule raises interesting questions about the effect of collisions with other molecules. Would CO_2 broadened coefficients be more like the air-broadened or self-broadened case? The answer to this question has some impact on the climates of Early Mars and Early Earth, which are often assumed to have had substantial amounts of CO_2 in their atmospheres. Unfortunately, laboratory measurements bearing on the subject are hard to come by.

Next we turn our attention to aspects of the absorption which govern the runaway greenhouse effect for a planet with a water-saturated atmosphere. Specifically, we revisit the question of the Kombayashi-Ingersoll limit, which is the limiting OLR such a planet can have in the limit of large surface temperature. As discussed for the grey gas case in Section 4.3.3, the limiting OLR is approximately determined by the temperature at the pressure level where the optical thickness between that level and the top of the atmosphere becomes unity. For a grey gas this characteristic pressure was independent of wavenumber, whereas for a real gas it is quite strongly wavenumber dependent. To keep things simple, we'll consider a saturated pure water vapor atmosphere. In this case, if the temperature at a given altitude is T, the corresponding pressure is $p_{sat}(T)$, determined by Clausius-Clapeyron. The equivalent path from this altitude to zero pressure is $a_{self}\frac{1}{2}p_{sat}(T)^2/(p_o g\cos\bar{\theta})$, where p_o is the standard reference pressure ($100mb$ for use with Figure 4.19) and a_{self} is the ratio of self-broadened to air-broadened absorption (about 6). By using this path together with the absorption coefficients in Figure 4.19, we can estimate the maximum effective radiating temperature as a function of wavenumber. Based on the median absorption in each band, the radiating temperature varies from about $245K$ at $100cm^{-1}$ to $278K$ at $500cm^{-1}$ to $350K$ at the valley of the window region near $1000cm^{-1}$. The high values of radiating temperature in the window region lead to large estimates of the Kombayashi-Ingersoll limit. As a crude estimate, if we assume that the planet radiates at $350K$ in the window regions between 544 and 1314 cm^{-1}, and on the high wavenumber side of $1950cm^{-1}$, then the OLR would be about $520W/m^2$ even if the planet didn't radiate at all in the rest of the spectrum. The Kombayashi-Ingersoll limit is strongly affected by the absorption in the window regions, and we will see in Section 4.4.8 that the absorption here is dominated by a continuum which is not captured by the nearby line contribution. The estimate of the Kombayashi-Ingersoll limit for H_2O (and CO_2) will be completed in Section 4.6, after we have discussed the continuum.

Methane

Figure 4.21 shows the standard-state absorption spectrum for methane. From the standpoint of OLR on Modern and Early Earth, and perhaps on Early Mars, the most important absorption feature is that near 1300 cm^{-1}, which occurs in a part of the spectrum where water vapor and CO_2 absorption are weak, but where the Planck function still has significant amplitude at Earthlike temperatures. In contrast with water vapor, the very long wave absorption (below 1000 cm^{-1}) is so weak that it does not significantly affect OLR in any atmosphere likely to have existed on an Earthlike or Marslike planet. Titan has extremely large amounts of methane in its atmosphere, which could in principle make the very longwave group important. Even there, however, the weakening of the absorption due to the very cold temperatures makes this absorption group fairly insignificant. For atmospheres containing appreciable amounts of oxygen, methane oxidizes rather quickly to CO_2, so it is hard to build up very large concentrations. The Earth's pre-industrial climate had about $1ppmv$ of methane in it, and the intensive agriculture of the past century may eventually come close to doubling this concentration. The associated equivalent paths are quite small – on the order of $.06kg/m^2$. For paths this small, Earth's atmosphere has an optical thickness of only .14 based on the median absorption coefficient occurring the $1300cm^{-1}$ peak. For methane paths typical of oxygenated atmospheres, one gets significant absorption only from the upper quartile of absorption coefficients, and a short distance from the dominant peak one only gets significant absorption very near the line centers. In this case, the ditch in OLR dug by methane is a very narrow feature centered on $1300cm^{-1}$. In an anoxic atmosphere, as Earth's is likely to have been earlier than about 2.7 billion years ago, the rate of methane destruction is much lower, and it is believed that production of methane by methanogenic (methane-producing) bacteria could have driven methane concentrations to quite large values. With $100ppmv$ of methane in an Earthlike atmosphere, the equivalent path is about $.6kg/m^2$, and based on the median absorption coefficient the atmosphere becomes optically thick from about $1200 - 1400cm^{-1}$. If the methane concentration builds up to 1% of the atmosphere, then the equivalent path is nearly $600kg/m^2$, and the atmosphere becomes optically thick from $1150 - 1750cm^{-1}$; for such high concentrations, the shoulder to the right of the absorption peak starts to become important. There are also some speculations that abiotic processes could have led to high Methane concentrations on Early Mars, or on the prebiotic Earth.

Beyond what is shown in Figure 4.21, Methane has strong absorption bands that extend well into the Solar near-IR. These are not terribly important at concentrations up to a few hundred $ppmv$ in a $1bar$ atmosphere, but when Methane makes up a percent or so of the atmosphere, it can absorb most of the incident solar energy between 2500 and 9000 cm^{-1}. At higher concentrations, significant absorption can extend even into the visible range.

It is often said that, molecule for molecule, CH_4 is a better greenhouse gas than CO_2. However, this is more a reflection of the relative abundances of CH_4 and CO_2 in the present Earth atmosphere than it is a statement about any intrinsic property of the gases; in fact, the absorption coefficients for the two gases are quite similar in magnitude, and CH_4 absorbs in a part of the spectrum that is less well placed to intercept outgoing terrestrial radiation than is the case for CO_2. The high effectiveness of CH_4 relative to CO_2 in the present atmosphere of Earth stems from the fact that currently there is rather a lot of CO_2 in the air ($380ppmv$ and rising) but rather little CH_4 ($1.7ppmv$ and also rising). In a situation like this, one has already depleted infrared of those frequencies that are most strongly absorbed by CO_2, so when adding CO_2 one is adding "new absorption" in spectral regions where the absorption is relatively weak. Hence, it takes a large amount of the gas to have much radiative effect. In contrast, when starting with a small amount of CH_4, when one adds more, one adds "new absorption" where the absorption

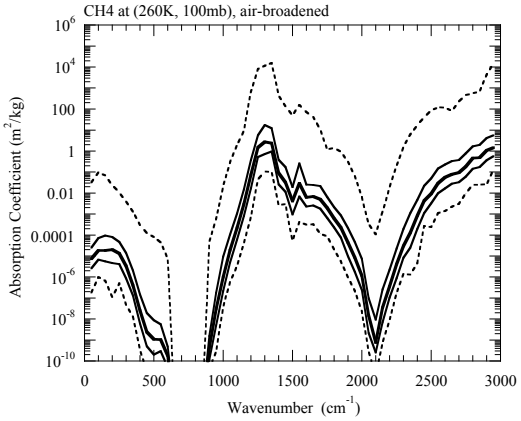

Figure 4.21: As in Figure 4.17, but for CH_4.

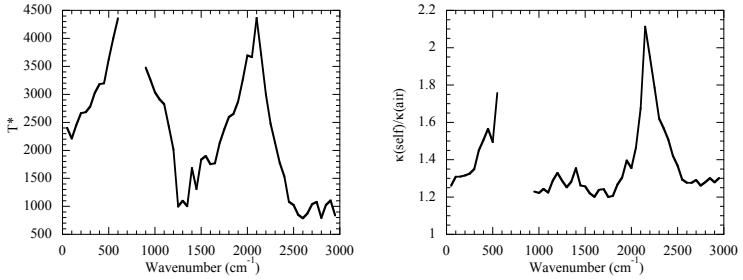

Figure 4.22: Left panel: Temperature dependence coefficient for air-broadened CH_4 at $100mb$, as defined in Fig. 4.18. Right panel: Ratio of self-broadened to air-broadened median absorption for CH_4 at $100mb$ and $260K$. As usual, medians are computed in bands of width $50cm^{-1}$.

coefficient is quite strong, since the strongly absorbing part of the spectrum is not yet depleted. This behavior depends crucially on the lack of significant overlap between the Methane and CO_2 absorption regions.

The temperature scaling coefficient for CH_4 is shown in Figure 4.22. It is in the same general range as for the cases discussed previously. The ratio of self to air broadening for CH_4 is similar to that for CO_2, throughout the spectral range of most importance for OLR. In the solar near-IR the self-broadened absorption can be nearly twice the air-broadened absorption. For Methane, the self-broadening is mostly of academic interest, since the methane concentration is too low for self-collisions to be significant in atmospheres encountered or envisioned so far. Titan and similar cryogenic atmospheres are potentially an exception to this remark, but there the absorption is dominated by a continuum that is not clearly related to the absorption lines under consideration here.

Because Titan has a surface temperature on the order of $100K$, the peak of the Planck function occurs at about the third of the wavenumber where the peak is for Earthlike temperatures. In consequence, there is little thermal emission in the vicinity of the dominant $1300cm^{-1}$ absorption group. It is only the longwave absorption group that is potentially of interest. The

lower atmosphere of Titan contains up to 20% CH_4, which with Titan's low surface gravity yields an actual (*not* equivalent) mass path of nearly $15000 kg/m^2$. However, due to the strong reduction of line strength with temperature, the median absorption is well under $10^{-6} m^2/kg$ in the longwave group, even when broadened by a pressure of $1.5 bar$. The radiative effect of Methane on Titan arises mainly from a collisional continuum of the sort described in Section 4.4.8.

4.4.8 Collisional continuum absorption

Diatomic molecules and general considerations

A Nitrogen molecule N_2 in isolation does not interact to any significant extent with infrared light; one might think that collisions do not change this picture, as N_2 has no lines to be broadened by collisions. Nonetheless, *during the time a collision is taking place* the pair of colliding molecules momentarily behaves somewhat like a more complex four-atom molecule, which has transitions that can indeed absorb and emit infrared radiation. This leads to *collision-induced absorption*, whose associated absorption coefficient is generally a smooth function of wavenumber. Because of the lack of line structure, such absorption is referred to as a *continuum*. There are many possible processes through which collisions can induce absorption. The collision can impart a temporary dipole moment to a rotation or vibration that ordinarily had none, allowing it to absorb or emit a photon. The collision can break a symmetry, allowing transitions that are otherwise "forbidden" by symmetry principles. Colliding molecules can form *dimers*, which are short-lived complexes which nevertheless persist long enough to have radiatively active transitions not present in the colliding molecules. Most of the "non-absorbing" diatomic molecules, including N_2 and H_2, exhibit significant collision-induced continuum absorption at sufficiently high densities. These are mostly associated with the induced-dipole mechanism, and therefore can to some extent be anticipated on the basis of the underlying transitions of the diatomic molecule.

Collision-induced absorption can be thought of as a trinary chemical reaction involving the two colliding molecules and a photon. The rate of "reaction" (i.e. absorption) is proportional to the product of the concentrations of the two colliding species with the photon concentration, the latter being proportional to the radiation flux. The absorption coefficient is the rate constant for the reaction. Unlike the case of collision-*broadened* line absorption, in collision-*induced* absorption there is no physical distinction between the "absorbing" molecule and the "perturbing" molecule. Both are equal partners in the process allowing absorption or emission of a photon. For this reason, it is most natural to describe collision-induced absorption in terms of a binary absorption coefficient, which expresses the proportionality between the product of the concentrations of the two colliding species and the rate of absorption of radiation. Nonetheless, in order to facilitate comparison with the previously defined line absorption coefficients, and in order to make it easier to incorporate collisional continuum absorption in radiation calculations which also take into account line absorption, it is convenient to characterize the collision-induced absorption by mass-specific absorption coefficients in which one of the colliding molecules is arbitrarily designated the "absorber," whose absorption is enhanced in proportion to the partial pressure of the "collider". For example, for an N_2-H_2 collision in a box of gas with uniform temperature T and uniform N_2 partial pressure p, the optical thickness can be expressed as

$$\tau = \frac{p_{N2}}{p_o} \kappa_{H2}(\nu, p_o, T) \ell_{H2} \tag{4.87}$$

where ℓ_{H2} is the mass path of Hydrogen in the box, in kg/m^2, and p_o is a standard pressure. The coefficient κ_{H2} has dimension m^2/kg and can be used in precisely the same way as the absorption

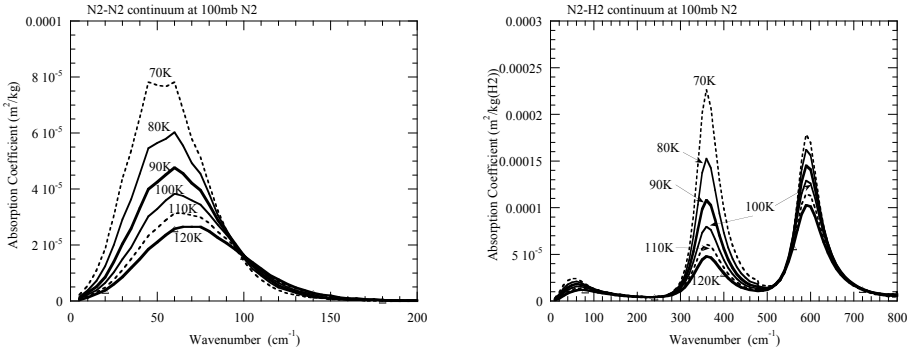

Figure 4.23: The $N_2 - N_2$ and $N_2 - H_2$ collision-induced continuum absorption coefficients as a function of temperature (indicated on curves). The coefficients are given at an N_2 partial pressure of $100mb$.

coefficients we defined earlier for use with the line spectrum. To relate this to the binary absorption coefficient commonly defined in the spectroscopy literature, let z be the length of the path in meters, so that $\ell_{H2} = z \cdot p_{H2}/(R_{H2}T)$. Then

$$\tau = \frac{\kappa_{\nu,p_o,H2}}{p_o R_{H2} T} p_{N2} p_{H2} z \equiv \kappa_{H2-N2}(\nu, T) p_{N2} p_{H2} z \tag{4.88}$$

which defines the binary coefficient in the case where the collider amounts are specified as partial pressures. It is more common to specify the collider amounts in terms of densities or molar densities, but the alternate forms can be readily derived from the preceding by use of the ideal gas law.

Continuum absorption may be difficult to understand from *a priori* physical principles, and difficult to measure accurately in the laboratory, but by definition the absorption coefficient for the continuum is a smoothly varying function of wavenumber. Therefore, it is relatively easy to incorporate into radiative transfer models. One only needs to determine the absorption coefficients and their pressure and temperature scaling in a set of relatively broad bands, and multiply the transmission computed from the line absorption (if any) by the corresponding exponential decay factor.

The continuum arising from diatomic molecule collisions becomes particularly important for dense, cold, massive atmospheres, of which Titan's is probably the best studied example. Figure 4.23 shows the $N_2 - N_2$ and $N_2 - H_2$ collision induced absorption coefficients in the temperature range prevailing in Titan's atmosphere. These coefficients are based on laboratory measurements made at somewhat higher temperatures, extrapolated to colder values using a theoretical model with a few empirical coefficients fit to the data. (See the paper by Courtin *et al.* listed in the Further Readings section of this chapter). The equivalent path for Titan based on N_2 partial pressure is about $10^6 kg/m^2$, which yields a peak optical thickness of 40 for temperatures near those prevailing at Titan's surface. The H_2 content of Titan's atmosphere is less well constrained, but plausible estimates suggest that this gas, too, can contribute significantly to the infrared opacity of Titan's atmosphere. Note that the absorption decreases sharply with increasing temperature; this is partly due to the decrease in density with temperature, but is also affected by the shorter duration of high-velocity collisions, which apparently are less effective at inducing a dipole moment. The $N_2 - N_2$ continuum is unimportant for Earthlike collisions because of the higher temperatures on Earth, and because Earth's atmosphere is much less massive than Titan's,per unit surface area.

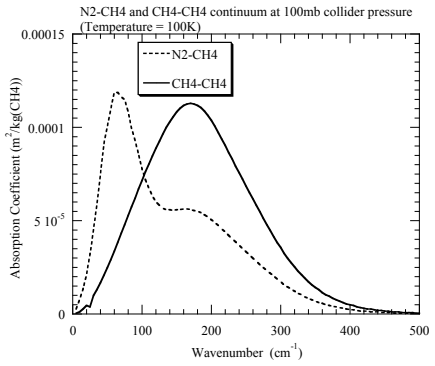

Figure 4.24: The $N_2 - CH_4$ and $CH_4 - CH_4$ collision-induced continuum absorption coefficients at $100K$, assuming a collider partial pressure of $100mb$. For $N_2 - N_2$ the "collider" is N_2, while for $CH_4 - CH_4$ the "collider" is CH_4.

H_2 also has a significant self-broadened continuum, which provides a great deal of the infrared opacity on the gas giant planets.

Methane Continuum on Titan

Methane also significantly affects the infrared opacity of Titan's atmosphere, though its effects on OLR are rather less than N_2 or H_2 since methane is concentrated in the warmer, lower layers of the atmosphere. Essentially all of the infrared absorption due to methane on Titan comes from a collision-induced continuum. The N_2-induced and self-induced absorption coefficients at $100K$ are shown in Figure 4.24. Like the other continuum coefficients, these too become weaker with increasing temperature. Assuming Methane to be in saturation with temperature given by the Methane-Nitrogen moist adiabat in Titan conditions, the pressure-weighted path for self-induced absorption is in excess of $40000kg/m^2$, while the equivalent path for N_2-induced absorption is over $170000kg/m^2$. The self-absorption yields a peak optical thickness of about 5, while the foreign absorption gives a peak optical thickness of about 20, dropping to 10 in the higher frequency shoulder near $200cm^{-1}$

Carbon Dioxide Continuum

Given the importance of the CO_2 window regions to the high-CO_2 climates of Venus and Early Mars, it is rather surprising that the CO_2 continuum has been so little studied. The coefficients in use in most models stem from limited experiments and there is little agreement on the theoretical basis for this continuum or its temperature scaling. At the time of writing, it appears that the subject has not been re-examined in laboratory experiments since the late 1970's. The discussion below is based on absorption coefficients reported in the literature cited in the Further Readings section of this chapter.

The measured CO_2 continuum absorption, rescaled to $100mb$ is shown in Figure 4.25. The values shown are for collisions of CO_2 in air; the self-induced continuum absorption is generally assumed to be 1.3 times that of the foreign-induced continuum. Referring to the equivalent paths in Table 4.1, we see that the continuum absorption is large enough to make the top one bar of

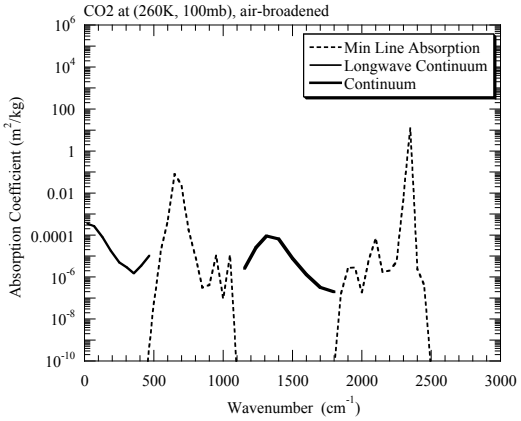

Figure 4.25: The air-induced CO_2 continuum absorption (solid lines) compared with the bandwise-minimum absorption computed from the line spectrum (dashed lines).

the atmosphere of Venus optically thick throughout most of the window region. The continuum absorption is strong enough to be important in the thick atmosphere of Early Mars, but only marginally so for the more moderate CO_2 levels present on Early Earth.

When incorporating the continuum in radiation models, it may be more convenient to work from a curve-fit rather than tabulated data. The absorption coefficient for the CO_2 continuum can be fit with the function

$$\kappa_{CO2}(\nu, 300K, 100mb) = \exp(-8.853 + 0.028534\nu - 0.00043194\nu^2 + 1.4349 \cdot 10^{-6}\nu^3 - 1.5539 \cdot 10^{-9}\nu^4) \tag{4.89}$$

from 25 to 450 cm^{-1} and by

$$\kappa_{CO2}(\nu, 300K, 100mb) = \exp(-537.09 + 1.0886\nu - 0.0007566\nu^2 + 1.8863 \cdot 10^{-7}\nu^3 - 8.2635 \cdot 10^{-12}\nu^4) \tag{4.90}$$

from 1150 to 1800 cm^{-1} where ν is measured in cm^{-1}. The continuum absorption coefficients *weaken* with increasing temperature, according to the empirical power law $(300./T)^n$, with $n = 1.7$ for wavenumbers greater than $190cm^{-1}$, increasing to 1.9 at 130 cm^{-1}, 2.2 at 70 cm^{-1} and 3.4 at 20 cm^{-1}.

The CO_2 continuum is poorly characterized experimentally, and not well understood theoretically. The continuum fits provided above are based on rather dated experiments, and there are some indications that a part of the above continuum may actually represent line spectra from asymmetric isotopologues of CO_2 such as $C^{16}O^{18}O$. This is an area where more experiments using modern techniques are badly needed.

Water vapor continuum

Since water vapor has absorption lines throughout the spectrum, it is hard to unambiguously define the continuum. Laboratory measurements clearly show, however, that in the window regions indicated in Figure 4.19 the net absorption is far in excess of what can be accounted for by the contributions of nearby lines. The prevailing view currently is that this excess absorption is due to the very far tails of the stronger absorption bands flanking the window regions, rather than dimers,

forbidden transitions, or collision-induced dipole moments. The theoretical and observational basis for this viewpoint is exceedingly weak, however. In the following we will confine ourselves to empirical descriptions of the laboratory measurements, without reference to underlying mechanisms. Comparisons with direct measurements of transmission in the Earth's atmosphere have confirmed that the laboratory measurements provide an adequate basis for modelling water vapor absorption in the window regions. The laboratory measurements show that the air-broadened or N_2 broadened water vapor continuum is very weak, so that the window region absorption is by far dominated by self-collisions of water vapor. The following discussion will therefore be limited to self-induced absorption; the characterization of foreign-induced absorption by CO_2 appears to be an open question at present, though it is potentially of importance to water-CO_2 atmospheres such as might have occurred on Early Mars.

There is some ambiguity in the spectroscopic literature as to how to define the water vapor continuum, given that in analyzing measurements one must be careful not to be thrown off by the strong absorption near the centers of individual lines in the window regions. Most useful definitions of the continuum amount to reading the absorption at the minima "between the lines." The results of such a measurement of the self-induced continuum are shown in Figure 4.26. The measurements were made for water vapor in saturation at $296K$, with water vapor partial pressure of about $28mb$, but have been scaled to a standard water vapor partial pressure of $100mb$ for the sake of discussion. We'll focus on the lower frequency of the two window regions, since that is by far the most important for planetary climate calculations. Similar data exists for the higher frequency window. From the figure, we see that the measured continuum absorption is several orders of magnitude stronger than the typical line contribution. To get an idea of the significance of the water vapor continuum, let's consider a layer of air of depth z, with uniform temperature T, within which the water vapor is at the saturation vapor pressure corresponding to T. Since the water vapor continuum is dominated by self-collisions, it matters little what the background air pressure is in this layer. The equivalent path for this layer is $(p_{sat}(T)/p_o)(p_{sat}(T)/(R_wT))z$; the first factor gives the degree of pressure-induced enhancement of absorption relative to the standard, while the secondd factor is the density of water vapor in the layer. Note that the equivalent path is *quadratic* in the water vapor partial pressure. For this reason, the optical thickness in the continuum region grows very rapidly with temperature. At $300K$, then, with a layer depth of $1km$ the equivalent path is $9.3kg/m^2$. Since the minimum absorption coefficient in the window regionis about $.1m^2/kg$, this path gives the layer an optical thickness of unity or more in the window region. Since the absorption is even stronger outside the window region, at the lowest layer of the Earth's atmosphere acts practically like an ideal blackbody at tropical temperatures. At $310K$ the equivalent path increases to $27kg/m^2$, so the window region closes off even more. This has profound consequences for the runaway greenhouse. In fact, there would be essentially no prospect for a runaway greenhouse even in Venusian conditions were it not for the water vapor continuum.

Over the range of wavenumbers shown in the figure, the water vapor continuum absorption can be fit by the polynomial

$$\kappa_{H2O}(\nu, 296K, 100mb) = \exp(12.167 - 0.050898\nu + 8.3207\cdot 10^{-5}\nu^2 - 7.0748\cdot 10 - 8\nu^3 + 2.3261\cdot 10^{-11}\nu^4) \tag{4.91}$$

Like the other continua, the water vapor continuum absorption becomes weaker as temperature increases. Data on the temperature dependence is sparse, but suggests a temperature dependence of the form $(296/T)^{4.25}$. For temperatures much colder than $300K$, the saturation vapor pressure is so low that the details of the temperature dependence are unimportant. As temperature increases beyond $300K$, the exponential growth of saturation vapor pressure is far more important to the optical thickness than the rather mild decline of the continuum absorption coefficient with temperature.

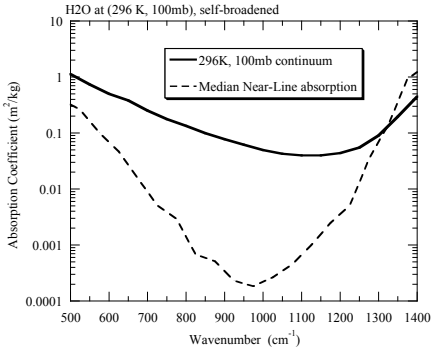

Figure 4.26: The water vapor self-induced continuum near 1000 cm^{-1}, compared with the median absorption coefficient computed from the Lorentz line contribution within 1000 line widths of the line centers in the `HITRAN` database. The continuum curve given is based on laboratory observations in saturation at $296K$, scaled up to what they would be at a water vapor partial pressure of $100mb$. See citations in the Further Readings section of this chapter for data sources.

Water vapor has another continuum region at shorter wavelengths, in the vicinity of $2500cm^{-1}$. This is not important for Earthlike temperatures, but it is a very significant factor for the hotter temperatures encountered in runaway greenhouse calculations. At temperatures much in excess of $320K$, there is enough emission in this region that it accounts for a significant part of the infrared cooling if the continuum is not included. The $2500cm^{-1}$ continuum, covering $2100cm^{-1} \leq \nu \leq 3000cm^{-1}$ can be represented by the polynomial fit

$$\kappa_{H2O}(\nu, 296K, 100mb) = \exp(-6.0055 - 0.0021363x + 6.4723 \cdot 10^{-7}x^2 - 1.493 \cdot 10 - 8x^3 + 2.5621 \cdot 10^{-11}x^4 + 7.328 \cdot 10^{-14}x^5) \tag{4.92}$$

where $x = \nu - 2500cm^{-1}$. The scaling in pressure and temperature can be taken to be the same as for the longer wavelength continuum, though the experimental support for the temperature dependence is somewhat weak.

The importance of the water vapor continuum to climate when temperatures approach or exceed $300K$ is demonstrated in Figure 4.27. Here we present calculations of the spectrum of OLR and of surface back-radiation using the exponential sum radiation code described previously, but modified to take into account the vertical variation of water vapor concentration and the continuum. With the continuum included, the low layer atmosphere radiates to the surface practically like a blackbody; in fact, if one increases the surface air temperature slightly, to $310K$, the back radiation becomes indistinguishable from the blackbody spectrum corresponding to the surface air temperature. In contrast, without the continuum, there is essentially no back-radiation in the window region, allowing the surface to cool strongly through the window. Likewise, without the continuum, the atmosphere can radiate to space very strongly through the window, whereas the cooling to space is very much reduced if the continuum absorption is included. These calculations were carried out for an Earthlike water-air atmosphere on a planet with $g = 9.8m/s^2$. On a planet with weaker surface gravity, the water vapor continuum would become important at lower temperatures, because a given partial pressure corresponds to a greater mass of water. Conversely, on a planet with stronger surface gravity, the water vapor window closes at higher temperatures.

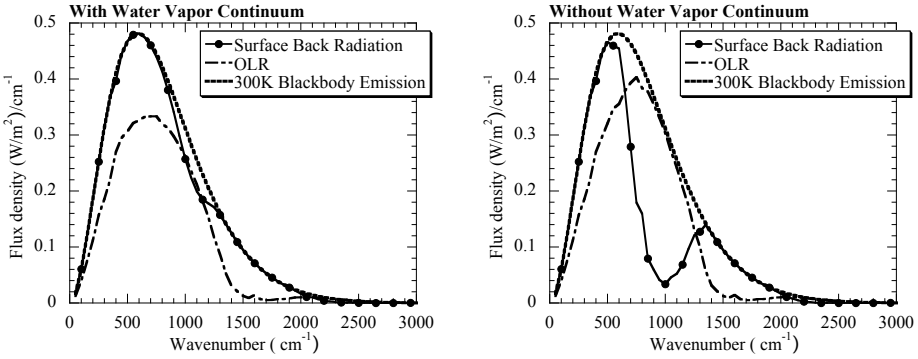

Figure 4.27: The spectra of surface back radiation and OLR computed using an exponential sum radiation code including the effects of the longwave water vapor continuum (left panel) and excluding the continuum (right panel). The calculation was done with the temperature profile on the water-air moist adiabat corresponding to $300K$ surface temperature, assuming the water vapor partial pressure to be saturated at all levels. This calculation does not take into account the temperature variation of absorption coefficients or the enhancement of self-induced absorption in the line contribution.

4.4.9 Condensed substances: Clouds

Clouds are made of particles of a condensed substance, which may be in a liquid or solid (e.g. ice) phase. The molecules of a condensed substance are in close proximity to one another, and at typical atmospheric temperatures the collisions are so frequent that no line structure survives in the spectrum. In consequence, the absorption coefficient for a condensed substance is generally a very smoothly varying function of wavenumber. Absorption by condensed substances behaves rather like the gaseous continuum absorption we discussed in the preceding section.

Water clouds are of particular interest, since they are by far the dominant type of cloud on Earth. They would also occur on any world habitable for life as we know it, since such a world would have a repository of liquid water somewhere, and condensation of water vapor somewhere in the atmosphere would then be practically inevitable. Water clouds would also form in the course of a runaway greenhouse on a world with a water ocean, such as the primordial Venus. The absorption coefficient for liquid water is shown over the infrared range in Figure 4.28. For comparison, we show the median absorption coefficient for water vapor at $100mb$ pressure and $260K$ temperature. Keep in mind that the absorption for liquid water is a true continuum, so that, unlike the median absorption curve shown for the vapor phase, the curve for liquid water displays the full wavenumber variability of the absorption.

We see that a kilogram of water in the liquid phase is a far better absorber than the same kilogram in the form of vapor. Near the peak absorption wavenumbers of water vapor, the difference can be as little as a factor of ten, but in the window regions liquid water has an absorption coefficient many thousands of times that of water vapor. The absorption coefficient for liquid water varies little enough that over the infrared range it can be quite well approximated as a grey gas (or more properly, a grey liquid). In fact, with a typical absorption coefficient of $100m^2/kg$, it takes a layer of liquid only 10^{-5} meters to have unit optical thickness and to begin to behave like a grey body. This is the depth of penetration of atmospheric infrared back-radiation into the surface of a lake or ocean, and it is the depth whose temperature directly determines the

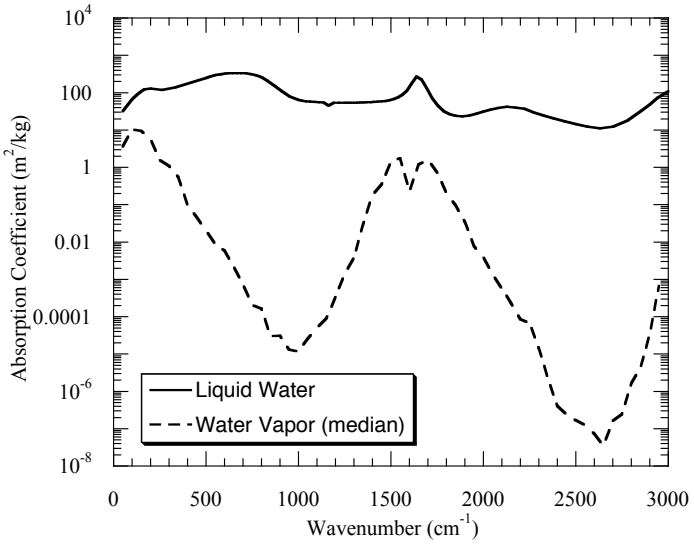

Figure 4.28: The absorption coefficient for liquid water. The median absorption coefficient for water vapor at $(100mb, 260K)$ is reproduced from Fig 4.19 for the purposes of comparison.

infrared radiation from the surface of a lake or ocean. Water ice is somewhat more transparent to infrared than liquid water, but the general features of the behavior still apply.

Because of the nonlinearity of the exponential function which determines the amount of absorption suffered by infrared as it passes through a number of particles, it matters greatly how the mass of water is distributed amongst particles of various sizes. For example, at 400 cm^{-1} liquid water has an absorption coefficient of 171 m^2/kg; since water has a density of $1000kg/m^3$ that means that a layer of liquid water of depth 5.8 μm has optical thickness of unity, and attenuates incident infrared by a factor of $1/e$. That means, roughly speaking, that a spherical droplet of radius r will remove essentially all the infrared hitting it – πr^2 times the incident flux – as long as r is 5 μm or more. A mere 10 grams of water is sufficient to make $1.9 \cdot 10^{10}$ particles of radius 5 μm, which would have a total cross section area of 1.5 m^2. Distributed randomly within a column of air having base 1 m^2 these particles would be sufficient to remove essentially all of the flux of infrared entering the base of the column. In other words, a mass path of liquid water as little as 10 grams per square meter is sufficient to make an optically thick cloud, provided the water takes the form of sufficiently small droplets. However, if we take the same mass of water and gather it up into a single drop of radius 1.3 cm – it would only intercept an insignificant .0005 m^2 of the incident light.

This estimate of the absorption by cloud droplets is not quite correct, because it fails to take into account the extent to which electromagnetic radiation pentrates the droplet as opposed to being diffracted around it. The calculation will be done more precisely in Chapter 5, but the simple estimate gives the right answer to within a factor of 2 or better.

The net result is that, for the typical droplet size found in Earth's water clouds, a cloud layer containing anything more than about 10 grams per square meter of condensed water acts essentially like a blackbody in the infrared. Water is not at all typical in this regard. Other condensed cloud-forming substances, including liquid methane and CO_2 ice, are far more transparent in the infrared,

and have a qualitatively different effect on planetary energy balance. The effect of such clouds will be taken up in Chapter 5.

4.5 Real gas OLR for all-troposphere atmospheres

Calculation of OLR is one of the most fundamental steps in determining a planet's climate. Now that we are equipped with an ability to compute the OLR for real gases, we can revisit some of our old favorite problems – Snowball Earth, the Faint Young Sun, Early Mars, and so forth – but this time relate the results to the actual atmospheric composition. In this section we present results for the all-troposphere model introduced in Section 4.3.2, occasionally limiting the upper air temperature drop by patching the adiabat to an isothermal stratosphere.

The homebrew exponential sum radiation model described in the preceding sections has the advantage of simplicity, generality and understandability. We will use it wherever it is sufficiently accurate to capture the main phenomena under discussion. However, professionally written terrestrial radiation codes are the product of a great deal of attention to detail, particularly with regard to temperature scaling and the simultaneous effects of multiple greenhouse gases. They can give highly accurate results provided one does not stray too far from the Earthlike conditions for which they have been optimized. In the following, and at various places in future chapters, we will have recourse to one of these standard radiation models, produced by the National Center for Atmospheric Research as part of the Community Climate Model effort. We'll refer to this model as the `ccm` radiation model. Although it uses a good many special tricks to achieve accuracy at high speed, and a lot of detailed bookkeeping to deal with the properties of a half dozen different greenhouse gases of interest on Earth, what is going on inside this rather massive piece of code is not fundamentally different from the Malkmus type band models and the exponential sum model described previously.

A detailed discussion of surface back-radiation for real gases will be deferred to Chapter 6. Some aspects of the infrared cooling profile for real gas atmospheres will be touched on in Section 4.7. The effect of clouds on OLR and on shortwave albedo will be discussed in Chapter 5.

4.5.1 CO_2 and dry air

First we'll compute the OLR for a mixture of CO_2 in dry air, with the temperature on the dry air adiabat $T(p) = T_g \cdot (p/p_s)^{2/7}$ and ground temperature equal to surface air temperature. This is a real-gas version of the calculation leading to Eq. 4.33; it amounts to a canonical OLR computation which serves as a simple basis for intercomparison of different radiation models. Performed for other greenhouse gases, it also can provide a basis for comparing the radiative effects of the gases. The results presented here are carried out with Earth gravity and 1 bar of air partial pressure, but can easily be scaled to other conditions. The CO_2 path is inversely proportional to gravity, so $100 ppmv$ of CO_2 on Earth is equivalent to $1000 ppmv$ on a planet with ten times the Earth's surface gravity or $10 ppmv$ on a planet with a tenth of the Earth's surface gravity. For fixed CO_2 concentration, surface pressure has a quadratic effect on the path, since the mass of CO_2 in the atmosphere (given fixed concentration) increases in proportion to pressure, but one gets an additional pressure factor in the equivalent path from pressure broadening. Thus, $100 ppmv$ of CO_2 in $1 bar$ of air is equivalent to $1 ppmv$ of CO_2 in $10 bars$ of air.

In Figure 4.29 we show how OLR varies with $CO2$ concentration for a fixed surface temperature of $273K$. This curve gives the amount of absorbed solar radiation needed to maintain the

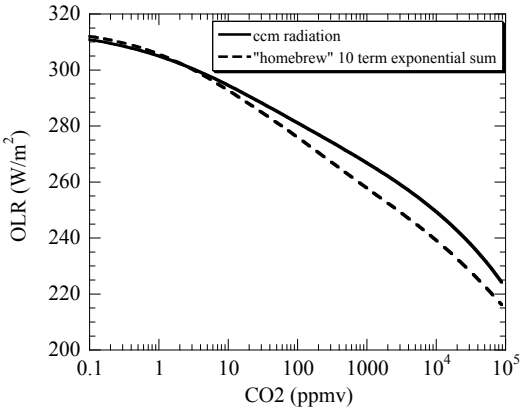

Figure 4.29: The OLR vs $CO2$ concentration (measured in $ppmv$) for CO_2 in a dry air atmosphere with temperature profile given by the dry adiabat. The surface air temperature and the ground temperature are both $273K$, and the acceleration of gravity is $9.8m/s^2$. Results are shown for a simple exponential sum radiation code without temperature weighting, and for the comprehensive `ccm` radiation code.

surface temperature at freezing. The CO_2 amount is expressed as molar concentration in $ppmv$, but for the range of concentrations considered, the difference between concentration and mixing ratio is not very significant. Results are shown for both the `ccm` model and the simplest form of the homebrew exponential sums radiation code. The homebrew calculation employed 10-term sums with coefficients computed at $260K$. The path was pressure-weighted to reflect collisional broadening, but temperature weighting was neglected. Over the range of CO_2 covered, only the principal absorption region centered on $650cm^{-1}$ needs to be taken into account. The homebrew calculation deviates by up to 10 W/m^2 from the more comprehensive `ccm` calculation, but this is quite good agreement in view of the fact that the homebrew code takes up barely a page and generalizes easily to any greenhouse gas, in contrast to the rather Earth-specific `ccm` code, which involves several thousand lines of rather unpretty `FORTRAN`. Most of the mismatch arises from the neglect of temperature scaling in the homebrew code. The homebrew code slightly overestimates OLR for low CO_2 where the radiating level is low in the atmosphere where warmer temperatures ought to increase the infrared opacity. It underestimates OLR because the higher, colder atmosphere is assumed more optically thick than it should be. If one complicates the homebrew code very slightly to incorporate a band-independent temperature weighting of the form $\exp -(T^*/T - T^*/T_o)$ in the path computation, then one can reduce the mismatch to under $2W/m^2$ with $T^* = 900K$.

As anticipated from the shape of the CO_2 absorption spectrum, the OLR goes down approximately in proportion to the logarithm of the CO_2 concentration. Between $10ppmv$ and $1000ppmv$ each doubling of CO_2 reduces OLR by about $4W/m^2$ based on the `ccm` model. At large CO_2 concentrations, the logarithmic slope becomes somewhat greater, as the weaker absorption bands begin to come into play. At $10000ppmv$, a doubling reduces OLR by $6.8W/m^2$. CO_2 becomes a somewhat more effective greenhouse gas at high concentrations, but never approaches the potency of a grey gas, for which each doubling would more than *halve* the OLR once the optically thick limit is reached. Were it not for the relatively gentle dependence of OLR on CO_2 caused by the highly frequency-selective nature of real-gas absorption, modest fluctuations in the atmosphere's CO_2 content would lead to wild swings of temperature and almost certainly render the planet

uninhabitable.

Calculations show that for fixed CO_2 concentration the OLR increases very nearly like the fourth power of temperature, just as in the grey gas result in Eq. 4.33. In effect, the radiating pressure remains nearly fixed as temperature varies. This makes it easy to do planetary temperature calculations. For example, let's compute what temperature the Earth would have if the CO_2 concentration were at the pre-industrial value of $280ppmv$, but there were no other greenhouse gases in the atmosphere. The OLR for a surface temperature of $273K$ is $267W/m^2$. Balancing against an absorbed solar radiation of $240W/m^2$, the temperature is determined by $267 \cdot (T/273)^4 = 240$, yielding $T = 263K$. Without the additional greenhouse effect of water vapor, Earth would be a very chilly place. According to Figure 4.29, for a dry Earth CO_2 would have to be increased all the way to $24000ppmv$ just to bring the temperature up to the freezing point.

Exercise 4.5.1 What temperature would Venus have if it had a $1bar$ air atmosphere mixed with $280ppmv$ of CO_2? Assume that the planetary albedo is 30%, like that of Earth. How does this temperature compare with the temperature Venus would have without any greenhouse gases in its atmosphere?

One can't get the Earth's temperature right without water vapor, but one can still make a decent estimate of the amount of CO_2 increase needed to offset the reduction of Solar forcing in the Faint Young Sun era, or due to a global Snowball Earth glaciation. A 25% reduction of Solar absorption at the Earth's orbit during the Faint Young Sun amounts to 60 W/m^2, assuming an albedo of .3. This is equivalent to the radiative forcing caused by increasing CO_2 from 100 $ppmv$ to $10^5 ppmv$ (about $100mb$ CO_2 partial pressure, or 10% of the atmosphere). From this we include that it's likely that it would take somewhat over $100mb$ of CO_2 to keep Earth unfrozen when the Sun was dim. Next let's take a look at what it takes to deglaciate a Snowball Earth. Let's suppose that for a Neoproterozoic solar constant, the tropics freeze over when the CO_2 is reduced to $100ppmv$ (this is not far off estimates based on comprehensive climate models). Icing over the Earth increases the albedo to about .6, leading to a reduction of almost $100W/m^2$ in absorbed solar radiation. Hence, restoring the Tropics to the melting point would require well in excess of $100mb$ of CO_2 in the atmosphere; calculations with the homebrew model at higher CO_2 concentrations than shown in Figure 4.29 indicate that fully a bar of CO_2 mixed with a bar of air would be needed. Deglaciation might not require restoring the full $100W/m^2$ of lost Solar forcing, but it is still clear that a great deal of CO_2 is needed to deglaciate a Snowball.

These estimates are crude, but they do get across one central idea: that because of the logarithmic dependence of OLR on greenhouse gas concentration, it takes a huge increase in the mass of CO_2 to make up for rather moderate changes in albedo or solar output. Aside from neglect of overlapping water vapor absorption, these estimates somewhat overstate the effect of CO_2 on OLR because they employ the dry adiabat rather than the less steep moist adiabat. We'll revisit the estimates shortly, after we bring water vapor into the picture.

4.5.2 Pure CO_2 atmospheres:Present and Early Mars, and Venus

Figure 4.30 shows the OLR as a function of surface pressure for a pure CO_2 atmosphere subject to Martian gravity. The results span the range of surface pressure from those similar to the thin atmosphere of present Mars up to the thick atmospheres commonly hypothesized for Early Mars [6]. The calculations were carried out for a fixed surface temperature of $270K$, since we are

[6]There is no strong reason to exclude the possibility of a substantial amount of N_2 in the Early Martian atmosphere. Addition of N_2 to the atmosphere would increase surface pressure and enhance CO_2 absorption.

primarily interested in the question of how much CO_2 there would have to be in order to warm Early Mars up to near the freezing point and permit the widespread liquid water at the surface that is seemingly demanded by the surface geology of the ancient Martian terrain. Results are shown for two different variants of the all-troposphere model. In the first, the atmosphere is on the dry CO_2 ideal gas adiabat throughout its depth. This profile is inconsistent at high surface pressure, however, since it becomes supersaturated aloft. For this reason, we also include results in which the temperature profile is on the one-component condensing CO_2 adiabat, which is on the dry adiabat where unsaturated but pinned to the Clausius-Clapeyron result when it becomes saturated (as in Fig. 2.6). With condensation, the surface pressure cannot be increased beyond $35.4bar$ at a surface temperature of $270K$, since the surface becomes saturated at that point and no further CO_2 can be added to the atmosphere without causing condensation. This does not pose a very significant constraint on the climate of Early Mars, however; a more important limitation is the amount of mass that could plausibly be lost from the primordial Martian atmosphere in the past four billion years. The effect of condensation aloft on OLR, in essence, increases the amount of CO_2 needed to warm Early Mars to the point where it is unclear that so much atmosphere could be lost.

At surface pressures comparable to that of Present Mars, the CO_2 greenhouse effect reduces the OLR by $35W/m^2$, and it would take $267W/m^2$ of absorbed solar radiation to maintain a surface temperature of $270K$. The required solar heating is well below the $440W/m^2$ solar forcing at the subsolar point. Assuming the OLR to scale with the fourth power of temperature for fixed surface pressure, this would support a temperature of $301K$ at the subsolar point, in contrast with a temperature of $292K$ in the absence of the atmospheric greenhouse effect. The atmosphere exerts only a modest warming effect on the surface temperature of Present Mars. To be sure, most of the planet is much colder; the absorbed solar flux averaged over the surface of the planet is only $110W/m^2$, which supports a temperature of $216K$ with the greenhouse effect and $210K$ without. Recall, however, that the planetary mean budget is not very meaningful on a planet like Present Mars with no ocean and little atmosphere to average out the diurnal variations. A thin layer of the rocky surface will be quite warm within a circle centered on the subsolar point, but the nightside surface falls to temperatures well below $216K$.

At higher CO_2, the OLR decreases, approximately logarithmically in surface pressure for pressures above 10^4Pa. At pressures above $1bar$, however, condensation becomes important, and the consequent increase in temperature aloft limits the decline of OLR. This limitation is quite important for the climate of Early Mars. Taking into account the relatively high albedo caused by molecular reflection from a thick CO_2 atmosphere, the absorbed solar radiation for Early Mars at a time when the solar flux is 30% reduced compared to today is about $70W/m^2$. This would be sufficient to sustain a $270K$ surface temperature with a surface pressure of $3.6bar$ if it weren't for the effects of condensation. When condensation is taken into account, however, fully 10 $bars$ of CO_2 are necessary to bring the surface temperature up to $270K$ [7]. In fact, given the increase in albedo associated with scattering of solar radiation in a $10bar$ atmosphere, even $10bars$ is likely to prove insufficient. The effect of atmospheric scattering on albedo will be quantified in Chapter 5. In that chapter we will also discuss the potential for the scattering greenhouse effect from CO_2 ice clouds to warm Early Mars. As time progresses toward the present, and the Sun gets brighter, it becomes progressively easier to warm Mars to the point where liquid water can persist at the surface. The climate history of Mars is a race between the brightening Sun and the loss of atmosphere, which seems to have been lost by the latter.

[7] The implications of CO_2 condensation for the gaseous greenhouse effect on Early Mars were first discussed in: Kasting JF 1991, *Icarus* **94**. The reader is also referred there for a more comprehensive treatment of the radiative transfer problem, including the effects of water vapor and a stratosphere. The conclusions are broadly similar to those based on our homebrew radiation model

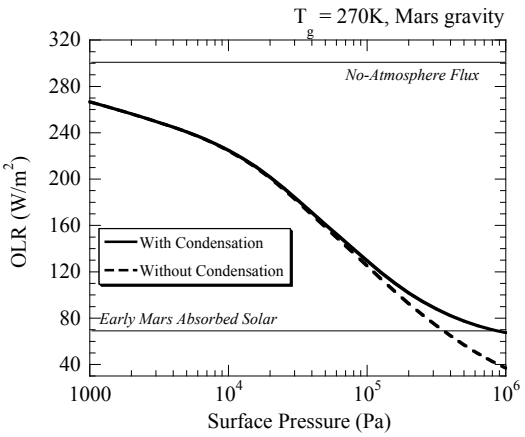

Figure 4.30: The OLR vs surface pressure for a pure $CO2$ atmosphere. The results were done with a 10-term exponential sum code based on a $100mb$ reference pressure, and including the temperature scaling of both the continuum and line absorption. Calculations were done for a ground temperature of $270K$, for Martian gravity. The dashed line shows the OLR for the case in which the temperature profile is on the dry ideal gas CO_2 adiabat, while the solid line incorporates the effect of condensation on the temperature profile.

The continuum absorption in the CO_2 window region is extremely important to these results. Without continuum absorption, the OLR for a $2bar$ atmosphere would be over $50W/m^2$ higher. The temperature scaling of the continuum affects the results by $10W/m^2$ or more. Whether or not one can account for prevalent liquid water on Early Mars by a gaseous CO_2 greenhouse effect hinges on a matter of $10W/m^2$ of flux or so, and therefore the importance of the continuum is disconcerting. To settle this question, one must get the CO_2 continuum right, and that is far from clear at this point, in view of the rather sparse experimental and theoretical results on the subject.

Now what about Venus? Are we finally equipped to say that we can account for the high surface temperature of Venus in terms of the CO_2 greenhouse effect? Unfortunately, not quite. The problem is that the high albedo of Venus means that the climate is maintained by a relative trickle of absorbed solar radiation, while the high surface temperature means that the infrared emission at wavenumbers higher than $2300cm^{-1}$ would exceed 3800 W/m^2 if the atmosphere were transparent in that spectral region. Unlike the Earthlike case, the atmospheric opacity there matters very much; however the `HITRAN` database does not include the weak absorption lines needed to accurately determine the atmospheric opacity at high wavenumbers. Extensions to the database suitable for use in Venusian conditions are described in the Further Readings at the end of this chapter. We will not pursue detailed calculations with the extended database, since one must, after all, stop somewhere. Instead, we will make use of a highly simplified treatment of the short wave thermal emission which at least tells us how close we are to being able to explain the temperature of Venus. Specifically, we use exponential sums based on `HITRAN` for the spectral region with lower wavenumbers than $2300cm^{-1}$, but represent the emission from higher wavenumbers by assuming that there is a radiating pressure p_{rad} such that the atmosphere radiates to space like a blackbody with temperature $T(p_{rad})$ throughout the high wavenumber spectral region. This is essentially the same approach as we took in formulating the simplest model of the greenhouse effect in Chapter 3, except that this time we apply the radiating-level concept only to the high wavenumber part of the emission. It is equivalent to stating that the absorptivity in the high wavenumber region is

sufficiently large to make the layer of the atmosphere between p_{rad} and the ground optically thick throughout the high wavenumber region. Because of the shape of the Planck function, as p_{rad} is made smaller and $T(p_{rad})$ is made colder, the peak emission shifts to lower wavenumbers and in consequence the shortwave emission is sharply curtailed.

With these approximations the OLR can be written as $OLR_<(T_g) + OLR_>(T_g, p_{rad})$ where $OLR_<$ is computed for the low wavenumber spectral region alone using exponential sums and

$$OLR_> = \int_{\nu_1}^{\infty} \pi B(\nu, T(p_{rad}, T_g))d\nu \tag{4.93}$$

where ν_1 is the frequency cutoff for the high wavenumber region, B is the Planck function and $T(p_{rad}, T_g)$ is computed using the dry CO_2 adiabat. The OLR thus computed must be balanced against the absorbed solar radiation to determine the surface temperature. At present, the solar radiation absorbed by Venus amounts to $163W/m^2$. Assuming a $93bar$ surface pressure, this is in balance with a surface temperature of $652K$ if there is no emission at all from the high wavenumber region, i.e. if $p_{rad} = 0$. This is a limiting case giving the maximum temperature that can be obtained with CO_2 alone regardless of how optically thick the high wavenumber region may be. The fact that this is still somewhat short of the observed $720K$ surface temperature of Venus means that a modest additional source of atmospheric opacity other than CO_2 is still needed to close the remaining gap in explaining the surface temperature. If p_{rad} is increased to 10 $bars$, the equilibrium temperature only drops to $633K$, so it is only necessary for CO_2 to be essentially opaque in the high wavenumber region at pressures of 10 $bars$ or greater. On the other hand, if CO_2 were really transparent at high wavenumbers, i.e. $p_{rad} = 93bar$, then the surface temperature would drop all the way to $461K$, which is well below the observed value. Detailed radiation modeling of the high wavenumber region is consistent with a value $p_{rad} \approx 10bar$, so it appears that the CO_2 greenhouse effect alone gets us almost all the way to explaining the high surface temperature. The remaining opacity needed to bring the surface temperature up to $720K$ is provided by Venus' high sulfuric acid clouds,the trace of water vapor in the atmosphere, and sulfur dioxide (in order of importance). The sulfur dioxide clouds are not good infrared absorbers, and exert their greenhouse effect through infrared scattering, as discussed in Section 5.

4.5.3 Water vapor feedback

For a planet like the Earth which has a substantial reservoir of condensed water at the surface (be it ocean or glacier), if left undisturbed for a sufficiently long time water vapor would enter the atmosphere until the atmosphere reached a state where the water vapor pressure was equal to the saturation vapor pressure at all points. In a case like this, if anything happens to increase the temperature of the atmosphere, then the water vapor content will eventually increase; since water vapor is a greenhouse gas, the additional water vapor will lead to an additional greenhouse gas, warming the planet further beyond the initial warming. Amplification of this sort is known as *water vapor feedback*, and works to amplify cooling influences as well. In this section we'll examine some quantitative models of real-gas water vapor feedback.

It turns out that atmospheric motions have a drying effect which keeps the atmosphere from reaching saturation. This is a very active subject of current research, but suffice it to say for the moment that comprehensive simulations of the Earth's atmosphere suggest that the situation can be reasonably well represented by keeping the relative humidity fixed at some subsaturated value as the climate warms or cools. That is the approach we shall adopt here, and it still yields an atmosphere whose water vapor content increases roughly exponentially with temperature. At present, there is no generally valid theory which allows one to determine the appropriate value of

relative humidity *a priori*, so one must have recourse to fully dynamic general circulation models or observations. Indeed, the relative humidity is not uniform but varies considerably both in the vertical and horizontal, so there is no single value that can be said to characterize the global humidity field. Similar considerations would apply to any condensible substance, for example methane on Titan.

In these all-troposphere models, we shall also assume that the temperature profile is given by the moist adiabat. Observations of the Earth's tropics show this to be a good description of the tropospheric temperature profile even where the atmosphere is unsaturated and not undergoing convection. (See Problem **??**). Evidently, the regions that are undergoing active moist convection control the lapse rate throughout the tropical troposphere; there are also theoretical reasons for believing this to be the case, but they require fluid dynamical arguments that are beyond the scope of the present volume. The situation in the midlatudes is rather less clear, but we will use the moist adiabat there as well, because it is hard to come up with something better in a model without any atmospheric circulation in it. Results are presented for Earth surface pressure and gravity, though we will make a few remarks on how the results scale to planets with greater or lesser gravity.

Since most of the big questions of Earth climate and climate of habitable Earthlike planets involve water vapor feedback in conjunction with one or more other greenhouse gases, it is in this section that we will for the first time get fairly realistic answers regarding the Faint Young Sun problem, and so forth.

The case of a saturated pure water vapor atmosphere will be treated in Section 4.6 as part of our treatment of the runaway greenhouse for real gas atmospheres. Here will will begin our discussion with water vapor mixed with Earth air, with the temperature profile on the moist adiabat for an air-water mixture. The OLR curve for this case is shown in Figure 4.31. Since there is no other greenhouse gas, the OLR for the dry case (zero relative humidity) is just σT_g^4. For ground temperatures below $240K$ there is so little water vapor in the air that the exact amount of water vapor has little effect on the OLR. At larger temperatures, the curves for different relative humidity begin to diverge. Even for relative humidity as low as 10%, the water vapor greenhouse effect is sufficient to nearly cancel the upward curvature of the dry case; at $320K$ the OLR is reduced by over $130W/m^2$ compared to the dry case. At larger relative humidites, the curvature reverses, and the OLR as a function of temperature shows signs of flattening at high temperature, in a fashion reminiscent of that we saw in our discussion of the Kombayashi-Ingersoll limit for grey gases. This is our first acquaintance with the essential implication of water vapor feedback: the increase of water vapor with temperature reduces the slope of the OLR vs. temperature curve, making the climate more sensitive to radiative forcing of all sort – whether it be changes in the Solar constant, changes in surface albedo, or changes in the concentration of CO_2. To take but one example, increasing the absorbed solar radiation from $346W/m^2$ to $366W/m^2$ increases the ground temperature from $280K$ to $283K$ in the dry case. When the relative humidity is 50%, it takes only $290W/m^2$ of absorbed solar energy to maintain the same $280K$ ground temperature, but now increasing the solar absorption by $20W/m^2$ increases the surface temperature to $288K$. Water vapor feedback has approximately doubled the climate sensitivity, which is a typical result for Earthlike conditions. The increase in climate sensitivity due to water vapor feedback plays a part in virtually any climate change phenomenon that can be contemplated on a planet with a liquid water ocean.

The influence of water vapor is strongest at tropical temperatures, but is still significant even at temperatures near freezing. It only becomes negligible at temperatures comparable to the polar winter.

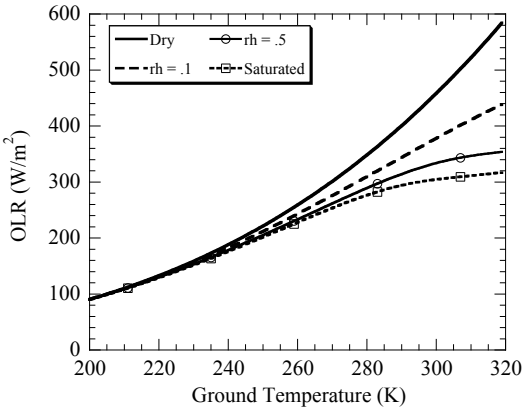

Figure 4.31: OLR vs. surface temperature for water vapor in air, with relative humidity held fixed. The surface air pressure is $1bar$, and Earth gravity is assumed. The temperature profile is the water/air moist adiabat. Calculations were carried out with the `ccm` radiation model.

One sometimes hears it remarked cavalierly that water vapor is the "most important" greenhouse gas in the Earth's atmosphere. The misleading nature of such statements can be inferred directly from Figure 4.31. Let's suppose the Earth's climate to be in equilibrium with $256W/m^2$ of absorbed solar radiation, averaged over the Earth's surface. This corresponds to an albedo of 25%, which we take to be somewhat smaller than the actual observed albedo to account crudely for the fact that part of the cloud albedo effect is canceled by cloud greenhouse effects which we do not take into account in the Figure. If water vapor were the only greenhouse gas in the Earth's atmosphere, the temperature would be a chilly $268K$, and that's even before taking ice albedo feedback into account, which would most likely cause the Earth to fall into a frigid Snowball state. We saw earlier that the Earth would also be uninhabitably cold if CO_2 were the only greenhouse gas in the atmosphere. With regard to Earth's habitability, it takes two to tango. In order to maintain a habitable temperature on Earth without the benefit of CO_2, the Sun would have to be 13% brighter. It will take well over a billion more years before the Sun will become this bright.

Now let's add some CO_2 to the atmosphere. Figure 4.32 shows how the OLR curve changes if we add in $300ppmv$ of CO_2 – slightly more than the Earth's pre-industrial value. The general pattern is similar to the water-only case, but shifted downward by an amount that varies with temperature. At 50% relative humidity, the addition of CO_2 reduces the OLR at $280K$ by a further $36W/m^2$ below what it was with water vapor alone. Because of the additional greenhouse effect of CO_2, the same $256W/m^2$ of absorbed solar radiation we considered previously can now support a temperature of $281K$ when the relative humidity is 50%. Note that without the action of CO_2, the atmosphere would be too cold to have much water vapor in it, so one would lose much of the greenhouse effect of water vapor as well. It appears that the actual surface temperature of Earth can be satisfactorily accounted for on the basis of the CO_2 greenhouse effect supported by water vapor feedback.

Exercise 4.5.2 For the four moisture conditions in Figure 4.32, determine how much absorbed solar radiation would be needed to support a surface temperature of $280K$. For each case, use the graph to estimate how much the surface temperature would increase if the absorbed solar radiation were increased were increased by $20W/m^2$ over the original value. How does the amplication due

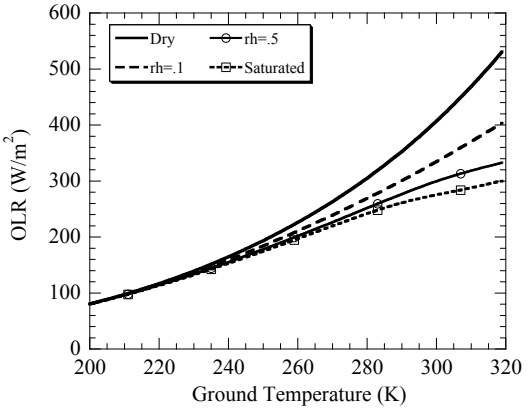

Figure 4.32: As in Figure 4.31, but with $300 ppmv$ of CO_2 included. Note that the "Dry" case excludes only the radiative effects of water vapor; the moist adiabat is still employed for the temperature profile.

to water vapor feedback compare with the results obtained without CO_2 in the atmosphere?

Next, let's take a look at how the OLR curve varies as a function of CO_2, with relative humidity held fixed at 50%. The results are shown in Fig. 4.33. The addition of CO_2 to the atmosphere lowers all the curves, and the more CO_2 you add, the lower the curves go. CO_2 is planetary insulation: adding CO_2 to a planet reduces the rate at which it loses energy, for any given surface temperature, just as adding fiberglass insulation to a house reduces the rate at which the house loses energy for a fixed interior temperature (thus reducing the fuel that must be burned in order to maintain the desired temperature). Another thing we note is that when the CO_2 concentration becomes very large, the curve loses its negative curvature and becomes concave upward, like σT^4. This happens because the CO_2 greenhouse effect starts to dominate the water vapor greenhouse effect, so that the flattening of the OLR curve due to the increase of water vapor with temperature becomes less effective over the temperature range shown. Even for very high CO_2, water vapor eventually would assert its dominance as temperatures are raised in excess of $320K$, causing the curve once more to flatten as the Kombayashi-Ingersoll limit is approached.

As an example of the use of the information in Fig. 4.33, let's start with a planet with $100 ppmv$ of CO_2 in its atmosphere, together with a sufficient supply of water to keep the atmosphere 50% saturated in the course of any climate change. From the graph, we see that absorbed solar radiation of 257 W/m^2 would be sufficient to maintain a mean surface temperature of $280K$. From the graph we can also see that if the absorbed solar radiation is held fixed, increasing CO_2 tenfold to $1000 ppmv$ would increase the temperature to $285K$ once equilibrium is re-established, and increasing it another tenfold to $10000 ppmv$ (about 1% of the atmosphere) would increase the temperature to $293K$. A further increase to $100000 ppmv$ (10% of the atmosphere) increases the temperature to $309K$. These represent substantial climate changes, but not nearly so extreme as they would have been if CO_2 were a grey gas. As another example of what the graph can tell us, let's ask how much CO_2 increase is needed to maintain the same $280K$ surface temperature with a dimmer sun. Reading vertically from the intersection with the line $T_g = 280K$, we find that this temperature can be maintained with an absorbed solar radiation of $195 W/m^2$ if the CO_2 concentration is $100000 ppmv$. Thus, an increase in CO_2 by a factor of 1000 can make up for a Sun

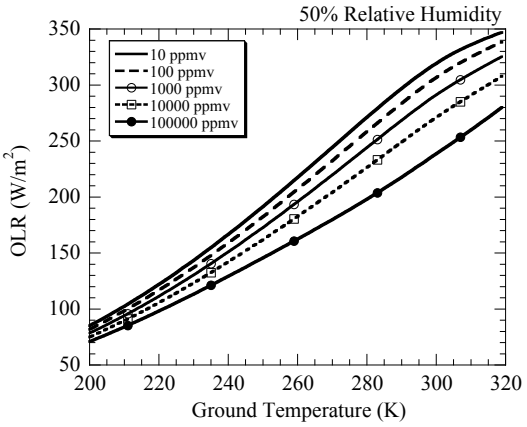

Figure 4.33: OLR vs surface temperature for various CO_2 concentrations, at a fixed relative humidity of 50%. Other conditions are the same as for Figure 4.31.

which is 25% dimmer than the base case.

Finally, we'll take a look at how the OLR varies with CO_2 for a fixed surface temperature (Fig. 4.34). This figure is a more Earthlike version of the results in Fig. 4.29, in that the effects of water vapor on radiation and the adiabat have been taken into account. As in the dry case, there is broad range of CO_2 concentrations – about $5ppmv$ to $5000ppmv$ – within which the OLR decreases very nearly like the logarithm of CO_2 concentration. The slope depends only weakly on the relative humidity, especially if one leaves out the completely dry case. This suggests that the effects of water vapor and CO_2 on the OLR are approximately additive in this range. We note further that the logarithmic slope of OLR vs CO_2 becomes steeper at very high CO_2, since one begins to engage more of the outlying absorption features of the CO_2 spectrum; again, CO_2 becomes an increasingly effective greenhouse gas at high concentrations. Conversely, at very low concentrations the logarithmic slope is reduced, as CO_2 absorption comes to be dominated by a relatively few dominant,narrow absorption features.

For any given CO_2 value, increasing the moisture content reduces the OLR, as would be expected from the fact that water vapor is a greenhouse gas. A little bit of water goes a long way. With $100ppmv$ of CO_2 in the atmosphere, going from the dry case to 10% relative humidity reduces the OLR by 36 W/m^2. To achieve the same reduction through an increase of CO_2, one would have to increase the CO_2 concentration all the way from $100ppmv$ to $10000ppmv$. Clearly, water vapor is a very important player in the radiation budget, though we have already seen that because of the thermodynamic control of water vapor in Earthlike conditions, CO_2 nonetheless remains important. As the atmosphere is made moister, the further effects of water vapor are less dramatic. Increasing the relative humidity from 10% to 50% only brings the OLR down by $17W/m^2$, and going all the way to a saturated atmosphere only brings OLR down by a further 12 W/m^2. Further, at very high CO_2 concentrations the OLR becomes somewhat more insensitive to humidity, as CO_2 begins to dominate the greenhouse effect.

The information presented graphically in Figures 4.32, 4.33 and 4.34 amount to a miniature climate model, allowing many interesting questions about climate to be addressed quantitatively without the need to perform detailed radiative and thermodynamic calculations; it's the kind of climate model that could be printed on a wallet-sized card and carried around everywhere.

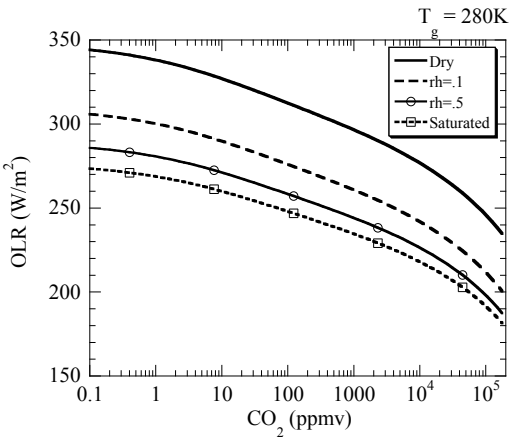

Figure 4.34: OLR vs CO_2 for a fixed surface temperature $T_g = 280K$, for various values of the relative humidity. Other conditions are the same as for Figure 4.31.

CO_2	a_o	a_1	a_2	a_3
10ppmv	258.56	2.5876	-0.0059165	-0.00013402
100ppmv	246.13	2.5056	-0.0034095	-0.00010672
1000ppmv	232.51	2.3815	-0.0015855	-8.3397e-05
10000ppmv	215.74	2.1915	0.00056634	-5.0508e-05
100000ppmv	189.06	1.8554	0.0044094	1.0735e-05

Table 4.2: Coefficients for polynomial fit $OLR = a_o + a_1 x + a_2 x^2 + a_3 x^3$, where $x = T_g - 275$. Calculation carried out with $rh = .5$.

Calculations using this information can be simplified by presenting the data as polynomial fits, which eliminates the tedium and inaccuracy of measuring quantities off graphs. Then, what we have is a miniature climate model that can be programmed into a pocket calculator. Polynomial fits allowing the OLR to be calculated as a function of temperature, CO_2 and relative humidity are tabulated in Tables 4.2 and 4.3. Values of OLR for parameters intermediate between the tabulated one can easily be obtained by interpolation.

Using these polynomial fits, we can put numbers to some of our old favorite climate questions by solving $OLR(T, CO_2) = S$ for T, given various assumptions about CO_2 and the absorbed solar radiation S. To wit –

- *Global Warming:* If we assume an albedo of 22.5%, the absorbed solar radiation is 265

rh	a_o	a_1	a_2	a_3
Dry	313.8	-6.275	-0.36107	-0.019467
.1	277.28	-5.9416	-0.35596	-0.020237
.5	259.52	-5.5332	-0.33915	-0.018932
Saturated	249.1	-5.183	-0.32187	-0.017367

Table 4.3: Coefficients for polynomial fit $OLR = a_o + a_1 x + a_2 x^2 + a_3 x^3$, where $x = \ln(CO_2/100)$, with CO_2 in $ppmv$. Calculation carried out with $T_g = 280K$ for the indicated moisture conditions.

W/m^2. For 50% relative humidity, and a pre-industrial CO_2 concentration of $280 ppmv$, the corresponding equilibrium temperature is $285K$, which is close to the pre-industrial global mean temperature. The albedo we needed to assume to get this base case is somewhat smaller than the Earth's observed albedo, because a portion of the cloud albedo is offset by the cloud greenhouse effect. Now, if we double the CO_2 to $560 ppmv$, the new temperature is $287K$ – a two degree warming. This is essentially the same answer as obtained by Manabe and Wetherald in their pioneering 1967 calculation, and was obtained by essentially the same kind of calculation we have employed. If we double CO_2 once more, to $1120 ppmv$, then the temperature rises to $289K$, a further two degrees of warming. The fact that each *doubling* of CO_2 gives a fixed additional increment of warming reflects the logarithmic dependence of OLR on CO_2; until one gets to extremely high concentrations, each doubling reduces the OLR by approximately $4W/m^2$.

- *Pleistocene Glacial-Interglacial Cycles:* In the depths of an ice age, the CO_2 drops to $180 ppmv$. Using the same base case as in the previous example, the temperature drops to $284K$, about a one degree cooling relative to the base case. This by no means accounts for the full amount of ice age cooling, but it is significant enough to imply that CO_2 is a major player. In the Southern Hemisphere midlatitudes, away from the direct influence of the growth of major Northern Hemisphere ice sheets, the CO_2 induced cooling is a half to a third of the total, indicating that either the ice sheet influence is propagated into the Southern Hemishere through the atmosphere or ocean, or that the cooling we have calculated has been further enhanced by feedbacks due to clouds or sea ice.

- *The PETM warming:* How much CO_2 would you have to dump into the ocean-atmosphere system to account for the PETM warming discussed in Section 1.9.1? One answers this by first deciding how much the atmospheric CO_2 concentration needs to increase, and then making use of information about the partitioning of carbon between the atmosphere and ocean. The PETM warming has been conservatively estimated at $4C$, and has nearly the same magnitude in the tropics as in the Arctic. The PETM event starts from an already warm hothouse climate, so for the sake of argument let's assume that the CO_2 starts at four times the pre-industrial value, yielding a starting temperature of $289K$. We need to increase the CO_2 from $280 ppmv$ to $1090 ppmv$ (just under two doublings) to achieve a warming to $293K$. This amounts to an addition of 1744 *gigatonnes* of C to the atmosphere as CO_2, which is comfortably within the limits imposed by the ^{13}C data, but not all carbon added to the atmosphere stays in the atmosphere. Over the course of a thousand years, approximately 80% of atmospheric carbon will be absorbed into the ocean, and over longer periods the ocean may be able to take up even more. Thus, to sustain the warming more than a millennium, one would need to add at least 5 times the nominal value, or 8720 *gigatonnes* to the ocean-atmosphere system. Can such a large addition of carbon be reconciled with the carbon isotopic record? This is the essential puzzle of the $PETM$, and may call for some kind of strong destabilizing feedback in the climate system.

- *Deglaciation of Snowball Earth:* If we increase the Earth's albedo to 60% (in accord with the reflectivity of ice) and reduce the solar constant by 6% (in accord with the Neoproterozoic value) the absorbed solar radiation is only 128 W/m^2. With CO_2 of 280 $ppmv$, the equilibrium global mean temperature is a chilly $228K$, more or less independent of what we assume about relative humidity. To determine the deglaciation threshold, we'll assume generously that the Equator is $20K$ warmer than the global mean, so that we need to warm the global mean to $254K$ to melt the tropics. Assuming 50% relative humidity, increasing CO_2 all the way to 200,000 $ppmv$ (about 20% of the atmosphere) still only brings the global mean up to $243K$, which is not enough to deglaciate. From this we conclude that without help from

some other feedback in the system, CO_2 would have to be increased to values in excess of 20% of the atmosphere to deglaciate. In fact, detailed climate model calculations indicate that it is even harder to deglaciate a snowball than this calculation suggests.

- *Temperature of the post-Snowball hothouse* Let's assume that somehow or other the Snowball does deglaciate when CO_2 builds up to 20%. After deglaciation, the albedo will revert to 22.5%, and the absorbed solar radiation to 249 W/m^2. When the planet re-establishes equilibrium, the temperature will have risen to $311K$. This is hot, and the tropics will be hotter than the global mean. However, the planet does not enter a runaway greenhouse and these temperature are well within the survival range of heat-tolerant organisms, especially since the polar regions would probably be no warmer than today's tropics.

- *Faint Young Sun:* Let's consider a time when the Sun was 25% fainter than today, reducing the absorbed solar radiation by 66 W/m^2. How much would CO_2 have to increase relative to pre-industrial values in order to keep the global mean temperature at $280K$ and prevent a freeze-out? We'll address this question by using the fit in Table 4.3. For 280 *ppmv* of CO_2 the OLR is 253 W/m^2 assuming 50% relative humidity. We need to bring this down by 66 W/m^2 to make up for the faint Sun. Using the fit, this can be done by increasing CO_2 to 240,000 *ppmv* (24% of the atmosphere) if we keep the relative humidity at 50%. If we on the other hand assume that for some reason the atmosphere becomes saturated with water vapor, then it is only necessary to increase the CO_2 to 15% of the atmosphere.

- *The Earth in one billion years:* According to Eq. 1.1, the solar constant will have increased to 1497 W/m^2, increasing the absorbed solar radiation per unit surface area to 290 W/m^2. If CO_2 is held fixed at the pre-industrial value, the Earth will warm to a global mean temperature of 296 K if relative humidity is held fixed at 50%. In order for silicate weathering to restore a temperature of $287K$, the weathering would have to bring the CO_2 all the way down to $10ppmv$, at which point photosynthesis as we know it would probably become impossible.

- *Temperature of Gliese 581c and 581d:* The planets Gliese 581c and 581d are in close orbits around a dim M-dwarf star. The redder spectrum of an M-dwarf would have some effect on the planetary energy balance, through changes in the proportion of solar energy absorbed directly in the atmosphere. Neglecting this effect, though we can estimate the temperatures of these planets assuming them to have an Earthlike atmosphere consisting of water vapor, CO_2 and N_2/O_2. Gliese 581c is in an orbit where it would absorb about 583 W/m^2, assuming the typical albedo of a rocky planet with an ocean. Gliese 581d would absorb only 50 W/m^2. Even if CO_2 were 20% of the atmosphere, the OLR would be 82 W/m^2, so Gliese 581d is likely to be an icy Snowball. On the other hand even with only $1ppmv$ of CO_2 in the atmosphere the OLR at $330K$ would be 351 W/m^2, far below the absorbed solar radiation for Gliese 581c. Thus, if Gliese 581c has an ocean, it is very likely to be in a runaway state – something we'll confirm when we re-examine the runaway greenhouse for real-gas atmospheres. There's an additional wrinkle to the Gliese system, though, in that these planets are more massive than Earth and have higher surface gravity. The higher gravity somewhat reduces the water vapor greenhouse effect, since for a given vapor pressure the corresponding amount of mass in the atmosphere is lower, according to the hydrostatic relation. This turns out to cool Gliese 581c somewhat, but still not enough to save it from a runaway.

Exercise 4.5.3 About how much carbon would need to be added to the atmosphere to achieve a $4K$ PETM warming if the initial CO_2 at the beginning of the event were only twice the pre-industrial value? If the initial CO_2 were eight times the pre-industrial value? (Note that in the

first case we are implicitly assuming some unknown process keeps the late Paleocene warm even with relatively little help from extra CO_2)

The above calculations include the effects of water vapor feedback, and also include the effects of changes in albedo due to ice cover, where explicitly mentioned. However, they do not incorporate any feedbacks due to changing cloud conditions. Cloud changes could either amplify or damp the climate change predicted on the basis of clear-sky physics, according to whether changes in the cloud greenhouse effect or the cloud albedo effect win out. Unfortunately, there is no simple thermodynamic prescription that does for cloud feedbacks what the assumption of fixed relative humidity does for water vapor feedbacks. We will learn more about the factors governing cloud radiative forcing in Chapter 5, but idealized conceptual models for prediction of cloud feedbacks remain elusive.

The problem of whether elevated CO_2 can account for hothouse climates such as the Cretaceous and Eocene is considerably more challenging than the other problems we have discussed above, since it requires one to answer a regional climate question: Under what conditions can we suppress the formation of polar ice? We already saw in Chapter 3 that a rather small change in radiation balance can make the difference between a planet with a small polar ice sheet and a planet which is globally ice-free. Suffice it to say at this point that an increase of CO_2 to 16 times pre-industrial values – the upper limit of what is plausibly consistent with proxy data – would yield a global mean warming of $10K$. Would this be enough to suppress formation of sea ice in the Arctic, and keep the mean Arctic temperatures around $10C$? Would the associated tropical temperatures be too hot to be compatible with available proxy data? We'll have to return to these questions in Chapter 7, where we discuss the regional and seasonal variations of climate.

4.5.4 Greenhouse effect of CO_2 vs CH_4

There is considerable interest in the idea that on the Early Earth methane may have taken over much of the role of CO_2 in offsetting the Faint Young Sun. In part this interest is due to rather sketchy geochemical evidence that at some times in the Archaean CO_2 concentrations may not have been high enough to do the trick, but regardless of whether the evidence actually *demands* a relatively low-CO_2 atmosphere, possibilities abound that in an anoxic atmosphere methane could build up to high concentrations. Even on an abiotic planet, there are possibilities for direct volcanic outgassing of methane, at a rate dependent on the state of oxygen in the planet's interior. Once biology comes on the scene, methanogens can convert volcanic CO_2 and H_2 to CH_4, or can make CH_4 by decomposing organic matter produced by anoxygenic photosynthesis.

Methane cannot build up to very high concentrations in a well-oxygenated atmosphere, but the relatively small amounts of methane in the atmosphere today (about 1.7 *ppmv*) nevertheless contribute significantly to global warming. There is, however, a widespread misconception that methane is in some sense an intrinsically better greenhouse gas than CO_2. A few simple calculations will serve to clarify the true state of affairs.

In order to compare the relative effects of CH_4 and CO_2 on a planet's radiation budget, we calculate the OLR for each case in what we have been calling the canonical atmosphere – a mixture of each gas into a dry atmosphere consisting of 1 bar of Earth air with temperature profile on the dry air adiabat, carried out with Earth's surface gravity. Results for a fixed surface temperature of $280K$, computed using the homebrew radiation model employing a constant temperature scaling coefficient $T^* = 900K$, are shown in Fig. 4.35. This graph in essence gives the amount of greenhouse gas needed to sustain a surface temperature of $280K$, given any specified amount of solar

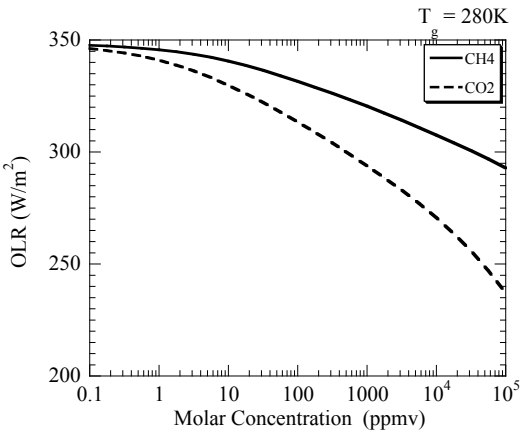

Figure 4.35: OLR vs. CO_2 and CH_4 concentration for each gas individually mixed with Earth air on the dry adiabat. The surface temperature is fixed at $280K$.

absorption. For example, with an absorbed solar radiation of $300W/m^2$ the surface temperature can be sustained with either $464ppmv$ of CO_2 or 35,600 $ppmv$ of CH_4 (3.56% of the atmosphere by mole fraction). These results somewhat overestimate the effect of each gas as compared to an actual moist atmosphere, since a moist atmosphere would be on the less steep moist adiabat and in a moist atmosphere the water vapor absorption would compete to some extent with the CO_2 and CH_4 absorption. Still, as an estimate of the relative effect of the two gases, the story is pretty clear. Methane is, intrinsically speaking, a considerably worse greenhouse gas than CO_2. The OLR curve for for methane is everywhere well above the curve for CO_2, so that it takes more methane than CO_2 to achieve a given reduction of OLR.

The common statement that methane is, molecule for molecule, a better greenhouse gas than CO_2 is true only for situations like the present where methane is present in far lower concentrations than CO_2. In this situation, the greater power of a molecule of CH_4 to reduce the OLR results simply from the fact that the greenhouse effect of both CH_4 and CO_2 are approximately logarithmic in concentration. Reading from Fig. 4.35, we see that for methane concentrations of around $1ppmv$, each doubling of methane reduces OLR by about $2W/m^2$. On the other hand, for CO_2 concentrations near 300 $ppmv$, each doubling of CO_2 reduces the OLR by about 6 W/m^2. Hence, to achieve the same OLR reduction as a doubling of CO_2 one needs three doublings of methane, but since methane starts from a concentration of only $1ppmv$, this only takes the concentration to $8ppmv$, and requires only $\frac{7}{300}$ as many molecules to bring about as was needed to achieve the same reduction using a doubling of CO_2. Equivalently, we can say that adding $1ppmv$ of methane yields as much reduction of OLR as adding $75ppmv$ of CO_2. The logarithmic slopes in this example are exaggerated compared to the appropriate values for Earth's actual atmosphere, because of the use of the dry adiabat and because of inaccuracies in the simple temperature scaling used in the homebrew radiation code. Using the `ccm` radiation code on the moist adiabat, with water vapor at 50% relative humidity, we find instead that each doubling of methane near $1ppmv$ reduces OLR by 0.77 W/m^2, while each doubling of CO_2 near $300ppmv$ reduces OLR by 4.3 W/m^2; in this case adding $1ppmv$ of methane reduces the OLR by as much as adding $38ppmv$ of CO_2. Nonetheless, the principle remains the same: If methane were the most abundant long-lived greenhouse gas in our atmosphere, and CO_2 were present only in very small concentrations, we would say instead that CO_2 is, molecule for molecule, the better greenhouse gas.

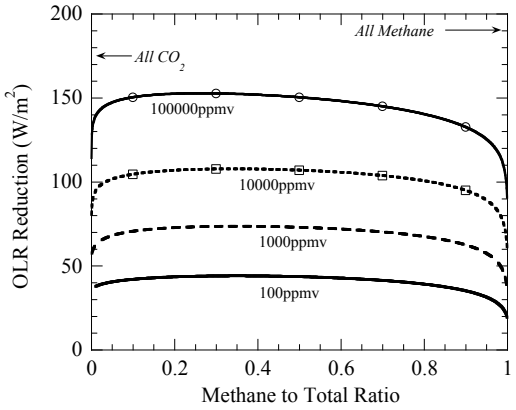

Figure 4.36: Total OLR reduction for the canonical atmosphere with a mixture of CH_4 and CO_2. Each curve gives the OLR reduction relative to a transparent atmophere for a fixed sum of CH_4 and CO_2 molar concentrations, indicated on the curve in units of $ppmv$. The results are plotted as a function of the CH_4 molar concentration to the total molar concentration for the two gases. The ratio is equal to the ratio of atmospheric carbon in the form of CH_4 to total atmospheric carbon. Larger values of the OLR reduction correspond to a stronger greenhouse effect.

Kirschvink and others have proposed that the Makganyene Snowball came about through a methane catastrophe, in which oxygenation converts methane to CO_2 and reduces the greenhouse effect sufficiently to precipitate a snowball. A methane crash, due to a reduction in methanogenic activity of some other mechanism, has sometimes been proposed as a trigger for the Neoproterozoic Snowballs as well. It is by no means easy to make these scenarios play out as they are supposed to, since methane contributes much less to the greenhouse effect than CO_2 when the two gases have similar abundances in the atmosphere. Conversion of methane to CO_2 will only reduce the greenhouse effect if methane is initially present in sufficiently small concentrations – but if there is too little methane present, the contribution of methane to the total greenhouse effect is too small to make much difference. Let's use the data in Fig. 4.35 to illustrate how the conversion of methane to CO_2 would affect climate in a few illustrative cases. The detailed numbers would change with a more accurate radiation model, or if the effect of water vapor were brought in, but the basic conclusions would remain much the same.

For example, suppose we started out with an atmosphere that contained 36,650 $ppmv$ of methane, which would be sufficient to maintain a $280K$ surface temperature given 300 W/m^2 of absorbed solar radiation. If this were all converted to CO_2 by oxidation, then, according to Fig. 4.35 the OLR would plunge to 255 W/m^2. In order to re-establish radiation balance, the planet would have to warm up to a temperature well in excess of the the initial $280K$. Far from causing a Snowball, in this case the oxidation of methane would cause a hot pulse, followed by gradual recovery to the original temperature as the CO_2 is drawn down by silicate weathering.

Next let's consider a more general situation, and identify the conditions necessary for a conversion of CH_4 to CO_2 to substantially reduce the greenhouse effect. Because the absorption features of CH_4 and CO_2 do not overlap significantly for the range of concentration under consideration, the combined effects of the two gases can be obtained by summing the OLR reduction ΔOLR for each of the gases taken in isolation. We wish to ask the question: if we have a given number of carbon atoms to use in supplying the atmosphere with greenhouse gases, how does the

net greenhouse effect depend on the way we divvy up those atoms between CH_4 and CO_2? Since each molecule has a single carbon, this question can be addresed by varying the molar concentration of CH_4 while keeping the sum of the molar concentrations of CH_4 and CO_2 fixed. Results of a calculation of this type are shown in Fig. 4.36. For any fixed total atmospheric carbon content, the OLR reduction has a broad maximum when plotted as a function of the methane ratio, and varies little except near the extremes of an all-methane or all-CO_2 atmosphere. The only case in which one can get a substantial reduction in greenhouse effect by oxidizing methane into CO_2 is when the initial CO_2 concentration is very high, the initial CH_4 fraction is between about 10% and 90% of the total, and the CH_4 is almost entirely converted to CO_2. For example, with a total carbon concentration of $10000 ppmv$, reducing the CH_4 concentration from $1000 ppmv$ to $1 ppmv$ reduces the greenhouse effect from $104 W/m^2$ to 80 W/m^2. Because the curve is so flat, starting from an atmosphere which is 80% methane works almost as well: in that case the greenhouse effect is reduced from 101 W/m^2 to 80 W/m^2. If we have a total of 100,000 $ppmv$ of carbon in the atmosphere, then the maximum greenhouse effect occurs for an atmosphere which is about 25% methane, and has a value of 151 W/m^2. Reducing the methane to 1 $ppmv$ brings down the greenhouse effect 36 W/m^2, to 115 W/m^2. In the Paleoproterozoic or Archaean, when the net greenhoue effect needed to be high to offset the Faint Young Sun, it is possible that a methane crash could have reduced the greenhouse effect enough to initiate a snowball, but it is essential that in a methane crash, the methane concentration be brought almost all the way down to zero; a reduction of methane from 50% of the atmosphere to 10% of the atmosphere would not do much to the greenhouse effect. By the time of the Neoproterozoic, when the solar luminosity is higher and less total greenhouse effect is needed to maintain open water conditions, it is far less likely that a methane catastrophe could have initiated glaciation. Some further remarks on atmospheric transitions that could initiate a Snowball will be given in Chapter 8.

4.6 Another look at the runaway greenhouse

We are now equipped to revisit the runaway greenhouse phenomenon, this time using the absorption spectrum of actual gases in place of the idealized grey gas employed in Section 4.3.3. The setup of the problem is essentially the same as in the grey gas case. We consider a condensible greenhouse gas, optionally mixed with a background gas which is transparent to infrared and noncondensing. A surface temperature T_g is specified, and the corresponding moist adiabat is computed. The temperature and the greenhouse gas concentration profiles provide the information necessary to compute the OLR, in the present instance using the homebrew exponential sums radiation model in place of the greygas OLR integral. As before, the OLR is plotted as a function of T_g for the saturated atmosphere, and the Kombayashi-Ingersoll limit is given by the asymptotic value of OLR at large surface temperature.

We'll begin with water vapor. Figure 4.37 shows the results for a pure water vapor atmosphere, computed for various values of the surfaced gravity. The overall behavior is very similar to the grey gas result shown in Fig. 4.3: the OLR attain a limiting value as temperature is increased, and the limit – defining the absorbed solar radiation above which the planet goes into a runaway state – becomes higher as the surface gravity is increased, and for precisely the same reasons as invoked in the grey gas case. The result, however, is now much easier to apply to actual planets, since with the real gas calculations we have the real numbers in hand for water vapor, and not for some mythical gas characterized by a single absorption coefficient. Several specific applications will be given shortly.

As in the grey-gas case the limiting OLR increases as surface gravity is increased. It would

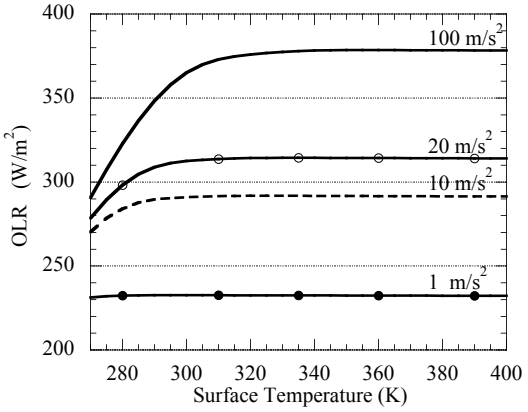

Figure 4.37: OLR vs surface temperature for a saturated pure water vapor atmosphere. The numbers on the curve indicate the planet's surface gravity. The calculation was done with the homebrew exponential sums radiation code, incorporating both the 1000 cm^{-1} and 2200 cm^{-1} continua, but neglecting temperature scaling of absorption outside the continua. 20 terms were used in the exponential sums, and wavenumbers out to 5000 cm^{-1} were included; the atmosphere was considered transparent to higher wavenumbers

be useful if the real gas result could be represented in terms of an equivalent grey gas, but first one must generalize Eq. 4.39 to incorporate the increase of absorption coefficient with pressure. This is done formally in Problem **??**, but the qualitative derivation runs as follows. We need to determine the pressure p_1 where the optical thickness to the top of the atmosphere is unity, and then evaluate the temperature at that point, along the one-component saturated adiabat. For linear pressure broadening or the continuum the optical thickness requirement implies $\frac{1}{2}\kappa_o p_1^2/p_o g = 1$, where p_o is the reference pressure to which the absorption is referred (generally $100mb$ for the data given in our survey of gaseous absorption properties). Substituting the resulting p_1 into the expression for $T(p)$ we infer an expression of the form

$$OLR_\infty = A'\sigma T(p_1)^4 = A'\frac{\sigma(L/R)^4}{[\ln(p^*/\sqrt{2p_o g/\kappa_o})]^4} \tag{4.94}$$

where p^* is defined as before and A' is an order unity constant which depends on L/R An examination of the g dependence of the calculated Kombayashi-Ingersoll limit in Figure 4.37 shows that over the range $1m/s^2 \leq g \leq 100m/s^2$, the numerically computed dependence can be fit almost exactly with this formula if we take $A' = .7344$ and $\kappa_o = .055$ (assuming $p_o = 10^4 Pa$). Though the pressure dependence of absorption causes the limiting OLR to vary more slowly with g than was the case for constant κ, the limit in the real gas case otherwise behaves very much like an equivalent grey gas with $\kappa_o = .055$. This is a surprising result, given the complexity of the real-gas absorption spectrum. The fact that the equivalent absorption is similar to that characterizing the 2500 cm^{-1} continuum suggests that the limiting OLR is being controlled primarily by this continuum. Thus, the behavior of this continuum is crucial to the runaway greenhouse phenomena (see Problem **??**). It cannot be ruled out that other continua may affect the OLR as temperature is increased to very high temperatures. For example, the total blackbody radiation at wavenumbers greater than $5000cm^{-1}$ is only $1.14W/m^2$ at $500K$, so it matters little what the absorption properties are in that part of the spectrum at $500K$ or cooler. However, when the temperature is reaised to $600K$ the shortwave emission is $15W/m^2$ and so the shortwave absorption begins to matter; by the

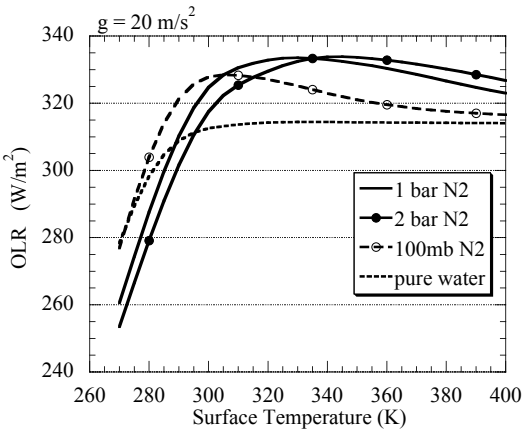

Figure 4.38: As for Fig. 4.37 but for a mixture of water vapor in N_2 on the saturated moist adiabat. Calculations were carried out with a surface gravity of $20m/s^2$, for the indicated values of N_2 partial pressure at the ground.

time T reaches $700K$ the shortwave blackbody emission is $106W/m^2$ and the shortwave emission properties are potentially important. On the other hand, at such temperatures there is so much water vapor in the atmosphere that a very feeble absorption would be sufficient to eliminate the contribution to the OLR.

Exercise 4.6.1 Verify the shortwave blackbody emission numbers given in the preceding paragraph by using numerical quadrature applied to the Planck function.

Exercise 4.6.2 Compute the Kombayashi-Ingersoll limit for water vapor on Mars, which has $g = 3.71m/s^2$. Compute the limit for Titan, which has $g = 1.35m/s^2$, and Europa which has $g = 1.31m/s^2$.

Now let's generalize the calculation, and introduce a noncondensing background gas which is transparent in the infrared; we use N_2 in this example, though the results are practically the same if we use any other diatomic molecule. The background gas affects the Kombayashi-Ingersol limit in two ways: First, the pressure broadening increases absorption, which should lower the limit. Secondly, the background gas shifts the lapse rate toward the dry adiabat, which is much steeper than the single-component saturated adiabat. The increase in lapse rate in principle could enhance the greenhouse effect, but given the condensible nature of water vapor, it actually reduces the greenhouse effect, because the low temperatures aloft sharply reduce the amount of water vapor there. If the background gas were itself a greenhouse gas, this effect might play out rather differently. At sufficiently high temperatures, water vapor will dominate the background gas and so the limiting OLR at high temperature will approach the pure water vapor limit shown in Figure 4.37. However, for intermediate temperatures, the background gas can modify the shape of the OLR curve.

Results for various amounts of N_2 are shown in Fig. 4.38. As expected., at large temperatures the limiting OLR asymptotes to the value for a pure water vapor atmosphere. This can be seen especially clearly for the case with only $100MB$ of N_2 in the atmosphere; with more N_2, one has to go to higher temperatures before the water vapor completely dominates the OLR, but the

trend is clear. A very important qualitative difference with the pure water vapor case is that the OLR curve for a binary mixture shows a distinct maximum at intermediate temperatures. This maximum arises because the foreign broadening of water vapor absorption features is relatively weak, while the presence of a noncondensing background gas steepens the lapse rate and reduces the amount of water vapor aloft. The hump in the curve means that the surface temperature exhibits *multiple equlibria* for a given absorbed solar radiation. For example, with $100mb$ of N_2, if the absorbed solar radiation is $320W/m^2$ there is a cool equilibrium with $T_g = 288K$ and a hot equilibrium with $T_g = 360K$. The latter is an unstable equilibrium; displacing the temperature in the cool direction will cause water vapor to condense and OLR to decrease, cooling the climate further until the system falls into the cool equilibrium. Conversely, displacing the temperature slightly to the warm side of the hot equilibrium will cause the climate to go into a runaway state. For these atmospheric parameters, the planet is in a metastable runaway state. The climate will persist in the cooler non-runaway state unless some transient event warms the planet enough to kick it over into the runaway regime. It is only when the absorbed solar radiation is increased to the maximum OLR at the peak ($328\ W/m^2$) that a runaway becomes *inevitable*. For future use, we'll note that a calculation with $g = 10m/s^2$ and 1 bar of N_2 has a peak OLR of $310\ W/m^2$ at $T_g = 325K$, while a case with the same gravity and 3 bars of N_2 likewise has a peak at $310W/m^2$ but the position of the peak is shifted to $360K$. The corresponding parameters for the slightly lower surface gravity of Venus differ little from these numbers. Both cases asymptote to the OLR for pure water vapor when the temperature is made much larger than the temperature at which the peak OLR occurs.

- *Runaway greenhouse on Earth*: With present absorbed solar radiation (adjusted for net cloud effects) of $265\ W/m^2$, the Earth at present is comfortably below the Kombayashi- Ingersoll limit for a planet of Earth's gravity. According to Eq. 1.1, as the solar luminosity continues to increase, the Earth will pass the $291\ W/m^2$ threshold where a runaway becomes possible in about 700 million years. In 1.7 billion years, it will pass the $310W/m^2$ threshold where a runaway becomes inevitable for an atmosphere with 1 bar of N_2 and no greenhouse gases other than water vapor.

- *Venus*: The present high albedo of Venus is due to sulfuric acid clouds that would almost certainly be absent in a less dry atmosphere. If we assume an Earthlike albedo of 30%, then very early in the history of the Solar System, the absorbed solar radiation of Venus would be $327\ W/m^2$. This is just barely in excess of the mandatory runaway threshold of $310W/m^2$ for a planet of Venus' surface gravity, assuming a *bar* or two of N_2 in the atmosphere and no greenhouse gases other than water vapor. It is thus possible that neglected effects (clouds, subsaturation, a higher albedo surface) could allow Venus to exist for a while in a hot, steamy but non-runaway state with a liquid ocean. The high water vapor content of the upper atmosphere would still allow an enhanced rate of photo-dissociation and escape of water to space. If Venus indeed started life with an ocean, however, it is plausible that it eventually succumbed to a runaway state, since with the present solar constant the absorbed solar radiation without sulfuric acid clouds would be $457W/m^2$, well in excess of the runaway threshold.

- *Gliese 581c*: We can now improve our earlier estimates of the conditions on the extrasolar planet Gliese 581c, which has an absorbed solar radiation of $583\ W/m^2$ assuming a rocky surface. This flux is well above the threshold of $334W/m^2$ for a mandatory runaway for a planet with twice Earth's surface gravity, even allowing for 2 *bar* of N_2 in the atmosphere. Thus, if Gliese 581c ever had an ocean it is likely to have gone into a runaway state; if the composition of the planet included a substantial amount of carbonate in the interior,

subsequent outgassing is likely to have turned it into a planet rather like Venus. It still remains, however, to assess the implications of the increased atmospheric absorption due to the greater proportion of infrared output of the M-dwarf host star.

- *Evaporation of icy moons in Earthlike orbit*: It has been suggested that icy moons like Europa or Titan could become habitable if the host planet were in an orbit implying Earthlike solar radiation. The low Kombayashi-Ingersoll limit for bodies with low surface gravity puts a severe constraint on this possibility, however. With the albedo of ice, such bodies could exist as Snowballs in an Earthlike orbit, but if the surface ever thawed, or failed to freeze in the firt place, the absorbed solar radiation corresponding to an albedo of 20% would be $274W/m^2$ – well above the runaway threshold of 232 W/m^2 for a body with surface gravity of 1 m/s^2. Small icy moons in Earthlike orbits are thus likely to evaporate away, unless they are locked in a Snowball state.

- *Lifetime of a post-impact steam atmosphere*: Suppose that in the Late Early Bombardment stage, enough asteroids and comets hit the Earth to evaporate 10 bars worth of the ocean and give the Earth a 10 bar atmosphere consisting of essentially pure water vapor (and a surface temperature in excess of $440K$, according to Clausius-Clapeyron). How long would it take for the steam atmosphere to rain out and the temperature to recover to normal? To do this problem, we assume that the atmosphere remains saturated as it cools, and loses heat at the maximunm rate given by the Kombayashi-Ingersoll limit for Earth; we also need to subtract the absorbed solar radiation from the heat loss. For Early Earth conditions, the net heat loss is about $100W/m^2$. On the other hand, the latent heat per square meter of the Earth's surface in a steam atmosphere with surface pressure p_s is Lp_s/g, or $2.5 \cdot 10^{11} J/m^2$ for the stipulated atmosphere. To remove this amount of energy at a rate of $100W/m^2$ would take $2.5 \cdot 10^9 s$, or 80 years. The rainfall rate during this time would be warm but gentle: $3.5(kg/m^2)/day$, or a mere $3.5mm/day$ based on the water density of $1000kg/m^3$. This is the average rainfall rate constrained by the rate of radiative cooling, but it is likely that at places the local rainfall rate could be orders of magnitude greater, owing to the lifting and condensation in storms and other large scale atmospheric circulations.

- *Freeze-out time of a magma ocean*:In Chapter 1 we introduced the problem of the freeze-out time of a magma ocean on the Early Earth, and estimated the time assuming a transparent atmosphere in Problem **??**. How long does it take for the magma ocean to freeze out if the planet is sufficiently water-rich that the atmosphere consists of essentially pure water vapor in saturation? The time is estimated in the same way as in Problem **??** except that the rate of heat loss is again taken to be the difference between the Kombayashi-Ingersoll limit – giving the maximum OLR – and the rate of absorption of solar energy. For the Early Earth this would be about $100W/m^2$, which is far smaller than the transparent atmosphere case, where the energy loss is nearly $100,000W/m^2$ based on σT^4 for the $2000K$ temperature of molten magma. As a result, the freeze-out time (using the same assumptions as in Problem **??**) increases to 3.5 million years.

The latter two estimates follow the line of thinking introduced by Norman Sleep of Stanford University, and again illustrate the principle that Big Ideas come from simple models.

Exercise 4.6.3 Estimate the lifetime of a post-impact pure water vapor atmosphere on Mars assuming that the planet absorbs $90W/m^2$ of solar radiation, per unit surface area. Estimate the precipitation rate, in mm of liquid water per day.

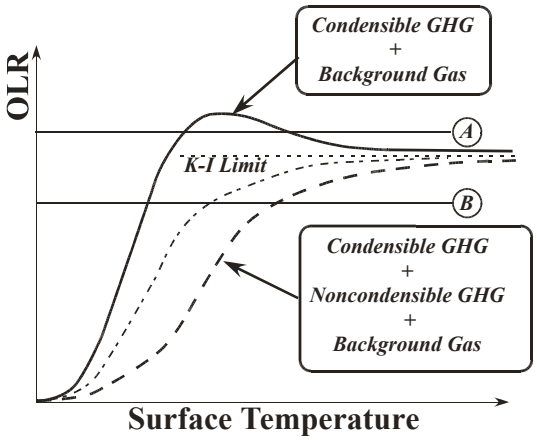

Figure 4.39: Qualitative influence of a noncondensible greenhouse gas on the shape of the OLR curves. The upper curve gives the OLR for an atmosphere consisting of a mixture of a saturated condensible greenhouse gas with a noncondensing transparent background gas, as in the N_2/H_2O case shown in Fig. 4.38, while the lower curve illustrates how the behavior would chance if a large amount of noncondensible greenhouse gas were added. The intermediate curve gives the OLR for a one-component saturated greenhouse gas atmosphere as in Fig. 4.37.

The above results presume that the saturated, condensible greenhouse gas is the only greenhouse gas present in the atmosphere. What happens if the atmosphere also contains a noncondensible greenhouse gas, whose total mass remains fixed as the surface temeprature increases? For Earthlike or Venuslike conditions, for example, one would typically need to consider atmospheres consisting of a mixture of condensible water vapor in saturation, noncondensing CO_2, and perhaps a transparent noncondensing background gas such as N_2. Can the addition of CO_2 in this situation trigger a runaway greenhouse when water vapor alone could not support a runaway? We will not pursue detailed radiative calculations of this sort, but some simple qualitative reasoning, summarized in the sketch in Fig. 4.39, suffices to map out the general behavior. The essential insight is that, at sufficiently high temperatures, the atmosphere is completely dominated by the condensible component, whose mass increases exponentially with temperature. Hence, the Kombayashi-Ingersoll limit will be unaffected by the addition of the noncondensible greenhouse gas. However, as more and more noncondensible greenhouse gas is added to the atmosphere, one must go to ever-higher temperatures before the limiting OLR is approached. At lower temperatures, the addition of a large amount of noncondensing greenhouse gas brings down the OLR below the Kombayashi-Ingersoll limit. Whether or not this triggers a runaway depends on the details of the situation. If the OLR curve without the noncondensible greenhouse gas is essentially monotonic in temperature, as in the one-component cases in Fig. 4.37, then the addition of the noncondensible greenhouse gas warms the planet, but does not trigger a runaway if the absorbed solar radiation is below the Kombayashi-Ingersoll limit. However, in a case like Fig 4.38, in which the OLR curve overshoots the limit and has a maximum, the addition of the noncondensible can eliminate the hump in the curve, eliminating the stable non-runaway state and forcing the system into a runaway. In the sketch, this situation is illustrated by the absorbed solar radiation line labeled "A". In that case, the addition of the noncondensing gas can indeed force the system into a runaway state. On the other hand, if the absorbed solar radiation is below the Kombayashi-Ingersoll limit, as in the line labeled "B", then the addition of the noncondensible greenhouse gas warms the equilibrium but does not trigger a runaway.

The concepts of runaway greenhouse and the Kombayashi-Ingersoll limit generalize to gases other than water vapor. For example, consider a planet with a reservoir of condensed CO_2 at the surface, which may take the form of a CO_2 glacier or a CO_2 ocean, according to the temperature of the planet. Specifically, if the surface temperature is above the triple point of $216.5K$ the condensible reservoir takes the form af a CO_2 ocean; otherwise it takes the form of a dry-ice glacier. If the atmosphere is in equilibrium with the surface reservoir and has no other gases in it besides the CO_2 which evaporates from the surface, then one can use the one-component adiabat with the homebrew radiation code to compute an OLR curve for the saturated CO_2 atmosphere which is analogous to the water vapor result shown in Fig. 4.37. Results, for various surface gravity, are shown in Fig. 4.40. The general behavior is very similar to that we saw for water vapor, but the whole system operates at a lower temperature and the OLR reaches its limiting value at a much lower temperature than was the case for water vapor.

The CO_2 runaway imposes some interesting contraints on the form in which CO_2 could exist on Mars, both present and past. For Martian surface gravity, the Kombayashi-Ingersoll limit for CO_2 is a bit over $63W/m^2$. In consequence, when the absorbed solar radiation exceeds this value, a permanent reservoir of condensed CO_2 cannot exist at the surface of the planet; it will sublimate or evaporate into the atmosphere, and continue to warm the planet until all the condensed reservoir has been converted to the gas phase. At present, the globally averaged solar absorption is about 110 W/m^2, so the planet is well above the runaway threshold for CO_2. From this we can conclude that Mars cannot have an appreciable permanent reservoir of condensed CO_2 which can exchange with the atmosphere. Note, however, that this does not preclude the temporary buildup of CO_2 snow at the surface. Such deposits can and do form near the winter poles, but sublimate back into the atmosphere as spring approaches. This situation can be thought of as arising from the fact that the *local* absorbed solar radiation near the winter pole is below the Kombayashi-Ingersoll limit for CO_2. The local reasoning applies because the thin atmosphere of present Mars cannot effectively transport heat from the summer hemisphere.

Even without the albedo due to a thick CO_2 atmosphere, Early Mars would have an absorbed solar radiation of only $77W/m^2$. This is still somewhat above the Kombayashi-Ingersoll limit for a pure CO_2 atmosphere, but allowing for the scattering effects of the atmosphere and perhaps also the influence of nitrogen in the atmosphere, Early Mars could well have sustained permanent CO_2 glaciers, given a sufficient supply of CO_2. Because the planet is so near a threshold, a more detailed calculation – probably involving consideration of atmospheric heat transports – would be needed to resolve the issue.

One can similarly compute a Kombayashi-Ingersoll limit for methane, using the continuum absorption properties described in Section 4.4.8. This calculation would determine whether a body could have a permanent methane ocean, swamp or glacier at the surface.

Any gas becomes condensible at sufficiently low temperatures or high pressures, and it is in fact the Kombayashi-Ingersoll limit that determines whether a volatile greenhouse gas outgassing from the interior of the planet accumulates in the atmosphere, or accumulates as a massive condensed reservoir (which may be a glacier or ocean) [8]. In the latter case, additional outgassing goes into the condensed reservoir, and the amount of volatile remaining in the atmosphere in the gas phase is determined by the the temperature of the planet. The condensed reservoir can form only if the absorbed solar radiation is below the Kombayashi-Ingersoll limit for the gas in question, and even then only if the total mass of volatiles available is sufficient to bring the atmosphere to a state of saturation. As an example of the latter constraint, let's suppose that Mars were in a

[8]There are other places an outgassed atmosphere can go; water can go into hydration of minerals, and CO_2 can be bound up as carbonate rocks.

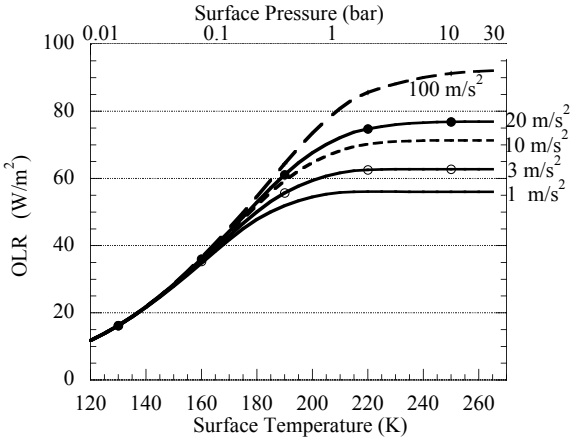

Figure 4.40: OLR vs surface temperature for a saturated pure CO_2 atmosphere. Calculations were performed with the values of surface gravity indicated on each curve. The scale at the top gives the surface pressure corresponding to the temperature on the lower scale.

more distant orbit, where the absorbed solar radiation were only $40 W/m^2$. Then, according to Fig. 4.40, the equilibrium surface temperature would be $165 K$ in saturation, and the corresponding surface pressure would be 5600 Pa. In order to reach this surface pressure it is necessary to outgas $5600/g$, or 1509 kg of CO_2 per square meter of the planet's surface. On Earth, outgassed water vapor accumulates in an ocean because the Earth is below the Kombayashi-Ingersoll limit for water vapor. With present solar luminosity Venus (without clouds) is well above the limit, so any outgassed water vapor would accumulate in the atmosphere (apart from the leakage to space). For CO_2, Earth, Mars and Venus are all above the Kombayashi-Ingersoll limit, so outgassed CO_2 accumulates in the atmosphere (apart from whatever gets bound up in mineral form). Even if you took away the water that allows CO_2 to be bound up as carbonate, Earth would not develop a CO_2 ocean; it would become a hot Venus-like planet instead, with a dense CO_2 atmosphere.

In parting, we must mention two serious limitations of our discussion of the runaway greenhouse phenomenon. First, in computing the Kombayashi-Ingersoll limit, it was assumed that the atmosphere was saturated with the condensible greenhouse gas. However, real atmospheres can be substantially undersaturated, though the dynamics determining the degree of undersaturation is intricate and difficult to capture in simplified models. Undersaturation is likely to raise the threshold solar radiation needed to trigger a runaway state. The second limitation is that the calculations were carried out for clear sky conditions. Clouds exert a cooling influence through their shortwave albedo, and a warming influence through their effect on OLR, and the balance is again hard to determine by means of any idealized calculation. Whether clouds have an inhibitory effect on the runaway greenhouse is one of the many remaining Big Questions.

4.7 Pure radiative equilibrium for real gas atmospheres

Pure radiative equilibrium amounts to an all-stratosphere model of an atmosphere, and is a counterpoint to the all-troposphere models we have been discussing. Real atmospheres sit between the two extremes, sometimes quite near one of the idealizations. In this section we will focus on pure infrared radiative equilibrium. The effects of solar absorption in real gases will be taken up in

Chapter 5.

From simple analytic solutions, we know essentially all there is to know about pure radiative equilibrium for grey gases. It is important to understand these things because the structure of atmospheres results from an interplay of convection and pure radiative equilibrium. A thorough understanding of pure radiative equilibrium provides the necessary underpinning for determining where the Stratosphere starts, and its thermal structure. We will now examine how the key elements of the behavior of pure radiative equilibrium differ for real gases. The specific issues to be addressed are:

- For grey gases in radiative equilbrium, the minimum temperature is the skin temperature based on OLR, and is found in the optically thinnest part of the atmosphere in the absence of atmospheric solar absorption. For real gases, can the radiative equilibrium temperature be much lower than the skin temperature?
- For a grey gas atmosphere with a given vertical distribution of absorbers, the radiative equilibrium temperature profile is uniquely determined once the OLR is specified. Specifically, one can determine the temperature profile of the radiative-equilibrium stratosphere without needing to know anything about the troposapheric temperature structure. To what extent is this also true for real gases?
- For a grey gas atmosphere with a given vertical distribution of absorbers, the normalized temperature profile $T(p)/T_g$ is independant of the ground temperature. This means that the radiative equilibrium temperature profile has the same shape, regardless of the magnitude of the solar radiation with which the atmosphere is in balance. How much does this result change for real gases?
- For a grey gas, the temperature jump at the ground is greatest when the atmosphere is optically thin, and vanishes as the atmosphere becomes optically thick. For real gases, the atmosphere is optically thick in some parts of the spectrum, and optically thin in others. What determines the temperature jump in these circumstances?
- Grey gases are most unstable to convection where they are optically thick. In the optically thick limit, the slope $d\ln T/d\ln p$ equals $\frac{1}{4}$ without pressure broadening, and $\frac{1}{2}$ with pressure broadening. Hence, a dry adiabat can go unstable near the ground only if $R/c_p < \frac{1}{4}$ without pressure broadening, or $R/c_p < \frac{1}{2}$ with pressure broadening. How do these thresholds differ for a real gas?

All of the issues except for the behavior of the static stability near the ground are well illustrated by the *semi-grey* model, which we referred to more poetically as "one-band Oobleck" in Section 4.4.1. In this model, we assume κ to be constant within a band of width Δ centered on frequency ν_o, and zero elsewhere. To keep the algebra simple we will make the additional assumption that Δ is small enough that B can be considered essentially frequency-independent within the band; this assumption can be easily dispensed with at the cost of a slightly more involved calculation. What makes the semigrey model tractable [9] is that the infrared heating is due solely to the flux within the absorbing band, so that one can still deal with a single optical thickness without any need to sum or integrate over frequency to get the net heating.

First, let's consider the semigrey radiative equilibrium in the optically thin limit. From the results in Section 4.2.2, the infrared radiative cooling at any level is simply $2\pi B(\nu_o, T)\Delta$, whereas

[9] Approximate solutions for the semigrey ("window-grey") model were presented by Sagan in 1969 (*Icarus* **10** 290-300.). The model has been rediscovered independently a number of times since.

the heating by absorption of upwelling infrared from the ground is $\pi B(\nu_o, T_g)\Delta$. Balancing the two, gives the equation determining the atmospheric temperature:

$$B(\nu_o, T) = \frac{1}{2} B(\nu_o, T_g) \tag{4.95}$$

Since the equation for T is independent of level, we conclude that in the optically thin case, the atmosphere is isothermal in infrared radiative equilibrium, just as it is for the optically thin grey case. The resulting atmospheric temperature is always lower than the ground temperature, and may be called the *semigrey skin temperature*. For a grey gas, $T_{skin}/T_g = 1/2^{1/4} \approx .84$, but for the semigrey case, the ratio depends on the frequency of the absorbing band. From the form of the Planck function, it can easily be shown that T_{skin}/T_g depends on frequency only through the ratio $h\nu_o/kT$. A simple analytic calculation shows that $T_{skin}/T_g \rightarrow \frac{1}{2}$ for small values of $h\nu_o/kT_g$ and $T_{skin}/T_g \rightarrow 1$ when $h\nu_o/kT_g$ becomes large; a simple numerical solution shows that the ratio increases monotonically between the two limiting cases (see Problem **??**). The grey-gas skin temperature ratio sits in the middle of this range, and indeed for frequencies near the peak of the Planck function, the semigrey ratio does not differ greatly from the grey value. For example, at $650cm^{-1}$ and a ground temperature of 280K, the semigrey skin temperature is $233K$, whereas the grey skin temperature is $235K$

When deriving the radiative equilibrium for grey gases, we specified the OLR and used that as a boundary condition for determining the thermal structure; the ground temperature is then computed from the resulting lower boundary fluxes. For real gases, it proves more convenient to specify the ground temperature T_g, and find the resulting temperature structure and OLR. We adopt this approach not only in our analytic solution of the semigrey case, but also in our numerical solutions for actual gases. When fixing T_g and increasing optical thickness, the OLR goes down, and so the amount of absorbed solar radiation needed to maintain the stated T_g decreases.

The derivation of the full infrared radiative equilibrium solution for the semigrey case is identical to that we used in the grey case, with the following substitutions: (1) The optical thickness τ is based on the value of κ in the absorbing band alone,(2) σT^4 is replaced by $\pi B(\nu_o, T)$ (3) The fluxes I_+ and I_- represent the flux integrated over the band Δ alone, and (4) The OLR appearing in the top boundary condition is no longer the net OLR emitted by the planet, but only the portion of OLR (call it OLR_Δ) emitted in the absorbing band. The assumption of infrared equilibrium still implies that the flux $I_+ - I_-$ is constant and equal to OLR_Δ, but since this flux is only the portion of OLR in the band, it is no longer determined *a priori* by planetary energy balance. Instead, it must be determined by making use of both boundary conditions, that $I_- = 0$ at the top of the atmosphere, and $I_+ = \pi B(\nu_o, T_g)$ at the bottom of the atmosphere (assuming the ground to have unit emissivity). The result of applying both boundary conditions is

$$I_+ - I_- = \frac{2\pi}{2 + \tau_\infty} B(\nu_o, T_g) \tag{4.96}$$

This is equal to the ground emission in the optically thin case, and approaches zero as the atmosphere is made more optically thick. Note, however, that it is only the OLR in the band that approaches zero; the net emission approaches a finite, and possibly quite large, lower bound, because the emission from the rest of the spectrum where the atmosphere is transparent is simply the ground emission. By substituting the expression for the net upward flux into the expression for $I_+ + I_-$, we find the following expression determining $T(\tau)$

$$B(\nu_o, T) = \frac{1 + \tau_\infty - \tau}{2 + \tau_\infty} B(\nu_o, T_g) \tag{4.97}$$

where τ is the optical thickness within the band where κ is nonzero. At the top of the atmosphere, this reduces to $B(\nu_T) = B(\nu_o, T_g)/(2 + \tau_\infty)$ Note that this is always less than the semi-grey skin

temperature, and becomes progressively colder in comparison the the skin temperature as τ_∞ is made large. For the semigrey case, then, the stratospheric temperature differs from the grey gas case in two important ways. First, as the atmosphere is made optically thick in the absorbing band, the stratospheric temperature approaches zero even though the net OLR remains finite (owing to the emission through the transparent part of the spectrum). Thus, the stratosperic temperature can be much colder than the grey-gas skin temperature, resolving the quandary raised in Chapter 3. Moreover, as the optical thickness of the atmosphere is increased, the stratosphere actually becomes colder than even the semigrey skin temperature; this contrasts with the grey case, where the temperature of the uppermost part of the atmosphere always approaches the skin temperature, regardless of how optically thick the rest of the atmosphere is. This important difference arises because, in the semigrey case, the optical thickness of the lower part of the atmosphere causes the upwelling spectrum illuminating the stratosphere to be depleted in those frequencies where the stratosphere absorbs best. Nonetheless, the stratosphere continues to *emit* effectively at those frequencies, leading to very cold temperature. This depletion also has the important consequence that ths stratospheric temperature is no longer independent of the existence of a troposphere, or of the tropopause height and thermal structure of the troposphere. This has the potential to affect the calculation of the tropopause height when we put radiative equilibrium together with convection.

The temperature jump at the ground is determined by $B(\nu_o, T_{sa}) = ((1+\tau_\infty)/(2+\tau_\infty))B(\nu_o, T_g)$, which has the same form as the corresponding expression for the grey gas case, save for the appearance of the Planck function in place of σT^4. As in the grey gas case, the jump is a maximum for optically thin atmospheres, and vanishes in the optically thick limit. The main difference here is that the jump can be made to vanish by making the atmosphere optically thick in just a limited part of the spectrum, even though the atmosphere is optically thin (in this case absolutely transparent) elsewhere. This feature will reappear in our discussion of radiative equilibrium for an atmosphere in which infrared absorption is provided by CO_2. As the atmosphere becomes optically thick in the absorption band, the unstable temperature jump at the ground diminishes, but what happens to the interior static stability of the air near the ground? To determing this, we take the derivative of Eq. 4.97 and multiply by p_s/T_{sa} to obtain the logarithmic radiative equilibrium lapse rate at the ground:

$$\frac{d\ln T}{d\ln p}|_{p_s} = \frac{B(\nu_o, T_g)}{T_{sa}\frac{dB}{dT}(\nu_o, T_{sa})}\frac{\kappa(p_s)p_s/g}{2+\tau_\infty} \tag{4.98}$$

In the optically thick limit, the second factor is unity without pressure broadening, or $\frac{1}{2}$ with pressure broadening. When the atmosphere is optically thick, the first factor can be evaluated with $T_{sa} = T_g$. With a little algebra, the first term can be re-written as $(\exp(u)-1)/(u\exp(u))$, where $u = h\nu_o/(kT_g)$; this term has a maximum value of unity at $u = 0$ (high frequencies or low temperatures) and decays to zero like $1/u$ at large u (low frequencies or high temperatures). For $u = 1$ which puts the absorption band near the maximum of the Planck function, the value is about .63. For the semi-grey model then, we conclude that the degree of instability of radiative equilibrium in the optically thick limit is bounded; for example, with $u = 1$ and incorporating pressure broadening, the radiative equilibrium near the ground is statically unstable when $R/c_p > .315$ (vs. a threshold value of .5 for a grey gas). Hence, the semigrey case has a somewhat enhanced instability at the ground as compared to the grey case, but the difference is not great. We'll see shortly that this is one regard in which the qualitative behavior of a real gas differs significantly from the semigrey case.

Finally, let's look at how the radiative equilibrium profile for the semigrey atmosphere changes if we change T_g and leave everything else fixed. For a grey gas, the function T/T_g is invariant because both the surface emission and the interior atmospheric emission increase by a

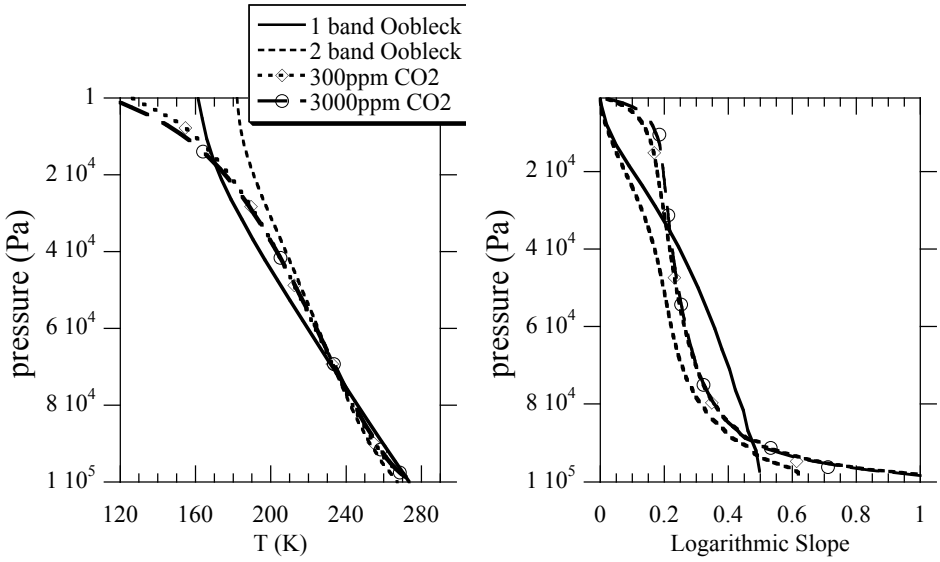

Figure 4.41: Left panel: Infrared radiative equilibrium temperature profiles for one and two-band Oobleck, and for a CO_2-air mixture at $300ppmv$ and $3000ppmv$, subject to a fixed $280K$ ground temperature. For the Oobleck cases, the absorption coefficients were chosen such that the optical depth of the atmosphere as a whole is 10 in the strong absorption band ($650-700cm^{-1}$) and 1 in the weak bands on the flanks ($600-650cm^{-1}$ and $700-750cm^{-1}$). Right panel: The corresponding logarithmic slope, $d\ln T/d\ln p$

factor b^4 if we replace T_g by $b\cdot T_g$. For the semigrey case, the emission is given by $B(\nu_o, T)$, which is no longer a simple power of T. This means that the ratio T/T_g is no longer independent of T_g. The nature of the dependance is left for the reader to explore in Problem **??**.

Working our way up the ladder to reality, we'll now present some numerical solutions for 2-band Oobleck and for mixtures of CO_2 in air. The latter are computed using our homebrew exponential sum radiation code, incorporating the effect of pressure broadening but without taking temperature scaling into account. This is sufficient to show how the extreme range of absorption coefficients in a typical real gas affect the radiative equilibrium. The equilibria were found by a simple time stepping method with fixed T_g. For any given initial temperature profile $T(p)$ one can calculate the infrared fluxes, and hence, by differencing in the vertical, the infrared heating rates. These are used to update $T(p)$, and the whole process is repeated until equilibrium is achieved, where the infrared heating is zero and the temperature no longer changes. Since we are only interested in the equilibrium and not the time course of the approach to equilibrium, we can afford to be somewhat sloppy in our time-stepping method, so long as it is stable enough to yield an equilibrium at the end. Figure 4.41 shows the resulting equilibrium profiles for two-band Oobleck with the secondary-band absorption coefficient one tenth the value of the primary, and for a CO_2-air mixture in an Earthlike atmosphere, at $300ppmv$ and $3000ppmv$. The left panel shows the temperature profile, while the right panel gives the logarithmic slope of temperature, which determines the static stability of the atmosphere. For comparison, the analytically derived results for one-band pressure-broadened Oobleck (the semigrey model) are also shown. These calculations were carried out with a ground temperature $T_g = 280K$. Some key characteristics of the results are summarized in Table 4.4.

Comparing the curves for 1-band and 2-band Oobleck shows that adding in the weakly absorbing bands reduces the vertical temperature gradient except in a thin layer near the ground. In this sense, most of the atmosphere acts as if it were made more optically thin, despite the fact that we have actually made the atmosphere more opaque to infrared by adding new absorption in additional bands without taking away absorption in the original band. Indeed, the OLR goes down from 333 W/m^2 in the 1-band case to 309 W/m^2 in the 2-band case, despite the warmer temperatures aloft in the latter. The key to this behavior was pointed out back at the end of Section 4.2.2: in an atmosphere which is optically thin in some parts of the spectrum but optically thick in others, the heating rate (which in turn determines the radiative equilibrium) is dominated by the parts of the spectrum where the optical thickness is nearest unity. This is the reason the weaker absorption bands control the behavior of the temperature in the interior of the atmosphere. The associated reduction in temperature gradient warms the atmosphere aloft, at pressures below $700mb$. In essence, the weak bands allow the upper reaches of the atmosphere to capture more of the infrared upwelling from below; spreading the absorption over a somewhat broader range of the spectrum in essence makes the problem a bit more like a grey gas, and makes the upper air temperature somewhat closer to the grey gas skin temperature.

While the addition of the weak bands reduces the temperature gradient aloft and hence stabilizes the atmosphere against convection, this comes at the expense of increasing the gradient near the ground, and destabilizing the layer there. This stabilization/destabilization pattern shows up clearly in the plot of the logarithmic temperature derivative shown in the right panel of Figure 4.41. Based on $R/c_p = 2/7$, the one-band Oobleck profile is unstable for pressures higher than $450mb$, whereas the two-band case is only unstable for pressures higher than $760mb$, though within the unstable layer the 2-band case is considerably more unstable than the one-band case. In the two-band case, more of the net vertical temperature contrast of the atmosphere is concentrated in a thin layer near the ground. Overall, the atmosphere acts as if it were optically thick near the ground, but relatively optically thin aloft. The behavior near the ground results from the extremely strong absorption near the center of the CO_2 absorption band. This spectral region captures the upwelling radiation from the ground, which is not yet depleted in the strongly absorbing wavenumbers. If there were much temperature discontinuity near the ground, the absorption would lead to strong radiative heating of the low level air; hence the only way to be in radiative equilibrium is for the air temperature to approach the ground temperature. More mathematically speaking, the phenomenon arises because the low level heating is controlled largely by the boundary terms in the flux integral in Eq 4.9, which in turn is dominated by the strongly absorbing spectral regions.

The real-gas CO_2 results have many features in common with 2-band Oobleck, notably the weak temperature gradient in the interior of the atmosphere and the enhancement of temperature gradient near the ground. The strong destabilization near the ground is even more pronounced for CO_2, because it has a far stronger peak absorption than the Oobleck case we considered, and because the absorption varies over a greater range of values. The associated optical thickness near the ground also keeps the unstable temperature jump between the ground and the overlying air small. An optically thin grey-gas would have a large unstable temperature jump at the ground. The strong absorption bands in a real gas smooth out this discontinuity and move it into a finite width layer in the interior of the atmosphere. From the standpoint of the convection produced, there is little physical difference between the two cases.

The main differences between the CO_2 cases and the Oobleck cases shows up in the upper atmosphere, where the CO_2 cases show steep declines of temperature with height – though not so steep as to destabilize the upper atmosphere. As at the bottom boundary, the culprit is the spectral region where absorption coefficients peak. These regions lead to very strong emission in the upper layers of the atmosphere, which are poorly compensated by absorption since the upwelling infrared

	OLR,W/m^2	T_{skin}	$T(0)$	$T(p/p_s = .5)$	$T_g - T(p_s)$
1-band Oobleck	333	232K	161K	206K	6.5K
2-band Oobleck	309	228K	182K	215K	12.75K
CO_2/air 300$ppmv$	295	226K	126K	214K	2.8K
CO_2/air 3000$ppmv$	275	222K	115K	214K	1.75K
Mars, CO_2,$p_s = 7mb$	303	227K	128K	215K	2.9K
Mars, CO_2,$p_s = 7mb$,$T_g = 250K$	192	203K	126K	196K	4.4K
Mars, CO_2,$p_s = 2bar$	86	166K	102K	208K	1.3K
Venus,CO_2,$p_s = 90bar$,$T_g = 700K$	55	148K	77K	500K	.1K

Table 4.4: Summary properties of infrared radiative equilibrium solutions. Calculations were done with $T_g = 280K$ unless otherwise noted.

reaching the upper atmosphere has been depleted in most wavenumbers that absorb at all well. The strong emission causes the temperature near the top to fall far below the skin temperature. Based on the net OLR, the grey gas skin temperature for the two CO_2 cases is somewhat above $220K$, whereas the actual temperature at the $1mb$ level is $126K$ in the 300$ppmv$ case and $115K$ in the more optically thick 3000$ppmv$ case. Finally, we note that increasing CO_2 by a factor of 10 has relatively little effect on the radiative equilibrium temperature profile, despite the fact that the increase lowers the OLR by nearly $20W/m^2$. The changes are principally seen near the ground, where the increased optical thickness in the wings has reduced the surface temperature jump. While the two-band Oobleck model reproduces the near-ground destablization present in the real CO_2 calculation, it is unable to simultaneously represent the temperature jump at the ground. Lacking the extremely strong peak absorption of CO_2, the addition of the weak absorption wings in 2-band Oobleck makes the surface budget act like an optically thinner atmosphere, increasing the surface jump.

To illustrate how the radiative equilibrium solution scales with T_g, we show the profiles of T/T_g for surface temperatures ranging from $240K$ to $320K$. Only the case of 300$ppmv$ CO_2 is shown, though the other atmospheres treated in Figure 4.41 yield similar results. For a grey gas, all the curves for a given atmospheric composition would collapse onto a single universal profile; for the reasons discussed in the semigrey case, this is no longer true for real gases. However, while the temperature aloft does not scale precisely with the ground temperature, the deviations are modest enough that one can still get useful intuition about the behavior of the system by assuming that radiative equilibrium temperature scales with the ground temperature. For CO_2, the actual temperature aloft is always somewhat colder than that which one would estimate by proportionately scaling the temperature upward from a colder to warmer ground temperature. Recall that these calculations are done without temperature scaling of the absorption coefficient, so the effect shown is purely due to the shape of the Planck function.

The general features encountered in the terrestrial calculations discussed above carry over to the pure CO_2 atmospheres characteristic of Present and (possibly) Early Mars. Martian radiative equilibrium solutions with a $7mb$ thin atmosphere or $2bar$ thick atmosphere are shown in Figure 4.43. The most striking feature of these solutions is that increasing the mass of the atmosphere by a factor of nearly 300 causes very little increase in the vertical temperature contrast. This stands in sharp contrast to the grey gas case, for which the enormous increase in optical depth going from the $7mb$ case to the $2bar$ case would cause the temperature in the latter to drop to nearly zero within a short distance of the ground. As before, the reason for the relative insensitivity of the temperature profile is that the radiative heating is determined largely by the part of the spectrum where the optical depth is of order unity. For the present Mars case, this occurs in the near wings

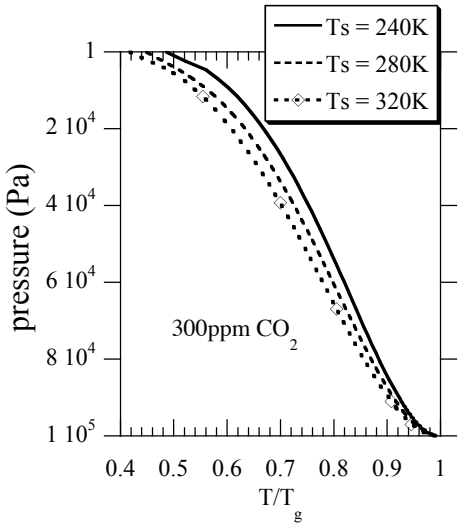

Figure 4.42: Variation of the shape of the temperature profile as a function of ground temperature. Results are shown for $300 ppmv$ of CO_2 in air.

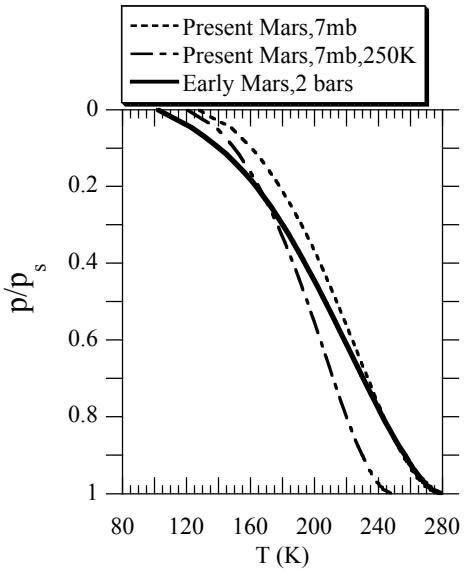

Figure 4.43: Infrared radiative equilibrium for pure CO_2 atmospheres with Martian gravity. Results are shown for an Early Mars case with a 2 bar surface pressure and $280K$ ground temperature, and for Present Mars cases with $700mb$ surface pressure and $250K$ and $280K$ ground temperatures. To make it possible to compare the cases, the temperatures have been plotted as a function of p/p_s. The $2bar$ case includes the CO_2 continuum absorption.

of the principle absorption peak, whereas for Early Mars it occurs within the continuum window region. The shift allows both cases to act roughly like a case with order unity optical depth, apart from the thin radiative boundary layers near the ground and the top. The temperature profiles are similar despite the fact that it would require $303 W/m^2$ of absorbed Solar radiation to maintain a surface temperature of $280K$ in the thin present Martian atmosphere, but only $86 W/m^2$ in a $2 bar$ atmosphere. For the Present Mars case, we have also included a calculation with with a realistically cold daytime surface temperature, the equilibrium temperature aloft is too cold in comparison with observations. This suggests once more an important role for solar absorption in determining the temperature structure of the present Martian atmosphere.

It is only when we go to massive atmospheres like that of Venus that the atmosphere becomes optically thick throughout the spectrum, allowing the vertical temperature contrast to increase dramatically. A simplified calculation without temperature scaling, and ignoring emission beyond $2300 cm^{-1}$, shows that with a $700K$ ground temperature, the radiative equilibrium temperature drops to $500K$ at the midpoint of the atmosphere and all the way to $80K$ at the $100 mb$ level. This yields a very strong greenhouse effect: it takes only a trickle of $55 W/m^2$ of absorbed solar radiation to maintain the torrid ground temperature. Note, however, that it is important that a fair amount of this trickle actually be absorbed at the ground, and not in the upper reaches of the atmosphere; otherwise the deep atmosphere becomes isothermal, and can in fact become as cold as the skin temperature in extreme cases, as discussed in Section 4.3.5. So far as the maintainence of its thermal structure is concerned, the troposphere of Venus is more like the Antarctic glacier than it is like the Earth's troposphere. The trickle of heat escaping the Earth's interior beneath the glacier – a mere 30 mW/m^2 – is sufficient to raise the basal ice temperature to the melting point and create subglacial Lake Vostok precisely because the diffusivity of heat in ice is so small. So it is, too, with the atmosphere of Venus; the extremely optically thick tropophere renders the radiative diffusivity of heat very small, and allows the tiny trickle of solar radiation reaching the surface to accumulate in the lower atmosphere and raise the temperature to extreme values. Unlike the glacier, however, when the lower atmosphere becomes hot enough, it can start to convect. Convection supplants the radiative heat flux, but also establishes the adiabat, allowing the surface to be much hotter than the radiating level.

Different noncondensible greenhouse gases differ somewhat in details of their radiative-equilibrium profiles, but the general picture does not differ greatly from what we have learned by looking at CO_2. When the greenhouse gas can condense near the ground, however,the situation becomes quite different. The case of water vapor in air provides a prime example. If the ground temperature is high enough that the amount of water vapor present in saturation makes the lower atmosphere optically thick, then the temperature will decline rapidly with height above the ground, because that is what optically thick atmospheres do in radiative equilibrium. Water vapor exhibits this effect particularly strongly, since it easily makes the atmosphere optically thick everywhere outside the window regions of the spectrum, and even the windows close off above $300K$. As the temperature decreases, however, the water vapor content decreases in accordance with the limits imposed by Clausius-Clapeyron. Within a small distance above the ground, the air is so cold that there is little water vapor left, and the atmosphere further aloft becomes optically thin. As a result, most of the variation in optical thickness of the atmosphere is concentrated into a thin, radiative boundary layer near the ground, and the optical thickness (and hence the temperature) varies greatly within this layer. Because of the strong temperature gradient and high optical thickness of the boundary layer, a strong greenhouse effect is generated entirely within the boundary layer, leading to low OLR. If one imposes equilibrium with an Earthlike absorbed solar radiation, the ground temperature must increase to temperature well in excess of $320K$ to achieve balance, and the steep increase of saturation vapor pressure with temperature further exacerbates the high temperature gradient in the radiative boundary layer. It is a bit as if the

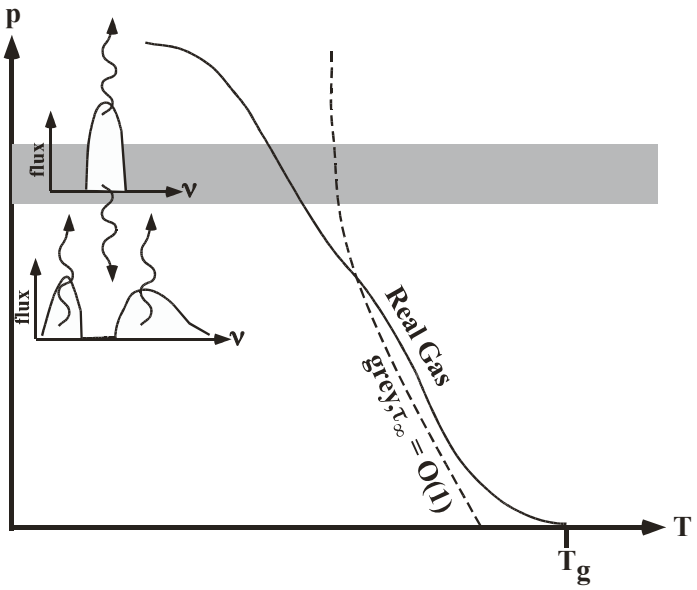

Figure 4.44: Basic features of real gas pure infrared radiative equilibrium.

.

entire optically thick atmosphere of Venus were squeezed into a boundary layer having a depth of a kilometer or less. Adding a noncondensing gas like CO_2 to the mix alters the temperature profile above the boundary layer, but does not eliminate the basic pathology of the situation. The equilibrium profile in this case is of little physical consequence because the slightest convection or other turbulent mixing would mix away the thin radiative boundary layer, warming and moistening a much deeper layer of the atmosphere The radiative equilibrium solutions we have been studying for the noncondensing case are worthy of protracted consideration because they provide some useful insight as to the stratospheric temperature for more realistic atmospheres in which there is some low level convection. The same cannot be said for pure radiative equilibrium in the condensing water-vapor/air system, which is an exercise in pathology having little or no bearing on the operation of real atmospheres.

Though real gas radiative equilibrium is not amenable to the kind of complete solution we enjoyed for the grey gas case, its behavior can be reasonably well captured by a few generalities, summarized in Figure 4.44. For real gases as for grey gases heated by infrared emission from the lower boundary, the temperature decreases with distance from the boundary. This can be viewed as a kind of thermal diffusion in the sense that heat transfer is down the temperature gradient, though the process is only described by a true local diffusion in the optically thick limit. Real gases behave as if they are optically thick near the ground, exhibiting strong convectively unstable temperature gradients there and little temperature jump between the ground and overlying air. The temperature gradient weakens in the interior, but there is generally a region of strong, though stable, temperature decline near the top of the atmosphere. The upper atmosphere is considerably colder than the grey gas skin temperature, since (by Kirchoff's law) the atmosphere radiates efficiently in the strongly absorbing parts of the spectrum, but the radiation illuminating the upper atmosphere is depleted in this portion of the spectrum. In contrast to the grey gas case, the contrast in temperature across the depth of the atmosphere is relatively insensitive to the amount of greenhouse gas in the atmosphere. As long as there are some spectral regions where

the atmosphere is optically thick, and some where the atmosphere is optically thin, the radiative cooling tends to be dominated by the intermediate spectral regions. As a result, temperature tends to drop by a factor of two to three between the ground and the upper atmosphere, with only slight increases even when the greenhouse gas content is increased by many orders of magnitude. This behavior persists until there is so much greenhouse gas present that the atmosphere becomes very optically thick throughout the thermal infrared spectrum, as is the case on Venus.

The upshot of all this is that atmospheres whose temperature is maintained by absorption of upwelling infrared from a blackbody surface will never exhibit pure radiative equilibrium. There will always be a layer near the surface which is unstable to convection. If the atmosphere is optically thin, the instability is generated by a temperature jump at the surface. If the atmosphere is optically thick and subject to pressure broadening, the instability is generated by strong temperature gradients in the interior of the atmosphere near the surface. This remark even applies qualitatively to gas giants which have no surface, as the deep atmosphere is dense enough that it can begin to act like a blackbody even though there is no distinct surface. An atmosphere can be stabilized throughout its depth, however, if it is subject to atmospheric solar heating which increases with altitude in a suitable fashion.

4.8 Tropopause height for real gas atmospheres

The radiative equilibrium solutions discussed in the preceding section are all unstable near the ground. As convection sets in, it will mix away the unstable layer and replace it by an adiabat; the well-mixed region is the troposphere. The change in lower level temperature profile, however, will alter the upward radiation which heats the stratosphere, and therefore cause temperature changes even above the layers reached directly by convection. When all this sorts itself out, how deep is the troposphere? This is is the problem of tropopause height, which we have already touched on briefly for grey gases. Here we will offer a taste of a few of the most important aspects of real gas behavior, and lay the physical basis the reader will need for further explorations with more comprehensive models. In this section we will assume, as in the all-troposphere model, that turbulent fluxes couple the ground so tightly to the overlying air that there is no discontinuity at the ground. This assumption will be relaxed in Chapters 6.

For a grey gas, the problem of finding the tropopause height is relatively simple. SInce the radiative equilibrium profile depends only on OLR – and that only via a simple formula – one starts with the radiative equilibrium profile for the desired OLR, picks a guess for the tropopause pressure, and then replaces the temperature between there and the ground with the adiabat for the gas under consideration. One then computes the actual OLR for the resulting profile, and generally will find that it is generally somewhat different from the OLR assumed in computing the radiative equlibrium. To make the solution consistent, one then adjusts the tropopause height until the computed OLR including the troposphere is the same as the target OLR within some desired accuracy. This is a simple problem in root-finding for a function of a single variable (the tropopause pressure), and can be solved by any number of means, Newton's Method and bisection being among the most commonly employed.

For a real gas,the radiative equilibrium in the upper atmosphere depends on the spectrum of the infrared upwelling from below, so we no longer have the luxury of assuming that the stratospheric temperature profile remains fixed as we vary the estimate of the tropopause height. Instead, one must simultaneously solve for both the tropopause height and the corresponding equilibrium profile aloft. This is most easily done by a modification of the time-stepping method we employed to compute the pure radiative equilibrium solutions. As in that case, it is somewhat awkward to

pick an OLR and find the corresponding ground temperature T_g. Instead, we fix T_g, and compute the corresponding OLR. This can be done for a range of ground temperatures, whereafter the ground temperature in equilibrium with any specified solar absorption can be determined. We are back in the familiar business of computing the $OLR(T_g)$ curve, much as we did for the all-troposphere model, but this time taking into account the effect of a self-consistent stratosphere on the OLR.

The general problem of representing convection in climate models is a very challenging one, about which entire volumes have been written. (See the Further Readings section of Chapter 2.) For the problem at hand, there are a number of simplifying assumptions which allow us to avoid some of the more subtle aspects of the subject. First, we will be content to assume that convection instantaneously resets the profile to an adiabatic profile. Next, given where instability occurs in the pure infrared radiative equilibrium profiles, it is safe to assume that convection occurs in a single layer extending from the ground to the tropopause height, without any possibility of multiple interleaved internal convecting and radiative equilibrium layers. Further, we will only seek an equilibrium solution, without attempting to accurately represent the approach to equilibrium. Finally, we carry out the calculation by holding T_g fixed and allowing the rest of the atmosphere to relax to the corresponding equilibrium. Under these circumstances, the elimination of unstable layers by convective mixing can be carried out through the following simple modification to the pure radiative equilibrium time-stepping algorithm: One calculates the adiabat $T_{ad(p)}$ corresponding to the ground temperature T_g and surface pressure p_s. Then at each timestep, wherever $T(p) < T_{ad}(p)$, the temperature is instantaneously reset to T_{ad}. The rationale for doing this is that convection is a much faster process than radiative relaxation, and that wherever the temperature is below the adiabatic temperature, air parcels starting at the ground have enough buoyancy to reach that level, mixing air all along the way. The procedure also assumes that the turbulent coupling of the ground to the overlying air is so strong that ground and air temperature remain essentially identical at all times. The adjustment to the adiabat with surface temperature T_g in general increases the static energy (moist or dry, as appropriate) of the adjusted layer of air. This is not a source of concern if we are only using time-stepping to find the equilibrium state; if we were instead trying to represent the actual time-course of approach to equilibrium, a more sophisticated adjustment approach conserving static energy would need to be employed. Conservative adjustments would transport heat vertically by cooling the lower levels at the same time they are warming the upper levels.

Results for an Earthlike air/CO_2 atmosphere are shown in Figure 4.45. The convection reaches to $380mb$, where the temperature is about $200K$; above that, the atmosphere is in radiative equilibrium, which defines the stratosphere. Note that the temperature continues to decline even above the maximum height reached by convection, because the infrared radiative equilibrium profile also has decreasing temperature, so long as part of the spectrum is optically thick. This is the case also for Earth's real stratosphere, even though ozone heating eventually causes the upper stratospheric temperature to turn around and begin to increase. Thus, one shouldn't take temperature decline as a signature of convection, and in a case where atmospheric heating causes upper stratospheric temperature to increase the temperature minimum will generally be above the top of the convective layer.

The OLR for all three calculations is similar: 296.4 W/m^2 for the radiative-convective model vs. 295 W/m^2 for pure radiative equilibrium and 296.6 W/m^2 for the all-troposphere model. Thus, if the atmosphere is maintained by 295 W/m^2 of absorbed solar radiation, neither the formation of a troposphere by convection nor the formation of a stratosphere by upper level radiative heating has much effect on the surface temperature, though the effects on the atmospheric profile are considerable.

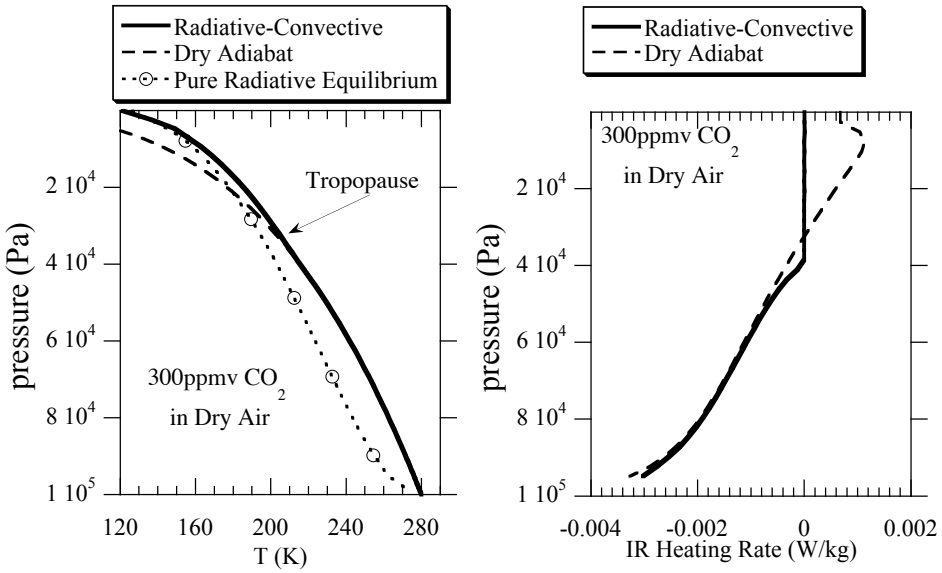

Figure 4.45: Radiative-convective equilibrium for an Earthlike dry atmosphere with $300 ppmv\ CO_2$. The left panel shows the temperature profile in comparison with pure radiative equilibrium and the dry adiabat. The right panel shows the radiative heating rates for the radiative-convective solution and for the all-troposphere model. In all cases the ground temperature is $280K$.

The heating rate profile shown in the right panel of Figure 4.45 sheds some light on the basic mechanism maintaining convection in the troposphere. The entire troposphere is subject to radiative cooling. This feature is guaranteed by the construction of the solution, since any positive heating would warm the atmosphere, causing it to exceed the adiabatic temperature and shutting off convection. Suppose convection has just occurred and reset the temperature to the adiabat. Then, in the next small interval of time, radiative cooling will cause the temperature to fall below the adiabat, triggering convection once more, which adds back the heat lost by radiative cooling and restores the adiabat. The heat is supplied by parcels of air that pick up heat from the ground, become buoyant, and carry the heat upward to the level where it is needed. The thermal balance in the troposphere is between radiative cooling and convective heating. The addition of atmospheric solar absorption to the troposphere would have no effect on the tropospheric temperature or the tropopause height, so long as it doesn't turn the net radiative cooling at any level into a net radiative heating. Short of that happening, the sole effect of tropospheric solar absorption is to reduce the convective heating, and hence the frequency or vigor of convection. However, since infrared cooling is weakest just below the tropopause, solar absorption near the tropopause level can easily move the tropopause downward.

Using Figure 4.45 we can compare two simple estimates of the tropopause height with the actual value. Looking at where the adiabat intersects the pure radiative equilibrium, we find an estimate of $205mb$. This is somewhat too high in altitude, since the formation of the troposphere changes the upward flux and warms the stratosphere. The other way to estimate tropopause height is to look at the heating profile computed for the dry adiabat, shown in the right panel of the Figure. Identifying the region of radiative heating with the stratosphere yields an estimate of $325mb$, which is closer to the true value, but still too high in altitude. In the light of the discussion surrounding Fig. 4.2, we can say that the real gas atmosphere behaves rather like a

pressure-broadened grey-gas atmosphere with optical depth somewhat greater than unity.

How does our computed tropopause height compare to the Earth's actual tropopause? We have defined the tropopause as the height reached by convection, and in comparing this with atmospheric soundings one needs to recall that even above the convective region, the temperature continues to decrease with height, because temperature goes down with height even in pure infrared radiative equilibrium; in the Earth's atmosphere, the temperature eventually begins to increase with height because of the effects of atmospheric solar absorption. Hence, the temperature minimum seen in Earth soundings, which is sometimes loosely called the tropopause, is always somewhat above the convectively defined tropopause (see Problem **??**). Still, if we take the position of the temperature minimum in tropical soundings as an estimate, we note that the tropopause estimated in the preceding calculation is considerably lower in altitude (higher in pressure) than the observed value of about $100mb$. What is it that makes the tropical tropopause so much higher?

The main factor governing the tropopause height is the lapse rate. If the lapse rate is weaker, then one has to go to higher altitudes in order to intersect the radiative equilibrium profile. In the warm tropics, the moist adiabat has significantly weaker gradient than the dry adiabat. In fact, a radiative-convective calculation based on the radiative effects of a dry CO_2/air atmosphere, but employing the moist adiabat in the temperature profile, yields a tropopause height of $130mb$ when the surface temperature is $300K$. This is quite consistent with the observed tropopical tropopause height. This suggests that the effects of moisture on lapse rate are more important than the radiative effects of tropospheric moisture in elevating the tropical tropopause. In other words, the main reason the Earth's present tropical tropopause is higher in altitude than the midlatitude tropopause is that the tropical lapse rate is weaker, owing to the greater influence of moisture for Earthlike tropical temperatures. The additional optical thickness due to the extra water vapor in the tropics plays at most a secondary role. For colder surface temperatures, the moist adiabat deviates less from the dry adiabat, so the tropopause height approaches the lower altitude found in the dry calculation. For example, with a surface temperature of $280K$, a calculation adjusting to the moist adiabat yieldes a tropopause pressure of $250mb$. This reasoning suggests that the tropopause height should be lower in the midlatitudes, and indeed observations show this to be the case. Optical thickness can indeed affect the tropopause height, but it is not the main player on Earth. The calculations referred to here are carried out in Problem **??**

4.9 The lesson learned

This has been a rather arduous chapter – certainly for the author, and no doubt for the reader as well, but hopefully to a lesser extent. The basic lesson, however, can be summed up in a few pithy remarks. The greenhouse effect relies on infrared optical thickness of the atmosphere and temperature decline with height. Real greenhouse gases do not make the atmosphere optically thick uniformly throughout the infrared spectrum. Rather, the optical thickness is concentrated in preferred gas-dependent spectral bands, and the main way the greenhouse effect gets stronger as the gas concentration increases is through the spread of the optically thick regions to every greater portions of the spectrum. There are two basic ways to get the temperature decline which is necessary to translate optical thickness into OLR reduction: radiative equilibrium in an atmosphere that is optically thick through at least part of the spectrum, or convection in an atmosphere where the radiative equilibrium is statically unstable either at the surface or internally. The tropopause height determines the blend of the two mechanisms in force. Both mechanisms can yield a surface temperature very much in excess of the no-atmosphere blackbody temperature, but a radiatively-dominated atmosphere is a very different place from a convectively-dominated

atmosphere, since the latter has vigorous vertical mixing that can give rise to a stew of small scale turbulent phenomena. A mostly radiative-equilibrium atmosphere is a more quiescent place, in which mixing is dominated by the more ponderous large scale fluid motions. A central remaining question is how the tropopause height behaves as an atmosphere is made more optically thick. When do we approach an all-troposphere atmosphere, and when do we approach an all-stratosphere atmosphere? These issues are explored in Problems **??** and **??**. We will return to the problem of tropopause height in Section 5.10.

4.10 For Further Reading

Finding appropriate spectroscopic data appropriate to a novel planetary atmosphere can be a real challenge. A wealth of specialized spectroscopic data can be found in the *Journal of Quantitatitive Spectroscopy and Radiative Transfer*. The reader is directed especially to their 2008 special issue on planetary atmospheres (see (Rothman L 2008, *JQSRT* doi:10.1016/j.jqsrt.2008.02.002). In addition, laboratory and theoretical spectroscopy related to planetary problems can often be found in the journal *Icarus*.

The `HITRAN` spectroscopic database is described in

- Rothman LS, *et al.* 2005: The HITRAN 2004 molecular spectroscopic database, *J. Quant. Spectroscopy and Radiative Transfer*, **96**, pp 139-204.

In a book that one hopes will stick around for a while, there is always some risk in referring to specific means of obtaining digital data. The `HITRAN` database is so valuable, however, that it is sure to be available in some form more or less indefinitely. At the time of writing, the `HITRAN` data can be obtained over the Internet at the URL `http://cfa-www.harvard.edu/hitran`. The 1970's era *Goody and Yung* book on atmospheric radiation refers to obtaining the data on "AFGL tapes," an no doubt earlier books made reference to things like "punch cards" or "paper tapes." No doubt, our reference to the "Internet" will seem similarly quaint within a few years.

The `HITRAN` database does not include the very weak CO_2 absorption lines that become important for extremely massive atmospheres such as that of Venus, and moreover, the temperature dependence data of the lines that are included becomes somewhat inaccurate at Venusian temperatures. There are two databases that extend the CO_2 absorption database to cover the Venusian regime, both of which use the same data format as `HITRAN`. The first is the `HITEMP` database. At the time of writing, there is neither a convenient published document describing the database nor a generally accessible download site, but an updated version of the `HITEMP` database and expanded documentation are expected to be made available through the `HITRAN` site in the near future. In the meantime, information about the existing database can be found in

- Rothman LS, *et al.* 1995: HITRAN, HAWKS, and HITEMP High-Temperature Molecular Database, *Proc.Soc.Photo-Optical Instrumentation Engineers* **2471** 105-111.

and the original version of the database can be downloaded by contacting the managers of the `HITRAN` site. A similar high-temperature,high-pressure database is described in

- Tashkun SA, *et al.* 2003: CDSD-1000, the high-temperature carbon dioxide spectroscopic databank, *J. Quantitative Spectroscopy and Radiative Transfer*, **82**, pp 165-196.

It is available online via `ftp` at `ftp.iao.ru/pub/CDSD-1000`.

Information on the CO_2 collision-induced continuum is very sparse. The modeling of the CO_2 continuum used throughout this book (and incorporated in the software supplement) is based on a polynomial fit to absorption coefficients described in

- Kasting JF, Pollack JB and Crisp D 1984: Effects of high C02 levels on the surface temperature and atmospheric oxidation state of the early Earth, *J. Atmos. Chem*, **1**, pp 403-428.

References to the laboratory measurements upon which the parameterization is based amount to one published paper, one NASA technical report, and one unpublished personal communication; these may be found in the above referenced article. Some theoretical developments, which have been incorporated in a few of the more recent representations of the far-infrared continuum, are described in

- Gruszka M and Borysow A 1997: Far Infrared Collision-Induced Absorption of CO2 for the Atmosphere of Venus at Temperatures from 200K to 800K, *Icarus*, **129**, pp 172-177.

but there seem to have been no new laboratory measurements since those discussed in the former paper.

The collision induced continua of H_2, CH_4 and N_2 relevant to the atmosphere of Titan are given in

- Courtin R 1988: Pressure-Induced Absorption Coefficients for Radiative Transfer Calculations in Titan's Atmosphere, *Icarus*, **75**, pp 245-254.

Radiative transfer on gas giants is discussed in

- Guillemot T, *et al* 1994: Are the Giant Planets Fully Convective?, *Icarus*, **112**, pp 337-353.

The water vapor continuum is described in the following two papers:

- Clough SA, Kneizys FX and Davies RW ,1989: Line shape and the water vapor continuum, *Atmospheric Research*, **23**, pp 229-241.
- Grant WB,1990: Water vapor absorption coefficients in the 8-12 μm spectral region: A critical review, *Applied Optics*, **29**, pp 451-462.

The first of these is considered the standard reference at time of writing, but one must take care in reading it, as there are a certain number of typographical errors and mislabeled figures.

The full-featured `ccm` radiation code is described in complete and somewhat intimidating detail as part of the general description of the NCAR Community Atmospheric Model (CAM) in NCAR Technical Note TN-464+STR, available at the time of writing at

- `http://www.ccsm.ucar.edu/models/atm-cam/docs/description/`.

An accessible overview of an earlier version of the radiation model can be found in

- Kiehl J and Briegleb B 1992: Comparison of the Observed and Calculated Clear Sky Greenhouse-Effect - Implications for Climate Studies, *J. Geophys. Res*, **97 (D9)**, 10037-10049.

A version of this radiation code with a simple `Python` user interface is distributed as part of the software supplement to this book. The `Python` interface makes it easy to use the code to compute OLR and heating rates within a `Python` script, and eliminates the need for any familiarity with the `FORTRAN` language in which the underlying computation is written.

The correlated-k refinement of the exponential-sums approximation to the transmission function is described in

- Lacis A and Oinas V, 1991: A description of the correlated k-distribution method for modeling non-grey gaseous absorption, thermal emission and multiple scattering in vertically inhomogeneous atmospheres. *J. Geophys. Res*, **96**, 90279063.

Chapter 5

Scattering

5.1 Overview

In the atmospheres considered so far, the blackbody source term adds new radiation to the atmosphere travelling in all directions, but once present in the atmosphere radiation travels in a fixed direction; it can be absorbed as it travels, but it does not scatter into other directions. In this class of problems, the two-stream approximation consists entirely in doing the calculation for a single equivalent propagation angle $\bar{\theta}$, but does not change the essential structure of the full problem. If one wanted more information about the angular distribution, one would simply do the same problem over several times, with different θ, and average the results to get the net upward and downward fluxes; each calculation is independent, and has the form of a simple first order ordinary differential equation for the flux. With scattering the situation is very different, as the scattering couples the flux at one angle with the fluxes at all other angles. The full problem now takes the form of a computationally demanding integro-differential equation, with the derivative of the flux at a given angle expressed as a weighted integral over the fluxes at all other angles.

Light with wavelengths in the near-infrared or shorter is significantly scattered from molecules, though molecules are too small to appreciably scatter thermal infrared or longer wavelengths. Many atmospheres (Earth's included) contain very fine aerosol particles with diameters on the order of a μm or less; they are typically made of mineral dust, or of condensed substances such as sulfuric acid or other sulfur compounds. They are very powerful scatterers of solar radiation, and therefore can significantly affect a planets's albedo even when the total mass of aerosols is quite small. Cloud particles made of various condensed substances have typical diameters of 10-100 μm. Because water clouds like Earth's absorb so strongly in the infrared, cloud scattering is often thought of primarily in terms of the solar spectrum. However, taking a broader view, cloud substances commonly found on other planets have a very important thermal infrared scattering effect. Water clouds are the exception, rather than the rule, but because of their importance on Earth, thermal infrared scattering by clouds is a far less developed subject than is shortwave scattering.

Clouds, in their many and varied manifestations, pose one of the greatest challenges to the understanding of Earth and planetary climate. On Earth, water clouds reflect a great deal of sunlight but also have a considerable greenhouse effect. The net cloud effect is a fairly small residual of two large and uncertain terms, and the way the two effects play out against each other plays a central role in climate change problems extending from the Early Earth to Cretaceous Warmth, to ice ages, to global warming, and the distant-future fate of our climate. The high

albedo of Venus is caused largely by clouds made of sulfuric acid droplets, but the very same clouds scatter and absorb infrared radiation, and help to increase the planet's greenhouse effect. On an Early Mars with a 2 *bar* CO_2 atmosphere, formation of clouds of CO_2 ice would play an important role in the planet's climate, both in the infrared nd solar spectrum. Titan's present-day methane clouds affect the satellite's radiation budget both through infrared and visible scattering; Neptune is cold enough that methane ice clouds can form in parts of its atmosphere. The swirling psychedelic colors of Jupiter and Saturn arise from clouds of a dozen or more different types, which no doubt also affect the radiation budget. It is hard to think of a planetary atmosphere in which clouds do not play a significant role.

Regardless of whether clouds exert a greenhouse effect by scattering or by the conventional absorption/emission process, the cross-cutting issue is that clouds affect the incoming and outgoing part of the planet's radiation budget in large but opposing ways. Clouds always reflect some incoming stellar radiation, and clouds at appreciable heights above the surface almost invariably reduce the outgoing infrared radiation. It will turn out that the net radiative effects of clouds are highly sensitive to the size of the particles of which they are composed. This leads to the disconcerting conclusion that the climate of an object as large as an entire planet can be strongly affected by poorly-understood processes happening on the scale of a few micrometers.

Scattering calculations play a critical role not only in determining planetary radiation balance, but also in interpreting a wide range of observations of the Earth, Solar System planets, and extrasolar planets. There is no doubt that if justice were to be served (and the reader had unlimited time) scattering should be given a treatment at least as in-depth as that which we have accorded to purely absorbing/emitting atmospheres. However, in order to maintain progress towards our primary goal of understanding the essentials of planetary climate, the treatment given in this chapter will be highly abbreviated, and focus on the minimal understanding of the subject needed to estimate planetary albedo, shortwave atmospheric heating, and the basic effect of clouds on outgoing infrared radiation and on solar absorption. In particular, we will leap directly into the two-stream approximation, without much discussion of the properties of the scattering equations in their full generality. Were it not for the position of clouds at the forefront of much research on planetary climate, we would be content to leave the discussion of scattering to a few brief remarks concerning planetary albedo.

Atmospheric absorption of the relatively short wave light from a planet's star is often, though not invariably, significantly affected by scattering. Hence, the absorption of stellar near-infrared, visible and ultraviolet radiation will be discussed in this chapter, together with a few implications for atmospheric structure. The effect of absorption of incident light on the temperature profile of the upper atmosphere was derived for grey gases in Chapter 4, but in the present chapter the reader will find results bearing on atmospheric heating due to absorption of incoming stellar radiation by CO_2, water vapor and methane in a variety of planetary contexts both within and outside the Solar System. The effect of ultraviolet absorption by ozone on the stratospheric temperature structure of Earth and Earthlike planets will also be discussed here.

5.2 Basic concepts

The atmosphere can be considered to be a mix of particles, some of which absorb, some of which scatter, and some of which do both. The particles could be molecules, or they could be macroscopic particles of a condensed substance, as in the case of cloud droplets or dust particles. One builds up the absorbing and scattering properties of the atmosphere as a whole from the absorbing and scattering properties of the individual particles. In keeping with the usage in the preceding

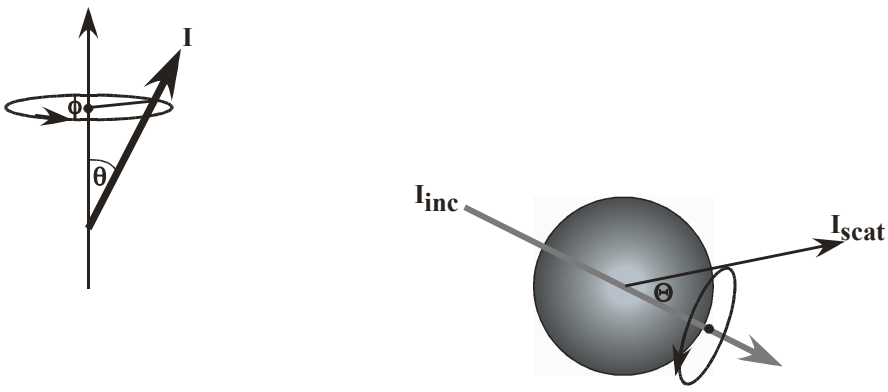

Figure 5.1: Definition of propagation angles (left) and scattering angle (right).

chapters, we will ultimately characterize the effect of atmospheric composition on radiation in terms of scattering properties per unit mass of atmosphere, just as we did for the absorption coefficient.

Consider a parallel, monochromatic (single-frequency) beam of light with flux F in W/m^2 travelling in some specific direction. When the beam encounters a particle of finite extent, a certain amount of the flux will be absorbed, and a certain amount will be scattered into other angles. The rate at which energy is taken out of the beam by absorption and scattering can be characterized in terms of coefficients with dimensions of area, which are known as *cross-sections*. The rate of energy absorption is $F\chi_{abs}$, where χ_{abs} is the *absorption cross-section*, and the rate of energy scattering into other directions is $F\chi_{sca}$, where χ_{sca} is the *scattering cross-section* [1]. The scattering and absorption cross-sections can be quite different from the actual cross-section area of the object. The cross-sections can be thought of as the cross-section areas of hypothetical equivalent objects which absorb or scatter all light hitting the object while leaving the rest of the beam to pass by undisturbed. The ratio of scattering cross-section to the actual cross-section area of the scatterer is called the *scattering efficiency*, Q_{sca}. For a spherical particle of radius r, $Q_{sca} = \chi_{sca}/(\pi r^2)$. The *absorption efficiency* is defined similarly. For spherical particles, the cross sections are independent of the angle at which the radiation is directed at the particle. For non-spherical particles the cross-sections for an individual particle depend on angle, but the typical physical situation involves scattering off of an ensemble of particles presented with random orientations. In this case, we can average over all orientations and represent the mean scattering or absorption in terms of the cross-section for an equivalent sphere. This approach can break down if particles are not randomly oriented, as can be the case for plate-like ice crystals that become oriented through frictional drag forces as they fall. The *single-scattering albedo* for a particle is the ratio of flux of the incident beam lost via scattering to net flux lost. Using the notation ω_{po} for the single-scattering albedo of an individual particle, we have $\omega_{po} = \chi_{sca}/(\chi_{sca} + \chi_{abs})$. Later we will introduce the single-scattering albedo for the medium as a whole. The cross sections for particles or molecules can be measured in the laboratory, and often can be computed from basic physical principles.

Since the radiation fields we will deal with are generally distributed over a range of frequencies and direction, instead of being monochromatic and unidirectional, we will write our equations

[1] The more usual notation for the cross-section is σ, but in our subject matter that symbol has been reserved for the Stefan-Boltzman constant. One can think of χ as standing for $\chi\rho\omega\sigma\sigma$-section.

in terms of the spectral irradiance I introduced in Chapters 3 and 4. Recall that if the spectral irradiance is $I((\theta,\phi),\nu)$ at a given point, then $Id\Omega d\nu$ is the flux of radiation in frequency band $d\nu$ with directions of travel within a solid angle $d\Omega$ about the direction (θ,ϕ), which passes through a plane perpendicular to the direction of travel. To apply the results of the preceding paragraph to smoothly distributed radiation, one needs only to substitute $Id\Omega d\nu$ for the incident flux F.

As in previous chapters, we'll make the plane-parallel assumption, and assume that I depends on position only through pressure. Suppose that in the vicinity of some pressure level p there are N scatterers of type i per unit mass of atmosphere, and that each scatterer has mass m_i. Suppose that the light impinging on the layer is traveling with angle θ to the vertical. Then, taking a layer of thickness dp which is small enough that multiple scattering can be neglected, the rate of energy lost by the incident beam due to absorption and due to scattering into different angles is

$$-\frac{dp}{g\cos\theta}N\cdot(\chi_{abs,i}+\chi_{sca,i})Id\Omega d\nu = -\frac{dp}{g\cos\theta}q_i\cdot(\frac{1}{m_i}\chi_{abs,i}+\frac{1}{m_i}\chi_{sca,i})Id\Omega d\nu \tag{5.1}$$

where q_i is the mass concentration of the particles in question. From this we can define the absorption coefficient of the substance $\kappa_i \equiv \chi_{abs,i}/m_i$ which has units of m^2/kg. This absorption coefficient is the same quantity we defined in Chapter 4 in connection with gaseous absorption. The additional term in the above equation characterizes the energy lost from the incident beam due to scattering. We won't introduce separate notation for this term since scattering is most commonly characterized in terms of the cross section itself.

If there is only one optically active substance i in the atmosphere, we define the optical depth in the vertical direction by the equation

$$\frac{d\tau^*}{dp} = -\frac{1}{g}(\kappa_i+\frac{1}{m_i}\chi_{sca,i})q_i \tag{5.2}$$

Because the absorbing and scattering properties typically depend on wavenumber, the optical depth is generally a function of wavenumber, though we will only append a wavenumber subscript to τ^* when we wish to call attention specifically to the wavenumber dependence. If there are many types of scatterers and absorbers – which could include particles of a single substance but with different sizes – then we define the optical depth by summing over all species. Thus

$$\frac{d\tau^*}{dp} = -\frac{1}{g}(\kappa+\sum_i\frac{1}{m_i}\chi_{sca,i}q_i) \tag{5.3}$$

where the net absorption coefficient is

$$\kappa \equiv \sum_i \kappa_i q_i \tag{5.4}$$

We then define the single-scattering albedo for the medium as a whole as

$$\omega_o \equiv \frac{\sum q_i\chi_{sca,i}/m_i}{\kappa+\sum q_i\chi_{sca,i}/m_i} \tag{5.5}$$

The pair (κ,ω_o) constitutes the basic description of the absorption and scattering properties of the medium. Both are typically functions of wavelength and altitude, and may also directly be functions of pressure and temperature. If the medium consists of only a single type of particle, and the gas in which the particles are suspended neither absorbs nor scatters, then $\omega_o = \omega_{po}$. In general, though, the single-scattering albedo of the medium depends on the mix of absorbers and scatterers. For example, an atmosphere may consist of a mix of cloud particles which are perfect

scatterers ($\omega_{po} = 1$) with a strong greenhouse gas which is an absorber. In this case, ω_o will go down as the greenhouse gas concentration increases, even if the cloud particle concentration is kept fixed.

Using the definition of optical depth, Eq 5.1 for the rate of energy loss from the beam can be rewritten as simply $dI = -Id\tau^*/\cos\theta$. Since the vertical component of flux is $I\cos\theta$, this expression can be recast as an expression for the rate of loss of vertical flux, namely

$$dI\cos\theta = -Id\tau^* \tag{5.6}$$

The proportion of this lost to scattering is ω_o while the proportion lost to absorption is $1 - \omega_o$. The fate of the energy lost to absorption is different from the fate of that lost due to scattering. The former disappears into the pool of atmospheric heat, whereas energy lost to scattering from one beam reappears as flux in a range of other directions, so we need to keep track of the two loss mechanisms separately. The beam loss in a given direction is offset by two source terms: one due to thermal emission, and one due to scattering from other directions. The thermal emission term is proportional to the Planck function, and can be treated in a fashion similar to that used in deriving the Schwartzschild equations. We'll leave the thermal emission out for now, and concentrate on scattering; the thermal emission term will be put back in in Section 5.5.

To understand better where the scattered flux goes, consider the energy budget for a box of thickness $d\tau^*$ in the vertical, shown from the side in Figure 5.2. Since the radiation field is independent of the horizontal dimensions, the flux entering the box from the side is the same as the flux leaving it from the side, and does not affect the budget. If the base of the box has area A, an amount $A \cdot I(\tau^*)\cos\theta$ enters the box from the bottom and a somewhat lesser amount $A \cdot I(\theta, \phi, \tau^* + d\tau^*)\cos\theta$ leaves the box from the top. Taking the difference gives the loss of energy from the beam per unit time, due to scattering and absorption. Using Eq. 5.6 this can be written as simply $A \cdot I(\theta, \phi, \tau^*)d\tau^*$; it doesn't matter whether I is evaluated at $d\tau^*$ or $\tau^* + d\tau^*$ in this expression, since $d\tau^*$ is presumed small. The energy per unit time scattered and restributed into all other directions is then $A \cdot \omega_o I(\theta, \phi, \tau^*)d\tau^*$. Now, to write an equation for how the vertical component of flux changes between τ^* and $\tau^* + d\tau^*$, we need to find how much flux is added to the direction (θ, ϕ) by scattering from all other directions of radiation impinging on the layer. We can do this by considering the incident radiation one direction at a time, and summing up. Consider a beam of light traveling with direction (θ', ϕ'), having radiance $I(\theta', \phi', \tau^*)$. The scattering contributed to direction (θ, ϕ) comes from the scatterers in the shaded parallelogram shown in Figure 5.2, which is greater than the amount of scatterer in a rectangular box by a factor of $1/\cos\theta$. Further, only a proportion of the radiation scattered from the contents of the parallelogram goes into the direction (θ, ϕ), We will write this proportion as $P/4\pi$, where P depends on both the incident and scattered directions. Thus, the radianced contributed to direction (θ, ϕ) by scattering is $A \cdot (P/4\pi)\omega_o I(\theta', \phi', \tau^*)d\tau^*/\cos\theta$, and the vertical component of this is obtained by multiplying by $\cos\theta$, yielding $A \cdot (P/4\pi)\omega_o I(\theta', \phi', \tau^*)d\tau^*$. This is the vertical flux contributed by scattering, and is added in to the flux leaving the top of the box. The scattering acts as a source of radiation in direction (θ, ϕ), which is added to the right hand side of Eq. 5.6. Dividing out the area of the base of the box, the flux balance for the box becomes

$$dI(\theta, \phi)\cos\theta = -I(\theta, \phi)d\tau^* + \frac{\omega_o}{4\pi}P(\theta, \phi, \theta', \phi')I(\theta', \phi')d\tau^* \tag{5.7}$$

if one considers only the flux contributed by scattering of a single direction (θ', ϕ') . To complete the equation, one must integrate over all incident angles (θ', ϕ'). To determine the radiation field in its full generality, it is necessary to satisfy the flux balance for each direction of propagation simultaneously. Before proceeding toward that goal, we'll check Eq. 5.7 to verify that the scattered

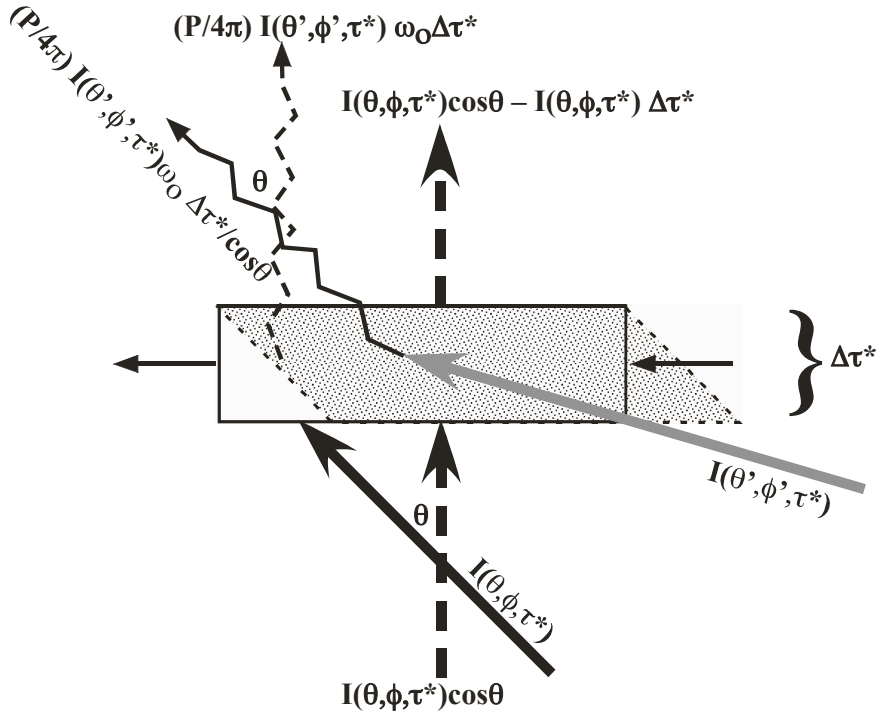

Figure 5.2: A scattering control volume, showing the flux added in the direction (θ, ϕ) due to scattering of an incident beam with direction (θ', ϕ'). Only the contribution from the slab of thickness $d\tau^*$ is considered. The incident beam illuminates the entire slab, but only the scatterers in the shaded parallelogram contribute to scattered radiation in the direction (θ, ϕ). The solid squiggly line represents scattered radiation, and the dashed squiggly line represents the vertical component of the scattered flux. The vertical straight,dashed arrows give the vertical component of the flux in the (θ, ϕ) direction, and show how it changes as the slab is traversed. Flux is lost from the (θ, ϕ) beam owing to absorption and scattering. The flux lost due to scattering shows up as scattererd radiation in all other directions; these are not shown in the diagram.

energy is conserved. Applying the control volume sketch to the *incident* beam direction, we infer that the incident beam traveling in direction (θ', ϕ') deposits energy in the control volume at a rate $I(\theta', \phi')d\tau^*$ (per unit area). A proportion $\omega_o P/4\pi$ of this should show up as an increase in the energy in the box propagating in direction (θ, ϕ), and that is precisely the source term appearing in Eq 5.7 The books are indeed balanced.

It is worth thinking quite hard about Figure 5.2, because the cosine terms that appear in such computations – and which are the source of most of the difficulties in writing two-stream approximations – can be quite confusing. The cosine weights play two quite different roles. In one guise, they express the number of scatterers or absorbers encountered along a slanted path, but in another guise they represent the projection of the flux on the vertical direction. Most confusion can be resolved by thinking hard about the energy budget of the control volume.

The quantity P introduced in Figure 5.2 is called the *phase function*, and describes how the scattered radiation is distributed over directions. For spherically symmetric scatterers, the

phase function depends only on the angle Θ between the incident beam and a scattered beam (as depicted in Fig. 5.1). The phase function is usually expressed as a function of $\cos\Theta$. If $\hat{n}_I$ is the unit vector in the direction of propagation of the incident beam, and $\hat{n}_{sca}$ is the unit vector in the direction of propagation of some scattered radiation, then

$$\cos\Theta = \hat{n}_I \cdot \hat{n}_{sca} = \cos\theta\cos\theta' + \sin\theta\sin\theta'\cos(\phi - \phi') \tag{5.8}$$

where θ and ϕ are the direction angles of the incident beam and θ' and ϕ' are the angles of the scattered beam under consideration. The phase function for the medium as a whole can be determined from the phase functions of the individual particles doing the scattering – remember that from ω_o and $d\tau^*$ we already know the amount of energy scattered out of a beam, so the phase function only needs to tell us how that energy is distributed amongst directions. The phase function for an individual particle is defined in such a way that the scattered flux within an element of solid angle $d\Omega'$ near direction (θ', ϕ') is $\chi_{sca} I(\theta, \phi) P(\cos\Theta(\theta, \phi, \theta', \phi')) d\Omega'/4\pi$. P is normalized such that $\int P d\Omega' = 4\pi$, so that integrating the scattered flux over all solid angles yields $\chi_{sca} I$. Note further that

$$\int P(\cos\Theta) d\Omega = 2\pi \int_{-1}^{1} P(\cos\Theta) d\cos\Theta = \int P(\cos\Theta) d\Omega' = 4\pi \tag{5.9}$$

where solid angle integrals without limits specified explicitly denote integration over the entire sphere. The final equality is a matter of definition and the other two equalities follow because one is free to rotate the coordinate system so as to define the angles with respect to any chosen axis, if one is integrating over the entire sphere. Isotropic scattering, in which the scattered radiation is distributed uniformly over all angles, is defined by $P = 1$.

If the scatterers in the atmosphere are all identical particles, then the phase function for the medium is the same as the phase function for an individual particle. If the phase functions differ from one particle to another, then the phase function for the medium is simply the average of the individual particle phase functions, weighted compatibly with Eq 5.3. The averaging is particularly important when the particles are non-spherical. Though the phase function for any individual particle is not a function of $\cos\Theta$ alone, the particles are generally oriented in random directions, and the average phase function for an ensemble of randomly oriented particles acts like the phase function for an equivalent sphere.

If one divides Eq 5.7 by $d\tau^*$ and integrates over all incident directions (θ', ϕ') the equation for the vertical component of the flux due to radiation traveling in direction (θ, ϕ) is found to be

$$\frac{d}{d\tau^*} I(\cos\theta, \phi)\cos\theta = -I(\cos\theta, \phi) + \frac{\omega_o}{4\pi} \int P(\cos\Theta) I(\cos\theta', \phi') d\Omega' \tag{5.10}$$

where $\cos\Theta$ is given in terms of $(\theta, \theta', \phi - \phi')$ by Eq. 5.8. Thermal emission would add an additional source term $B(\nu, T(\tau^*))$ to the right hand side, but we shall leave that out for now. This is the full equation whose solutions give the radiation field. The integral couples together all directions of propagation; if one approximated the integral by a sum over 100 angles, for example, the equation would be the equivalent of solving a system of 100 coupled ordinary differential equations. While, with modern computers, this is not so overwhelming a task as it once might have seemed, it is still intractible in typical climate calculations, where one is doing the calculation for each of a large array of wavenumbers, at each time step of a radiative-convective model, and perhaps for each latitude and longitude grid point in a general circulation model as well. Moreover, it is always helpful to have a simplified form in hand if one's goal is understanding and not merely computing a number. Hence, our emphasis will be on reduction of the equation to an approximate set of equations for two streams of radiation, which may be thought of as the upward and downward streams. In this

section we'll derive some exact constraints, which will be used to obtain two-stream closures of the problem in Section 5.5.

We first need to define the upward and downward fluxes, which are

$$\begin{aligned} I_+ &\equiv \int_{\Omega^+} I(\cos\theta,\phi)\cos\theta d\Omega = \int_{\cos\theta=0}^{1}\int_0^{2\pi} I(\cos\theta,\phi)\cos\theta d\phi d\cos\theta \\ I_- &\equiv -\int_{\Omega^-} I(\cos\theta,\phi)\cos\theta d\Omega = -\int_{\cos\theta=-1}^{0}\int_0^{2\pi} I(\cos\theta,\phi)\cos\theta d\phi d\cos\theta \end{aligned} \tag{5.11}$$

The fluxes are defined in such a way that both are positive numbers. Given that $d\Omega$ can be written as $d\cos\theta \cdot d\phi$ it is convenient to write all the fluxes as a function of $\cos\theta$, as we have done here. Henceforth we shall use Ω^+ and Ω^- as shorthand for integral over the upward or downward hemisphere, respectively. With these definitions, the net vertical flux (positive upward) is $I_+ - I_- = \int I\cos\theta d\Omega$, the integral being taken over the full sphere.

Solar radiation enters the top of the atmosphere in the form of a nearly parallel beam of radiation, characterized by an essentially unique angle of propagation. It is gradually converted by scattering into radiation that is continuously distributed over angles. Because the incoming solar radiation has an angular distribution concentrated on a single direction of propagation, it is useful to divide the radiation up into a *direct beam* component propagating exactly in this direction, and a *diffuse* component, travelling over all angles. You can see the Sun as a sharply defined disk in clear sky, which shows that the direct-beam solar radiation isn't completely converted into diffuse radiation by scattering, except perhaps in heavily cloudy conditions. To define the direct beam flux, let $L_\circledast$ be the solar constant and ζ be the angle between the vertical and the line pointing toward the Sun; ζ is called the *zenith angle*. By convention, the zenith angle is defined as the angle of the vector pointing *toward* the Sun, rather than the direction of the rays coming *from* the Sun. Thus, if θ_{dir} is the angle of the direct-beam radiation in our usual angle coordinate system, the zenith angle is $\zeta = \pi - \theta_{dir}$. The azimuth angle of the direct beam radiation ϕ_{dir} is defined in the usual coordinate system.

Now, since the direct beam flux is concentrated in a single direction, there is essentially zero probability of any scattered flux contributing back into the exact direct beam direction. That would be like the exactly hitting an infinitesimal dot on a dartboard. Therefore, flux is scattered out of the direct beam but is never added into it, and the direct beam decays exponentially. Making use of the slant path, the direct beam flux is then $L_\circledast \exp(-(\tau^*_\infty - \tau^*)/\cos\zeta)$. We rewrite the flux as the sum of a diffuse component and the direct beam:

$$I(\cos\theta,\phi) = I_{diff}(\cos\theta,\phi) + L_\circledast \exp(-(\tau^*_\infty - \tau^*)/\cos\zeta)\delta(\theta - (\pi-\zeta))\delta(\phi - \phi_{dir}) \tag{5.12}$$

where I_{diff} is the diffuse flux and δ is the Dirac delta-function. From now on, for economy of notation we'll drop the "diff" subscript on the diffuse radiation and simply write I for the diffuse component. In typical situations, the top-of-atmosphere boundary condition states that the radiance of all downward-directed angles of the diffuse component must vanish.

Substituting into Eq 5.10, the equation for the diffuse flux becomes

$$\begin{aligned} \frac{d}{d\tau^*} I(\cos\theta,\phi)\cos\theta = &- I(\cos\theta,\phi) + \frac{\omega_o}{4\pi}\int P(\cos\Theta) I(\cos\theta',\phi') d\Omega' \\ &+ L_\circledast \frac{\omega_o}{4\pi} P(\cos\Theta(-\cos\zeta,\cos\theta,\phi-\phi_{dir})\exp\left(-(\tau^*_\infty - \tau^*)/\cos\zeta\right) \end{aligned} \tag{5.13}$$

The scattering from the direct beam acts as a source term for the diffuse radiation. Integrating

over all angles yields the following exact expression for the net vertical diffuse flux

$$\frac{d}{d\tau^*}(I_+ - I_-) = -(1-\omega_o)\int I(\cos\theta',\phi')d\Omega' + \omega_o L_\circledast \exp\left(-(\tau^*_\infty - \tau^*)/\cos\zeta\right) \tag{5.14}$$

since $\int P(\cos\Theta)d\Omega = 4\pi$. In this expression, I_+ and I_- now represent just the diffuse part of the flux. *Conservative scattering* – that is, scattering without absorption – is defined by $\omega_o = 1$. For conservative scattering the first term on the right hand side of Eq. 5.14 vanishes. Integrating the direct beam term with respect to τ^* just multiplies it by $\cos\zeta$, whence we are left with the result that $I_+ - I_- - L_\circledast \cos\zeta \exp\left(-(\tau^*_\infty - \tau^*)/\cos\zeta\right)$ is a constant. Thus, for conservative scattering the sum of the direct beam vertical flux – which is negative because it is downward – with the diffuse flux is independent of height. As the direct beam is depleted, the flux lost goes completely into the diffuse component. This is as it should be, because, in conservative scattering, the flux lost has no place else to go.

Eq. 5.14 provides the first of the two constraints needed to derive the two-stream approximations. The second constraint is provided by multiplying Eq 5.13 by a function $H(\cos\theta)$ which is antisymmetric between the upward and downward hemispheres, and then performing the angle integral. The rationale for multiplying by an antisymmetric function is that we already know something about $I_+ - I_-$ from the first constraint, and weighting by an antisymmetric functions gives us some information about $I_+ + I_-$. Multiplying by H and carrying out the angle integral, we get

$$\begin{aligned}\frac{d}{d\tau^*}\int I(\cos\theta,\phi)H(\cos\theta)\cos\theta d\Omega = &-\int IH(\cos\theta)d\Omega + \omega_o\int G(\cos\theta')I(\cos\theta',\phi')d\Omega' \\ &+ \omega_o L_\circledast G(-\cos\zeta)\exp\left(-(\tau^*_\infty - \tau^*)/\cos\zeta\right)\end{aligned} \tag{5.15}$$

where

$$G(\cos\theta') = \frac{1}{4\pi}\int H(\cos\theta)P(\cos\Theta(\cos\theta,\cos\theta',\cos\phi))d\Omega \tag{5.16}$$

We were free to replace $\cos(\phi - \phi')$ in this expression by $\cos\phi$, since the integral is taken over all angles ϕ and so a constant shift of azimuth angle does not change the value of the integral. Sinced H is assumed antisymmetric, the function $G(\cos\theta')$ characterizes the up-down asymmetry of scattering of a beam coming in with angle θ'. The symmetry properties of $\cos\Theta$ imply that $G(-\cos\theta') = -G(\cos\theta')$.

Exercise 5.2.1 Derive the claimed antisymmetry property of G.

If the phase function satisfies $P(\cos\Theta) = P(-\cos\Theta)$ the scattering is said to be *symmetric*. For symmetric scatterers, there is no difference between scattering in the forward and backward directions. From Eq. 5.8 it follows that $\cos\Theta(\cos\theta,\cos\theta',\phi) = -\cos\Theta(-\cos\theta,\cos\theta',\phi+\pi)$. The antisymmetry of $H(\cos\theta)$ then implies that G vanishes if P is symmetric, since the contribution to the integral from $(\cos\theta,\phi)$ cancels the contribution from $(-\cos\theta, phi + \pi)$. For symmetric scattering, Eq. 5.15 takes on a particularly simple form, since both terms proportional to ω_o vanish. The physical content of this result is that symmetric scattering does not directly affect the asymmetric component of the diffuse radiation, since equal amounts are scattered into the upward and downward directions.

When the scattering isn't symmetric, the terms involving G do not vanish, and we need a way to characterize the asymmetry of the phase function. The most common measure of asymmetry is the cosine-weighted average of the phase function

$$\tilde{g} \equiv \frac{1}{2}\int_{\cos\Theta=-1}^{1} P(\cos\Theta)\cos\Theta d\cos\Theta \tag{5.17}$$

which goes simply by the name of the *asymmetry factor*. The asymmetry factor vanishes for symmetric scattering. All radiation is backscattered in the limit $\tilde{g} = -1$, as if the scattering particles were little mirrors. When $\tilde{g} = 1$ there is no back-scatter at all, and all rays continue in the forward direction, though their direction of travel is altered by the particles, much as if they were little lenses.

The asymmetry factor $\tilde{g}$ characterized forward-backward scattering asymmetry relative to the direction of travel of the incident beam, but some tedious manipulations with Eq. 5.8 allow one to show that the same factor characterizes cosine-weighted asymmetry in the upward-downward direction, regardless of the direction of the incident beam. Specifically, if the incident beam has direction (ϕ', θ'), then

$$\frac{1}{4\pi} \int P(\cos\Theta(\cos\theta, \cos\theta', \cos\phi)) \cos\theta d\Omega = \tilde{g} \cos\theta' \tag{5.18}$$

where $d\Omega = d\phi \cdot d\cos\theta$ as usual. This leads to a particularly tidy result if we choose $H(\cos\theta) = \cos\theta$ in Eq. 5.15, since then $G(\cos\theta') = \tilde{g}\cos\theta'$ and the antisymmetric projection of the scattering equation becomes

$$\frac{d}{d\tau^*} \int I(\cos\theta, \phi) \cos^2\theta d\Omega = -(1 - \omega_o\tilde{g})(I_+ - I_-) + \omega_o L_\circledast \tilde{g} \cos\zeta \exp\left(-(\tau^*_\infty - \tau^*)/\cos\zeta\right) \tag{5.19}$$

The integral appearing on the left hand side is sensitive only to the symmetric component of the radiance field. In order to obtain a two-stream closure, it is necessary to express the integral in terms of $I_+ + I_-$, which requires making an assumption about the angular distribution of the radiation. The same assumption applied to the right hand side of Eq. 5.14 allows one to estimate $\int I d\Omega$ in terms of $I_+ + I_-$. The different forms of two-stream approximations we shall encounter correspond to different assumptions about the angular distribution of radiance.

For other forms of H the asymmetry function $G(\cos\theta')$ has more complicated behavior that is not so simple to characterize. The other form of H we shall have occasion to deal with is

$$H(\cos\theta) = \begin{cases} 1 & \text{for } \cos\theta \geq 0, \\ -1 & \text{for } \cos\theta < 0, \end{cases} \tag{5.20}$$

which is used to derive the hemispherically-isotropic form of the two-stream equations. This choice is convenient because the left hand side of Eq. 5.15 reduces to the derivative of $I_+ + I_-$, but it is inconvenient because G no longer has a simple cosinusoidal dependence on the incident angle. One could simply compute G from the phase function for the medium and use this to form the weights in the scattering equation, but given the inaccuracies we already accept in reducing the problem to two streams, it is hardly worth the effort. Instead, we will *approximate* G as having a cosinusoidal dependance as it does in the previous case. This approximation is exact if the phase function has the form $P = 1 - \frac{1}{3}b + a\cos\Theta + b\cos^2\Theta$, and one can add a third and fourth order term without very seriously compromising the representation. By carrying out the integral defining the asymmetry factor, we find that $\tilde{g} = \frac{1}{3}a$. Then, evaluating G for the assumed form of phase function we find $G = \frac{1}{2}a\cos\theta' = \frac{3}{2}\tilde{g}\cos\theta'$. With this result the antisymmetric scattering equation projection becomes

$$\frac{d}{d\tau^*}(I_+ + I_-) = -(1 - \omega_o\frac{3}{2}\tilde{g})(I_+ - I_-) + \omega_o L_\circledast \frac{3}{2}\tilde{g} \cos\zeta \exp\left(-(\tau^*_\infty - \tau^*)/\cos\zeta\right) \tag{5.21}$$

The right hand side becomes precisely the same as Eq. 5.19 if we redefine the asymmetry factor to be $\frac{3}{2}\tilde{g}$. Eq. 5.21 is already written in terms of the upward and downward stream, and needs no

further approximation in order to be used to derive a two-stream approximation. To complete the derivation of the hemispherically isotropic two-stream equation, one need only write the integral $\int I d\Omega$ appearing in Eq. 5.14 in terms of $I_+ + I_-$ using the assumed angular distribution of radiance. If I is assumed hemispherically isotropic in forward and backward directions separately, then this integral is in fact $2(I_+ + I_-)$, which completes the closure of the problem.

There is one last basic quantity we need to define, namely the *index of refraction*, which characterizes the effect of a medium on the propagation of electromagnetic radiation. It will turn out that the index of refraction amounts to an alternate way of representing the information already present in the scattering and absorption cross-sections. For a broad class of materials – including all that are of significance in planetary climate – the propagation of electromagnetic radiation in the material is described by equations that are identical to Maxwell's electromagnetic equations, save for a change in the constant that determines the speed of propagation (the "speed of light"). In particular, the equations remain linear, so that the superposition of any two solutions to the wave equations is also a solution, allowing complicated solutions to be built up from solutions of more elementary form. The reduction in speed of light in a medium comes about because the electric field of an imposed wave induces a dipole moment in the molecules making up the material, which in turn gives rise to an electric field which modifies that of the imposed wave. The equations remain linear because the induced dipole moment for non-exotic materials is simply proportional to the imposed electric field. When the medium is nonabsorbing, the ratio of the speed in a vacuum to the speed in the medium is a real number known as the *index of refraction*.

The physical import of the index of refraction is that, at a discontinuity in the index such as occurs at the surface of a cloud particle suspended in an atmosphere, the jump in the propagation speed leads to partial reflection of light hitting the interface, and deflection (*refraction*) of the transmitted light relative to the original direction of travel. The larger the jump in the index of refraction, the larger is the reflection and refraction. To a considerable extent, the refraction of light upon hitting an interface can be understood in terms of a particle viewpoint. If one represents a parallel beam of light as a set of parallel streams of particles all moving at speed c_1 in the outer medium, then if the streams hit an interface with a medium where the speed it $c_2 < c_1$, then the streams that hit first will be slowed down first, meaning that the wave front will tilt and the direction of propagation of the beam will be deflected toward the normal, as shown in Figure 5.3. The classic analogy is with a column of soldiers marching in line, who encounter the edge of a muddy field which slows the rate of march. If Θ_1 is the angle of incidence relative to the normal to the interface, and Θ_2 is the angle of the refracted beam on the other side of the interface, then the deflection due to the change in speed is described by Snell's Law, which states $c_2 \sin\Theta_1 = c_1 \sin\Theta_2$, or equivalently $\sin\Theta_2 = (n_1/n_2)\sin\Theta_1$. Now, if a beam is traveling within the medium at angle Θ_2 and exits into a medium with lower index of refraction (e.g. glass to air), then the angle of the exiting beam is given by $\sin\Theta_1 = (n_2/n_1)\sin\Theta_2$; hence, the beam is deflected away from the normal, as indicated in the sketch. At such an interface, if $(n_2/n_1)\sin\Theta_2 > 1$ then there is no transmitted beam and the ray is refracted so much that it is totally reflected back into the medium – a phenomenon known as *total internal reflection*. In reality, there is always some partial reflection at an interface. Partial reflection, as well as many other phenomena we shall encounter, depends on the wave nature of light as described by Maxwell's equations, and cannot be captured by the "corpuscular" viewpoint. This was as much of a groundbreaking conceptual challenge for early optical theorists as blackbody radiation was for investigators presiding over the dawn of quantum theory.

The concept of index of refraction can be extended to absorbing media. Suppose that a plane wave propagating through the medium has spatial dependence $exp(2\pi ikx)$, where x is the distance measured in the direction of propagation. Then, the expression for the speed of the wave

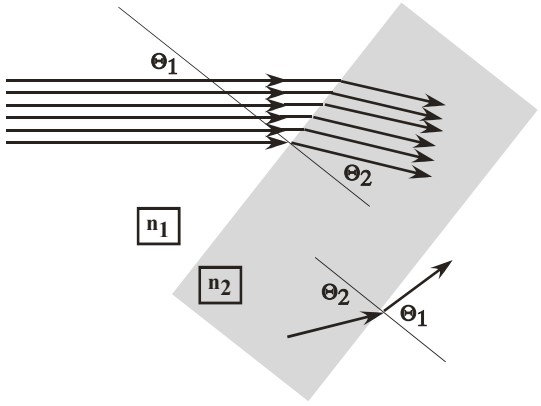

Figure 5.3: Refraction of a beam of light at an interface between a medium with index of refraction n_1 and a medium with index of refraction n_2. In the sketch, $n_2 > n_1$ so the speed of light is slower in the medium than in the surroundings, as is the case for glass or water in air.

in terms of its frequency and wavenumber becomes

$$\frac{\nu}{k} = \frac{1}{n}c \tag{5.22}$$

where c is the speed of light in a vacuum. Thus $k = (\nu/c)n$. Note that ν/c is the vacuum wavenumber we have been using all along to characterize radiation. For real n, k is the wavenumber in the medium, which is larger than the vacuum wavenumber by a factor of n. If we allow n to be complex, its imaginary part characterizes the absorption properties of the medium. To see this, write

$$k_R + ik_I = \frac{\nu}{c}n_R + i\frac{\nu}{c}n_I \tag{5.23}$$

Since the wave has spatial dependence $\exp(2\pi ikx) = \exp(w\pi ik_r x)\exp(-2\pi k_i x)$, the coefficient $2\pi k_i = 2\pi n_I(\nu/c)$ gives the attenuation of the light by absorption, per unit distance travelled. Note that because of the factor ν/c, the quantity $2\pi n_I$ gives the attenuation of the beam after it has traveled by a distance equal to one wavelength of the light. Hence $n_I = 1$ corresponds to an extremely strong absorption. Visible light traveling through such a medium, for example, would be almost completely absorbed by the time it had traveled one μm.

The absorption coefficient k_I is proportional to the absorption cross section per unit mass we introduced in Chapter 4, and which reappeared above in the context of absorption by particles. If the density of the medium is ρ, the corresponding mass absorption coefficient is $\kappa = 2\pi k_I/\rho = n\cdot(\nu/c)/\rho$. In mks units, this quantity has units of m^2/kg, and is thus an absorption cross section per unit mass.

The real index of refraction for some common cloud-forming substances is given in Table 5.1. The index of refraction for these and similar substances lies approximately in the range 1.25 to 1.5, and is only weakly dependent on wavenumber; data also shows the index of refraction to depend only weakly on temperature. The weak dependence of index of refraction on wavelength does give rise to a number of readily observable phenomena, such as separation of colors by a prism or the droplets that give rise to rainbows, but such phenomena, beautiful as they are, are of little importance to planetary energy balance. The one exception to the typically gradual variation

	Thermal-IR	Near-IR	Solar	UV-B
Liquid water	1.40	1.31	1.33	1.43
Water ice	1.53	1.29	1.31	1.39
CO_2 ice	1.45	1.40	1.41	1.54
Liquid CH_4	1.28	1.27	1.27	1.49
H_2SO_4 38%	1.56	1.36	1.38	1.53
H_2SO_4 81%	1.41	1.51	1.44	1.58

Table 5.1: Real part of the index of refraction for selected condensed substances. Thermal-IR data is at 600 cm^{-1}, Near-IR is at 6000 cm^{-1}, Solar at 17000 cm^{-1} (.59 μm), and UV-B at 50000 cm^{-1} (.2 μm). Liquid water data was taken at a temperature of $293K$, water ice at $273K$, CO_2 ice at $100K$, liquid methane at $112K$ and H_2SO_4 at approximately $270K$. The percentage concentrations given for the latter are in weight percent.

of the real index of refraction occurs near spectrally localized absorption features; the real index also has strong variations in the vicinity of such points. In considering the scattering of light by particles suspenden in an atmospheric gas, the index of refraction of the gas can generally be set to unity without much loss of accuracy. A vacuum has $n = 1$, and gases at most densities we'll consider are not much different. Specifically, for a gas $n - 1$ is proportional to the density. At $293K$ and $1bar$, Earth air has an index of refraction of 1.0003 in the visible spectrum. CO_2 in the same conditions has an index of 1.0004, and even at the 90 *bar* surface pressure of Venus has an index of only 1.016. The resultant refraction by the atmospheric gas can be useful in determining the properties of an atmosphere through observations of refraction from the visible through radio spectrum, but it has little effect on scattering by cloud particles.

Insofar as the real index of refraction goes, it would appear that it matters very little what substance a cloud is made of. The minor differences seen in Table 5.1 are far less important than the effects of cloud particle size and the mass of condensed substance in a cloud. The absorption properties, on the other hand, vary substantially from one substance to another, and these can have profound consequences for the effect of clouds on the planetary energy budget. The behavior of the imaginary index for liquid water, water ice, and CO_2 ice is shown in Figure 5.4. Water and water-ice clouds are nearly transparent throughout most of the solar spectrum; for these substances, n_I is less than 10^{-6} for wavenumbers between $10000cm^{-1}$ and $48000cm^{-1}$ (wavelengths between 1 μm and .2 μm, though the absorption increases sharply as one moves into the far ultraviolet. In the thermal infrared spectrum, however, water and water-ice are very good absorbers, having n_I in excess of .1 between wavenumbers of 50 and 1000 cm^{-1}. Such a large value of n_I implies that most thermal infrared flux would be absorbed when passing through a cloud particle having a diameter of 10 μm. For this reason, infrared scattering by water and water-ice clouds can be safely neglected, such clouds being treated as pure absorbers and emitters of infrared. This is not the case for clouds made of CO_2 ice (important on Early Mars and perhaps Snowball Earth) or liquid CH_4 (important on Titan). CO_2 ice clouds are still quite transparent in the solar spectrum, apart from strong absorption in the far ultraviolet. In constrast to water clouds, however, they are largely transparent to thermal infrared. For CO_2 ice clouds, n_I is under 10^{-4} between 1000 and 2000 cm^{-1}, and even between 500 and 1000 cm^{-1} n_I is generally below .01 except for a strong,narrow absorption feature near 600 cm^{-1}. Likewise, liquid methane has n_I well under .001 between 10 and 1200 cm^{-1}. In both cases the infrared scattering effect of clouds can have an important effect on the OLR, leading to a novel form of greenhouse effect. Concentrated sulfuric acid, which makes up aerosols on Earth and the clouds of Venus, is quite transparent for wavenumbers larger than 4000 cm^{-1} but the imaginary index of refraction increases greatly at smaller wavenumbers. In much of the thermal infrared spectrum sulfuric acid absorbs nearly as well as water. Nonetheless,

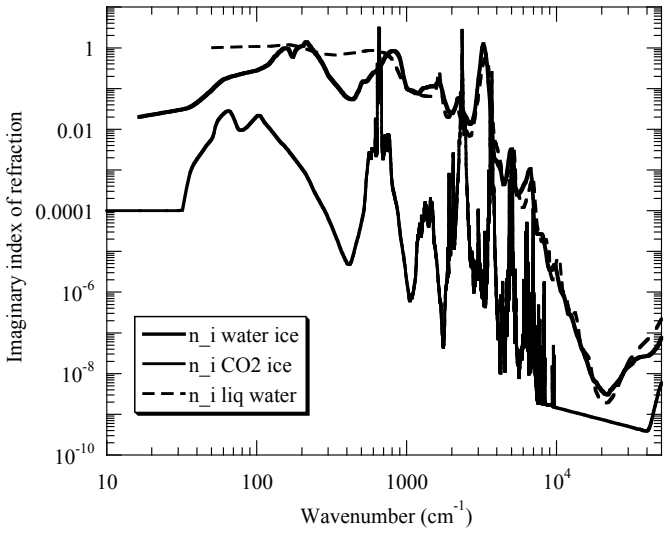

Figure 5.4: The imaginary index of refraction for liquid water, water ice, and CO_2 ice.

the scattering by sulfuric acid clouds has a significant effect on the OLR of Venus, in part because Venus is so hot that it has considerable thermal emission at wavenumbers greater than $4000 cm^{-1}$.

The strongly reflecting character of sulfate aerosols explains the volcanic cooling of the troposphere seen in the temperature time series of Figure 1.17, but what accounts for the accompanying stratospheric warming? Since the aerosols are largely transparent in the visible and solar near-infrared, the answer must lie in the the thermal infrared effect. This seems paradoxical, since we already know that increasing the infrared opacity of the stratosphere by adding CO_2 *cools* the stratosphere. The resolution to this paradox is found in the difference in absorption spectrum between CO_2 and the aerosols. CO_2 absorbs and emits very selectively and we saw in Chapter 4 that this leads to a stratospheric temperature that is considerably colder than the grey-body skin temperature. Sulfate aerosols, in contrast, act much more like a grey body. Therefore, they *raise* the stratospheric temperature towards the grey-body skin temperature. Any aerosol that absorbs broadly in the thermal infrared should behave similarly.

As a general rule of thumb, typical cloud-forming condensates tend to be very transparent in the visible and near-ultraviolet and quite transparent in the near-infrared, but vary considerably in their absorption properties in the thermal infrared. Most substances – whether gaseous or condensed – are very good absorbers in the very shortwave part of the ultraviolet spectrum, with wavelengths below .1 μm. For this reason, this part of the UV spectrum is often referred to as "vacuum UV," because it is essentially only present in the hard vacuum of outer space.

5.3 Scattering by molecules: Rayleigh scattering

Rayleigh scattering theory is a classical (i.e. non-quantum) electromagnetic scattering theory which began life as a theory for scattering of an electromagnetic plane wave from a small sphere with real index of refraction n. "Small" in this context means small compared to the wavelength of the light

being scattered. The scattering calculation is quite simple in the Rayleigh limit because the incident electric field is nearly constant over the particle, which makes it simple to compute the induced electromagnetic field within the particle. In essence, the electric field of the incident wave causes charges within the particle to migrate so that positive charge accumulates on one side and negative charge on the other, leading to a dipole moment which oscillates with the same frequency as that of the incident wave. The index of refraction is in fact a measure of the polarizability of the medium – the proportionality between the strength of the electric field and the strength of the dipole moment induced. The scattered wave in the Rayleigh limit is then simply the electromagnetic radiation emitted by an oscillating dipole, which is one of the more elementary calculations that can be done in electromagnetic theory.

Perhaps surprisingly, the Rayleigh theory works quite well as a description for scattering of light from molecules, even though molecules are not dielectric spheres. It's true that the typical size of a molecule (e.g .0003 μm for N_2) is much smaller than the wavelength of visible or even ultraviolet light, but one might have thought that the quantum response of the molecule might substantially affect the scattering. Certainly, Rayleigh theory does not provide a suitable basis for computing molecular absorption of radiation, which, as we have seen in Chapter 4, is inextricably linked to the quantum nature of the molecule. We will not go further into the reasons that a classical theory works so much better for molecular scattering than for molecular absorption, but it is indeed a convenient turn of events. In practice, it works fine to use spectroscopically measured absorption coefficients to compute gaseous absorption, together with Rayleigh scattering to compute gaseous scattering.

For a spherically symmetric scatterer, Rayleigh theory yields the following formula for the scattering cross-section:

$$\chi_{sca} = \frac{8\pi}{3}(\frac{2\pi}{\lambda})^4\alpha_p^2 \tag{5.24}$$

where λ is the wavelength of the incident light in vacuum, and α_p is the polarizability constant of the scatterer, which expresses the proportionality between the electric field and the induced dipole moment. In practice, the polarizability constant is inferred from measurements of the scattering cross section itself. It is only a weak function of wavelength. The very strong dependence of Rayleigh scattering cross section on wavelength is notable; short waves (high wavenumbers) scatter much more strongly than long waves (low wavenumbers). The explanation of the blue skies of Earth is perhaps the most famous application of Rayleigh scattering: blue through violet light has shorter wavelength than the rest of the visible spectrum, and therefore dominates the diffuse radiation caused by scattering of the solar beam from air molecules. For scatterers that are not spherically symmetric – and this includes all the polyatomic molecules like N_2, H_2 and CO_2 present in most of the atmospheres we have been considering – the dipole moment is not in the same direction as the imposed electric field, and this effect slightly alters the expression for the scattering cross section. Molecules in a gas are randomly oriented, and it can be shown that, averaged over all orientations, the modified cross section consists of the symmetric cross section in Eq 5.24 multiplied by $3(2+\delta)/(6-7\delta)$, where δ is the *depolarization factor*, which is a property of the molecule. The depolarization factor is zero for a spherically symmetric scatterer. For our purposes, the effect of the depolarization factor is not very consequential. It has a value of .054 for O_2, of .0305 for N_2 and .0805 for CO_2. These lead to only a minor increase in the scattering cross section.

Using Maxwell's equations, it can also be inferred that the index of refraction is related to the polarizability of the molecules making up the medium via the relation

$$n = 1 + 2\pi N\alpha_p \tag{5.25}$$

where N is the number of molecules per unit volume. This is a very useful relation, as it allows

	H_2	He	air	N_2	O_2	CO_2	H_2O	NH_3	CH_4
χ_{sca}	1.	.0641	4.4459	4.6035	3.8634	10.5611	3.3690	7.3427	10.1509
χ_{sca}/m	1.	0.0321	0.3066	0.3288	0.2415	0.4800	0.3743	0.8638	1.2689
τ_{ray}	0.40653	0.01305	0.12464	0.13367	0.09818	0.19513	0.15216	0.35116	0.51585

Table 5.2: Rayleigh scattering cross sections, and cross sections per unit mass, relative to H_2. These results are based on observations of the index of refraction, and do not take into account variations in the polarization factors. The scattering cross sections and cross sections per unit mass for H_2 are $1.4 \cdot 10^{-38} m^2$ ($4.215 \cdot 10^{-12} m^2/kg$) at a wavelength of 10 μm, $8.270 \cdot 10^{-33} m^2$ ($2.490 \cdot 10^{-6} m^2/kg$) at a wavelength of 1 μm, and $3.704 \cdot 10^{-28} m^2$ (.11 m^2/kg) at a wavelength of .1 μm. τ_{ray} is the optical depth due to Rayleigh scattering at $\frac{1}{2}\mu m$ for a 1 bar atmosphere of the indicated gas under Earth gravity.

one to determine Rayleigh scattering cross sections through simple measurements of the refractive index, which can be carried out by straightforward measurement of the angle of deflection of light as it moves from a transparent solid container (e.g. glass) into the gas.

Table 5.2 gives the measured Rayleigh scattering cross section relative to H_2 for a number of common atmospheric gases, as well as the corresponding cross-section per unit mass. The absolute value of the cross section for H_2 is given for a number of wavelengths in the caption, allowing the actual cross sections for the other molecules to be readily computed; the values given for H_2 in the caption deviate somewhat from the $1/\lambda^4$ wavelength scaling because of the slight dependence of index of refraction on wavelength but generally speaking it is adequate to extrapolate to other wavelengths using the fourth-power law. He stands out as an exceptionally weak scatterer. Most of the rest of the molecules have scattering cross-sections per unit mass which are moderately smaller than H_2, with the exception of CH_4,which is moderately larger.

To get an idea of how important the scattering is in various contexts, we can use the cross sections per unit mass to determine the optical depth of the entire column of an atmosphere. When the optical depth is small, the atmosphere scatters hardly at all, but when optical depth becomes large a significant amount of radiation will be scattered; in the case of the incident Solar radiation, this means a lot of the incident beam will be reflected back to space. The last line of Table 5.2 gives the optical depth for an atmosphere consisting of 1 *bar* of the given gas under Earth gravity. One can scale this up to other planets by multiplying by the appropriate surface pressure, and dividing by the planets' gravity relative to Earth gravity. The optical depth values are given at a wavelength of $\frac{1}{2}\mu m$, in the center of the visible spectrum, which is also near the peak of the Solar spectrum. For Earth's present atmosphere, the optical depth is small, but not insignificant; Rayleigh scattering affects about 12% of the incident beam. For an Early Mars having 2 *bar* of CO_2 in its atmosphere, the Rayleigh scattering is quite strong, owing to the somewhat elevated scattering cross section of CO_2 relative to air, to the low gravity and to the extra surface pressure. The Rayleigh optical depth for Early Mars would be 1.03 in the visible. The associated reflection of solar radiation is a significant impediment to warming Early Mars with a gaseous CO_2 greenhouse effect. If one took away the reflective clouds of Venus, the CO_2 Rayleigh scattering would still make Venus quite reflective, since the optical depth of a 90 *bar* CO_2 atmosphere on Venus is nearly 20. A 1.5 *bar* N_2 atmosphere on Titan would have an optical depth of 1.45 in the absence of clouds. The top 10 bars of Jupiter's mostly H_2 atmosphere would have an optical depth of 1.6, and likewise scatter significantly.

The optical depths for other wavelengths can be obtained by scaling these results according to $1/\lambda^4$. Thus, at thermal infrared wavelengths which are 5 times or more greater than the one we have been considering, the optical depth is at least 625 times smaller; Rayleigh scattering is

insignificant at these wavelengths, which is why it is safe to neglect gaseous scattering when doing computations of OLR. On the other hand, the Rayleigh scattering optical depths are at least 16 times greater for ultraviolet. We will learn how to turn these optical depth values into planetary albedos in Section 5.6.

Rayleigh scattering is not isotropic, but the phase function is symmetric between the forward and backward direction. Within the two-stream approximation, then, we do not really need any information beyond the scattering cross section. It is worth having a look at the phase function anyway, if only to get an idea of how close to isotropic it is. Given the depolarization factor δ, the Rayleigh phase function is

$$P(\cos\Theta) = \frac{3}{2(2+\delta)}(1+\delta+(1-\delta)\cos^2\Theta) \tag{5.26}$$

From this we see that Rayleigh scattering is mildly anisotropic, with stronger scattering in the forward and backward direction than in the direction perpendicular to the incident beam. For $\delta = 0$ the scattering is twice as strong in the forward and backward directions ($\Theta = 0, \pi$) as it is in the side lobes ($\Theta = \pi/2$). Increasing δ reduces the anisotropy. In fact, laboratory measurements of the intensity of side vs. forward scattering provide a convenient way to estimate the depolarization factor.

5.4 Scattering by particles

Rayleigh theory tells us everything we need to know about scattering from the gas making up an atmosphere, but to deal with cloud and aerosol particles, we need to know about scattering from objects that are not small compared to a wavelength, and indeed could be considerably larger than a wavelength, as is the case for visible light scattering from water or ice clouds on Earth. The answer is provided by Mie theory [2], which is a general solution for scattering of an electromagnetic wave from a spherical particle having uniform complex index of refraction. Mie theory reduces to Rayleigh theory in the limit of small non-absorbing particles. Ice crystals and dust particles are not spherical, and while numerical solutions are available for complex particles,for our purposes it will prove adequate to treat such cases in terms of equivalent spheres.

The Mie solution is a solution to Maxwell's electromagnetic equations which is asymptotic to a plane wave at large distances from the particle, and satisfies appropriate continuity conditions on the electromagnetic field at the particle boundary, where the index of refraction is discontinuous. Since Maxwell's equations are linear, the solution can be built up from more elementary solutions to the equations, and this is how Mie theory proceeds. It furnishes the solution in terms of an infinite sum over spherical Bessel functions, and is a real tour-de-force of early 20^{th} century applied mathematics. The formula is very convenient for evaluation of scattering parameters on a computer, but it is too complicated to yield any insight as to the nature of the solution. For that reason, we do not bother to reproduce the formula here; it is derived and discussed in the references section for this chapter, and a routine for evaluating the Mie solution is provided as part of the software supplement. Here we will only present some key results needed to provide a basis for cloud and aerosol scattering in the solar spectrum and infrared.

Let n_o, assumed real, be the index of refraction of the medium in which the particle is suspended. Define the Mie parameter $\bar{r} = n_o r/\lambda$, where r is the particle radius and λ is the

[2]The theory is named for Gustav Mie (1869-1957), who published the solution in 1908, while he was a professor at Greifswald University in Germany.

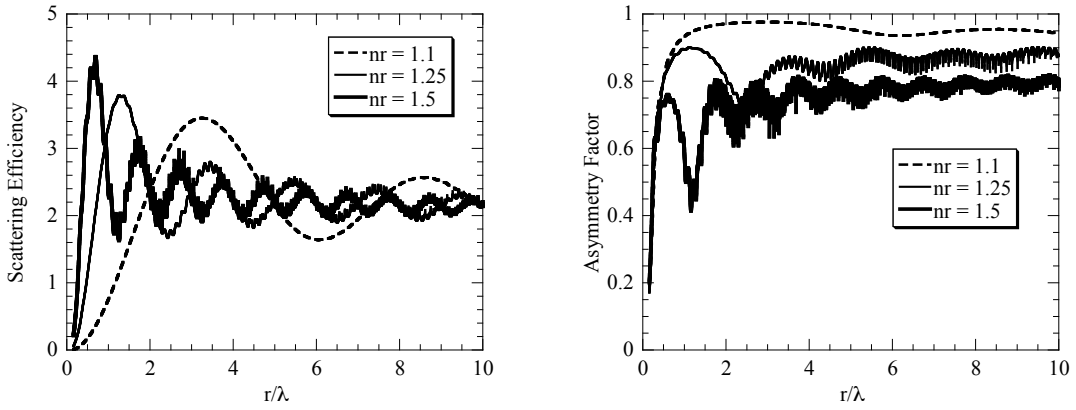

Figure 5.5: Scattering efficiency (left panel) and asymmetry factor (right panel) for Mie scattering from a non-absorbing sphere. The index of refraction of the medium is unity. r is the radius of the particle, and λ is the wavelength of the incident light measured in the same units as r.

wavelength of the incident light in vacuum, measured in the same units as r. The relative index of refraction is $n_{rel} = n/n_o$, where n is the index of refraction of the substance making up the particle. The main things we wish to compute from the Mie solution are the phase function, the scattering efficiency, the absorption efficiency, and the asymmetry factor. These are all non-dimensional quantities, and depend only on $\bar{r}$ and n_{rel}.

Let's first take a look at some scattering properties in the conservative case, $n_I = 0$. In this case, $Q_{abs} = 0$. Fig. 5.5 shows the scattering efficiency and asymmetry factor as a function of r/λ for several different values of the real index of refraction of the scatterer. Since $1/\lambda$ is the wavenumber, the graph can be thought of as displaying the scattering properties for increasing particle size with fixed wavenumber, or for increasing wavenumber with fixed particle size. Note that in the limit $n_R \to 1$ there should be no scattering at all, since in that case the particle is not optically distinct from the surrounding medium.

For any given n_R, the scattering efficiency becomes small when r/λ is sufficiently small; this is the Rayleigh limit. The scattering efficiency reaches its first peak at an order unity value of r/λ, and the position of the first peak gets closer to zero as $n_R - 1$ is made larger and the particle is made more refractive. In this sense, for a given size and wavelength, particles made of more refractive substances like CO_2 ice or concentrated H_2SO_4 act like smaller particles than particles made of less refractive substances like water or liquid CH_4. The first peak represents the optimal conditions for scattering. At the first peak, the scattering cross section can be 4 times or more the actual cross-section area of the particle.

As r/λ is increased past the first peak, the scattering efficiency oscillates between values somewhat below 2 to somewhat above 2 through a number of oscillations of decreasing amplitude, asymptoting at a value of 2 when the particle is large compared to the wavelength. The limit of large r/λ is the *geometric optics* limit, familiar from schoolbook depictions of how lenses work. In the geometrics optic limit, a beam of light is represented as a bundle of independent parallel rays, each of which travels in a straight line unless deflected from its course by an encounter with the interface between the particle and the medium – once upon entering the particle, and once upon leaving it. It is surprising that, in this limit, the scattering efficiency should asymptote to 2, since

one would be quite reasonable in thinking that rays that do not encounter the object would be unaffected, implying $Q_{sca} = 1$. What is missing from the geometric optics picture is *diffraction*. Light is indeed a wave, and this has consequences that cannot be captured by the ray-tracing on which geometric optics is based. The light encountering the sphere is a plane electromagnetic wave, and the scattering takes a circular chunk out of it; a wave with a "hole" in it is simply not a solution to Maxwell's equations, and as one proceeds past the obstacle the hole fills in with parts of the beam that never directly encountered the obstacle. It is the diversion of this part of the incident beam that accounts for the "extra" scattering cross section. The nature of diffraction is far more easily understood through examination of some elementary solutions to Maxwell's equations than it is through this rather cryptic explanation, and the reader in pursuit of deeper understanding is encouraged to study the treatment in the textbooks listed in the references to this chapter.

The main thing to take away from the preceding discussion is that refractive particles made of common cloud-forming substances become very good scatterers when their radius is comparable to or exceeds the wavelength of the light being scattered. The scattering cross section is two or more times the cross section area, and for particles whose radius is more than a few times the wavelength, the scattering efficiency is close to 2, more or less independent of what the particle is made of, and more or less independent of wavelength. This limit applies to visible light scattering off of typical cloud droplets, which have radii of 5 to 10 μm. For infrared light scattering off of cloud particles (e.g. methane clouds on Titan or dry-ice clouds on Early Mars), or for visible light scattering from micrometer-sized aerosol particles, the wavelength is comparable to the particle size, and one needs to take both wavelength and index of refraction into account in order to see how much the scattering efficiency is enhanced over the geometric-optic limit.

Turning to the right-hand panel of Figure 5.5, we see that the scattering becomes symmetric for particles small compared to a wavelength, but becomes extremely forward-peaked for particles of radius comparable to or larger than a wavelength. For $n_R = 1.25$, somewhat smaller than the value for water, the asymmetry factor is on the order of .85 for large particles. The reason for the strong forward bias is that the large particles are rather like spherical lenses, and can bend light somewhat from the oncoming path, but cannot easily reflect it into the backward direction. This feature of scattering is very important in the treatment of cloud effects on the radiation budget. The forward bias in scattering reduces the effectiveness of clouds as scatterers, and reduces their albedos well below what one would have for layers of symmetric scatterers having the same optical thickness. Without the forward scattering bias, clouds would be much more reflective of solar radiation, and planets would be much colder.

A better appreciation of the strongly forward-peaked nature of Mie scattering from large particles can be obtained by examining the phase functions shown in Figure 5.6. Almost nothing is back-scattered, and there is a sharp spike near $\Theta = 0$ which becomes sharper as the particle is made larger. For $r/\lambda = 4$, 40% of the scattered flux is in the forward peak with $\Theta < .1 radians$. When the particle is comparable in size to a wavelength, the scattering is still forward-peaked, but much less so, with appreciable amounts of flux being scattered a half radian or more from the original direction. Multiple scattering in this case allows a considerable amount of light to be back-scattered relative to its original direction, though it can take several bounces before the light turns the corner. This effect allows carbon dioxide ice clouds composed of particles of size comparable to an infrared wavelength to be quite good reflectors of infrared light trying to escape to space. Visible light scattering from such clouds is much more forward-peaked, and correspondingly less efficient.

Now let's compute the optical thickness of some typical cloud and aerosol layers. Suppose that the layer is made up of nonabsorbing particles with a scattering efficiency Q_{sca}. Recall that the scattering efficiency is close to 2 for particles large compared to the wavelength, and can be

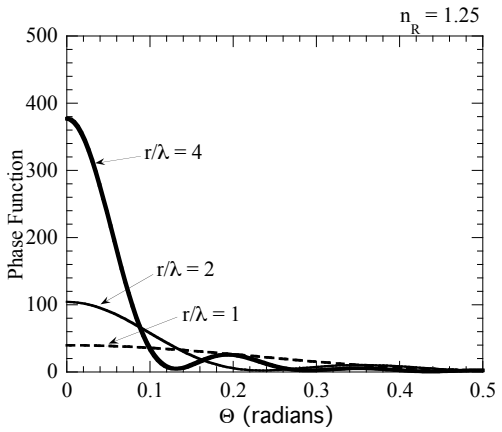

Figure 5.6: Phase function for conservative Mie scattering, under the conditions of Fig. 5.5.

as large as 4 for particles with diameter comparable to a wavelength, but falls rapidly to zero as the particle size is made smaller. For light in the solar spectrum, particles with diameters of a half μm or more are very efficient scatterers, with particles at the small end of this range being the most efficient. If the density of the substance making up the particle is ρ and the particle radius is r, then $\chi_{sca}/m = \pi r^2 Q_{sca}/(\frac{4}{3}\pi r^3 \rho) = \frac{3}{4} Q_{sca}/(\rho r)$. This is the factor by which one multiplies the mass path of scatterer to get the optical depth. The formula implies that, for a given mass of scatterers, small particles lead to much more scattering than large particles. 1 kg of 1 μm sulfate aerosol particles in a column of atmosphere yields as much scattering as 10 kg of $10\mu m$ cloud droplets. The difference in index of refraction between sulfuric acid aerosols and water droplets is of far less consequence than the difference in particle size. To proceed, assume $Q_{sca} = 2$ and $\rho = 1000 kg/m^3$. Then, for $1\mu m$ particles, $\chi_{sca}/m = 1500$, so it takes a mere $\frac{1}{1500}$ of a kilogram – *two thirds of a gram* – of aerosol particles added to a column of atmosphere with a one square meter base to bring the optical depth up to unity. This is the reason that tiny amounts of aerosol forming compounds can have significant effects on planetary climate. The small particle size also tends to make the albedo effect of aerosols dominate their greenhouse effect, despite the fairly strong absorption coefficient of sulfuric acid in the thermal infrared. For cloud droplets with a radius of 10 μm, it would take about 7 grams of water to achieve the same optical thickness as for the smaller aerosol particles, but this is still a tiny fraction of the water content of the atmosphere. A 1 km column of air in saturation at $280K$ at Earth surface pressure contains 7.8 kg of water vapor per square meter, for example. A cloud need not weigh much in order to have a profound effect on the albedo of a planet!

Now we'll turn our attention to absorbing particles. Figure 5.7 shows the scattering and absorption efficiencies for particles with $n_R = 1$ having various nonzero values if n_I. For $n_I = .1$ and $n_I = .01$ the absorption efficiency increases monotonically with particle size, and approaches unity from below (for the latter of these cases, the absorption efficiency is .89 at $r/\lambda = 100$ and .997 at $r/\lambda = 500$, outside the range plotted in the graph). This is typical behavior for particles with n_I appreciably less than unity, for which the light takes several wavelengths to decay upon encountering the particle. Larger particles absorb more simply because the light travels a longer distance within the particle, and has more opportunity to decay. For similar reasons, when n_I is made smaller, the particle must be made larger in order for there to be appreciable decay. In any event, when the particle is made large enough, it essentially absorbs a portion of the incident

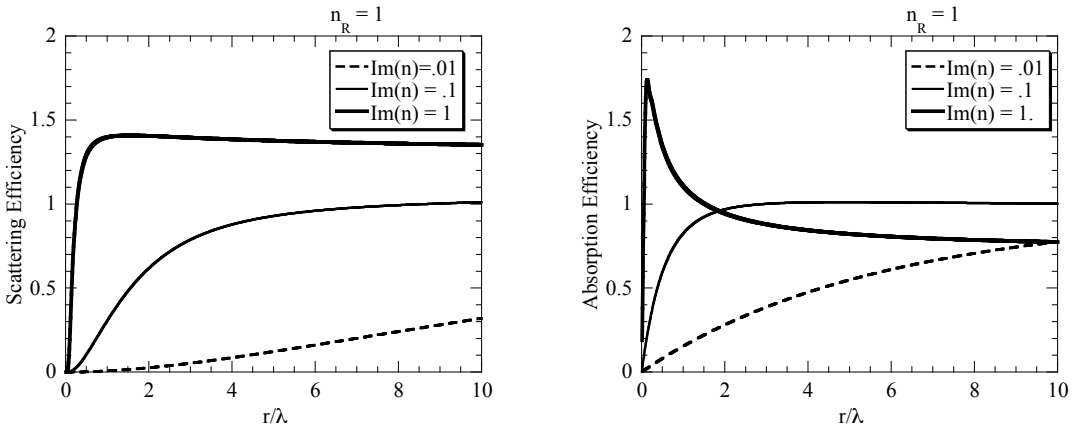

Figure 5.7: Scattering efficiency (left panel) and absorption efficiency (right panel) for Mie scattering from a partially absorbing sphere. The real part of the index of refraction is held fixed at unity while the imaginary part is varied as indicated for each curve in the figure. Other parameters are defined as in Fig. 5.5.

beam with area equal to the cross section area of the particle. Perhaps more surprisingly, though, the scattering efficiency also rises with particle size – absorbing particle don't just absorb; they also deflect light from the incident direction. This is due to the same diffraction phenomenon we encountered previously. Taking a disk out of the incident beam inevitably causes the remaining part of the beam to be deflected from its original path. For large particles the scattering efficiency and absorption efficiency sum to 2, with half the intercepted beam being absorbed and the other half scattered.

When the particles are very strongly absorbing the behavior is somewhat different,as typified by the curve for $n_I = 1$ in Fig. 5.7. In this case the absorption efficiency actually overshoots unity for particles somewhat smaller than a wavelength. The particle is able to sweep up and absorb radiation from an area larger than its cross-section, owing to the distortion of the electromagnetic field caused by the particle itself. On the other hand, as the particle is made larger, the absorption efficiency goes down and asymptotes to a value somewhat less than unity, because the incident wave is not able to penetrate deeply into the particle, and instead skirts along near its surface. In compensation, the scattering efficiency for large particles becomes greater than unity.

The above results were for particles with $n_R = 1$, in which case there is neither scattering nor absorption when n_I approaches zero. The behavior for $n_R > 1$ is explored in Problem **??**. When $n_R > 1$ then the scattering cross section behavior resembles the conservative case when n_I is small, but even a small n_I damps out the ripples in $q_{sca}(r)$ for sufficiently large particles. The asymptotic value of q_{abs} is still close to unity for sufficiently large particles throughout the range of n_I, but the overshoot properties differ somewhat from the case $n_I = 1$.

As a rule of thumb, then, we can say that when particles become larger than the characteristic decay length $\lambda/2\pi n_I$, they absorb essentially everything within a disk of area equal to the cross section area of the particle. The absorption efficiency is somewhat reduced for particles with $n_I > .1$, but even for $n_I = 1$, which is the largest value likely to be encountered, the reduction in efficiency is rather modest. Liquid water has $n_I > .1$ throughout the infrared, so any liquid water cloud droplet larger than about 10 μm in radius will absorb nearly all the infrared it encounters.

In fact, for liquid water $n_I \approx 1$ throughout most of the infrared, so that even quite small particles are efficient absorbers. Water ice can have n_I as low as .05 in some parts of the infrared spectrum, so the particles need to be twice as large to be equally good absorbers in that part of the spectrum, but ice clouds observed on Earth do tend to have larger particle sizes than water clouds.

5.5 The two-stream equations with scattering

The two stream approximations to the full scattering equation are derived from Eq. 5.14 and Eq. 5.15 by constraining the angular distribution of the radiation in such a way as to allow all integrals appearing in these equations to be written in terms of either $I_+ + I_-$ or $I_+ - I_-$. In the resulting equations, flux in the upward stream is absorbed, or scattered into the downward stream, at a rate proportional to the upward stream intensity, and similarly for the downward stream. The two-stream approximations are an instance of what physicists euphemistically like to call "uncontrolled approximations," in that they are not actually exact in any useful limit but are nonetheless physically justifiable and perform reasonably well in comparison to more precise calculations. The two-stream approximiations have inevitable inaccuracies because it is not, in fact, possible to precisely determine the scattering or the absorption from knowledge of the upward and downward fluxes alone. The two-stream approximations can be thought of as the first term in a sequence of N-stream approximations which become exact as N gets large. Fortunately, $N = 2$ proves sufficiently accurate for most climate problems.

The general form of a two-stream approximation for diffuse radiation is

$$\begin{aligned}\frac{d}{d\tau^*}I_+ &= -\gamma_1 I_+ + \gamma_2 I_- + \gamma_B \pi B(\nu, T(\tau^*_\nu)) + \gamma_+ L_\circledast \exp(-(\tau^*_\infty - \tau^*)/\cos\zeta) \\ \frac{d}{d\tau^*}I_- &= \gamma_1 I_- - \gamma_2 I_+ - \gamma_B \pi B(\nu, T(\tau^*_\nu)) - \gamma_- L_\circledast \exp(-(\tau^*_\infty - \tau^*)/\cos\zeta)\end{aligned} \tag{5.27}$$

where τ^* is the optical depth in the vertical direction (increasing upward) including scattering loss. The coefficients γ_j depend on frequency, on the properties of the scatterers, and on the particular assumption about angular distribution of radiation that was made in order to derive an approximate two-stream form from the full angular-resolved equation. We recover the hemispherically isotropic Schwartzschild equations used in previous chapters by taking $\gamma_2 = \gamma_+ = \gamma_- = 0$ and $\gamma_1 = \gamma_B = 2$. The terms proportional to γ_B represent the source due to thermal emission of radiation, while the terms proportional to γ_+ and γ_- represent the source of diffuse radiation caused by scattering of the direct beam. The direct beam is assumed to have flux $L_\circledast$ (generally the solar constant) in the direction of travel, and to travel at an angle ζ relative to the vertical. There is no upward direct beam term because it is assumed that all direct beam flux scattered from the ground scatters into diffuse radiation.

The symmetry between the coefficients multiplying I_+ and I_- is dictated by the requirement that the equations be invariant in form when one exchanges the upward and downward directions. γ_1 gives the rate at which flux is lost from the upward or downward radiation, while γ_2 gives the rate of conversion between upward and downward radiation by scattering. We can derive additional constraints on the γ_j. Subtracting the two equations gives us the equation for net vertical flux

$$\frac{d}{d\tau^*}(I_+ - I_-) = -(\gamma_1 - \gamma_2)(I_+ + I_-) + 2\gamma_B \pi B + +(\gamma_+ + \gamma_-)L_\circledast \exp(-(\tau^*_\infty - \tau^*)/\cos\zeta) \tag{5.28}$$

First, we demand that in the absence of a direct-beam source, the fluxes reduce to black body radiation in the limit of an infinite isothermal medium. Since B is constant and $L_\circledast = 0$ in

this case, we may assume the derivative on the left hand side to vanish. Since $I_+ = I_- = \pi B$ for blackbody radiation, we find $\gamma_B = \gamma_1 - \gamma_2$. Comparing with Eq. 5.14 we also find that $\gamma_+ + \gamma_- = \omega_o$. To further exploit Eq. 5.14 we must approximate $\int I d\Omega$ as being proportional to $I_+ + I_-$. The constant of proportionality, which we shall call $2\gamma'$, depends on the angular distribution of radiation assumed. With this approximation it follows that $\gamma_1 - \gamma_2 = 2\gamma'(1 - \omega_o)$.

Exercise 5.5.1 As a check on the above reasoning, show that in the conservative scattering limit $\omega_o = 1$ the sum of the diffuse vertical flux with the direct beam vertical flux is constant.

Next, we sum the equations for I_+ and I_- to obtain

$$\frac{d}{d\tau^*}(I_+ + I_-) = -(\gamma_1 + \gamma_2)(I_+ - I_-) + (\gamma_+ - \gamma_-)L_\circledast \exp(-(\tau^*_\infty - \tau^*)/\cos\zeta) \tag{5.29}$$

This can be compared to the symmetric flux projection given by Eq. 5.15. Making an assumption about the angular distribution allows one to approximate $\int IH(\cos\theta)\cos\theta d\Omega$ as proportional to $I_+ + I_-$, and $\int IH d\Omega$ and $\int IH d\Omega$ each as being proportional to $I_+ - I_-$. In consequence, $\gamma_1 + \gamma_2 = 2\gamma\cdot(1 - \hat{g}\omega_o)$, where γ is related to the proportionality coefficient and $\hat{g}$ is a coefficient characterizing the asymmetry of the scattering. If $H = \cos\theta$, then $\hat{g}$ is in fact the asymmetry factor $\tilde{g}$ defined by Eq. 5.17, but other forms of H yield somewhat different asymmetry factors, though these tend to be reasonably close to $\tilde{g}$. For example, with the form of H given by Eq. 5.20, we showed that $\hat{g} = \frac{3}{2}\tilde{g}$ for phase functions that are truncated to their first three Fourier components. Finally, under circumstances when we can write $G = \hat{g}\cos\theta$, it follows that $\gamma_+ - \gamma_- = -2\gamma\omega_o\hat{g}$. This relation holds exactly when $H = \cos\theta$, and imposing it for other forms of H introduces errors that are no worse than other errors that are inevitable in reducing the full scattering equation down to two streams.

The general form of the set of two-stream coefficients satisfying all the above constraints is then

$$\begin{aligned}
\gamma_1 &= \gamma \cdot (1 - \hat{g}\omega_o) + \gamma' \cdot (1 - \omega_o) \\
\gamma_2 &= \gamma \cdot (1 - \hat{g}\omega_o) - \gamma' \cdot (1 - \omega_o) \\
\gamma_B &= 2\gamma' \cdot (1 - \omega_o) \\
\gamma_+ &= \frac{1}{2}\omega_o - \gamma\omega_o\hat{g}\cos\zeta \\
\gamma_- &= \frac{1}{2}\omega_o + \gamma\omega_o\hat{g}\cos\zeta
\end{aligned} \tag{5.30}$$

The coefficients γ and γ' are purely numerical factors that depend on the assumption about the angular distribution of radiation which is used to close the two-stream problem. All vertical dependence then comes in through ω_o, and possibly through $\hat{g}$ if the asymmetry properties of scattering particles vary with height. There are three common closures in use. The first is the hemispherically isotropic closure, which we used earlier in deriving the two-stream equations without scattering. In this closure, it is assumed that the flux is isotropic (i.e. I is constant) in each of the upward and downward hemispheres, but with a different value in each hemisphere. The hemi-isotropic closure is derived by using the weighting function H defined by Eq. 5.20 and making use of Eq. 5.21. Given the isotropy of the blackbody source term, it is generally believed that the hemi-isotropic approximation is most appropriate for thermal infrared problems, with or without scattering. Another widely used closure is the *Eddington approximation*. The Eddington closure is obtained by taking $H = \cos\theta$ and making use of Eq. 5.19. To complete the closure, $\int I\cos^2\theta d\Omega$ is written in terms of $I_+ + I_-$ by assuming that the flux is truncated to the first two Fourier components, so

	γ	γ'
Hemi-isotropic	1	1
Quadrature	$\frac{1}{2}\sqrt{3}$	$\frac{1}{2}\sqrt{3}$
Eddington	$\frac{3}{4}$	1

Table 5.3: Coefficients for various two-stream approximations

$I = a + b\cos\theta$. This is probably the most widely used closure for dealing with solar radiation. It is generally believed that this closure is a good choice for dealing with both Rayleigh scattering and the highly forward-peaked scattering due to cloud particles, though the mathematical justification for this belief is not very firm. The *quadrature* approximation is similar, except that $\int I\cos^2\theta d\Omega$ is evaluated using a technique known as *Gaussian quadrature*, which yields a different proportionality constant from the Eddington closure. The defining coefficients for the three closures are given in Table 5.3.

When $\omega_o = 0$ there is no scattering and so the upward and downward streams should become uncoupled. From Eq. 5.30 we see that this decoupling happens only if $\gamma = \gamma'$, a requirement that is satisfied for the hemi-isotropic and quadrature approximations but not for the Eddington approximation. It follows that the Eddington approximation can incur serious errors when the scattering is weak, though it can nonetheless outperform the other approximations when scattering is comparable to or dominant over absorption.

The two-stream equations form a coupled system of ordinary differential equations in two dependent variables. Therefore they require two boundary conditions. At the top of the atmosphere, there is generally no incoming diffuse radiation, so the boundary condition there is simply $I_- = 0$. At the bottom boundary we require that the upward diffuse radiation be the sum of the upward emission from the ground with the reflected direct beam and downward diffuse radiation. In general, the direct beam reflection might in part yield a reflected direct beam (as in reflection from a mirror-like smooth surface), but in the following we'll assume that all reflection from the bottom boundary is diffuse. Thus, the boundary condition at $\tau^* = 0$ is

$$I_+(0) = e_g\pi B(\nu, T_g) + \alpha_g L_\circledast \cos\zeta \exp(-\tau^*_\infty/\cos\zeta) + \alpha_g I_-(0) \tag{5.31}$$

where e_g is the emissivity of the ground and α_g is the albedo of the ground, both of which vary with ν; Kirchoff's law implies that $e_g = (1 - \alpha_g)$ for any given frequency.

5.6 Some basic solutions

When the scattering and absorption properties of the atmosphere are independent of τ^*, the two stream equations have simple exponential solutions. We'll begin with an elementary solution for conservative scattering, which provide quite useful estimates of the effect of clear-sky and cloudy atmospheres on the solar-spectrum albedo of planets. We shall specify an incoming direct-beam flux of solar radiation, and we seek the outgoing reflected flux at the same wavenumber. For conservative scattering, $\omega_o = 1$. In that case $\gamma_1 = \gamma_2$. Then, Eq 5.28 becomes

$$\frac{d}{d\tau^*}(I_+ - I_-) = L_\circledast \exp(-(\tau^*_\infty - \tau^*)/\cos\zeta) \tag{5.32}$$

so

$$I_+ - I_- - L_\circledast \cos\zeta \exp(-(\tau^*_\infty - \tau^*)/\cos\zeta) = C \tag{5.33}$$

where C is a constant. This equation states that, for conservative scattering, the net of the diffuse flux and the vertical component of the surviving direct-beam flux is constant. As the upper boundary condition we require that the diffuse incoming radiation be zero; hence $C = I_{+,\infty} - L_\circledast \cos\zeta$, and

$$I_+ - I_- = I_{+,\infty} + L_\circledast \cos\zeta(\exp(-(\tau^*_\infty - \tau^*)/\cos\zeta) - 1) \tag{5.34}$$

Deep in the atmosphere, $I_+ - I_-$ becomes constant, and is equal to the difference between the top-of-atmosphere incoming minus outgoing flux, i.e. $I_{+,\infty} - L_\circledast \cos\zeta$. When the atmosphere is optically thick, or when $\cos\zeta$ is small (i.e. when the sun is close to the horizon) the exponential term is significant only near the top of the atmosphere, It represents the conversion of the direct beam into diffuse radiation by scattering, which occurs within a conversion layer of depth $1/\cos\zeta$ in optical depth units.

$I_{+,\infty}$ is the reflected flux we wish to determine, and we must close the problem by applying a boundary condition at the ground. For this we need $I_+(\tau^*)$, which we obtain by using the equation for $I_+ + I_-$. Let's restrict attention to the symmetric scattering case, $g = 0$, for which $\gamma_+ = \gamma_-$ and $\gamma_1 = \gamma_2 = \gamma$. Then

$$\frac{d}{d\tau^*}(I_+ + I_-) = -2\gamma(I_+ - I_-) = -2\gamma(I_{+,\infty} - L_\circledast \cos\zeta + L_\circledast \cos\zeta \exp(-(\tau^*_\infty - \tau^*)/\cos\zeta)) \tag{5.35}$$

The solution which satisfies $I_- = 0$ at the top of the atmosphere, is

$$I_+ + I_- = I_{+,\infty} + 2\gamma(I_{+,\infty} - L_\circledast \cos\zeta)(\tau^*_\infty - \tau^*) + 2\gamma L_\circledast \cos^2\zeta(1 - \exp(-(\tau^*_\infty - \tau^*)/\cos\zeta)) \tag{5.36}$$

Let's suppose that the ground is perfectly absorbing. This calculation characterizes the albedo of the atmosphere alone. Later, we'll compute how much the surface albedo enhances the planetary albedo. If the ground is perfectly absorbing, then we require $I_+ = 0$ at $\tau^* = 0$. To apply the boundary condition, we add Eqns. 5.34 and 5.36 and evaluate the result at the ground. Applying the boundary condition and solving for $I_{+,\infty}$ we find

$$I_+, \infty = \frac{(\frac{1}{2} - \gamma\cos\zeta)\beta_\circledast + \gamma\tau^*_\infty}{1 + \gamma\tau^*_\infty} L_\circledast \cos\zeta \equiv \alpha_a L_\circledast \cos\zeta \tag{5.37}$$

where $\beta_\circledast = 1 - \exp(-\tau^*_\infty/\cos\zeta)$; this quantity is the proportion of the direct solar beam that has been lost to scattering by the time the beam reaches the ground. The fraction multiplying $L_\circledast \cos\zeta$ in Eq. 5.37 is the planetary albedo. Since any flux reaching the ground is absorbed completely, this albedo is in fact the albedo of the atmosphere alone, which we will call α_a. In the optically thin limit, $\beta_\circledast \approx \tau^*_\infty/\cos\zeta$, so the albedo approaches zero like $\frac{1}{2}\tau^*_\infty/\cos\zeta$. Half of the small amount of flux scattered by the atmosphere exits the top of the atmosphere, but the other half is scattered into the ground, where it is absorbed.

As the atmosphere is made more optically thick, the albedo increases in two stages. The first stage is an exponential adjustment, as the direct beam is converted to diffuse radiation. Some simple algebra shows that the numerator of Eq. 5.37 always increases with τ^*, regardless of the value of $\cos\zeta$. However, when the incident beam is relatively near the horizon, so that $\gamma\cos\zeta < \frac{1}{2}$, the conversion term leads to an exponential increase of albedo with τ^*. The effect becomes more pronounced when the Sun is more nearly on the horizon. This effect comes from the direct scatter of the incident beam to space. When τ^*_∞ becomes appreciably larger than $1/\cos\zeta$, however, the direct beam has been completely converted, $\beta_\circledast \approx 1$, and the albedo no longer varies exponentially. For large τ^*_∞, the albedo approaches unity like $1 - (\frac{1}{2}\gamma^{-1} + \cos\zeta)/\tau^*_\infty$. The rather slow approach to a state of complete reflection is due to multiple scattering. In contrast to the exponential decay

	Earth	Early Mars	Venus	Titan	Water cloud	Sulfate aerosol haze
τ^*_∞	0.12	1.03	20.	1.45	3.	1.25
$\hat{g}$	0.	0	0	0	.87	.76
albedo, hemi-isotropic	.08	.43	.94	.52	.13	.10
albedo, quadrature	.08	.43	.94	.51	.17	.13
albedo, Eddington	.08	.42	.93	.51	.20	.16

Table 5.4: Albedos for purely scattering atmospheres. Values given for Earth, Early Mars, Venus and Titan are for hypothetical clear-sky atmospheres consisting of 1 bar of air for Earth, 2bar of CO_2 for Early Mars, 90bar of CO_2 for Venus and 1.5bar of N_2 for Titan. The water cloud case assumes a path of 20 grams of water per square meter in droplets of radius 10 μm, having a scattering efficiency of 2. The sulfate aerosol case assumes a path of 1 gram per square meter in sulfuric acid droplets of radius 1μm, having a scattering efficiency of 3. Albedos are computed at a wavelength of .5μm, with a zenith angle of 45^o for the direct beam.

of the direct beam, the diffuse radiation surviving to be absorbed at the surface decays only like $1/\tau^*_\infty$ because much of the radiation scattered upward is latter scattered back downward. The exponential decay of the direct beam just represents a conversion to diffuse radiation,and therefore does not materially alter the conclusion that scattering is a relatively ineffective way of preventing radiation from reaching the surface. That is why one cannot rely on Rayleigh scattering alone to shield life at the surface from harmful ultraviolet radiation, despite the fact that the Rayleigh optical thickness of an Earthlike atmosphere is quite high in the ultraviolet.

The simple albedo formula given above has many physically important ramifications. Before computing the albedo for various conservatively scattering atmospheres, though, it is necessary to bring in the effect of asymmetric scattering if we are to deal with clouds, as scattering from cloud particles is strongly forward-peaked. A nonzero asymmetry factor simply adds a direct beam source term to the equation for $d(I_+ + I_-)/d\tau^*$, since $\gamma_+ - \gamma_-$ no longer vanishes. Some straightforward algebra shows that, allowing for a nonzero asymmetry factor, the albedo formula becomes

$$\alpha_a = \frac{(\frac{1}{2} - \gamma\cos\zeta)\beta_\circledast + (1-\hat{g})\gamma\tau^*_\infty}{1 + (1-\hat{g})\gamma\tau^*_\infty} \tag{5.38}$$

This differs from the symmetric scattering form only in that the optical thickness is multiplied by $1-\hat{g}$, which reduces the effective optical thickness when $\hat{g} > 0$. Thus, in the context of the two stream equations, the effect of asymmetric scattering is not very surprising or subtle. Since forward scattering just adds back into the forward radiation as if scattering had not occured at all, forward-dominated scattering simply has the effect of reducing the optical thickness of the atmosphere. The only reason one can't get by with simply redefining the optical depth is the presence of the direct beam; the direct beam transmission factor $\beta_\circledast$ is computed using the unmodified optical depth, rather than the rescaled optical depth, because even forward-scattered radiation is transferred out of the direct beam and into the diffuse component. The behavior of the albedo formula is explored in Problems **??** and **??**.

In Table 5.4 we use Eq.5.38 to compute the albedos of a number of clear-sky and cloudy planetary atmosphere, based on optical depths computed from the Rayleigh scattering or Mie-scattering cross-sections in the visible spectrum. Results are shown for both the hemispherically isotropic and the Eddington approximations; the results differ little between the approximations for the symmetric Rayleigh scattering cases. In clear sky conditions, Earth's atmosphere reflects about 8% of the incoming solar energy back to space. This represents nearly a third of Earth's observed albedo, and is a significant player in the energy budget. The thick CO_2 atmosphere

postulated for Early Mars has an even more significant effect on albedo, reflecting fully 43% of the solar energy. Further increases in CO_2 lead to even greater reflection, making it hard to warm Early Mars with the gaseous CO_2 greenhouse effect alone. The case of Venus is particularly interesting. We see from the table that if the thick clouds of Venus were removed, Rayleigh scattering alone would be sufficient to keep the albedo high. The value in the table is an overestimate of the albedo Venus would have in the no-cloud case, since it ignores the solar absorption by CO_2, but it suffices to show that virtually all of what escapes absorption would be scattered back to space by Rayleigh scattering. This is important to the evolution of Venus-like planets, which might not have atmospheric chemistry that supports sulfuric acid clouds like those of Venus at present; for that matter, it is not completely certain that thick clouds are a perennial feature of our own Venus. When one factors in the fact that rather little solar radiation penetrates the present thick clouds of Venus, it is evident that the strong Rayleigh scattering of the surviving flux implies that only a trickle of solar radiation reaches the surface of such a planet. It is only a trickle, but as we have seen in Chapter 4, it is a very important trickle, since the surface could not be so hot if all of the solar energy were absorbed aloft.

The cloud cases in Table 5.4 could really apply to any planet with condensible water or sulfer compounds. The cloud parameters chosen are quite typical of Earth conditions. Scattering of solar radiation from cloud particles is quite different from Rayleigh scattering because of the strong asymmetry, which makes the albedo considerably lower than one would expect on the basis of optical thickness. Nonetheless, a small mass of cloud water, or a still smaller mass of sulfate aerosol in the form of micrometer-sized droplets, leads to a very significant albedo. To put the mass path of sulfate aerosol into perspective, we note that if we assume a 10 day lifetime for aerosol in the atmosphere, the assumed mass path is equivalent to a world wide sulfur emission of about 8 $megatonnes/da$, allowing for the proportion of S in H_2SO_4. Actual worldwide sulfur emissions for 1990 are estimated to have been more like $\frac{1}{3}$ $megatonne/da$, which is why the albedo of Earth's sulfate aerosol haze is lower than the estimate in the table, though still a significant player in the radiation budget.

Note also that when the asymmetry is as pronounced as it is for cloud particles, the albedo predicted by the Eddington approximation is significantly greater than that for the hemispherically isotropic case. Although the asymmetry factor reduces the attenuation of diffuse radiation by the cloud, the decay of the direct beam is exponential in the optical depth itself, leading to near total attenuation of the direct beam by the water cloud. This is why a thick cloud looks bright, though you cannot easily discern the disk of the Sun. It is also why it is possible to get quite thoroughly sunburned on a cloudy day.

In the preceding calculation we assumed that the ground was perfectly absorbing. If the ground is instead partially reflecting, having albedo α_g, it will reflect some of the light reaching the surface back upward. Some of this light reflected from the surface will make it through the atmosphere and escape out the top, increasing the planetary albedo. How does the albedo of the atmosphere combine with the albedo of the ground to make up the planetary albedo? Since we are assuming that there is no atmospheric absorption, if a proportion α_a of incoming light is reflected by the atmosphere, then a proportion $(1-\alpha_a)$ reaches the ground. The proportion of this reflected back upward is $(1-\alpha_a)\alpha_g$, and if we were to simply add this upward proportion to the part reflected directly by the atmosphere we'd get a planetary albedo of $\alpha_a + (1-\alpha_a)\alpha_g$. This simple estimate already illustrates the important result that putting a reflective (e.g. cloudy) atmosphere over an already reflective surface changes the planetary albedo much less than putting such an atmosphere over a dark surface – you can't make the planet whiter than white, as it were. The simple estimate overestimates the planetary albedo, though, since some of the upward radiation from the ground bounces back from the atmosphere whereafter some of the remainder is absorbed at the ground.

The rest is reflected back upwards, and ever diminishing proportions remain to multiply scatter back and forth between the atmosphere and the ground. One doesn't need to actually sum an infinite series to solve the problem; all we need to do is to correctly specify the boundary condition on the upward radiation at the ground, which requires in turn a specification of the proportion of upward radiation which is reflected back to the surface by the atmosphere. The upward radiation reflected from the ground is, by definition, purely diffuse, so as a preliminary to this calculation we need the albedo of the atmosphere for upward-directed diffuse radiation. Since the radiation is purely diffuse, the form of this atmospheric albedo is somewhat simpler than the formula for incoming solar radiation. It is derived in Problem **??** , and is

$$\alpha'_a = \frac{(1-\hat{g})\gamma\tau^*_\infty}{1+(1-\hat{g})\gamma\tau^*_\infty} \tag{5.39}$$

Note that this has the same form as the expression for α_a, except that the direct beam term in the numerator, proportional to $\beta_\circledast$, has been dropped. In terms of α'_a, the boundary condition on upward radiation at the ground is

$$I_+(0) = \alpha_g \cdot (I_-(0) + L_\circledast \cos\zeta \exp(-\tau^*_\infty/\cos\zeta)) = \alpha_g \cdot ((1-\alpha_a)L_\circledast \cos\zeta + \alpha'_a I_+(0)) \tag{5.40}$$

which can be solved for $I_+(0)$. When this boundary condition is applied to the conservative two-stream equations, the resulting expression for $I_{+,\infty}$ yields the following expression for the planetary albedo:

$$\begin{aligned} \alpha &= \alpha_a + \frac{(1-\alpha'_a)(1-\alpha_a)}{1-\alpha_g\alpha'_a}\alpha_g \\ &= 1 - \frac{(1-\alpha_g)(1-\alpha_a)}{(1-\alpha_g)\alpha'_a + (1-\alpha'_a)} \end{aligned} \tag{5.41}$$

The details are carried out in Problem **??**. When all three albedos, α_a, α'_a and α_g are small, the expression reduces to the sum of the albedos $\alpha_a + \alpha_g$. Eq. 5.41 has very important consequences for the net effect of clouds on the planetary radiation budget. Clouds have both a warming and a cooling effect on climate. High-altitude clouds have a warming effect, since they strongly reduce OLR either by absorption and emission, or by scattering, of infrared radiation. Clouds at any altitude have a cooling influence, through increasing the planetary albedo in the solar spectrum. The net effect of clouds depends on how the competition between these two factors plays out. Eq. 5.41 shows that clouds increase the planetary albedo rather little, if they are put over a highly reflective surface (such as ice), or if they are put into an atmosphere which is already quite reflective (such as the dense atmosphere of Early Mars). In either case, introduction of clouds will tend to have a strong net warming effect, because the cloud greenhouse effect is relatively uncompensated by the cloud albedo effect. It is for this reason that clouds can greatly facilitate the deglaciation of a Snowball Earth, and that clouds of either water or CO_2 can very significantly warm Early Mars.

Now let's do an infrared scattering problem, one that illustrates the scattering greenhouse effect in its simplest form. Consider an atmosphere made of a gas that is completely transparent (hence also non-emitting) in the infrared. Suspended in the atmosphere is a cloud made of a substance such as CO_2 ice or liquid methane that is almost non-absorbing in the infrared; we'll idealize it as being exactly non-absorbing, and assume the scattering to be symmetric. Since neither the gas nor the cloud emit infrared, the temperature profile of the atmosphere is immaterial. This atmosphere lies above a blackbody surface with temperature T. What is the OLR? This problem is also a case of conservative scattering ($\omega_o = 1$), but with different boundary conditions. Since there is no incoming infrared, the upper boundary condition is $I_- = 0$. At the ground, the upward flux boundary condition is $I_+ = \pi B(\nu, T)$. Without any direct beam or blackbody source term,

$I_+ - I_-$ is a constant, which is equal to the outgoing radiation $I_{+,\infty}$. at the frequency under consideration. The equation for $d(I_+ + I_-)/d\tau^*$ then tells us that $I_+ + I_- = 2\pi B - (1+\gamma\tau^*)I_{+,\infty}$. Finally, imposing the boundary condition that $I_- = 0$ at $\tau^* = \tau^*_\infty$, we conclude that

$$I_{+,\infty} = \pi B/(2 + \gamma\tau^*_\infty) \tag{5.42}$$

Hence, the infrared scattering reduces the outgoing infrared by a factor $1/(2 + \gamma\tau^*_\infty)$ relative to what it would be in the absence of an atmosphere. This increases the surface temperature of the planet in the same fashion as the OLR reduction from the more conventional absorption/emission greenhouse effect. However, the scattering greenhouse effect works quite differently, since it reduces the OLR regardless of whether the atmospheric temperature goes down with height.

The scattering greenhouse effect exerts an important warming influence on planets which form non-emissive clouds. This includes the case of CO_2 ice clouds on Early Mars and on a cold Snowball Earth. Whether the *net* effect of the clouds is to warm or cool the planet depends on how much of the scattering greenhouse effect is offset by additional reflection in the incoming stellar spectrum. The following factors tend to tilt the balance in favor of net warming:

- If the particles have a size on the order of 10 μm, then the Mie scattering efficiency is enhanced in the thermal infrared range, compared to what it is for shorter wavelengths.
- Unless the particles are very small, the asymmetry factor is much greater for the incoming short wave radiation, which leads to inefficient scattering.
- If the planet has a high albedo to begin with, as in the case of Rayleigh scattering from a thick CO_2 atmosphere or the high albedo of a Snowball Earth, then the effect of the shortwave cloud albedo on absorption of incoming solar radiation is reduced.
- The clouds will exert a pronounced net warming effect in circumstances such as high latitude winter, in which there is little incoming solar radiation available for reflection.

Detailed calculations indicate a net warming effect on Early Mars and on Titan. The net influence of the sulfuric acid clouds of Venus presents a particularly interesting problem, because they are dynamically maintained by the sulfur cycle of the the planet; it could well be that Venus has gone through periods when these clouds were absent. Would such a Venus be hotter or cooler? Venus clouds act on the OLR through a mix of absorption and scattering, and the estimate of their net effect depends moreover on what the solar-spectrum albedo of Venus would be if you took the clouds away. The few calculations that have been done on this problem tend to suggest that the solar albedo effect wins in this case, and Venus would become considerably hotter without clouds. This is a problem that involves many subtleties and would repay further study, particularly in view of the fact that Venus represents an archetype for the climate evolution of hot, dry planets. Further explorations of the scattering greenhouse effect and of the competition between cloud greenhouse and cloud albedo effects are pursued throughout the Workbook. See especially Problems **??**,**??**, **??** and **??**.

Next we'll extend the preceding scattering greenhouse problem to allow $\omega_o < 1$, so the atmosphere can absorb and emit. We'll assume the atmosphere to be isothermal at the same temperature T as the ground. In this case, $I_+ = I_- = \pi B(\nu, T)$ is a particular solution satisfying the boundary condition on I_+ at the ground, though it does not satisfy the boundary condition $I_- = 0$ at the top of the atmosphere. We must add a homogenous solution to the particular solution, which cancels I_- at the top of the atmosphere for the particular solution, but leaves the bottom boundary condition intact. The homogeneous equation is obtained by taking the derivative

of Eq. 5.28 and substituting the derivative of $I_+ + I_-$ using Eq. 5.29, dropping the source terms from both. Assuming ω_o and g to be independent of height, the homogeneous equation is then

$$\begin{aligned}\frac{d^2}{d\tau^{*2}}(I_+ - I_-) &= -(\gamma_1 - \gamma_2)(\gamma_1 + \gamma_2)(I_+ - I_-) \\ &= -4\gamma\gamma'(1 - \hat{g}\omega_o)(1 - \omega_o)(I_+ - I_-)\end{aligned} \tag{5.43}$$

The general solution to this is $a\exp(-K \cdot (\tau^*_\infty - \tau^*)) + b\exp(K \cdot (\tau^*_\infty - \tau^*))$, where

$$K = 2\sqrt{\gamma\gamma'(1 - \hat{g}\omega_o)(1 - \omega_o)} \tag{5.44}$$

One term grows exponentially in optical depth, while the other decays exponentially. The solution for $I_+ + I_-$ is then obtained from the solution for $I_+ - I_-$ using Eq. 5.28, allowing us to obtain the two fluxes individually for use in applying the boundary conditions. First, the homogeneous solution we add in must not disturb I_+ at the ground, since the particular solution already satisfies the boundary condition there. To keep the algebra simple, let's assume $\tau^*_\infty \gg 1$. In that case, we approximately satisfy the boundary condition at the ground by taking the solution which decays toward the ground, i.e. $b = 0$. The boundary condition on $I_-(0)$ then determines the value of the coefficient a. Carrying out the algebra and adding the homogeneous to the particular solution, we find the outgoing radiation to be

$$I_{+,\infty} = 2\frac{\gamma'(1 - \omega_o)}{\gamma'(1 - \omega_o) + \sqrt{\gamma\gamma'(1 - \hat{g}\omega_o)(1 - \omega_o)}}\pi B \tag{5.45}$$

In this equation, πB is the emission the planet would have in the absence of an atmosphere, and the the coefficient multiplying it gives the reduction in emission due to the atmosphere. Note that for a non-scattering atmosphere, the isothermal atmosphere assumed would have no effect whatever on the outgoing radiation. In contrast to the conservative scattering case described by Eq 5.42, the emission in the partially absorbing case does not approach zero in the optically thick limit, but rather approaches the nonzero value given by Eq. 5.45. When there is no scattering, i.e. $\omega_o = 0$, the atmosphere should have no effect on emission, and this limit shows the shortcomings of the Eddington approximation. For $\omega_o = 0$, the factor reducing the emission is $2\gamma'/(\gamma' + \sqrt{\gamma\gamma'})$, which reduces to unity only when $\gamma = \gamma'$. Both the quadrature and the hemispherically isotropic assumptions satisfy this requirement, but as we have seen before, the Eddington approximation gives the wrong answer when scattering is weak, and is not suitable for such cases. For any of the approximations, as the scattering is made stronger relative to absorption, $\omega_o \to 1$ and the emission goes to zero in proportion to $\sqrt{(\gamma/\gamma')(1 - \omega_o)/(1 - \hat{g})}$.

The basic lesson here is that sufficiently strong scattering can kill off emission from the atmosphere. When the scattering becomes strong, infrared can escape to space only from a thin layer near the top of the atmosphere; radiation from deeper layers is scattered back downwards and absorbed before it can escape.

Finally, we'll use an elementary solution to show how scattering affects the vertical distribution of solar absorption. We'll suppose that thermal emission is negligible at the frequency under consideration, as is the case for solar radiation on planets at Earthlike (or even Venus-llke) temperatures. The basic idea is that scattering increases the net path traveled by radiation in going from one altitude to another, because the radiation bounces back and forth many times rather than proceeding in a straight line. This allows more radiation to be absorbed within a thinner layer, as compared to the no-scattering case. At the same time, however, scattering reflects some radiation back to space before it has any opportunity to be absorbed at all. As will be shown in the forthcoming derivation, the net result is to reduce the solar absorption while at the same time

concentrating it more in the upper atmosphere, as compared to a no-scattering case with the same distribution of absorbers.

The full problem with arbitrary surface albedo, optical depth, and asymmetry factor is analytically tractable so long as ω_o is constant. The full solution is somewhat unwieldy, however, so we'll now make a few simplifying assumptions. To keep the algebra simple, in the present discussion we'll assume τ^*_∞ to be very large, so that the lower boundary does not affect the solution. The atmosphere is effectively semi-infinite (a top but no bottom) in this solution. We'll also assume that the asymmetry factor vanishes. The solution begins with taking the derivative of Eq 5.29 and substituting from Eq. 5.28, which gives us

$$\frac{d^2}{d\tau^{*2}}(I_+ + I_-) = -K^2(I_+ + I_-) - 2\gamma\omega_o L_\circledast \exp(-(\tau^*_\infty - \tau^*)/\cos\zeta)) \tag{5.46}$$

where K is defined by Eq. 5.44 with $\hat{g}$ set to zero. A particular solution to this equation is

$$I_+ + I_- = -\frac{2\gamma\omega_o L_\circledast \cos^2\zeta}{1 - K^2\cos^2\zeta} \tag{5.47}$$

to which we have to add superpositions of the two homogeneous solutions $\exp(\pm K(\tau^*_\infty - \tau^*))$ so as to satisfy the boundary conditions at the top and bottom of the atmosphere. So far the only assumptions we have used are that ω_o is constant, the thermal emission is neglected, and $\hat{g} = 0$. Since the atmosphere is semi-infinite, the only admissible homogeneous solution is $a\exp(-K(\tau^*_\infty - \tau^*))$, since the other solution blows up deep in the atmosphere. It remains only to determine a, which is done by applying the condition $I_- = 0$ at the top of the atmosphere. To do this we need $I_+ - I_-$. This is obtained from Eq 5.29, which takes on a particularly simple form when $\hat{g} = 0$. Using the value of a thus obtained, the net vertical diffuse flux is found to be

$$\begin{aligned} I_+ - I_- = [&- \frac{K}{2\gamma + K}\frac{1 + 2\gamma\cos\zeta}{1 - K^2\cos^2\zeta}\exp(-K(\tau^*_\infty - \tau^*)) \\ &+ \frac{1}{1 - K^2\cos^2\zeta}\exp(-(\tau^*_\infty - \tau^*)/\cos\zeta)]\omega_o L_\circledast \cos\zeta \end{aligned} \tag{5.48}$$

Since $I_- = 0$ at the top of the atmosphere, the albedo is obtained by evaluating this expression at $\tau^* = \tau^*_\infty$ and taking the coefficient of the incoming flux $L_\circledast \cos\zeta$. Thus,

$$\begin{aligned} \alpha &= \frac{2\gamma\omega_o}{(1 + K\cos\zeta)(2\gamma + K)} \\ &= \frac{2\gamma\omega_o}{(1 + 2\sqrt{\gamma\gamma'(1-\omega_o)}\cos\zeta)(2\gamma + 2\sqrt{\gamma\gamma'(1-\omega_o)})} \end{aligned} \tag{5.49}$$

In the absence of scattering, all the incident flux should be absorbed no matter how low the concentration of absorbers, since the atmosphere is assumed infinitely deep. Consistently with this reasoning the above albedo approaches zero as $\omega_o \to 0$. For small ω_o, the albedo increases linearly with ω_o. It continues to increase monotonically as ω_o is further increased. In the limit $\omega_o \to 1$ where scattering becomes very strong, $\alpha \to 1$ and the atmosphere becomes perfectly reflecting; radiation is scattered back to space before it has much opportunity to be absorbed in the atmosphere.

Though strong scattering reduces the opportunity for absorption, it also reduces the depth scale over which the small amount of absorbed radiation is deposited in the atmosphere. The reason is that scattering increases absorption through multiple reflections that increase the path length. To get a better handle on what is going on, we need to examine the vertical profile of

the flux as $\omega_o \to 1$ while holding the concentration of absorbers fixed. Taking the limit this way would correspond, for example, to looking at ultraviolet absorption as we increase the amount of conservatively scattering cloud particles in an atmosphere while keeping the amount of ultraviolet-absorbing ozone fixed. This is equivalent to writing $\tau^*_\infty - \tau^* = (\kappa/(1 - omega_o))(p/g)$, if we neglect pressure broadening, since $(1 - \omega_o)\Delta\tau$ gives the absorption in a layer of thickness $\Delta\tau$. It follows from this expression for optical thickness that all the direct beam flux is converted to diffuse flux in a very thin conversion layer if $\omega_o \to 1$. Below the conversion layer all flux is diffuse, and the net vertical flux is then

$$\begin{aligned} I_+ - I_- &= -\frac{K}{2\gamma + K}\frac{1 + 2\gamma\cos\zeta}{1 - K^2\cos\zeta}\exp(-K\frac{1}{1-\omega_o}\frac{\kappa p}{g})\omega_o L_\circledast \cos\zeta \\ &\approx \sqrt{\frac{\gamma}{\gamma'}}(1 - 2\gamma\cos\zeta)\sqrt{1-\omega_o}\exp(-2\sqrt{\gamma\gamma'}\frac{1}{\sqrt{1-\omega_o}}\frac{\kappa p}{g})L_\circledast\cos\zeta \end{aligned} \tag{5.50}$$

Hence the flux which manages to penetrate into the atmosphere is absorbed over a layer depth which scales with $\sqrt{1-\omega_o}$, and approaches zero as $\omega_o \to 1$. It follows also from Eq. 5.50 that the heating rate $d(I_+ - I_-)/dp$ remains order unity in this limit, though it becomes concentrated in a thinner and thinner layer near the top of the atmosphere.

5.7 Numerical solution of the two-stream equations

Eq. 5.27 is a coupled, linear system of ordinary differential equations requiring two boundary conditions. It takes the form of a *two-point boundary value problem*, because one specifies a boundary condition on I_+ at $\tau^* = 0$ and on I_- (generally that it vanish) at $\tau^* = \tau_\infty$. In vector form, the system can be written

$$\frac{d}{d\tau^*}\mathbf{V} = \mathbf{M}(\tau^*)\cdot\mathbf{V} + \mathbf{F}(\tau^*) \tag{5.51}$$

where

$$\mathbf{V} \equiv \begin{bmatrix} I_+ \\ I_- \end{bmatrix}, \mathbf{M} \equiv \begin{bmatrix} -\gamma_1 & \gamma_2 \\ -\gamma_2 & \gamma_1 \end{bmatrix}, \mathbf{F} \equiv \begin{bmatrix} \gamma_B \\ -\gamma_B \end{bmatrix}\pi B(\nu, T(\tau^*)) + \begin{bmatrix} \gamma_+ \\ -\gamma_- \end{bmatrix}\exp(-(\tau^*_\infty - \tau^*)/\cos\zeta). \tag{5.52}$$

The matrix $\mathbf{M}$ varies with τ^* because the single scattering albedo ω_o and also the asymmetry parameter in general will be functions of altitude. The forcing $\mathbf{F}$ depends on τ^* on account of the variation of T with altitude and the exponential factor in the direct-beam term. Because of these variations, the equations must generally be solved numerically. In most applications, the profiles of T and scatterer properties are given as functions of p, and often as tabulated values on a fixed grid rather than as functions. In order to carry out the needed integrations, the profiles must be re-expressed as functions of τ^*. The practicalities of how one goes about doing this will be discussed somewhat later. In this section we will show how to get the solution for a single frequency ν, for which the absorption is characterized by a single absorption coefficient at each pressure level. The extension to the band-averaged case where we have to deal with a distribution of absorption coefficients will be dealt with in Section 5.9.

Because the system described by Eq. 5.51 is linear, the solution satisfying the boundary conditions can, in principle, be built up from a suitable superposition of solutions obtained by numerically integrating the system starting from the lower boundary. Using a numerical differential equation integrator, one first constructs the following three solutions:

- A *particular solution* $\mathbf{V}_{part}$ which satisfies Eq. 5.51 subject to values of I_+ and I_- at $\tau^* = 0$ specified in whatever way proves convenient. We will assume $I_+ = I_- = 0$ for the particular solution at the lower boundary, but almost any other choice would do as well.

- A pair of *homogeneous solutions* $\mathbf{V_{hom,o}}$ and $\mathbf{V_{hom,1}}$ which satisfy Eq. 5.51 with the forcing term $\mathbf{F}$ set equal to zero. We will construct these solutions by integrating the homogeneous form of the equation subject to lower boundary conditions $I_+ = 1, I_- = 0$ for $\mathbf{V_{hom,o}}$ and $I_+ = 0, I_- = 1$ for $\mathbf{V_{hom,1}}$

The general solution to the original inhomogeneous system is then

$$\mathbf{V} = \mathbf{V}_{part} + a_o\mathbf{V_{hom,o}} + a_1\mathbf{V_{hom,1}} \tag{5.53}$$

where a_o and a_1 are any real numbers. At the upper boundary we require that the downward diffuse component vanish, i.e. $I_-(\tau^*_\infty) = 0$. At the lower boundary, we require that Eq. 5.31 be satisfied. The superposition of basic solutions that satisfies both boundary conditions is determined by solving the system

$$\begin{aligned} 0 &= I_{-,part}(\tau^*_\infty) + I_{-,hom,o}((\tau^*_\infty))a_o + I_{-,hom,1}((\tau^*_\infty))a_1 \\ e_g\pi B(\nu, T_g) + \alpha_g L_\circledast \cos\zeta \exp(-\tau^*_\infty/\cos\zeta) &= (I_{+,part}(0) - \alpha_g I_{-,part}(0)) \\ &\quad + (I_{+,hom,o}(0) - \alpha_g I_{-,hom,o}(0))a_o \\ &\quad + (I_{+,hom,1}(0) - \alpha_g I_{-,hom,1}(0))a_1 \end{aligned} \tag{5.54}$$

for the coefficients a_o and a_1. The conditions define a 2×2 linear system of equations which in general has a unique solution. Once the coefficients a_o and a_1 are known, the solution to the problem is complete. The procedure can equally well be carried out by starting at the top and integrating downward instead.

This approach works well and is very efficient as long as the atmosphere is not too optically thick. However, when τ^*_∞ becomes large, the method breaks down for the following reasons. The homogeneous version of Eq. 5.51 has solutions which grow or decay exponentially, with a local exponential growth or decay rate of $\pm\sqrt{\gamma_1^2 - \gamma_2^2}$. In the pure scattering limit $\gamma_2 = \gamma_1$ and the growth becomes merely algebraic, as evidenced by the analytic solutions considered previously. However, in the general case where $\gamma_2 < \gamma_1$, the exponentially growing solution causes numerical difficulties in the optically thick limit because the exponential growth will cause an overflow error when one attempts to integrate from the lower boundary to the upper boundary (or vice-versa). A related problem is that it becomes impossible to find two linearly independent homogeneous solutions because the the exponentially growing solution comes to dominate both homogenous solutions as one integrates sufficiently far from the boundary. As a simple example of this, consider that $\exp(\tau^*) + a\exp(-\tau^*)$ eventually converges on $\exp(\tau^*)$ to within computer roundoff error for sufficiently large τ^*, regardless of the value of a. Exactly how large τ^*_∞ needs to be before this problem becomes serious depends on the vertical profile of $\sqrt{\gamma_1^2 - \gamma_2^2}$, but except when gaseous absorption is weak and scattering is dominant this quantity is order unity, and so the solution method tends to break down when $\tau^*_\infty \approx 10$. Because of the exponential growth which is at the root of the problem, increasing the precision of the computer arithmetic only modestly extends this value.

Fortunately, there are a number of simple resolutions to the problem. First of all, if one only wants the OLR, then it is not necessary to integrate the equations all the way from the top of the atmosphere down to the ground. At a frequency where the atmosphere is optically thick by virtue

of strong absorption, radiation from lower levels of the atmosphere is exponentially attenuated and does not significantly affect the OLR. In these circumstances, one starts the integration at a height $\tau^*_\infty - \tau^*_1$ for some suitably chosen $\tau^*_1 < tau^*_\infty$. At $\tau^*_\infty - \tau^*_1$ one can impose any order unity boundary condition on I_+ that proves convenient, since the boundary condition there won't affect the OLR as long as the fluxes are not made exponentially large. The lower boundary condition $I_+ = 0$ would suffice, though one could squeeze out a little more accuracy by using the optically thick limit for the boundary condition (i.e. $I_+ = \pi B(\nu, T(\tau^*_\infty - \tau^*_1)$). One carries out the rest of the procedure exactly as before, except for applying the lower boundary condition at the artificial lower boundary instead of $\tau^* = 0$. The choice of τ^*_1 depends on the profile of $\sqrt{\gamma_1^2 - \gamma_2^2}$. If the typical value of this quantity is A, then we would take $\tau^* \approx 10/A$, based on the notion that $e^{10} + e^{-10}$ is accurately distinguishable from e^{10} with computer arithmetic having precision of at least twelve or thirteen decimal places. When $\sqrt{\gamma_1^2 - \gamma_2^2}$ varies greatly in the vertical it can be a bit tricky determining an appropriate "typical value," and so it is generally better to start the integration from the top, in which case one can simply integrate the two homogeneous solutions downward until one or the other grows by a factor of at least e^{10}, whereupon one stops and applies the artificial lower boundary condition.

A similar idea can be applied if one wants to compute the fluxes deep in the interior of the atmosphere. The typical application of such a calculation would be to determine the radiative heating profile for use in a radiative-convective equilibrium calculation. Even in an all-troposphere model, for which the climate can be solved for knowing the OLR alone, one might wish to compute the interior infrared cooling rate so as to determine the magnitude of the convective heat transport required to balance radiative cooling.

Suppose one wishes to compute the fluxes in the vicinity of some level τ^*_o. The solution method exploits the fact that the fluxes near τ^*_o will be insensitive to conditions many optical thicknesses *above or below* τ^*_o, since the fluxes from distant layers will be exponentially attenuated by the time they reach τ^*_o. One then builds the solution from solutions constructed by integrating both upwards and downwards from tau^*_o as follows:

- Integrating upward, one constructs a particular solution $V_{>,part}$ and a homogeneous solution $V_{>,hom}$ which satisfy the upper boundary condition $I_- = 0$. The upper boundary condition is applied at τ^*_∞ if the true upper boundary is reached before the exponentially growing mode dominates, but otherwise the integration is halted and the upper boundary condition is applied when the exponential dominance criterion is met. The particular solution satisfying the upper boundary condition is constructed from a superposition of an arbitrary particular solution and two independent homogeneous solutions, all obtained by upward integration starting from τ^*_o. Similarly, the homogeneous solution satisfying the upper boundary condition is obtained as a suitable superposition of two independent homogeneous solutions.

- Integrating downward , one constructs a particular solution $V_{<,part}$ and a homogeneous solution $V_{<,hom}$ which satisfy the appropriate lower boundary condition. If the true lower boundary $\tau^* = 0$ is reached before the exponentially growing mode dominates, then the physical boundary condition in Eq. 5.31 is applied. Otherwise, an artificial boundary condition $I_+ = 0$ is applied. As for the upward integration, the solutions are constructed through suitable superpositions of independent particular and homogeneous solutions obtained by downward integration.

The general solution satisfying the upper and lower boundary condition is then $V_{>,part} + a_o V_{>,hom}$ for $\tau^* > \tau^*_o$ and $V_{<,part} + a_1 V_{<,hom}$ for $\tau^* < \tau^*_o$. By imposing the requirement that the fluxes I_+ and I_- be continuous across τ^*_o, one obtains a 2×2 linear system for the coefficients a_o and a_1

much as we had before. This uniquely determines the solution. When the physical boundary isn't reached, then on the side of τ_o^* where this happens the solution will only be valid within a few optical thicknesses of τ_o^*, because the artificial boundary condition will begin to affect the solution as the artificial boundary is approached. To map out the entire flux profile, one only needs to carry out the procedure for a list of τ_o^* dense enough that the regions where the solutions are valid overlap. We'll refer to this solution method as the *piecewise ODE method* (ODE being shorthand for Ordinary Differential Equation). Note that in the optically thick limit, the fluxes typically vary very little over a unit optical thickness, because the fluxes are mostly determined by the local values of temperature and optical constants, which vary slowly when written in optical thickness coordinates. This means that it is usually possible to compute the flux at a rather coarse grid of τ_o^* and just use interpolation to get intermediate values if they are needed.

Once the fluxes have been determined, it is easy to get the heating rate, which is proportional to $d(I_+ - I_-)/d\tau$. Rather than numerically differentiating the fluxes, one can obtain the necessary derivatives by simply evaluating the right hand side of Eq. 5.51.

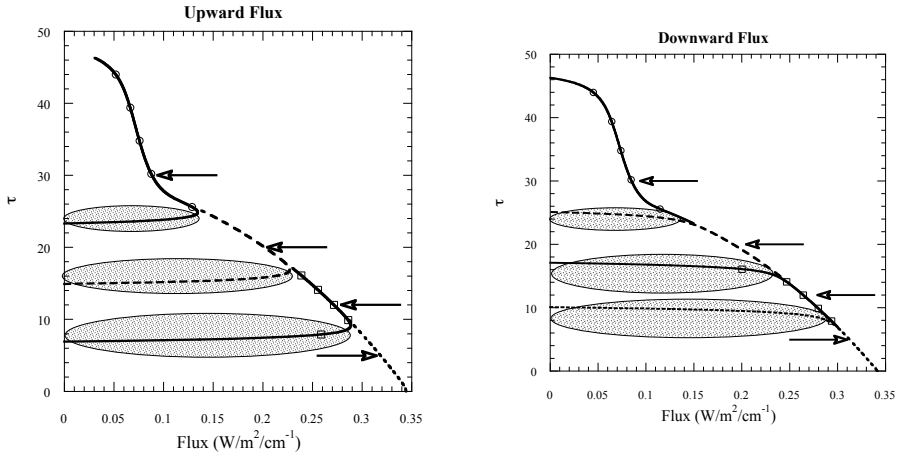

Figure 5.8: An example of the results of computing the upward and downward flux using the piecewise ODE method in the optically thick case. The arrows show the values of τ_o^* for which the upward and downward integrations were performed. The shaded ellipses show the regions where the artificial boundary condition has significantly affected the solution. This calculation was carried out for a wavenumber of 600 cm^{-1} assuming the temperature profile to lie on the dry air adiabat with a ground temperature and surface air temperature of $270K$. The surface pressure is 2 bar, the absorption coefficient is $1.5 \cdot 10^{-4} m^2/kg$ at $100mb$ and linearly pressure-broadened elsewhere. An optically thick purely scattering cloud is included, centered at an optical thickness value of 35 . The calculation assumes a gravitational acceleration of 10 m/s^2.

The procedure may sound complicated, but it is actually rather simple to implement on the computer, since it is built from the same basic integration operation carried out several times over, supplemented by a numerical routine that solves a 2×2 linear system. An example showing the results of this procedure is given in Fig. 5.8 . The solution is carried out for a single infrared wavenumber (600 cm^{-1}) with gaseous absorption as described in the figure caption. In addition, a purely scattering cloud has been placed by the top of the atmosphere. For this profile, carrying out the procedure for four suitably chosen τ_o^* is sufficient to map out the flux profile over the entire domain. Note that the portion of the subintegrations which are contaminated by artificial

boundary conditions are different for the upward and downward fluxes; for I_+, it is the vicinity of the artificial lower boundary that is most affected, whereas for I_- it is the vicinity of the upper boundary. Note also that the subintegration carried out nearest the top of the atmosphere covers a wider range of optical thickness than the others. This is because the optical thickness near the top is dominated by the purely scattering cloud, and pure scattering does not yield solutions with exponentially growing character.

In the calculations shown in Fig. 5.8 we used $I_+ = 0$ or $I_- = 0$ for the artifical boundary conditions, so as to show more clearly the regions where the artificial conditions affect the solution. One can achieve considerably better accuracy close to the artificial boundaries by instead using the approximate optically thick solution $I_+ = I_- = \pi B(\nu, T)$ as the boundary condition, where T is evaluated at the boundary.

The piecewise ODE method is flexible, accurate and easy to implement, and it deserves to be more widely used. A more customary approach to dealing with the numerical issues in the optically thick case is to turn the problem into a discrete set of linear equations and solve it using numerical linear algebra algorithms. To do this we set up a grid of discrete values $\tau_j = j\Delta\tau$ for $j = 0, 1, ...N$, and approximate the derivative in Eq. 5.51 as $(dV/d\tau^*)_j \approx (V_{j+1} - V_{j-1})/\Delta\tau$, where the integer subscripts stand for the value of the corresponding quantity at $\tau^* = j\Delta\tau$. With this approximation, Eq. 5.51 can be written as

$$V_{j+1} - V_{j-1} - \Delta\tau \mathbf{M}_j \cdot \mathbf{V}_j = \Delta\tau \mathbf{F}_j \tag{5.55}$$

where $j = 1, 2, ...N - 1$. The $j = 0$ and $j = N$ lines must be left out because evaluation of the approximate derivative would require data off the end of the array. These two missing lines are replaced by the statement of the boundary condition at the bottom and top of the atmosphere respectively. This results in a $2N \times 2N$ linear system if the upward and downward fluxes are merged into a single solution vector of the form $[(I_{+,o}, I_{-,o}, I_{+,1}, I_{-,1}, I_{+,2}, I_{-,2}, ...]$. The corresponding matrix defining the coefficients of the system is *band-diagonal*, and has nonzero entries only within two spaces to the left or right of the diagonal. Such systems can be solved efficiently using standard methods of linear algebra available in virtually all numerical libraries [3]. We'll refer to this solution technique a the *matrix method*.

Exercise 5.7.1 Write out the first three lines and the last three lines of the matrix problem outlined above, assuming ω_o to be constant.

Resolving the exponential growth or decay of the homogeneous solutions requires $\Delta\tau < 1$ at least, so it might be thought that a disadvantage of the matrix method is that it requires dealing with very large matrices when the τ^*_∞ is large – and with strong gaseous absorption, optical thicknesses of several thousand are not at all uncommon. However, the required resolution is not really determined by the exponential behavior, since, as noted previously, the fluxes are largely determined by local properties in the optically thick limit, but local properties such as temperature usually vary slowly when written as a function of τ^* in the optically thick case. Thus, the fluxes vary little over a unit optical depth, and the derivative of the flux can be accurately computed even if $\Delta\tau$ is quite large. Another way of seeing this is to note that when $\Delta\tau >> 1$ the term $V_{j+1} - V_{j-1}$ in Eq. 5.55 is negligible compared to the remainder of the terms, and the solution of what is left gives the correct lowest-order solution in the optically-thick limit. (Recall that the optically thick local solution is determined by treating the temperature and single-scatter albedo as if they were constant). Corrections to this approximate solution are of order $1/\Delta\tau$, and will be properly computed in the linear solution algorithm. The only caveat needed is to note that

[3] See, for example the routine `bandec` in *Numerical Recipes*.

besides the correct optically thick solution the linear system also has a spurious homogeneous two-gridpoint oscillating solution of the form $I_+(j) = [1, -1, 1, -1, ...]$ and similarly for I_-. This spurious solution replaces the unresolved exponential homogenous solution, and care must be taken to use a linear system solver that correctly suppresses contamination by the spurious mode.

There are two other small technical details that need to be taken care of when using the matrix method in the optically thick limit. The first is that the conversion layer at the top of the atmosphere is not resolved, so that the direct-beam term is not correctly transformed into diffuse radiation. This is easily dealt with by eliminating the direct beam term and dumping the corresponding energy directly into $I_-(\tau^*_\infty$ as an upper boundary condition, which explicitly captures the conversion of direct beam energy at the top of the atmosphere. Finally, if there is a significant temperature discontinuity between the ground temperature and the overlying air temperature, the exponential homogeneous solution comes into play in a kind of radiative boundary layer, and something must be done to explicitly resolve this layer, either analytically or by increasing the resolution near the boundary. When there is a temperature discontinuity, there is a strong, shallow radiative heating or cooling within a unit optical depth of the ground, and this will not be resolved by the matrix method using large $\Delta\tau$. With the ODE method, the radiative boundary layer is automatically resolved, but when using the matrix method one must generally use an analytical exponential solution to get the radiative fluxes near the ground.

Whichever method one chooses, it is often most convenient to work in optical thickness coordinates rather than transforming the equations to pressure or log-pressure space before performing the solution. Since the mapping between pressure and optical thickness is usually frequency dependent, it is necessary to express the fluxes as a function of pressure before summing up the fluxes and heating rates across frequencies. The easiest way to re-express the results is to use numerical integration to compute a suitable list of $\tau^*(p)$ from the differential equation defining τ^*, and then to use this list to define a numerical interpolation function giving $p(\tau^*)$ [4]. Using this function a list of triples (I_+, I_-, τ^*) can be transformed to the list of triples (I_+, I_-, p) at the corresponding pressure levels. This list can then be used to interpolate the fluxes to a standard grid of pressure values; heating profiles are treated similarly. The interpolation function $p(\tau^*)$ fulfills an additional role when carrying out the integration for the fluxes. Namely, the temperature profile T and the scattering parameters are generally given as functions of p, or as tabulated values for a list of pressure levels. Using the interpolation function, however, one simply evaluates $T(p(\tau^*))$ and so forth to get the required expression. If $T(p)$ is specified as a table of values rather than a function, then one transforms the pressure entry to optical depth using the interpolation function, and then writes an interpolation function to get $T(\tau^*)$ from this table. When using the matrix method, it may not be necessary to write functions for T and the scattering parameters since tabulated values will generally be sufficient. When using the differential equation, method, however, it is usually necessary to provide functions, since the typical high order numerical integrator needs to be able to evaluate the right hand side of Eq. 5.51 at an arbitrary value of the vertical coordinate.

Clearly, in order to carry out the above procedure, it is necessary to have at hand an efficient, accurate and easy-to-use interpolation function. Some useful advice on interpolation methods has already been given in connection with the numerical analysis tutorial problems in the Workbook section of Chapter 1.

[4] By "interpolation function" we mean a computer-implemented function that takes a list of N values (x_j, y_j) and an arbitrary argument x, and returns an interpolated estimate of the value of y corresponding to x. By putting in a list of (y_j, x_j) instead, the same function can be used to obtain $x(y)$

5.8 Water and ice clouds

Now we'll take a closer look at the way Earth's water and water-ice clouds effect the radiation budget, taking account of the balance between the shortwave albedo effect of clouds which act to cool the planet and the longwave cloud greenhouse effects which act to warm the planet. Much of the general behavior in evidence on Earth applies equally well to water clouds on other planets, or for that matter to any cloud-forming substance which is strongly absorbing in the infrared but fairly transparent in the solar spectrum.

To set the stage, we'll first discuss some calculations with the `ccm` radiation code which show how high clouds affect the albedo and OLR under typical tropical conditions. The results are shown in Fig. 5.9. These calculations include most of the radiative effects operating in the real Tropics, including solar and infrared absorption by water vapor and CO_2, though we have left ozone out of the picture. The surface albedo has been set to zero, so as to focus on the reflective effect of the cloudy atmosphere itself. In this calculation, the tropospheric temperature profile is on the moist adiabat, and we place a geometrically thin cloud with specified water content in the upper troposphere, where the pressure is 283 mb and the temperature T_c of the cloud is 243 K. At these temperatures, the cloud is primarily composed of ice, and the `ccm` radiation model makes use of the complex index of refraction appropriate to water-ice particles. Results are given as a function of the condensed water path of the cloud. As long as the cloud is geometrically thin enough to be essentially isothermal, the actual geometrical thickness of the cloud is irrelevant to this calculation.

For 30 μm particles, which are typical of actual tropical ice clouds, it only takes a path of 50 g/m^2 to make the cloud act like a blackbody. The OLR is about 20 W/m^2 below the blackbody emission σT_c^4 of the cloud itself, because there is some water vapor and CO_2 greenhouse effect in the colder air above the cloud. This effect would disappear if the cloud were placed at the minimum temperature part of the atmosphere, and would increase if the cloud were lower. In essence, a cloud that is optically thick in the infrared acts like a new "ground", radiating upward into the upper part of the atmosphere with a blackbody temperature T_c. If the particles are made smaller, it takes less cloud water in order to make the cloud optically thick, because the same mass of water yields more aggregate cross-section area of cloud particles. As a practical matter, most high clouds occuring in the vicinity of deep convection in the Tropics can be considered optically thick in the infrared. The associated cloud greenhouse effect is enormous, and would lead to an uninhabitably hot planet if not compensated by shortwave albedo effects that are of similar magnitude.

The albedo effect of the cloud also increases monotonically with cloud water content, but at a much slower rate than the greenhouse effect, for the reasons discussed in Section 5.6. For 30 μm particles, the planetary albedo has only reached .2 when the cloud water path is $50 g/m^2$. The albedo doesn't reach .5 until the cloud water content approaches 200 g/m^2. On the other hand, the failure of cloud albedo to saturate until large cloud water paths means that the particle size can have a very important influence on albedo; reducing the particle size to 10 μm increases the albedo to .7 for a cloud containing 200 g/m^2 of ice. As for compensation between shortwave and longwave cloud effects, taking a cloud with 30 μm particles and 100 g/m^2, the albedo is about .4, which yields 170 W/m^2 reduction in solar absorption based on typical annual average tropical insolation. This compares with a cloud greenhouse effect of $120 W/m^2$, so such clouds have a moderate net cooling effect. If the cloud only had a water content of 50 g/m^2, though, the cloud greenhouse effect would be nearly the same but the cloud albedo effect is reduced by nearly half, and the cloud would have a net warming effect. Similarly, if the ground were partially reflecting (owing to vegetation cover or low-lying clouds), the change in albedo due to high clouds would be reduced, shifting the balance again in favor of net cloud warming. On the other hand, reducing the particle

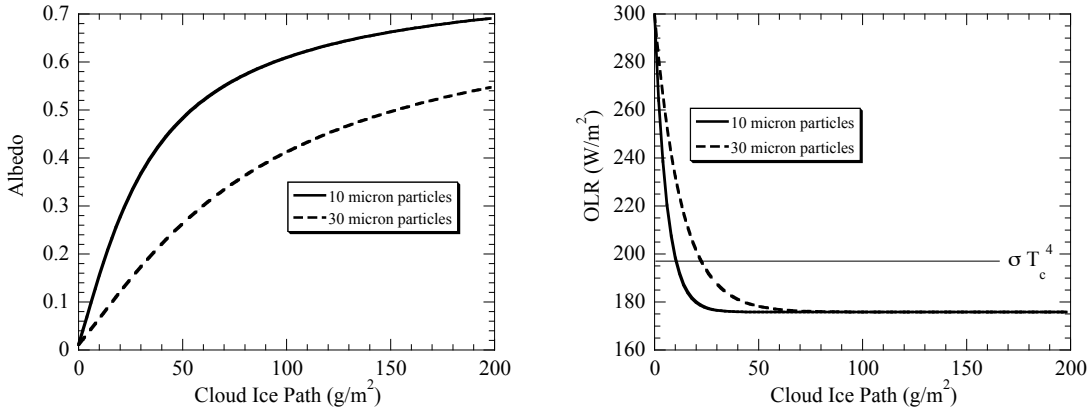

Figure 5.9: Albedo and OLR as a function of cloud condensed water path, for a high ice cloud with temperature $T_c = 242K$ at a pressure $p_c = 283mb$. The temperature profile is on the moist adiabat corresponding to a surface temperature of $300K$, patched to an isothermal $180K$ stratosphere. The relative humidity is 50% and the CO_2 concentration is $300ppmv$, but there is no ozone in the atmosphere. Calculations were done with the `ccm` radiation code, and results are shown for both 30 μm and 10 μm particles.

size of the clouds makes them much brighter, making it easier for the clouds to have a net cooling effect.

For fixed particle size, the cloud altitude has relatively little effect on albedo for a given cloud condensed water path. Low altitude liquid water clouds tend to have smaller particles than ice clouds as well as larger water content (because there is more water around to condense), and are correspondingly more reflective. Because of this effect, the balance of power for mid-level and low-level clouds shifts decidedly toward a net cooling effect on the planet.

High clouds have both a warming and a cooling effect, and which one wins depends on the detailed of the cloud properties, including cloud temperature, particle size, and condensed water content. Colder cloud temperatures tend to favor net warming. Making particles smaller or increasing cloud water enhances the cooling effect except for very thin clouds, since it takes little cloud water to make the cloud act like a blackbody whereas the albedo continues to increase with cloud water increase or particle size decrease, for the reasons discussed in Section 5.6. The balance between albedo effect and greenhouse effect also depends on the intensity of solar radiation, since albedo matters not at all if there is no sunlight. In the polar night, clouds have an unambiguous warming effect as long as they are not right at the surface. Similarly, the cloud albedo effect depends on the albedo of the underlying surface; clouds over a reflective surface like ice (or surface clouds!) will tend to have a warming effect, as also discussed in Section 5.6. As the cloud is made lower, the cloud greenhouse effect is attenuated, because the cloud temperature is closer to the ground temperature and also because (especially in moist regions) the greenhouse effect of the clear air above the clouds masks the longwave radiative effect of the cloud itself.

Although the large and competing effects of clouds on OLR and albedo pose similar challenges on any planet whose atmosphere contains a condensible substance, Earth is the only case at present for which we have good observations of the net radiative effect of clouds. The first satellite mission to do this accurately was the Earth Radiation Budget Experiment (ERBE), and

subsequent missions have taken a similar approach. We discussed some ERBE clear-sky results in Chapter 3, and now we will see what ERBE has to tell us about cloud effects. The ERBE mission measured the Earth's radiation budget using two sets of highly accurate broadband radiometers borne on satellites – one in the infrared spectrum and one in the shortwave (i.e. solar) spectrum. Moreover, the processing algorithm made use of the patchiness of Earth's cloud cover in order to estimate the effect of clouds on the longwave and shortwave radiation. Within each scene examined (think of a scene as a $50km$ square patch of the Earth's surface) the algorithm identified those pixels which represented cloud-free clear-sky conditions, and defined "clear sky" longwave and shortwave flux as the value the flux over the scene would have if the flux of *all* pixels in the scene were replaced by the average of the clear-sky pixels. In the longwave, for example, the ERBE retrieval reports the all-sky OLR, called OLR_{all} and the clear-sky OLR, called OLR_{clear}. The cloud longwave forcing is then defined as $OLR_{clear} - OLR_{all}$. Since clouds reduce the OLR by making the upper troposphere more optically thick, the cloud longwave forcing is positive, and represents a warming effect. Similarly, the cloud shortwave forcing is defined as $S_{abs,all} - S_{abs,clear}$, where S_{abs} is the top-of-atmosphere absorbed solar radiation – the difference between incoming and reflected solar radiation. Clouds reduce the solar absorption by increasing the albedo, and so the cloud shortwave forcing thus defined is generally negative, representing a cooling effect. The sum of the cloud longwave and cloud shortwave forcings is the net cloud forcing, with positive values representing a warming tendency and negative values representing a cooling tendency. Clear sky and all-sky albedo can be defined similarly.

Results for clear and cloudy albedo, and for the cloud radiative forcing are shown for the year 1988 in Figure 5.10. Other years show a similar pattern. The ERBE dataset contains information of this sort for each month, reported on a latitude-longitude grid. Here we show only annual-mean results averaged along latitude circles. The full monthly-mean dataset for all available years is provided as part of the dataset collection in the supplementary materials for this book. Turning attention first to the clear-sky albedo, we see that without clouds the albedo varies in a narrow range of .11 to .16 from 60S latitude to 42N. Poleward of 60S, the albedo increases sharply owing to the high albedo Antarctic ice. Still, the values indicate that the albedo of the partially snow-covered Antarctic ice must exceed .7, since the atmospheric absorption makes the planetary albedo lower than the surface albedo. The clear-sky estimates near the pole are somewhat unreliable, since it is hard to distinguish between low clouds and ice. Going towards the North pole from 42N the albedo increases somewhat more gently, owing to the patchy distribution of sea ice and its seasonal fluctuations; the rest of the Northern high-latitude albedo increase is due to winter snow-cover over land. Clouds have a strong reflective effect, approximately doubling the tropical albedo and increasing the midlatitude albedo to .4 or more. The area-weighted mean albedo is .19 for clear sky and .33 including cloud effects. Area-weighting doesn't take into account the seasonal and latitudinal distribution of sunlight though; a more appropriate mean albedo is based on taking the ratio of global net reflected solar radiation to the incident radiation. This estimate yields somewhat smaller values: a mean clear-sky albedo of .16 and a mean all-sky albedo of .30.

If uncompensated by the cloud greenhouse effect, the high albedo of clouds would probably be sufficient to throw the Earth into a Snowball state. In reality, the reduction of OLR by clouds cancels most of the cloud cooling effect, as shown in the right hand panel of Figure 5.10. The cloud longwave forcing – i.e. the reduction in OLR due to clouds – is anti-correlated with the cloud shortwave forcing, and has sufficient magnitude to cancel most of the cloud shortwave forcing. The distribution of cloud forcing takes us into some consideration of aspects of the general circulation we have not introduced previously. Somewhat North of the Equator there is a region of deep convection yielding deep, thick clouds, which is manifest in the Figure as a peak in both the cloud longwave and shortwave forcing (marked "ITCZ" in the figure for *Inter-Tropical Convergence Zone*, in honor of the winds which converge moisture into this region and feed the convection). The ITCZ

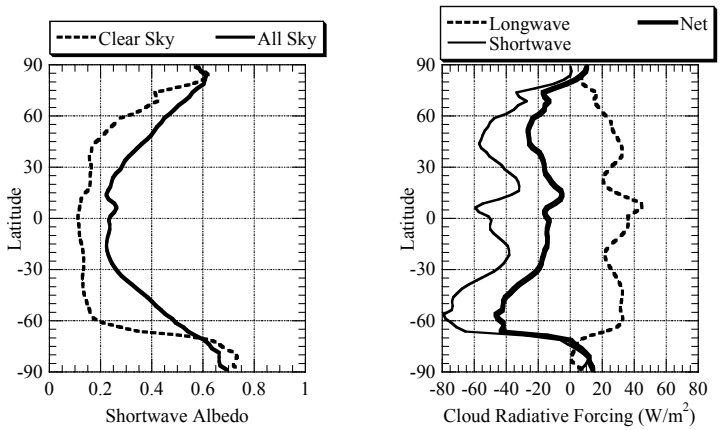

Figure 5.10: Zonally averaged annual mean clear and cloudy sky albedo(left panel) and cloud radiative forcing (right panel) measured by ERBE for the year 1988.

is flanked by two subtropical regions where convection is suppressed by downward motions in the atmosphere, and is shallow or absent. Here one encounters local minima in both the cloud longwave and shortwave forcing. Throughout the tropics, the two terms sum to a net cooling effect of about $-20W/m^2$, which is stronger in the subtropics than near the ITCZ. The subtropical cloud cooling is in part due to near-surface clouds which are associated with the boundary layer rather than deep convection. In the midlatitudes, there is another region of deep cloud activity. This one is associated with the large scale organized storm tracks, which loft water from the subtropical ocean and move it poleward and upward. The albedo effect of clouds more strongly dominates the greenhouse effect in this region, and even more so towards the Antarctic region, where there is strong cloud shortwave forcing associated with low-lying marine stratus clouds. As a result there is strong net cooling in the midlatitude and polar regions.

The area-weighted global mean cloud longwave forcing is $28W/m^2$, while the mean cloud shortwave forcing is $-47W/m^2$, which nets out to a cooling influence of $-19W/m^2$. Using a sensitivity factof of $2.2W/m^2K$ from Section 4.5, we conclude that the Earth would be about $8.6K$ warmer if there were no clouds.

In circumstances under which the clear-sky regions do not absorb much solar radiation, high clouds have a potent net warming effect, though there must be enough convection around to loft water to sufficiently high altitudes to make a high optically thick cloud. As we have already mentioned, clouds can contribute significantly to deglaciation of the high-albedo Snowball Earth, though the main question there is whether it is possible to make sufficiently high clouds with sufficiently great water content in an atmosphere with low water content (because of low temperature) and sluggish convection (because of low solar absorption). Another situation in which the clear-sky solar absorption is low is the high-latitude winter. Here, there is little solar radiation to reflect from clouds simply because it is night or twilight most of the time, and so if there are high clouds they will have a pronounced winter warming effect, and perhaps even inhibit the formation of sea ice in open water conditions. This effect may play a role in the Arctic during the Cretaceous hothouse climate, since there is open water in the Arctic Ocean which can maintain a supply of relatively warm water throughout the winter to feed deep convection. It seems plausible that this mechanism would help explain the mysterious low-gradient climate of Cretaceous and

similar hothouse climates, described in Section 1.9.1. General circulation models to date do not support a sufficiently strong cloud effect for clouds to be the answer to the Cretaceous puzzle, but there is much remaining to be learned about clouds, so the last word has not by any means been uttered on this topic. This potential mechanism is only viable when there is open water in the polar ocean. Over a polar continent, such as Antarctica, the ground would cool off rapidly in the Winter, foreclosing any serious possibility of deep convection and the associated deep clouds.

The effect of clouds on the water-vapor runaway greenhouse represents one of the most vexing and important unresolved issues in planetary climate. The observed behavior of Earth clouds can provide little guidance as to cloud effects in a much warmer atmosphere in which water vapor is the dominant component. Given the availability of water throughout the depth of the atmosphere, and how little water it takes to make a highly reflective cloud, it seems almost inevitable that the albedo will become very high. There is no simple physics, however, that can be employed to estimate the fraction of the atmosphere which will be cloudy; this is an intrinsically dynamical question. High clouds can also reduce the OLR, however, which has the potential to offset the albedo effect. A strong cloud greenhouse effect is likely, given that the top 100 mb of a near-runaway steam atmosphere contains far more water vapor than Earth's entire atmosphere. In order for clouds to make the radiating temperature cold enough to offset the strong cloud albedo increase, one would need optically thick clouds in regions of the atmosphere which are very cold. This is not impossible, though: a thick cloud at an upper atmospheric temperature of 200 K would reduce the OLR to 91 W/m^2, which is just about the same as the solar radiation Early Venus would absorb if it had an albedo of 80 %. In Earth's atmosphere, optically thick tropical cirrus clouds occur at similar temperatures, so one can hardly rule them out for Venus. The question of whether the cloud greenhouse or cloud albedo effect wins out in a runaway situation simply cannot be answered by back-of-the-envelope calculations, even if we have a rather large envelope. A definitive answer must await attainment of a far better understanding of convection, cloud microphysics and cloud fraction under near-runaway conditions.

5.9 Things that go bump in the night: Infrared-scattering with gaseous absorption

As we have already noted, infrared scattering needs to be taken into account when the planet's clouds are made of a substance that is nearly transparent to infrared in significant parts of the spectrum. This is the case for CO_2 ice clouds, liquid or solid CH_4 clouds, N_2 ice clouds and concentrated sulfuric acid clouds. The CO_2 cloud scattering is evident even in spectra of Present Mars, but it has the potential to have a major impact on the climate of an Early Mars with a thick CO_2 atmosphere. The CH_4 case is relevant to Titan, and the N_2 case would be highly important there if Titan were a bit colder. The sulfuric acid case is relevant to dry hot planets like the present Venus. All these scattering clouds, and probably more, are present to some extent in the atmospheres of gas and ice giants. This is an important cloud regime. In most of these cases, the scattering of infrared radiation by clouds acts jointly with the full complexities of gaseous absorption.

Traditionally, the problem of dealing with the joint effects of scattering and gaseous absorption has been considered to be one of the scariest problems in radiative transfer. The problem appears scary only if one intends to mount an attack on it by some kind of modification of the band-averaged transmission function approach. The problem here stems from the fact that band-averaged transmission functions do not satisfy the multiplicative property, so that the path that one feeds to the transmission function involves the entire past history of the radiation between

being the time it is emitted (or injected by solar radiation) and the time it leaves the atmosphere. When there is no scattering, the path is simple, but multiple reflection leads to an ensemble of very complex paths. For example, consider an isothermal layer of a nongrey gas sandwiched between a perfectly reflecting ground and a cloud that reflects half of all energy incident on it from below. Suppose the layer has a mass path ℓ. Upward radiation emitted from near the center of the layer will be attenuated according to a path length $\frac{1}{2}\ell$ by the time it reaches the cloud. Half of this will escape to space, but the other half will be reflected downward. To get out, it reaches the bottom boundary, where it is reflected upward. By the time it hits the top again, the path length is $\frac{3}{2}\ell$, and the beam has been attenuated accordingly. Half of this escapes while the other half is reflected downward, and the process continues until there is essentially no beam left. If $\mathfrak{T}$ is the transmission function written as a function of the path, the escaping flux is

$$\frac{1}{2}I_o\mathfrak{T}(\frac{1}{2}\ell) + \frac{1}{4}I_o\mathfrak{T}(\frac{3}{2}\ell) + \frac{1}{8}I_o\mathfrak{T}(\frac{5}{2}\ell) + ... \tag{5.56}$$

where I_o is the initial upward radiation emitted from the center of the layer. If the transmission function had the multiplicative property, then we would have

$$\mathfrak{T}((\frac{1}{2}+n+1)\ell) = \mathfrak{T}((\frac{1}{2}+n)\ell)\mathfrak{T}(\ell) \tag{5.57}$$

and the problem could be done as an iteration in which the fate of each reflection of flux hitting the cloud is independent of how many bounces it took before it got there. For band averaged transmission functions, which do not retain the multiplicative property, this is not the case and one needs to perform the sum over all past histories of paths taken by radiation. For the two-reflector case this is not too bad, but when one considers a more realistic problem in which absorption and scattering occur between a continuous array of pairs of layers of the atmosphere, one does indeed begin to quake in ones boots, if just a bit.

Most of the fear can be dispelled, however, through use of the exponential sums approach and it variants. It is only the band averaging that creates the problem. For any given value of the absorption coefficient, the multiplicative property holds, which is in fact what allows us to write the two-stream equations as a differential equation. Since the exponential sum calculation is built from a number of non-averaged calculations for the individual absorption coefficients going into the sum, the essential difficulty is circumvented.

Within each band, the calculation is identical to either of the two approaches described in Section 5.7, except that one needs to do the calculation over many times for various values of the absorption coefficient, and then sum the results with the appropriate weighting. Specifically, as in the exponential sums method without scattering, we adopt a distribution function $H(\kappa_o)$ for the absorption coefficient at a reference pressure and temperatuer p_o and T_o. For each of the κ_o in the table, we scale $\kappa(p)$ over the rest of the profile according to pressure and temperature scaling factors, perform the numerical calculation of the flux, and then sum the results weighted according to H. Note that in order to perform this sum, the fluxes for each κ_o must be written on a common pressure grid, by interpolation or by other means. Finally, to get the net flux, the result is summed over all the bands that contribute.

To illustrate the application of the technique, we will carry out a simplified version of the Early Mars case. We assume a pure CO_2 atmosphere with a 2 *bar* surface pressure under the influence of Mars gravity. The temperature is taken to be on the dry CO_2 adiabat corresponding to a surface temperature T_g, until the threshold for CO_2 condensation is reached, whereafter the temperature follows the dew-point formula for the single-component saturated CO_2 adiabat. The calculations below were performed for $T_g = 275K$, since we are principally interested in how much

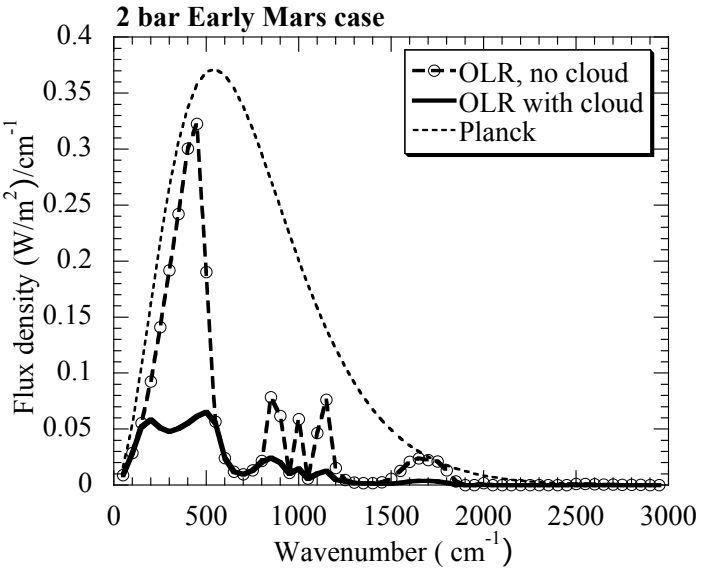

Figure 5.11: The spectrum of OLR for an Early Mars case with a 2 bar CO_2 atmosphere and a surface temperature of $275K$. The temperature profile is on the one-component adiabat, which noncondensing for pressures greater than 400 mb and condensing at higher altitudes. Results are shown both for clear sky conditions and with a high-altitude CO_2 ice cloud with peak condensate mixing ratio of 10^{-5}.

absorbed solar flux is needed to maintain liquid water at the surface. With this value of T_g the atmosphere hits the condensation point at a pressure level of $400mb$. We will look at how the OLR changes when we introduce an idealized upper level CO_2 ice cloud in the vicinity of the condensing region. Specifically we assume that the profile of CO_2 ice mass concentration in the cloud is

$$q(p) = q_m \exp(\frac{(p - p_m)^2}{\Delta p^2}) \tag{5.58}$$

We further assume that the mass-specific cross section of the scatterers is $\chi = 100m^2/kg$, which is approximately the value appropriate to $10\mu m$ CO_2 ice particles. This calculation ignores the wavenumber dependence due to Mie scattering properties and the variation of index of refraction with wavenumber. It also ignores the absorption of infrared by CO_2 ice. The incorporation of these additional effects presents no special technical difficulties, but they have been left out for the sake of simplicity.

Results for $p_m = 500mb$ and $q_m = 10^{-5}$ are shown in Fig. 5.11. These results show the spectrum of outgoing radiation with and without the cloud. Note that converting only a tiny proportion of the ambient atmosphere into condensed form has a profound effect on the radiation. The cloud powerfully reduces the OLR, particularly in the region below 500 cm^{-1} where CO_2 is fairly transparent, even including the continuum. Averaging over the spectrum, introducing the cloud reduces the OLR from 110 W/m^2 to a mere 37 W/m^2. Recall, however, that the warming effect is partly offset by increased reflection of solar radiation, as was discussed in Section 5.6.

In dealing with multiple greenhouse gases within this framework, one no longer has the option of implementing the random overlap assumption by multiplying the transmission functions

for the individual gases, since the transmission function is never explicitly computed. Instead, random overlap is implemented by forming the distribution $H(\kappa_o)$ for the mixture of gases as a suitably weighted convolution of the distributions for the individual gases. One can do even better and eliminate the random-overlap assumption by computing $H(\kappa_o)$ from the sum of the actual spectrally resolved absorption coefficient data for the mixture, which is a weighted sum of the absorption coefficients for the the individual species. This procedure can be slow if one does it on the fly each time the distribution for a new mixture is needed (as in the case of a highly variable substance like condensible water vapor), but the procedure can be sped up considerably by precomputing the distributions needed for a range of mixtures, and interpolating between the tabulated distributions. When many different greenhouse gases are involved, this can involve a considerable amount of storage, but computer memory is abundant and cheap, so this approach is gradually taking hold.

As with all approaches based on the exponential sums concept, the main vulnerability is the validity of the scaling assumption $\kappa(p,T) = \kappa_o F(p,T)$. The shortcomings of scaling can be to some extent overcome using the correlated-k variant of exponential sums, but the mathematical justification for the use of correlated-k when scaling is invaled is on even more shakey grounds in the presence of scattering than it is in the non-scattering case. When dealing with novel radiative transfer problems, it is always judicious to compare selected cases with the results of line-by-line calculations.

In the above we have emphasized the joint effects of infrared scattering with gaseous absorption, but many of the same issues apply in the visible and ultraviolet range. In that case, the scattering is even important for liquid water and water ice, and and Rayleigh scattering becomes significant as well. The problems are somewhat less challenging than for the infrared case because visible and ultraviolet absorption spectra tend to have less intricate fine structure than is the case for infrared Nonetheless the problem presents similar challenges, which can similarly be dealt with using exponential sums or its variants just as was done for the infrared case.

5.10 Effects of atmospheric solar absorption

On the present Earth the idealized picture of climate in which all solar absorption occurs at the ground is useful, but even for the present Earth about 20% of solar radiation is absorbed within the atmosphere. For other atmospheres, the proportion absorbed in the atmosphere could be much greater. The effect of this absorption on climate depends very much on the vertical distribution of the absorption, and that is what we will explore here for selected real gases.

The two key questions we have in mind for this section are the effect of solar absorption on the stratospheric temperature profile and the effect of solar absorption on surface temperature. When is solar absorption strong enough to inhibit convection and chill the surface? Even when the primary effects are stratospheric, it should be kept in mind that indirect feedbacks through stratospheric chemistry can still have an important influence on tropospheric climate. In particular, it takes very little mass of cloud to substantially affect the planet's radiation budget, and the temperature of the stratosphere can influence the kind of clouds that form there. A prime example of this phenomenon in the current Solar System is Titan, whose stratospheric organic haze clouds are a key player in the radiation budget. It has been suggested that such clouds may have played a role in a methane-rich anoxic Early Earth atmosphere, and indeed similar phenomena may be widespread in the universe.

For gas or ice giants, the problem of atmospheric solar absorption is even more critical, since

it is the *whole* story with regard to the external energy supply of the atmosphere, there being no distinct liquid or solid surface at which to absorb solar radiation. The profile of absorption can determine whether the solar driving hinders or helps convection, and (together with heat flux from the interior) determines where the troposphere is located. These planets contain a diverse variety of condensible substances and clouds, many of which are known to absorb near-IR, visible and UV radiation. However, little is known about the vertical distribution of absorbers, or even which ones are dominant in giant planets in our own Solar System. This is a very unsettled area about which we will have little to say; a few pointers into the literature are given in the Further Readings section. The situation with regard to stellar absorption in extrasolar gas or ice giants is of course even more unsettled, but offers wide scope for exploration of hypothetical atmospheres.

5.10.1 Near-IR and visible absorption

We'll begin with an overview of the near-IR and visible absorption characteristics of CO_2, CH_4 and water vapor, shown in Fig. 5.12 and Fig. 5.13. These are drawn from data in the `HITRAN` database. As was the case for thermal IR, the absorption coefficients have an intricate line structure leading to fine-scale variations with wavenumber. The figures summarize the properties by showing only the median absorption in bins of width 50 cm^{-1}. This is sufficient to provide a general idea of where the gases absorb strongly and where they are largely transparent. Results are given at a standard pressure of 100 mb, and can be scaled (approximately) linearly to other pressures as discussed in Chapter 4.

CO_2 has a dense forest of absorption features in the near-IR below 8000 cm^{-1}, and sparser, weaker absorption features at higher wavenumbers. We see immediately that the spectral class of the planet's star matters very much to the importance of near-IR absorption. For a cool, red M star, the stellar spectrum overlaps very considerably with the absorption features of CO_2, whereas a much lower (though not by any means insignificant) proportion of a hotter G star output lead to absorption. The same remark applies to virtually all infrared-active gases.

Which of the CO_2 absorption features come into play depends on how much CO_2 there is in the planet's atmosphere, and for this we turn to the pressure-adjusted path, defined in Section 4.2.1. For 300 $ppmv$ of CO_2 in 1 bar of Earth air under Earth gravity, the adjusted path based on the reference pressure used in Fig. 5.12 is about 20 kg/m^2, so it is only the three spikes that rise above an absorption coefficient of .01 m^2/kg that contribute significantly to atmospheric heating. (The adjusted path of the present thin Martian atmosphere is similar, at 13 kg/m^2, so the near-IR absorption picture for present Mars is rather similar to Earth.) All three spikes overlap significantly with the output of an M-star, but it is only the two higher wavenumber features that contribute significantly for a G star like the Sun. If the CO_2 on Earth increases to 20% (mole fraction) of the atmosphere, then the path is over 15000 kg/m^2, so all absorption features stronger than $10^{-5}m^2/kg$ come into play, which takes a fairly sizable chunk out of the incoming radiation for a G star and even more so for an M star. CO_2 levels of this magnitude are typical of the Faint Young Sun period on Earth, and are also similar to the levels required to enable deglaciation of a Snowball state; in such circumstances, near-IR absorption due to CO_2 becomes a significant player in climate. If we go to a 2 bar pure CO_2 atmosphere, the adjusted path is about $2 \cdot 10^5$ for Earth gravity, around three times as much for Mars gravity, and about half as much for a typical massive Super-Earth. At these values, the gaps where the atmosphere is transparent narrow considerably, and the atmosphere becomes optically thick in most of the spectral region below 6000 cm^{-1}. A 100 bar Venus-like atmosphere would absorb nearly all the incoming stellar radiation within all the absorption regions shown in Fig. 5.12, though the net increase in absorbed flux would be modest since the atmosphere will have already absorbed most of what it can absorb in the top 2 bar.

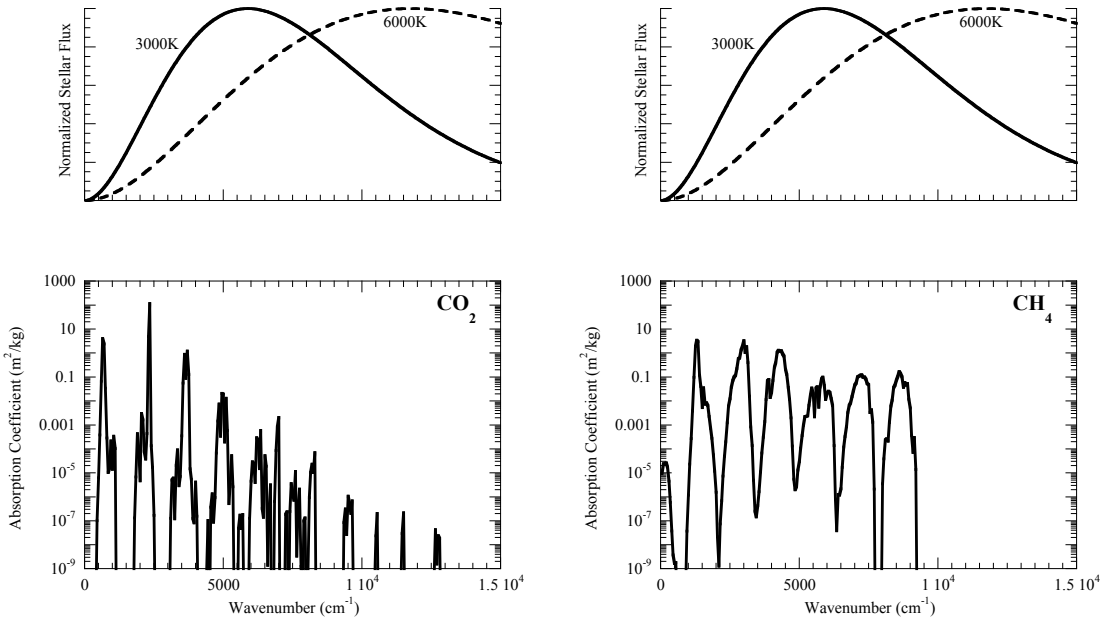

Figure 5.12: The lower panels show the median absorption coefficient in 50 cm^{-1} bins for CO_2 (left panel) and CH_4 (right panel). The absorption coefficients were computed with $T = 260K$ and $p = 100mb$. The 75^{th} percentile coefficients are typically about an order of magnitude above the median. The upper panels show the distribution of incoming stellar radiation for a typical cool, red M dwarf, and for a hotter yellow G-dwarf like the Sun.

CH_4 is a much more potent near-IR absorber than CO_2, and has fewer transparent window regions. This limits the potential of CH_4 as a greenhouse gas at high concentrations, since the anti-greenhouse effect arising from heating of the upper atmosphere partly offsets the surface warming due to thermal-IR opacity. At present Earth methane concentrations concentrations of around 2 $ppmv$, the adjusted path is only 0.06 kg/m^2, so given the typical magnitude of the absorption coefficients, the near-IR absorption can be neglected. However, if the concentration rises to 1000 $ppmv$, as it easly could in an anoxic atmosphere, CH_4 absorbs a considerable portion of the stellar flux below 10000 cm^{-1}. As before, the implications for climate are even more consequential for an M star than for a G star.

Water vapor has strong absorption features extending well into the visible range, though it is also fairly well supplied with relatively transparent window regions. There are three distinct archetypical planetary situations to be thought about with regard to water vapor. First, in Earth-like conditions, water vapor is a minor and condensible constituent, which is concentrated in the lower atmosphere. For example, in a saturated 100 mb thick near-ground layer at $300K$ there are 22 kg/m^2 of water vapor, yielding a pressure-adjusted path of about 200 kg/m^2. There are several absorption peaks below 6500 cm^{-1} that are effective for a path this large. Because water vapor in the Earthlike regime absorbs largely near the ground, it acts almost the same as reducing the ground albedo. However, if the ground would have absorbed the near-IR anyway, the net effect on climate would be minimal. Over a high albedo surface, the absorption would be more consequential. On a Snowball Earth with $250K$ near surface air temperatures, the same layer still has an adjusted path of almost 5 kg/m^2, and referring to Fig. 5.13 we see that there would still

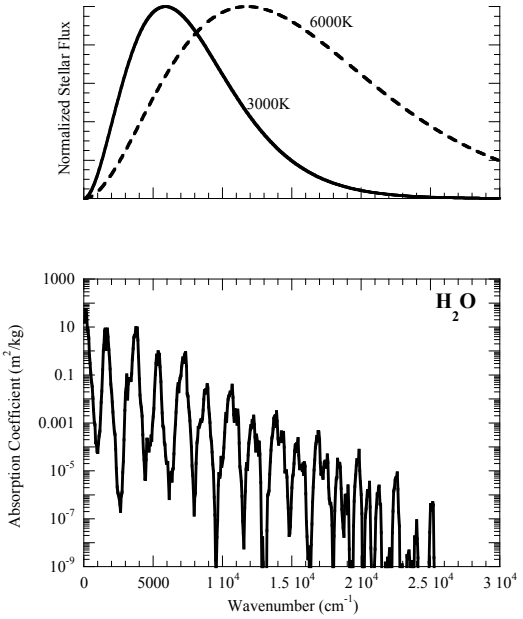

Figure 5.13: As for Fig. 5.12, but for water vapor. Note that the spectral range shown is twice that for CO_2 and CH_4, since water vapor absorbs strongly out to higher wavenumbers.

be considerable near-surface absorption.

The second regime to consider is Venus-like, in which water vapor is a noncondensible well-mixed trace gas. 20 $ppmv$ of water vapor in the atmosphere of Venus yields a pressure-adjusted path of over 3000 kg/m^2, which would yield strong absorption all the way out to 15000 cm^{-1}, with the exception of a few narrow window regions. Water vapor thus plays a very significant role in solar absorption on Venus, just as it does in the Venus greenhouse effect. By proportion, there is not much water in the atmosphere of Venus, but because the atmosphere is so massive the amount of water adds up to a considerable value, and its absorption is further strengthened by the high-pressure environment.

The final regime to consider is a runaway greenhouse steam atmosphere. The top bar of such an atmosphere under Earth gravity has a pressure-adjusted path of 50000 kg/m^2, and so nearly all of the star's near-IR output would be absorbed in that layer, considerably heating it and affecting the planet's energy balance. For an M star, this spectral region contains most of the star's output, and so the near-IR absorption has the potential to play a key role in the climate of a runaway greenhouse atmosphere for planets orbiting M stars.

The spectral overview we have just provided does not tell us very precisely how much incoming flux is absorbed and how the absorption is distributed in the vertical. For this, we need to compute the flux profile taking into account the full variability of the absorption coefficient. We shall do this by using the exponential sum method to compute the transmission function from the top of the atmosphere to level p in each of a set of bands covering the spectrum, and then summing up the transmission weighted by the incoming stellar flux in each band. This is in essence a simplified subset of the calculation done in the homebrew radiation code discussed in Chapter 4. In this calculation, the thermal emission from the atmosphere is negligible, so one only needs

to compute the transmission of radiation entering from the top. To focus on the most essential features, we'll also assume that all flux reaching the bottom boundary is absorbed there, so we need not consider the transmission of upward-reflected stellar radiation. In most circumstances, this is a minor influence, since the parts of the spectrum where the atmosphere absorbs well are mostly depleted by the time the ground is reached. As a further simplification, we'll assume the atmosphere to be transparent outside the spectral region covered in the spectral survey, and neglect Rayleigh scattering. Rayleigh scattering is fairly weak in the near-IR, but Rayleigh scattering of visible and ultraviolet light would keep some of the incoming radiation from being absorbed at the ground. Finally, the calculations are carried out with a fixed zenith angle having $\cos\zeta = \frac{1}{2}$, rather than averaging the zenith angle over day and season at some given latitude. The profiles were all computed subject to a net downward stellar radiation of 350 W/m^2 coming in at the top of the atmosphere, but as the problem is linear, the flux can be readily scaled to any other value. In each case, we did the calculation for an M star with photospheric temperature of $3000K$ and for a G star at $6000K$. The calculations were carried out with a gravitational acceleration of $10m/s^2$.

Results for a pure CO_2 atmosphere, a CO_2-air mixture, and a pure H_2O atmosphere are shown in Fig. 5.14. These were computed for an isothermal $260K$ atmosphere, but the transmission is not terribly sensitive to the temperature profile. When examining fluxes plotted against a logarithmic pressure axis, it is good to keep in mind that if F is the flux, then the heating rate per unit mass is proportional to the slope dF/dp, which is $p^{-1}dF/d\ln p$. Thus, a given slope in log coordinates corresponds to a greater heating rate at low pressures than it does at high pressures. For well-mixed greenhouse gases, one typically finds high heating rates aloft, since the upper atmosphere gets the first chance to absorb the part of the spectrum that is absorbed very strongly.

In all cases, the absorption for the M star case is substantially greater than that for the G star case, as expected. A 2 *bar* CO_2 atmosphere absorbs 30 W/m^2 of the incident 350 W/m^2 for the G star, but 100 W/m^2 for the M star. For both kinds of stars, about a third of the total flux is deposited at pressures lower than 100 mb, which will lead to intense heating in the upper atmosphere. Still, a great deal of the flux is absorbed in the lower atmosphere. Over a highly absorbing surface like an ocean or dark land, it matters little to the tropospheric temperature whether the energy is absorbed in the troposphere or at the surface. Over a highly reflective surface, as in a Snowball state, the effect of the lower atmospheric absorption is to decrease the effective albedo, which will lead to a warming of the troposphere. In any event, the vertical distribution of absorption for CO_2 does not suggest a pronounced anti-greenhouse effect.

The situation for a mixture of 20% CO_2 in air is quite similar, and in fact leads to only slightly less total absorption, because the additional CO_2 in the pure CO_2 case is only able to absorb parts of the spectrum where CO_2 is a relatively poor absorber. This is another instance of the typical logarithmic dependence of radiative properties on greenhouse gas concentrations. The general implication is that stellar absorption by CO_2 should have only a minor effect on tropospheric climate when the surface has low albedo. Over a high albedo surface, as in a snowball, we expect some modest tropospheric warming, which can help in deglaciation. For example, assuming a surface pressure of 1000 mb, the atmospheric absorption is 6 W/m^2 for a G star between the 400 mb level and the ground, or about twice as much for an M star. If the near-IR albedo of the surface is 50%, then half of the amount absorbed in the troposphere would have been absorbed at the surface anyway, so the additional radiative forcing due to tropospheric absorption is only half the stated values. Assuming a 60% surface albedo averaged over the entire solar spectrum, the surface solar absorption for a transparent atmosphere would be 140 W/m^2. Hence, the additional radiative forcing amounts to between 2.5% and 5% of the energy budget. This is not overwhelming, and would be partially offset by the cooling due to stratospheric absorption. Still,

it is a factor that should be taken into consideration in determining the conditions for deglaciating a Snowball state.

Now let's turn to the pure water vapor case for a layer extending to a pressure of 2 *bar*. This can be thought of as the top 2 *bar* of an atmosphere undergoing a runaway greenhouse, or alternately the whole atmosphere for a world with a substantial ocean having a surface temperature of around 395 K. Since the absorption coefficients in this example are computed with a fixed temperature of 280 K, the opacity of the lower part of the atmosphere has been underestimated, but the example nonetheless suffices to demonstrate just how powerfully water vapor absorbs in the near-IR. Even on a logarithmic plot the slope at 50 mb is greater than the slope at 500 mb, indicating an extremely intense upper level heating which should lead to pronounced warming of the upper atmosphere. Rather little stellar flux makes it to the 2 *bar* level – only 150 W/m^2 in the G star case and 40 W/m^2 in the M star case. Calculations with thicker atmospheres (not shown) reveal that the attenuation continues as the surface pressure is further increased. For example, in the M star case, only 15 W/m^2 reaches the ground for a 20 *bar* atmosphere and 5 W/m^2 for a 200 *bar* atmosphere. For a hot atmosphere, temperature scaling of line strengths would further reduce the penetration, though the precise effect takes us into unknown territory with regard to temperature scaling of the water vapor continua at high temperatures. Further, the pressure scaling of absorption coefficients in the exponential sum code probably underestimates the true influence of line broadening at very high pressures, so most probably the flux would be further reduced in a precise calculation. For a G star there would be less absorption, but this would be offset by greater Rayleigh scattering owing to the shorter wavelengths in the incident stellar radiation. A runaway greenhouse would be a very dark place at the surface, with hardly any radiation penetrating to the ground.

With regard to the climate implications of the limited flux penetrating the atmosphere, however, it should be kept in mind that this atmosphere is also very optically thick in the thermal infrared, so that even in the 2 *bar* M star case, the lower atmosphere would have to achieve a very high temperature in order to be able to lose 40 W/m^2 by radiative diffusion through the highly opaque atmosphere. If the tropopause rises so high as a result of this heating that it engulfs the stellar heating region aloft, the anti-greenhouse effect will be suppressed. Even in pure radiative equilibrium one must reckon with the fact that if H_2O is a good absorber of stellar near-IR, it is also a good emitter of thermal IR, and the outcome of the competition between these two factors is not easy to resolve *a priori*. The most that can be said at this point is that near-IR stellar absorption must be considered as a serious factor in steam atmospheres, all the more so in the M star case.

Now let's take a look at the Earthlike case of saturated water vapor mixed with air, shown in Fig 5.15. In this case, the temperature profile matters a great deal, since it determines the vertical distribution of water vapor. These calculations were done for an atmosphere on the moist adiabat. Results are shown for a tropical case with surface temperature of $300K$ and a Snowball case with a surface temperature of 250 K. In contrast to the case of a pure steam atmosphere, the absorption is trapped in the lower atmosphere for the Earthlike case. This adds significantly to the heating of the troposphere, but again, over a strongly absorbing surface the absorption is only acting to radiatively deposit energy directly in the troposphere, which otherwise would have been absorbed at the surface and communicated to the troposphere by convection. As we discussed for the CO_2 case, the absorption is more consequential if the surface is highly reflective. In this regard, it is important to note that the low level absorption is substantial even when the surface temperature is only $250K$. In both the G and M star cases, the absorption is markedly greater than the corresponding low level absorption in the CO_2 case. Since tropical temperatures during a Snowball can easily reach $250K$, solar absorption by water vapor can significantly assist

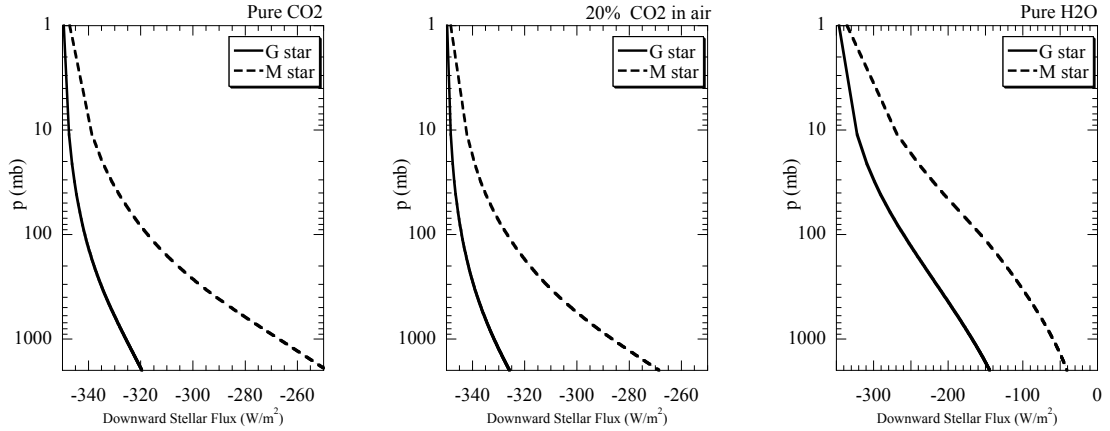

Figure 5.14: Profiles of incoming stellar flux computed using an exponential-sum transmission function for three different atmospheres: $2bar$ pure CO_2 (left), 20% molar CO_2 in Earth air (middle) and $2bar$ pure H_2O (right). The incoming vertical flux at the top of the atmosphere is 350 W/m^2 in all cases, and results are shown for both a G star and M star spectrum.

deglaciation, by lowering the effective surface albedo. Any process which warms the atmosphere will further increase the atmospheric water content and thus further increase the solar absorption. This constitutes a novel sort of water vapor feedback, which operates via the effect of water vapor on the solar spectrum instead of the thermal infrared effect.

For a complete appreciation of the effect of stellar absorption on the temperature profile, one must compute the radiative-convective equilibrium in the presence of absorption. We will discuss a few such calculations now, for the case of pure CO_2 atmospheres. In Section 4.8 we carried out thermal infrared radiative-convective solutions by fixing the ground temperature T_g and finding the atmosphere that was in equilibrium with the upwelling radiation from the ground. If one then wanted to know what T_g would be supported by a given amount of absorbed stellar radiation, it was only necessary to vary T_g until the desired OLR was achieved. This procedure will not do in the presence of atmospheric stellar absorption, since the incoming flux must be known in order to compute the temperature profile. Hence, in the calculations to follow, we adopt a somewhat different procedure, specifying the incoming stellar radiation, time-stepping the temperature profile as before, but this time adjusting T_g until the top-of-atmosphere energy balance is satisfied. Where the atmosphere is statically unstable relative to the ground temperature, the temperature profile is reset to the adiabat as before.

There are various ways to approach the problem of adjusting T_g. If one were interested in reproducing the actual time evolution of the system, it would be necessary to use the surface downwelling stellar radiation and thermal infrared in order to determing the surface temperature change; then, turbulent and radiative heat fluxes would heat the low lying air and if instability results, the heat would be mixed upward by convection. In our case, we are only interested in getting the equilibrium state, so any procedure that converges to the equilibrium will do. The following simple iteration works quite well in practice. The general idea is that if the net top-of-atmosphere radiation (incoming stellar minus OLR) is downward, then T_g needs to be increased in order to bring the atmosphere closer to balance; conversely, if the net is upwards, T_g needs to be

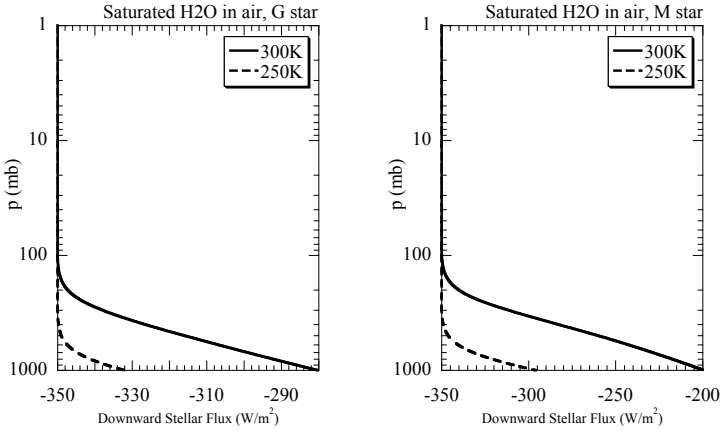

Figure 5.15: Profiles of incoming stellar flux computed using an exponential-sum transmission function for saturated water vapor in Earth air on the moist adiabat. The temperature labels indicate surface temperature. The left panel shows results for a G star spectrum, while the right shows M star results. These calculations take into account the effect of the temperature profile on water vapor, but do not incorporate the temperature-scaling of absorption coefficients.

increased. The stellar absorption is not very sensitive to temperature, so the change in T_g mainly affects the OLR. Thus, the main problem is to figure out the climate sensitivity, $dOLR/dT_g$. If the atmosphere is optically thick, then increasing T_g with the atmospheric temperature fixed would not change OLR, but the process we envision is that increasing T_g warms the rest of the atmosphere by convection and radiation, and this ultimately leads to an increase in OLR. The key simplification in the iteration is to estimate $dOLR/dT_g$ as if the atmosphere were a grey gas. Specifically, we compute the radiating temperature T_{rad} from $\sigma T_{rad}^4 = OLR$, since we know the OLR from the radiation calculation. Then, the climate sensitivity is estimated as $dOLR/dT_g \approx 4\sigma T_{rad}^4$. Adopting the convention that a net downward flux is negative, the iteration to be performed at each time step is

$$T_g \rightarrow T_g - \epsilon \frac{F_{toa}}{4\sigma T_{rad}^4} \tag{5.59}$$

where F_{toa} is the net top-of-atmosphere flux and ϵ is an under-relaxation factor that adjusts the ground temperature just part of the way towards the target, which improves the stability of the iteration. The following calculations were performed with $\epsilon = .05$, which provides a reasonable compromise between stability and rate of convergence. This iteration procedure is rather *ad hoc*, and no doubt there are more sophisticated schemes which rest on firmer ground. However, we have found it to serve quite well over a range of situations.

Figures 5.16 and 5.17 show radiative-convective equilibrium results for pure CO_2 atmospheres, carried out using this procedure. As in the flux profiles shown previously, the surface was assumed to be completely absorbing. The results in Fig. 5.16 are in equilibrium with 350 W/m^2 of incoming stellar radiation, on a planet having a surface gravity of $10m/s^2$. In the 2 *bar* G star case, the near-IR absorption moderately warms the stratosphere relative to the no-absorption case. The effect of absorption on surface temperature is hardly detectable in the figure. It amounts to a cooling of about $4K$. For the 2 bar M star case, the warming aloft is much more pronounced, and there is a significant lowering of the tropopause. In this case the surface cooling caused by

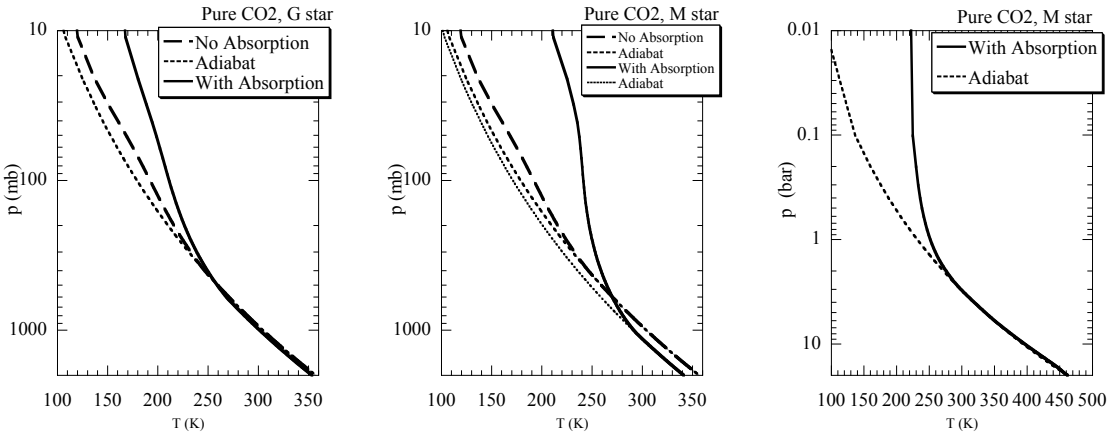

Figure 5.16: Radiative-convective equilibrium subject to an incoming stellar flux of 350 W/m^2 for pure CO_2 atmospheres, for a planet with 10 m/s^2 surface gravity. The first two panels have surface pressure of 2 *bar* and the rightmost panel has surface pressure of 20 *bar*. The surface is assumed completely absorbing. In the 2 *bar* cases, calculations without atmospheric stellar absorption are shown for comparison. The G star cases assume a blackbody spectrum of incoming radiation with a temperature of $6000K$, whereas the M star cases assume $3000K$.

absorption is 15 K, though given the high surface temperature this is hardly a very consequential effect. When the surface pressure is increased to 20 *bar* in the M star case, the surface temperature increases dramatically, but the stratospheric temperature changes little, with the main effect being a slight warming of the highest stratosphere, which leaves the stratosphere quite isothermal. The no-absorption case for the 20 *bar* atmosphere (not shown) very closely follows the adiabat and has a very high tropopause. The surface cooling caused by absorption in this case increases to 22 K, but this offsets little of the additional greenhouse warming which raises the surface temperature to 460 K.

Finally, to give an example of the situation for thin atmospheres, we show a calculation carried out in the Present Mars regime in Fig. 5.17. This calculation was done with Mars gravity, and subject to G star illumination. We see that, as in the dense atmosphere cases, the G stellar absorption causes only a moderate warming of the stratosphere. The observed stratospheric temperature is notably warmer than the calculation, which suggests that near-IR absorption alone is not able to fully account for the Martian stratospheric temperature seen in this sounding. Absorption due to dust is a likely culprit, but effects due to the global scale stratospheric circulation may be playing a role as well.

The summary situation for pure CO_2 atmospheres is that stellar near-IR absorption causes moderate stratospheric warming for planets about G stars case and more pronounced stratospheric warming for M stars, but in neither case does the stellar absorption lead to a stratospheric temperature inversion; the temperature is monotonically decreasing everywhere. The effect of absorption on the tropospheric temperature is a modest cooling. Thus, the anti-greenhouse effect is not terribly consequential for pure CO_2 atmospheres. We'll note also that all of the stratospheres in the above results are optically thin in comparison with the tropospheres, which implies the happy result that reasonable estimates of surface temperature can be made on the basis of the simple and

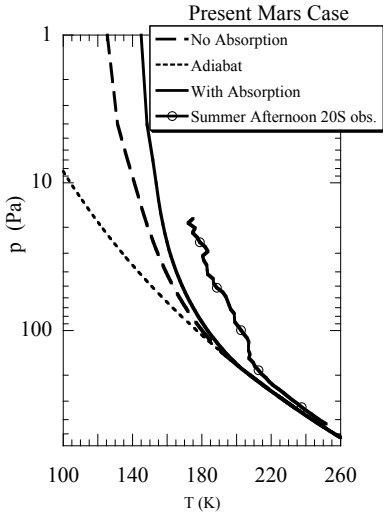

Figure 5.17: Radiative-convective equilibrium for Present Mars. The incoming Solar flux is 250 W/m^2 and has been chosen to yield a tropospheric temperature similar to the observation shown in the figure. The observation shown for comparison is from a summer afternoon tropical sounding, from the Mars Global Surveyor radio occultation dataset.

swift all-troposphere OLR model.

The radiative-convective behavior of thick water vapor atmospheres presents considerably more challenge because of the extreme optical thickness of water vapor both in the near-IR and thermal-IR. This is a very rich problem which entails consideration of the delicate balance between very slow radiative cooling and the small amount of stellar radiation reaching the surface, as well as the effects of condensation on the adiabat and the increase of surface pressure with temperature for an atmosphere in equilibrium with an ocean. For the hot steam atmospheres which are of greatest interest, one also comes up against the largely unknown behavior of the water vapor continua at high temperatures and pressures. We will be content to leave this deep and interesting problem as a subject for research. It is an important problem because the stellar absorption has the potential to significantly increase the threshold illumination needed to trigger a runaway greenhouse. The reader is now in posession of all the tools necessary to carry out an inquiry of this sort.

5.10.2 Ultraviolet absorption

Because of its importance to near-surface life on Earth, ozone (O_3) is probably the most familiar of all ultraviolet-absorbing gases. The interest stems from the fact that the shorter wave and more energetic forms of UV radiation wreak havoc with key biological molecules of life as we know it, and in particular genetic information encoded in DNA. It is a highly Earth-centric view to think that an ozone shield is necessary to protect complex life in general from deady UV radiation, but notwithstanding that issue, ozone has some very profound effects on the stratospheric temperature structure that play a key role in the prospects for detecting O_2 (and presumably oxygenic photosynthetic life) on extrasolar planets.

UV wavelengths are customarily measured in *nanometers* (nm, or $10^{-9}m$). The radiation

begins to become harmful to Earth life at 320 nm, and wavelengths shorter than 300 nm cause extreme damage. Ozone plays a distinguished role in shielding Earth life from UVB (320-280 nm) and UVC (280-100 nm) radiation. Shorter UV wavelengths, to say nothing of Solar X-rays, are even more deadly, but there are many molecules which efficiently absorb wavelengths shorter than 100 nm. For example, CO_2, which is far more abundant than O_3 in Earth's atmosphere, is every bit as absorbent as O_3 in the vicinity of 140 nm. In contrast O_3 is a potent absorber between 200 and 300 nm, whereas CO_2 and most other reasonably abundant atmospheric gases are nearly transparent there. Ozone also absorbs significantly in the 400 to 700 nm range. These wavelengths are not particularly damaging to life, but because the stellar output is abundant in this range for G-class and hotter stars, the effect on atmospheric heating is significant.

Ozone is a feature of atmospheres rich in free O_2, bombarded by UV radiation. So far, Earth's is the only known example of such an atmosphere. Ozone is a highly reactive substance with a short lifetime. Therefore, it is very inhomogeneous. In particular, in Earth's present atmosphere ozone is concentrated in a stratospheric layer near the altitude where its production rate is strongest. At earlier times, when there was less O_2 around, the ozone layer was probably found at a lower altitude. At present, maximum ozone values occur at about the 20 mb level in the tropics, and reach values on the order of 2 $ppmv$. The concentration drops by two orders of magnitude as the tropopause is approached. Even at such low concentrations, ozone is a very effective absorber. A concentration of $1ppmv$ in the layer of atmosphere above 20 mb gives this layer an optical thickness exceeding 4.0 at 250 nm, which is sufficient to exhaust virtually all of the UV flux at that wavelength. Detailed data on the UV absorption spectrum of O_3 and other gases can be found in the resources listed in the Further Readings section.

Heating due to UV absorption by ozone has an important effect on the stratospheric temperature profiel, but ozone is also a very powerful absorber in the thermal infrared range. It is the only abundant infrared absorber which is concentrated in the stratosphere, and that that sense provides a counterpoint to water vapor, which is concentrated in the troposphere.

We will now carry out some calculations with the `ccm` radiation model which illustrate the key effects of ozone in an Earthlike setting. We adopt an idealized ozone profile of the form

$$\eta_{o3} = \eta_m \exp[-(p - 20mb)^2/(50mb)^2] \tag{5.60}$$

where η_{o3} is the molar mixing ratio of ozone and η_m is the peak value, which we take to be 2 $ppmv$ in the following calculations. The left panel of Fig. 5.18 shows the net downward flux for an atmosphere that contains Earth air and ozone, but no other gases. The atmosphere is illuminated with 400 W/m^2 of incoming radiation, 25 W/m^2 of which is reflected back by Rayleigh scattering. The ground is assumed to be perfectly absorbing. The temperature profile is immaterial, since UV absorption is nearly independent of temperature for typical planetary temperatures. Of the sunlight that is not scattered, about 15 W/m^2 is absorbed within the ozone layer. This includes nearly all of the harmful UVB and UVC radiation. Given the low mass of this region of the atmosphere, the absorption gives rise to a considerable heating, which should substantially warm the stratosphere. There is some weak absorption within the troposphere, which is due to O_2.

A key question is whether ozone causes the stratospheric temperature to increase with height in the stratosphere, as is seen in observations of the Earth's atmosphere. The right panel of Fig. 5.18 shows radiative-convective equilibrium calculations using the `ccm` code, computed with a dry atmosphere containing 300 $ppmv$ of CO_2, in which convectively unstable layers are adjusted to the dry air adiabat. The profiles shown are in equilibrium with 400 W/m^2 of incident solar radiation. Results with ozone are compared with a control case for a solar-transparent CO_2/air mixture. In the control case, the radiative-convective equilibrium temperature decreases monotonically with height, just as seen in the simulations of Chapter 4.

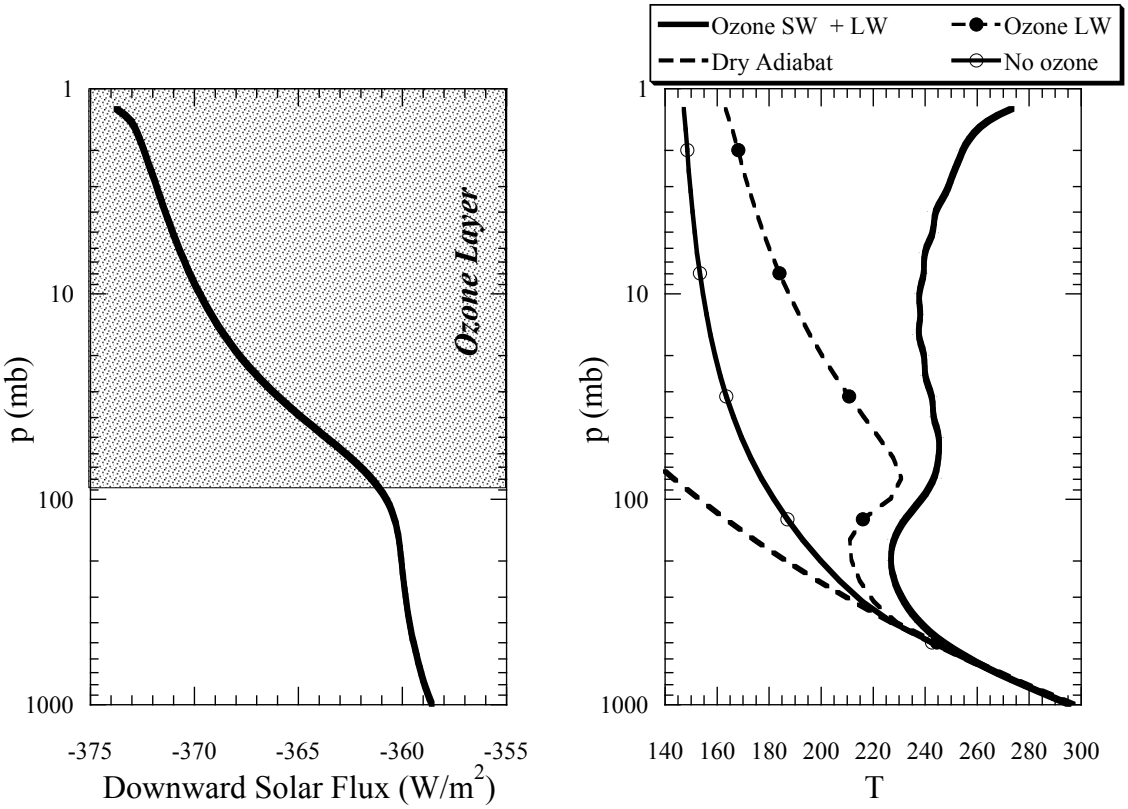

Figure 5.18: Left panel: Net downward solar flux for an atmosphere with the ozone profile described in the text. The calculation was performed for an isothermal $300K$ atmosphere, but the results are essentially insensitive to temperature. The incoming solar radiation is $400W/m^2$, some of which is scattered back by Rayleigh scattering. Right panel: Radiative-convective equilibrium for a dry atmosphere containing 300 *ppmv* of CO_2, computed for three cases as follows. Thin solid curve with open circles – no ozone or solar absorption; Dashed curve with filled circles – ozone thermal infrared effects incorporated; Thick solid curve – ozone infrared and solar absorption incorporated. The plain dashed curve gives the adiabat for the third case. The ground temperatures differ slightly between the cases, but the difference is not visible in the figure. All calculations were performed with the `ccm` radiation model.

Because of the dual role of ozone as an infrared and UV absorber, its effect on the temperature profile is complex. In the circumstances of this particular simulation, introducing just the effect of ozone on infrared absorption introduces a temperature increase with height in the lower stratosphere. This arises because the ozone is concentrated aloft, and absorbs at wavelengths that escape the CO_2 effects in the troposphere. This leads to an intense heating layer, which must warm up until it comes into equilibrium. Without solar absorption by ozone, however, the upper stratospheric temperature still declines sharply with height. Introducing the solar absorption warms the upper stratosphere considerably, and causes it to increase with height. It also results in a pronounced lowering of the tropopause.

The effect of ozone on stratospheric temperature is profound, but its effect on surface temperature is modest and largely invisible in Fig. 5.18. For the control case, the surface temperature is $295.48K$. It rises to $297.8K$ when ozone infrared effects are introduced, owing to the greenhouse effect of ozone. However, when the ozone solar absorption is brought into the picture, the warming of the stratosphere allows the upper atmosphere to radiate better to space, and this brings the surface temperature back down to $295.22K$. Thus, the main climatic effects of ozone are in the stratosphere, though it is quite possible that the lowering of the tropopause would have repercussions for tropospheric climate. Further, the absence of ozone in the anoxic Early Earth atmosphere would have led to a much colder stratosphere, which is important to take into account in working out the chemistry of Titan-like stratospheric haze clouds.

Water vapor, CO_2 and CH_4 all absorb ultraviolet quite strongly for wavelengths shorter than 180 nm, but the only common atmospheric constituent that competes with ozone at longer wavelengths is SO_2. This gas is abundant in volcanic outgassing, but in oxygenated atmospheres it forms sulfates which are removed by rainout if the planet supports liquid water. On dry planets or planets without oxygen, SO_2 can build up to higher concentrations, but its status as an ultraviolet shield still hinges on atmospheric chemistry. The very fact that it absorbs ultraviolet so well tends to dissociate the molecule, and the question then is whether there are chemical pathways that can restore it. There is no question, however, that SO_2 is a molecule that holds very interesting prospects as a mediator of planetary climate evolution, especially in view of the fact that it is also a potent greenhouse gas.

5.11 Albedo of snow and ice

To emphasize the generality and power of the physics of scattering that has been the central theme of this chapter, we will remark in closing that the very same physical principles account for the high albedo of snow and ice. The point of commonality with scattering from cloud droplets is that any discontinuity in index of refraction will lead to scattering. In the case of snow, the discontinuity is between the crystals of particles and the voids between them. Given how much denser most solids are than gases, it matters little whether the voids are filled with air as on Earth, or filled with near-vacuum as they would be on Europa. Similarly, as long as the snow is made of a mostly transparent solid, it matters little just what it is made of. There are variations in index of refraction amongst different ices, but all are significantly different from unity. All snow is highly reflective, whether it be N_2 snow on Triton or CO_2 snow on Mars. For ice, the scatterers are air bubbles or brine pockets, and here it matters a bit more what the composition of the freezing fluid may be. To make gas bubbles in the frozen liquid, there must be a significant amount of some gas dissolved in the fluid, and to make brine pockets there must be some suitable solute (salt in the Earth case). The rate of freezing also makes a difference, to the albedo of ice, since slow freezing allows gas to be rejected before bubbles form, leading to clear, low-albedo ice. Things get even more interesting

if one allows for an admixture of absorbing particles (dust or soot) with the snow or ice.

Since albedo has such an important effect on planetary radiation budgets, the physics of snow and ice albedo is a critical field of play for radiative transfer. It can be treated using essentially the same techniques that have been introduced in this chapter.

5.12 For Further Reading

For general background on scattering of electromagnetic radiation, including basic concepts of refraction and diffraction, the reader is referred to

- Jackson JD 1998: *Classical Electrodynamics*. Wiley

A complete discussion of Rayleigh scattering, including the derivation of polarizability factors, can be found in Section 7.3 of

- Goody RM and Yung Y 1995: *Atmospheric Radiation*, Oxford University Press.

The derivation of the infinite series describing Mie scattering is very intricate and involves considerable facility with manipulation of special functions and vector spherical harmonics. It is a masterful solution, but the effort required to fully understand the derivation is not commensurate with the rather limited number of additional problems one can solve using the techniques. A reasonable compromise is to read through the much simpler case of scattering from a conducting sphere, which involves many of the same contexts in a less challenging setting. This calculation is given in *Classical Electrodynamics*, referenced above. For the truly devoted, or at least those looking for a usable description of the series solution, the full Mie scattering calculation is worked out in Chapter 4 of

- Bohrens CF and Huffman DR 2004: *Absorption and scattering of light by small particles*. Wiley-VCH . (available online through Wiley Interscience)

Appendix A of this book gives the full text of a computer program for evaluating the solution. Many implementations of this algorithm can be found on the web by searching for `bhmie`. Chapters 5 and 6 have a useful discussin of Rayleigh scattering, which can be consulted as an alternative to Goody and Yung.

For optical properties of condensed substances see

- **Liquid water, real refractive index** – Schiebener P *et al.* 1990: Refractive-index of water and steam as function of wavelength, temperature and density *J. Phys. Chem. Ref. Data* **19** 677-717.
- **Liquid water, absorption properties** – Hale GH and Querry MR 1973: Optical constants of water in the 200nm to 200micron wavelength region. *Appl. Opt.* **12** 555–563.
- **Water ice** – Warren SG 1984: Optical constants of ice from the ultraviolet to the microwave *Appl. Opt.*, **23**, 1026–1225.
- CO_2 **ice** – Warren, SG 1986: Optical constants of carbon dioxide ice, *Appl. Opt.* **25**, 26502674.

- CO_2 **ice (updates Warren** – Hansen GB 2005: Ultraviolet to near-infrared absorption spectrum of carbon dioxide ice from 0.174 to 1.8μm. *J. Geophys. Res* **110**, doi:10.1029/2005JE002531.

- H_2SO_4 – Myhre, CE *et al.* 2003: Spectroscopic Study of Aqueous H2SO4 at Different Temperatures and Compositions: Variations in Dissociation and Optical Properties. *J. Phys. Chem. A*, **107**, 1979-1991.
  Data via `http://www.kjemi.uio.no/09_spekt/Atmosfaere/OPA/OPA.html`

- H_2SO_4**,esp. on Venus** – Palmer KF and Williams D 1975: Optical constants of sulfuric acid; Application ot the clous of Venus? *Appl. Opt.* **14**,208-219.

- **Liquid** CH_4 – Martonchik JV and Orton GS 1994: Optical constants of liquid and solid methane. *Appl. Opt.* **33**, 8306-8317.

Many of these sources provide functional fits to the data that are convenient to use in computations. Most also give data in tabulated form, which can in some cases be found in digital form through various web sites or from the authors. A selection of the data has been provided in the Workbook datasets online supplement for this chapter.

The importance of the scattering greenhouse effect for Early Mars climate was first discussed in

- Forget, F and Pierrehumbert RT 1997: Warming Early Mars with carbon dioxide clouds that scatter infrared radiation. *Science* **278** , 1273 - 1276.

The effect of sulfuric acid cloud cycles on Venus climate evolution has been discussed in

- Bullock MA and Grinspoon DH 2001: The recent evolution of climate on Venus. *Icarus* **150**, 19-37.

The treatment of the radiative effect of removing the clouds of Venus is highly simplified in this work, and there is much room for further study.

The *MPI-Mainz-UV-VIS Spectral Atlas of Gaseous Molecules* contains a comprehensive database of UV and visible absorption cross sections. Data and documentation can be downloaded from the site

- `www.atmosphere.mpg.de/spectral-atlas-mainz`

Some aspects of solar absorption relevant to gas giants are discussed in Radiative transfer on gas giants is discussed in

- Guillemot T, *et al* 1994: Are the Giant Planets Fully Convective?, *Icarus*, **112**, pp 337-353.

Chapter 6

The Surface Energy Balance

6.1 Overview

This results of this chapter are pertinent to a planet with a distinct surface, which may defined as an interface across which the density increases substantially and discontinously. The typical interface would be between a gaseous atmosphere and a solid or liquid surface. In the Solar system, there are only four examples of bodies having both a distinct surface and a thick enough atmosphere to significantly affect the surface temperature. These are Venus, Earth, Titan and Mars; among these, the present Martian atmosphere is so thin that it only marginally affects the surface temperature, though this situation was probably different early in the planet's history when the atmosphere may have been thicker. Although thin atmospheres have little effect on the surface temperature, the atmosphere itself can still have interesting behavior, and the flux of energy from the surface to the atmosphere provides a crucial part of the forcing which drives the atmospheric circulation. This is the case for example, for the thin Nitrogen atmosphere of Neptune's moon Triton. Apart from the examples we know, it is worth thinking of the surface balance in general terms, because of the light it sheds on the possible nature of the climates of extrasolar planets already detected or awaiting discovery.

The exchange of energy between the surface and the overlying atmosphere determines the surface temperature relative to the air temperature. It also turns out that it determines the exchange of mass between the surface and the atmosphere (as in sublimation from a glacier or evaporation from an ocean, lake or swamp). Because outer space is essentially a vacuum, the only energy exchange terms at the top of the atmosphere are radiative. At the surface, energy can be exchanged by means of fluid motions as well as by radiation.

The atmospheric gas in direct contact with the surface must have the same velocity as the surface; because the surface material is so much denser (and in the case of a solid so much more rigid) than the atmosphere, the atmospheric flow must typically adjust to the presence of the surface over a rather short distance. The resulting strong shears lead to random-seeming complex turbulent motions sustained by the kinetic energy of the shear flow near the boundary. We may subdivide the atmosphere into the *free atmosphere* – which is sufficiently far above the surface to be little affected by turbulence stirred up at the surface, and the *planetary boundary layer* (*PBL*, for short) where the transfer of heat, chemical substances, and momentum is strongly affected by surface-driven turbulence. We may further identify the *surface layer*, which is the thin portion of the PBL near the ground within which all the vertical fluxes may be considered independent of

height.

Given that the whole troposphere is created by convection – which is a form of buoyancy-driven turbulence – it is not at once clear why the PBL should exist as a distinct entity from the troposphere in general. The main reason one can typically distinguish the PBL is that mechanically driven turbulence is more trapped near the surface than is buoyancy driven turbulence, and also has distinct time and space scales. On the present Earth, the effect of moisture is also important in maintaining the distinction, since moisture gives deep convection an intermittent character: most of the troposphere-forming mixing takes place in rare convective events, while most of the troposphere remains quiescent most of the time. Because dry (i.e. noncondensing) convection is typically shallower than moist convection, in planets which have both forms the dry convection can often be treated as part of the boundary layer. This is the case for Earth, and likely for other planets with a surface and an atmosphere in which latent heat release is important (Titan and perhaps Early Mars being the only other known examples so far). For planets like Present Mars or Venus, where dry convection is the *only* form of convection, it is less clear that the PBL can be productively distinguished from the troposphere in general. Even in such cases, though, one can identify a constant-flux surface layer; the depth of the surface layer typically range from a few meters to a few tens of meters.

As in previous chapters, we let T_g be the temperature of the planet's surface. Previously, we used T_{sa} to denote the temperature of the air in immediate contact with the ground, but now we modify the definition somewhat, and allow T_{sa} to be the temperature at the top of the surface layer, assuming the air at the bottom of the surface layer (which is in contact with the ground) has the same temperature as the ground itself. A model of the PBL is necessary to connect T_{sa} to the temperature of the lowest part of the free troposphere. For many purposes, we can dispense with the PBL and patch the surface layer directly to the free troposphere. We shall adhere to this expedient in most of the following discussion.

Now let's discuss, in general terms, how the surface budget affects the climate. The state of the atmosphere and the ground must adjust so that the top-of-atmosphere and surface budgets are simultaneously satisfied. If the atmosphere is optically thick in the longwave spectrum, the top-of-atmosphere budget becomes decoupled from the surface budget, since radiation from the ground and lower portions of the atmosphere is absorbed before it escapes to space. In this case, the determination of T_g can be decomposed into two stages carried out in sequence. First one determines T_{sa} by adjusting this temperature until the top of atmosphere balance is satisfied, assuming that the rest of the troposphere is related to T_{sa} through the appropriate dry or moist adiabat. Then, once T_{sa} is known, one makes use of a model of the surface flux terms to determine the value of T_g which balances the surface budget with T_{sa} fixed at the previously determined value. This can be done without reference to the top-of-atmosphere budget, since the OLR is independent of T_g in the optically thick limit.

If the atmosphere is very optically thin in the longwave spectrum, the OLR is determined entirely by the ground temperature and ground emissivity. Further, since an optically thin atmosphere radiates very little, the only way the atmosphere itself loses energy is through turbulent exchange with the surface. Suppose first that the atmosphere is transparent to solar radiation. In that case, *in equilibrium* the net turbulent exchange between atmosphere and surface must vanish, since otherwise the atmospheric temperature would rise or fall, there being nothing to balance a net exchange. In consequence, the ground temperature will be just what it would have been without an atmosphere despite the presence of turbulence. In this case, one determines the ground temperature as if the planet were in a vacuum, the top of atmosphere budget is automatically satisfied, and then, once T_g is known, the surface budget is used to determine T_{sa}, and (via an adiabat) the rest of the atmospheric structure. It is exactly the inverse of the process used in

the optically thick case. In fact, the basic picture is little altered even if the atmosphere absorbs solar radiation. In that case, the requirement that the atmosphere be in equilibrium implies that any solar radiation absorbed in the atmosphere be passed on to the surface by turbulent fluxes. The result is much the same as if the solar radiation were absorbed directly by the surface; one does the ground temperature calculation as before, but simply remembers to add the atmospheric absorption to the solar energy directly absorbed by the ground. It should be kept in mind that these considerations apply only in equilibrium. Even an optically thin atmosphere can affect the transient behavior of the surface (e.g. in the diurnal or seasonal cycle), as will be discussed in Chapter 7.

In the intermediate case, where the atmosphere is neither optically thick nor thin, one must solve for T_{sa} and T_g simultaneously, so as to find the values that satisfy both the top-of-atmosphere and surface energy budgets. We'll do this crudely in the present chapter through the introduction of atmospheric transparency factors. Generally speaking, though, when the atmosphere is not too optically thin, the surface budget will have some effect on the temperature of the ground. For Earth this temperature is of interest because the ground is where people live and where much of the biosphere resides as well; for a broad range of planets actual or hypothetical the ground temperature also affects chemical processes which determine atmospheric composition, as well as the melting of ices at the surface. We shall see, however, that it is a fairly common circumstance that the surface fluxes effectively constrain the ground temperature to be nearly equal to the overlying air temperature, so that the climate can be determined without detailed reference to how the surface balance works out.

6.2 Radiative exchange

6.2.1 Shortwave radiation

The surface receives radiant energy in the form of shortwave (solar) and longwave (thermal infrared) flux. The shortwave flux incident on the surface is equal to the shortwave flux incident at the top of the atmosphere, diminished by whatever proportion is absorbed in the atmosphere or scattered back to space. We will call the shortwave flux incident on the ground S_g. The shortwave flux absorbed at the surface is then $(1-\alpha_g)S_g$, where α_g is the albedo of the ground. S_g is affected by clouds, atmospheric absorption and atmospheric Rayleigh scattering.

6.2.2 The behavior of the longwave back-radiation

The longwave radiation striking the surface is the infrared *back radiation* emitted by the atmosphere, which was discussed in Chapter 4. The back radiation depends on both the greenhouse gas content of the atmosphere – which determines its emissivity – and the temperature profile. When the atmosphere is optically thick in the infrared, most of the back radiation comes from the portions of the atmosphere near the ground, whereas in an optically thinner atmosphere the back radiation comes from higher – and generally colder – parts of the atmosphere, and is correspondingly weaker. If the atmosphere is very optically thin, the back radiation will be weak regardless of the atmospheric temperature profile, simply because an optically thin atmosphere radiates very little. As in Chapter 4, $I_{-,s}$ will denote the back radiation integrated over all longwave frequencies. The absorbed infrared flux is then $e_g I_{-,s}$, where e_g is the longwave emissivity of the ground. The ground loses energy by upward radiation at a rate $e_g \sigma T_g^4$. Thus, the net infrared cooling of the

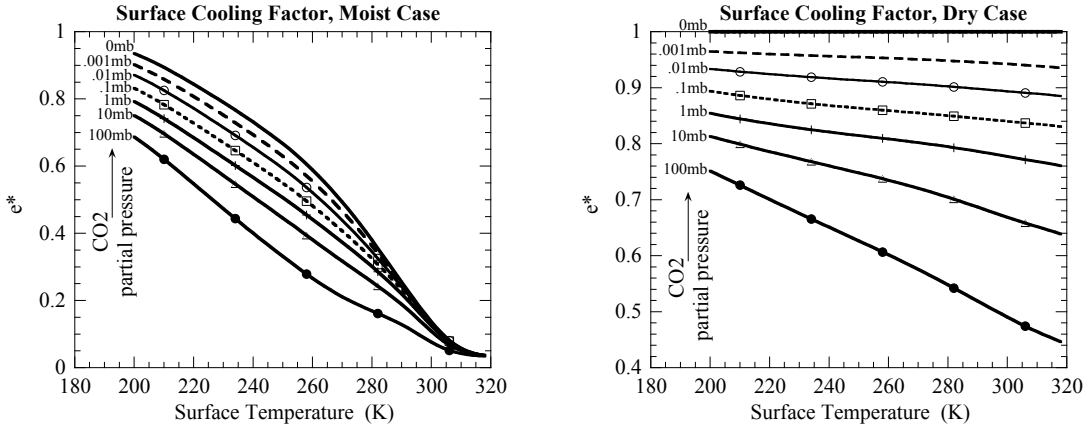

Figure 6.1: Surface cooling factor e^* for a $1bar$ Nitrogen-Oxygen atmosphere with water vapor and CO_2. The surface gravity is that for Earth. In the left panel, the calculations were done with free tropospheric relative humidity set to 50%, and low-level relative humidity set to 80%. Results in the right panel are for a dry atmosphere (zero relative humidity, but with the temperature profile kept the same as in the moist case). In both cases, the numbers on the curves indicate the partial pressure of CO_2 in mb.

ground is

$$F_{g,ir} = e_g \cdot (\sigma T_g^4 - I_{-,s}) \tag{6.1}$$

According to Eq. 4.21, $I_{-,s}$ approaches σT_{sa}^4 when the atmosphere is optically thick throughout the infrared. In order to characterize the optical thickness of the atmosphere, we introduce the effective low level atmospheric emissivity e_a, defined so that $I_{-,s} = e_a \sigma T_{sa}^4$. e_a depends on the temperature profile as well as the optical thickness, as illustrated by Eq. 4.21 in the optically thick limit. When $T_g = T_{sa}$ the surface cooling becomes $e_g \cdot (1 - e_a)\sigma T_g^4$, which vanishes in the optically thick limit where $e_a \to 1$. Let $e^* = (1 - e_a)$; this is the effective emissivity of the ground when the air temperature equals the ground temperature. If the air temperature is not too different from the ground temperature, we may linearize the term σT_g^4 about $T_g = T_{sa}$, which results in

$$F_{g,ir} = e_g \cdot e^* \sigma T_{sa}^4 + (4\sigma T_g^3 e_g)(T_g - T_{sa}) \tag{6.2}$$

From this equation we can define the infrared coupling coefficient, $b_{ir} = 4\sigma T_g^3 e_g$. When b_{ir} is large, a small temperature difference leads to a large radiative imbalance, and it is correspondingly hard for the ground temperature to differ much from the overlying air temperature. Later, we will derive analogous coupling coefficients for the turbulent transfers.

Figure 6.1 shows how e^* varies with temperature for an Earthlike atmosphere in which the only greenhouse gases are water vapor and CO_2, with the water vapor relative humidity held fixed as temperature is changed. In the moist case (left panel), e^* rapidly approaches zero as the temperature increases; this is because of the increasing optical thickness caused by the increase of water vapor content with temperature (owing to the fixed *relative* humidity). Increasing the CO_2 content also increases the optical thickness, correspondingly reducing e^*. At low temperatures, the CO_2 effect dominates, because there is little water in the atmosphere. However, by the time Earthlike tropical temperatures ($300K$) are reached, water vapor is sufficient to make e^* essentially zero all on its own without any help from CO_2. To underscore the relative role of CO_2 and water

vapor, results for a dry atmosphere are given in the right hand panel of Figure 6.1. e^* still goes down with temperature, because temperature affects the opacity of CO_2; however the decline is much less pronounced than it is in the moist case. Even with $100mb$ of CO_2 in the atmosphere, e^* falls only to about .4 at $320K$, and significant infrared cooling of the surface is possible. In sum, CO_2 by itself is relatively ineffective at limiting surface cooling, but the opacity of water vapor can practically eliminate surface infrared cooling at temperatures above $300K$, unless the ground temperature significantly exceeds the air temperature.

Though the results of Fig. 6.1 were were computed for Earth conditions, they give a fair indication of the extent of surface radiative cooling on other planets whose atmospheres consist of an infrared-transparent background gas mixed with CO_2 and with water vapor fed through exchange with a condensed reservoir. Through the hydrostatic relation, the surface gravity g affects the mass of greenhouse gas represented by a given partial pressure; the lower the g the greater the mass (and hence the greater the optical thickness), and conversely. This is especially important in the case of water vapor, since in that case the partial pressure is set by temperature, through the Clausius-Clapeyron relation. Thus, for a "large Earth" with high g, it takes a higher temperature to make the lower atmosphere optically thick. For example, calculations of the sort used to make Figure 6.1 show that with $1mb$ of CO_2 in a moist atmosphere having temperature $280K$, increasing g to $100m/s^2$ increases e^* to .507 (vs .303 for $g = 10m/s^2$). In the same atmospheric conditions, e^* falls to .102 for a "mini-Earth" with $g = 1m/s^2$. Increasing the pressure of the transparent background gas makes the greenhouse gases more optically thick through pressure broadening. With $g = 10m/s^2$, increasing the background air pressure to $10bar$ has a very profound effect, lowering e^* to .094. Reducing the air pressure below $1000mb$ should in principle increase e^*, but in fact it is found to very slightly reduce it, to .299. It appears that the reduction in opacity from less pressure broadening is offset by the changes in the moist adiabat that occur when the air pressure is reduced: the latent heat of condensation is spread over less background gas, so the temperature aloft is greater and hence the air aloft contains more water.

Without water vapor, it takes an enormous amount of CO_2 to make the lower atmosphere optically thick. This case is relevant to Venus and Venus-like planets, which may be defined to be planets having a dry rocky surface and a thick, dry CO_2 atmosphere. The near surface radiative properties can be determined using the homebrew exponential-sum radiation code; for better accuracy, we did this based on exponential-sum tables computed for the surface temperature and pressure conditions under consideration, so as to minimize errors due to pressure and temperature scaling of absorption coefficients. Line parameters in the `HITRAN` database were used to compute the absorption coefficients. For a 1 *bar* pure CO_2 atmosphere on a planet with the gravity of Venus, this calculation yields $e^* = .43$ when the surface temperature is $300K$, falling further to under 10^{-6} for pressures of 10 *bar* or more. The sharp decline in surface cooling between 1 *bar* and 10 *bar* arises from line-broadening, which fills in the window regions in the CO_2 absorption spectrum. At the $727K$ surface temperature of Venus, the surface emission shifts toward higher wavenumbers where CO_2 doesn't absorb as well, but the high temperature increases the line strengths while the high pressure causes the absorption to further spill over into the windows. Hence, at 92 *bar* and $727K$ the calculation still yields a value of e^* that is under 10^{-5}, even assuming CO_2 to be completely transparent for wavenumbers higher than 10000 cm^{-1}. The estimates of e^* at high pressure should be viewed with some caution, however. At high pressures, the contribution of each line to spectral distances far removed from the line center is considerable, and there is much uncertainty about the appropriate form of line shape to be used in computing this far-field contribution. We'll see shortly that it actually makes some difference to the climate of Venus whether e^* is zero or 0.05 .

In the opposite extreme, atmospheres like the thin Martian atmosphere have very little effect on the surface radiative cooling. For a Martian CO_2 atmosphere on the dry adiabat with $7mb$ of

surface pressure, $e^* = .9$ at $220K$, falling only modestly to .86 at $280K$. Recall that, per square meter of surface, Mars actually has vastly more CO_2 in it's atmosphere than the Earth has at present; allowing for the difference in gravity, a $7mb$ pure CO_2 atmosphere on Mars has as much CO_2 per unit area as an Earth atmosphere with a CO_2 partial pressure of $18.5mb$ at the ground. In comparison, the present Earth's atmosphere has a partial CO_2 pressure of a mere $.38mb$ (in 2006). The weak emission of the Martian atmosphere is due to the low total pressure, which yields little collisional broadening of the emission lines. If the same amount of CO_2 on Mars at present were mixed into a $1bar$ atmosphere of N_2, the effective surface emissivity e^* falls to .75 at $230K$ and .69 at $280K$.

Among common greenhouse gases, water vapor appears unique in its ability to make the lower atmosphere nearly opaque to infrared, even at concentrations as low as a few percent.

Clouds made of an infrared-absorbing substance such as water act just like a very effective greenhouse gas in making the lower atmosphere optically thick (making e_a close to unity). It takes very little cloud water to make the lower atmosphere act essentially like a blackbody. Infrared-scattering clouds in the surface layer, like those made of methane or CO_2, have a very different effect on the back-radiation. First, they shield the surface from back-radiation coming down from the upper atmosphere by reflecting it, rather than absorbing it; hence the shielding is accomplished without the cloud layer heating up in response to absorption. More importantly, the downwelling radiation from a reflective cloud is determined by the upwelling ground radiation incident upon it; the resulting back radiation is then determined by the ground temperature, and is independent of the cloud temperature. As a result, the surface cannot increase its longwave cooling by warming up until it is substantially warmer than the atmosphere. This gives a scattering cloud great potency to increase the ground temperature, if it allows sufficient solar radiation to get through to the ground. Either IR-reflecting or absorbing clouds are different from a greenhouse gas, in that they also strongly increase the shortwave albedo.

6.2.3 Radiatively driven ground-air temperature difference

Now we consider the equilibrium temperature difference between the ground and the overlying air that would be attained in the absence of turbulent heat exchange. This temperature difference is important in determining the extent to which convection is driven from below, by positive buoyancy generated near the ground. We have already discussed this issue for the case in which the atmosphere itself is in pure radiative equilibrium (See 3.6,4.3.4 and 4.7). Our concern now is with what happens once convection has set in and altered the atmospheric temperature profile.

If the only heat exchange is radiative, the surface budget reads

$$(1 - \alpha_g)S_g + \sigma e_a e_g T_{sa}^4 = \sigma e_g T_g^4 \tag{6.3}$$

Since the second term on the left hand side is positive, the infrared back-radiation always drives T_g to exceed its no-atmosphere value. However, this value might be more or less than T_{sa}. To examine this difference, we linearize the surface radiation budget about T_{sa}, which results in

$$(1 - \alpha_g)S_g = \sigma e^* e_g T_{sa}^4 + b_{ir} \cdot (T_g - T_{sa}) \tag{6.4}$$

The linearized form can be immediately solved for the ground-air temperature difference. Substi-

tuting the expression for b_{ir}, we find

$$\begin{aligned}(T_g - T_{sa}) &= \frac{1}{4}\frac{(1-\alpha_g)S_g}{\sigma e_g T_{sa}^4}T_{sa} - \frac{1}{4}e^* T_{sa} \\ &= \frac{1}{4}((\frac{T_o}{T_{sa}})^4 - e^*)T_{sa}\end{aligned} \tag{6.5}$$

where T_o is the no-atmosphere ground temperature, which satisfies $\sigma e_g T_o^4 = (1-\alpha_g)S_g$. For planets with an optically thick lower atmosphere, the ground temperature can get extraordinarily hot relative to the air temperature if there are no turbulent fluxes to help carry away the heat. The first term on the right hand side of Eq. 6.5 is large in tropical Earth conditions. For $(1-\alpha_g)S_g = 300W/m^2$ and $T_{sa} = 300K$ with $e_g = 1$, it has the value $49K$. But in tropical Earth conditions, e^* is on the order of .1, so the second term subtracts little ($15K$ for $T_{sa} = 300K$). Thus, the ground temperature is $34K$ warmer than the overlying air temperature, or $334K$. In reality, the sea surface temperature hardly ever gets more than a few degrees warmer than the free-air temperature in the Earth's tropics.

Ironically, for planets which have such a strong greenhouse effect that the low level air temperature is much larger than the no-atmosphere value, $T_g - T_{sa}$ can be quite small even if the lower atmosphere is optically thick enough to make $e^* \approx 0$, and even in the absence of turbulent heat fluxes. This conclusion is readily deduced from the factor multiplying T_{sa} in the second line of Eq. 6.5. For example, Venus has a small T_o because of the highly reflective clouds which keep sunlight from reaching the surface, yet has a high T_{sa} because of its strong greenhouse effect. In consequence, this factor is only .0024 for Venus in the limit $e^* = 0$, whence $T_g - T_{sa} \approx 1.8K$. If some CO_2 window region not reproduced by the procedure we used to estimate the surface radiative cooling on Venus allowed e^* to increase modestly to 0.05, then the ground temperature would actually become slightly *cooler* than the overlying air temperature, leading to a low-level temperature inversion and cutting off near-surface convection. For planets like Venus, the surface radiation budget is dominated by infrared back-radiation, and the comparatively feeble sunlight has little power to drive the ground temperature to values much greater than the overlying air temperature. It is situations like the Earth's tropics, which combine an optically thick lower atmosphere (due to water vapor in our case) with a rather modest greenhouse effect, where the radiation budget tends to drive the ground temperature to large values relative to that of the overlying air.

When the lower atmosphere is optically thin, as in the case of present Mars, the ground-air temperature difference cannot be determined without considering the top-of-atmosphere balance simultaneously with the surface balance. For an optically thin atmosphere, Eq. 6.3 tells us that T_g is just slightly greater than its no-atmosphere value, but it does not by itself tell us how T_g relates to T_{sa}. The general idea for an optically thin atmosphere is that the ground temperature is close to what it would be without an atmosphere, while the atmosphere cools down until the energy it loses by emission is equal to the energy gained by absorption of infrared upwelling from the ground (plus atmospheric solar absorption, if there is any). This generally leaves the low level air temperature much colder than the ground, since the atmosphere loses energy by radiating out of *both* its top and its bottom. The most straightforward way to make this more precise is to consider the radiative energy budget of the atmosphere, which is the difference between top-of-atmosphere and surface energy budget.

The net infrared radiative flux into the bottom of the atmosphere is $e_g\sigma T_g^4 - e_a\sigma T_{sa}^4$, while the infrared flux out of the top of the atmosphere is the OLR. As discussed in Chapter 4, the OLR is the sum of the emission from the atmosphere itself and the portion of the upward emission from the ground which is transmitted by the atmosphere. Let a_+ be the proportion of upward radiation

from the ground which is absorbed by the full depth of the atmosphere, and express the upward atmospheric emission escaping the top of the atmosphere in the form $e_{a,top}\sigma T_{sa}^4$. Then

$$OLR = e_{a,top}\sigma T_{sa}^4 + (1 - a_+)e_g\sigma T_g^4 \tag{6.6}$$

Let's assume for the moment that the atmosphere does not absorb any solar radiation. Then, in the absence of turbulent heat fluxes the atmospheric energy budget reads

$$0 = OLR - (e_g\sigma T_g^4 - e_a\sigma T_{sa}^4) = a_+e_g\sigma T_g^4 - (e_{a,top} + e_a)\sigma T_{sa}^4 \tag{6.7}$$

whence

$$T_{sa} = \left(\frac{a_+e_g}{e_{a,top} + e_a}\right)^{\frac{1}{4}} T_g \tag{6.8}$$

Note that we have not yet made use of the assumption that the atmosphere is optically thin. For an optically thick atmosphere with a very strong greenhouse effect (like Venus), $a_+ \approx e_a \approx 1$ and $e_{a,top} \approx 0$, and so we recover our previous result that $T_{sa} \approx T_g$ for such an atmosphere,provided the emissivity of the ground is close to unity. For an optically thin atmosphere, a_+, $e_{a,top}$ and e_a are all small, so one needs to know precisely how small the absorption coefficient is relative to the two emission coefficients. For an isothermal atmosphere –whether grey or not– Eq. 4.9 implies $e_{a,top} = e_a$. For a grey atmosphere, it follows in addition that $a_+ = e_{a,top} = e_a$. In this case $T_{sa} = T_g/2^{1/4}$, reproducing the result of Section 3.6. When the atmosphere is not grey, the absorption coefficient differs somewhat from atmospheric emission coefficient, because the spectrum of the upwelling radiation from the ground is different from that of the atmospheric emission (by virtue of the difference between ground temperature and air temperature). However, the deviation from the grey gas result is typically modest for an isothermal atmosphere. For example, a $7mb$ Marslike pure CO_2 atmosphere with a uniform temperature of $230K$ has $a_+ = e_{a,top} = e_a \approx .14$

However, introduction of a vertical temperature gradient strongly affects the relative magnitude of the three coefficients. If we take the same Marslike atmosphere with the same ground temperature and pressure, but stipulate that the temperature is on the dry adiabat rather than isothermal, then a_+ and e_a are reduced slightly (to .116 and .106, respectively), but are still approximately equal. In contrast, $e_{a,top}$ is substantially reduced, to .043. In consequence, the temperature jump at the ground is $T_{sa} = T_g/1.28^{1/4}$ – substantially weaker than the isothermal case, but still quite unstable. Results for a dry Earth, with 300 $ppmv$ of CO_2 in a $1bar$ N_2/O_2 atmosphere having $300K$ surface temperature, are similar: $a_+ \approx e_a \approx .14$ while $e_{a,top} \approx .04$. What is happening in both cases is that the atmosphere appears optically thin when averaged over all wavenumbers, but is really quite optically thick in a narrow band of wavenumbers near the principal CO_2 absorption band. The optical thickness in this range introduces a strong asymmetry in the upward and downward radiation, and also weights the absorption towards the bottom of the atmosphere (which is also where a disproportionate amount of the infrared back-radiation is coming from. A rule of thumb for such cases is that a_+ and e_a will have similar magnitudes, while $e_{a,top}$ will be smaller (but, in the optically thin case, still non-negligible); it follows that the surface temperature jump is weaker than the isothermal case, but still unstable. For an optically thin grey gas the situation is different. In that case, $e_a = e_{a,top}$ and both are less than a_+; nonetheless, the relative magnitudes are such that an unstable temperature jump can generally be sustained at the surface even if the lower atmosphere is on a dry adiabat (see Problem **??**).

The upshot of the preceding discussion is that, in the absence of atmospheric solar absorption, the radiative balance in an optically thin atmosphere almost always drives the surface to be notably warmer than the overlying atmosphere, even if convection has established an adiabat in the atmosphere. This provides a source of buoyancy that can maintain the convection which

stirs the troposphere and maintains the adiabat. A moist adiabat is more isothermal than the dry adiabat, so our conclusion is even firmer in that case. Atmospheric solar absorption, on the other hand, would warm the atmosphere relative to the surface, weakening or even eliminating the unstable surface jump.

Moving on, let's consider the temperature the ground of a planet would have in radiative equilibrium at night-time, when $S_g = 0$. In this case, there is little to be gained by linearizing the surface budget, as it reduces to simply $\sigma T_g^4 = e_a \sigma T_{sa}^4$, whence $T_g/T_{sa} = (e_a)^{1/4}$. For an optically thick lower atmosphere, the infrared back radiation keeps the ground temperature nearly equal to the air temperature. However, when the lower atmosphere is not optically thick, the ground temperature plummets at night, or would do so if it had time to reach equilibrium. Cold climates tend to be comparatively optically thin because they cannot hold much water vapor even in saturation. For example, using the moist case in Figure 6.1, we find that when $T_{sa} = 240K$, $e_a \approx .3$ with $.1mb$ of CO_2 in the atmosphere. This implies that at night the ground temperature plunges toward the fearsomely cold value $T_g = 177K$. Liquid surfaces like oceans cannot generally cool down rapidly enough to approach the night-time equilibrium temperature, because turbulent motions in the fluid bring heat to the surface which keeps it warm. Solid surfaces like snow, ice, sand or rock can cool down very quickly, though, and do indeed plunge to very low temperatures at night. This situation applies to Snowball Earth and to the present-day Arctic and Antarctic. Very cold climates are of necessity dry, because of the limitations imposed by Clausius-Clapeyron. However, even relatively warm climates can be dry if the moisture source is lacking. This is why deserts can go from being unsurvivably hot in the daytime to uncomfortably cold at night. Turbulent fluxes can bring additional heat to the ground and moderate the night-time cooling somewhat, but these fluxes tend to be weak in the situation just described, because turbulent eddies must expend a great deal of energy to lift cold dense air from the ground to the outer edge of the surface layer (a matter taken up in more detail in Section 6.4)

The preceding discussion technically applies whether or not T_{sa} itself drops substantially at night, but is most meaningful in the situation where the atmosphere cools slowly enough that the atmosphere remains relatively warm as the night-time ground temperature drops. This is a fair description of the situation in the massive atmospheres of Titan, Earth and Venus, except to some extent during the long polar night on Titan and Earth. The tenuous atmosphere of present Mars, in contrast, cools substantially throughout its depth during the night, even at midlatitudes. In this situation, the relative temperature of air and ground at night is determined by the relative rates of cooling of the two media, rather than radiative equilibrium. We will take up the issue of thermal response time in detail in Chapter 7.

6.3 Basic models of turbulent exchange

Anybody who has watched dry leaves or dust blow around on a windy day has noticed that where the air comes up against the surface there arises a complex mass of turbulent eddies. In comparison, the interior of planetary atmospheres are fairly quiescent places, except in the immediate vicinity of rapidly rising buoyant plumes and active cloud systems. The turbulent fluid motions near the planetary surface exchange energy between the surface and the atmosphere, both in the form of sensible heat (energy corresponding to the change of temperature in a mass) and latent heat (energy associated with the change of phase of a condensible substance, with fixed temperature). Representing the effects of turbulence is not like representing radiation, where we can write down some basic physical principles then proceed through a set of systematic approximations until we arrive at a set of equations we can solve. When it comes to turbulence, the state of physics is not

yet up to that challenge, and may never be. Instead, one must take a largely empirical approach from the outset, constrained by some fairly broad principles such as conservation of energy.

In this section we will derive the so-called *bulk exchange* formulae describing the flux of a quantity from the surface to the overlying atmosphere. The general idea is the same whether the quantity is a chemical tracer, sensible heat (associated with temperature fluctuation) or latent heat, so we will first present the formulae for a general tracer. The calculation will be introduce using simple physically-based scaling arguments, and then will be revisited in a more precise and systematic fahion in Section 6.4.

Let c be the specific concentration of some substance, and c' be the fluctuating or "turbulent" part, usually thought of as a deviation from a time or space mean over some suitable interval. Further, let w' be the fluctuating vertical velocity at the top of the surface layer. Then, the flux of the substance, in kg/m^2, is

$$F_c = \overline{\rho w' c'} \approx \rho_s \overline{w' c'} \tag{6.9}$$

where the overbar represents a time or space average and ρ is the total density of the gas making up the atmosphere. We assume further that the surface layer is thin enough that the variation in pressure and temperature across it is small enough that the variations in density can be neglected. Thus, the density factor can be replaced by a constant typical surface density, ρ_s, and taken outside the average. The ideal gas law states that $\rho = p/RT$. If the surface layer has a thickness of a few tens of meters or less, then the hydrostatic law typically guarantees that the contribution of pressure to the density variations is small. It is not inconceivable, however, that the temperature difference across the surface layer could reach 10% of the mean, leading to corresponding changes in the density. With a little more work, the effect of these fluctuations can be brought into the picture, but we will not pursue this refinement as the effects are probably overwhelmed by the uncertainties in the representation of turbulence itself.

Next, we must estimate the correlation $\overline{w' c'}$. We build this estimate from a typical vertical velocity δw, a typical concentration fluctuation δc, and a non-dimensional factor $0 < a < 1$ describing the degree of correlation. Thus, we write $\overline{w' c'} = a \cdot \delta w \cdot \delta c$. Next, we assume that δw is proportional to the mean horizontal wind speed U at the top of the surface layer, so $\delta w = s \cdot U$. The constant of proportionality s can be thought of as a typical slope characterizing the turbulent eddies, which is in turn roughly related to the roughness of the surface. Note that U is the wind *speed*, and is therefore positive. We then assume that the typical concentration fluctuation scales with the concentration difference between the air in contact with the ground and the edge of the surface layer, so $\delta c = f \cdot (c_g - c_{sa})$, where c_{sa} is the concentration at the edge of the surface layer, c_g is that at the ground, and f is a nondimensional constant of proportionality. Putting it all together and lumping the proportionality constants into the *drag coefficient* $C_D \equiv a \cdot s \cdot f$, we write

$$F_c = \rho_s C_D U (c_g - c_{sa}) \tag{6.10}$$

C_D is called the *drag coefficient* because when c is taken to be the turbulent velocity itself, the flux formula gives the flux of momentum, and hence the drag force on the surface. In writing the flux in the form of Eq. 6.10, we have adopted the convention that a positive flux represents a transfer of substance from the ground to the atmosphere. The turbulent flux acts like a diffusion, transferring substance from regions of higher concentration to regions of lower concentration. It is like a bucket-brigade, with partly empty buckets being handed downstairs from the top of the surface layer to the ground, where they are filled and sent back upstairs again (or with full buckets sent downstairs to be partly dumped out on the ground). The mass of substance in a bucket being carried upstairs is proportional to $\rho_s c_g$, while the mass of substance in a bucket going downstairs

is proportional to $\rho_s c_{sa}$, while $C_D U$ gives the rate at which buckets are being handed up or down the stairs.

6.3.1 Sensible heat flux

To obtain the sensible heat flux, we take $c_p T$ to be our tracer. This is essentially the dry static energy (see Eq. 2.23), since the surface layer is thin enough that the height z can be taken to be nearly constant. With this choice of tracer, Eq. 6.10 becomes

$$F_{sens} = c_p \rho_s C_D U (T_g - T_{sa}) \tag{6.11}$$

If the ground is warmer than the air, heat is carried away from the ground at a rate proportional to the temperature difference. If the ground is cooler than the air, the sensible heat flux instead acts to warm the ground.

If C_D is independent of temperature, then F_{sens} is exactly linear in the difference between the ground temperature and air temperature. Hence the coupling coefficient b_{sens} – analogous to b_{ir} – is simply $b_{sens} = c_p \rho_s C_D U$. When the surface layer becomes stably stratified, however, C_D can be driven nearly to zero because the energy of turbulence is expended in mixing dense air upward. This effect will be quantified in Section 6.4. The consequent temperature dependence of C_D would alter the linearized coupling coefficent.

Note that the sensible heat flux becomes small when the atmosphere has low density. The "wind-chill" factor on present Mars would be exceedingly weak! Conversely, very dense atmospheres like those of Venus or Titan can very effectively exchange heat between the surface and the atmosphere. With $C_D = .001$,$U = 10 m/s$ and $T_g - T_{sa} = 1K$ the sensible heat flux is $.13 W/m^2$ on present Mars, $11 W/m^2$ on Earth, $55 W/m^2$ on Titan, and a whopping $540 W/m^2$ on Venus. It is for similar reasons that immersion in near-freezing water is far more life-threatening than walking about scantily clad in air of the same temperature – water is about 1000 times denser than Earth air. One must take care to distinguish thickness of an atmosphere (in terms of density) from optical thickness. An atmosphere can be thick (i.e. dense) while being optically thin, and conversely a thin (low density) atmosphere can nonetheless by optically thick if the greenhouse gas it is made of is sufficiently effective.

Now let's suppose that the sensible heat flux dominates the surface energy budget. By "dominates," we mean that the sensible heat flux due to a small departure from equilibrium (considering the sensible heat flux alone) overwhelms the other terms in the surface energy balance. This would be true if the wind speed and density were large, provided that the ground and atmosphere are dry enough that evaporation remains small. Sensible heat flux vanishes when $T_g = T_{sa}$, so this is the state that the system is driven to when sensible heat flux dominates. Taking the radiative and latent fluxes into account would cause a small deviation from this limit.

6.3.2 Latent heat flux

Whatever the condensed substance making up the surface, some of the condensed substance will transform into the vapor phase in the atmosphere contacting the surface, until it reaches the saturation vapor pressure determined by Clausius-Clapeyron. If the winds then carry away this vapor-laden air and replace it with unsaturated air, more mass will evaporate or sublimate from the surface. Since the phase change involves latent heat, a flux of mass away from the surface cools the surface by carrying away latent heat. Conversely, a flux of mass from vapor into the condensed

surface will warm the atmosphere where condensation occurs. All substances will evaporate or sublimate to some extent, and whether the latent heat flux is significant is a matter of how big the saturation vapor pressure is at the typical temperature of the surface. For water ice on Titan at $95K$, the vapor pressure is under $10^{-15}Pa$, so the latent heat flux of water is utterly negligible. The situation is the same for basalt at $300K$ on Earth, or even at $750K$ on Venus. However, the vapor pressure of CO_2 on present Mars, of liquid water or water ice on Earth, and of methane on Titan are all high enough to allow substantial latent heat flux. Whatever the condensible substance in question we will use terms like "humidity" by analogy with the archetypal case of water vapor on Earth. Also, for the sake of verbal economy we will often refer simply to "evaporation" in situations where the actually process might be either evaporation or sublimation.

In dealing with latent heat flux, it is more convenient to deal with the mass mixing ratio of the condensible to dry air, rather than specific humidity. This makes it somewhat easier to treat cases where the condensible makes up a substantial part of the total mass. Thus, we use the mass mixing ratio r_w as the tracer in Eq. 6.10. If ρ_a is the density of dry air in the surface layer, then the mass of condensible per unit volume is $\rho_a r_w$ and this mass carries a latent heat $L\rho_w r_w$. we can write the mixing ratio r_{sa} at the edge of the surface layer as $h_{sa} r_{sat}(T_{sa})$, where where h_{sa} is the relative humidity at the outer edge of the surface layer and $r_{sat}(T)$ is the saturation mass mixing ratio. In terms of saturation vapor pressure, the saturation mass mixing ratio is $(M_w/M_a)(p_{sat}(T)/p_a$, with p_a being the partial pressure of dry air in the surface layer. Now suppose that at the ground there is a reservoir of a condensed phase of the substance "w" – an ocean, lake, swamp, snow field, glacier or the surface of an icy moon. In this case, the vapor pressure in the air in contact with the surface must be in equilibrium with the condensed phase, and must therefore follow the Clausius-Clapeyron relation evaluated at the temperature of the ground. Equivalently, we can say that $r_g = r_{sat}(T_g)$. Using the two mixing ratios, the latent heat flux becomes

$$F_L = L\rho_a C_D U(r_{sat}(T_g) - h_{sa} \cdot r_{sat}(T_a)) \tag{6.12}$$

Alternately, using the definition of the mixing ratios and assuming the partial pressure of dry air to be approximately constant within the boundary layer, Eq. 6.12 can be written

$$F_L = \frac{L}{R_w T_{sa}} C_D U(p_{sat}(T_g) - h_{sa} \cdot p_{sat}(T_{sa})) \tag{6.13}$$

The latter form of the latent heat flux demonstrates that the flux is in fact unaffected by the presence of the dry air. Assuming temperature and wind to be held constant, the evaporation from the Earth's ocean would remain unchanged even if all the N_2 were taken out of the atmosphere. This conclusion would no longer be valid if the gases in question had substantial non-ideal behavior, for then the law of partial pressures would no longer hold.

Exercise 6.3.1 Derive Eq. 6.13. What do you have to assume about the air temperature within the surface layer?

In situations where a major constituent of the atmosphere can condense out onto the surface or sublimate or evaporate from it, a constraint on the temperature change across the surface layer enters the problem in a significant way. The constraint arises from the fact that, since the surface layer is thin, the pressure must be nearly constant within the layer. The implications of this constraint are easiest to see when the atmosphere consists of a single condensible component; a concrete example of this situation is provided by the state of the surface layer over seasonal CO_2 frost layers on Mars. Let's suppose that the system has a layer of condensate of the atmospheric substance at the surface – an ocean or glacier. Then, since the atmosphere consists of only the one

constituent, the surface pressure is fixed in terms of the ground temperature by Clausius-Clapeyron, namely $p_s = p_s(T_g)$. The pressure at the upper edge of the surface layer must be very nearly equal to this value, otherwise there would be a large unbalanced pressure gradient which would drive a strong flow that would soon transport enough mass to equalize the situation. It follows that *if the atmosphere at the upper edge of the surface layer is saturated*, we must have $T_{sa} \approx T_g$. In other words, in saturated conditions, the temperature at the ground and the temperature at the upper edge of the surface layer must adjust nearly instantaneously so as to keep the two equal. Under what circumstances can the upper edge of the surface layer be considered saturated? First note that if T_{sa} were colder than T_g, then the pressure contininuity condition would require the air to be supersaturated. This situation cannot persist for long, so in a case where the free atmosphere is cooling or the ground is heating up, T_{sa} would adjust nearly instantaneously to remain equal to T_g. This adjustment does involve a transfer of latent heat, which alters the thermal response time of the system. One could treat this transfer in terms of an strong enhancement of C_D in such conditions, but there are more natural ways to deal with essentially instantaneous adjustments. The implications for the seasonal cycle of condensible atmospheres will be considered in Section 7.7.5, where such an alternate approach will be illustrated. On the other hand, a situation with $T_g < T_{sa}$ is perfectly consistent if the atmosphere aloft is subsaturated. In such situations, the transfer of latent heat flux is governed by Eq. 6.13 as usual. The transfer would act both to cool the surface, and to add mass to the atmosphere bringing it closer to saturation. However, in situtations where the atmosphere remains saturated as the system cool down, the previous temperature continuity constraint applies.

From Eq. 6.13 we observe that latent heat flux carries heat away from the ground when the saturation mixing ratio at the ground is less than the mixing ratio of the surface layer. Since typically $h_{sa} < 1$, this can happen even if the ground is colder than the overlying air. We also note that the latent heat flux becomes insignificant at sufficiently cold temperatures, since both saturation vapor pressures in the equation become small in that limit.

Sensible and radiative heat transport carry no mass away from the surface, but latent heat transport is of necessity accompanied by mass transfer. The mass flux into or out of the ground is simply F_L/L. The mass flux is needed for calculating the rate of ablation of glaciers by sublimation, the drying out of lakes or soil by evaporation, and the rate of salinity change at the surface of an ocean (since evaporation carries away the condensible but not the solute).

Now let's look at how the fluxes behave when the temperature difference between the ground and the outer edge of the surface layer is small. Carrying out a Taylor series expansion of the flux about $T_g = T_{sa}$, as we did for the infrared cooling case, we write

$$F_L = E_o + b_L \cdot (T_g - T_{sa}) \tag{6.14}$$

Defining the characteristic flux $F_L^* \equiv C_D U p_{sat}(T_{sa})$, we find

$$E_o = (1 - h_{sa}) \frac{L}{R_w T_{sa}} F_L^*, b_L = \frac{1}{T_{sa}} (\frac{L}{R_w T_{sa}})^2 F_L^* \tag{6.15}$$

where R_w is the gas constant for the condensible. The Clausius-Clapeyron relation has been used to substitute for dp_{sat}/dT in the expression for b_L. E_o is the heat flux due to evaporation or sublimation that would occur with $T_g = T_{sa}$; it vanishes if the surface layer is saturated ($h_{sa} = 1$), but is positive otherwise. Both E_o and b_L are proportional to the characteristic flux F_L^*, which vanishes vanishes as $T_{sa} \to 0$, since the saturation vapor pressure vanishes like $\exp(-L/R_w T)$ in this limit. As one might expect, latent heat flux becomes negligible at sufficiently low temperatures. How low one must go for this to be the case depends on the gas in question. As temperature increases, the characteristic flux becomes large, and hence E_o and b_L become large as well. The

increase is abetted by the fact that L/R_wT is a large number at typical planetary temperatures (e.g. 18.06 for water vapor at $300K$, or 10.3 for methane at $95K$). For temperatures high enough that b_L becomes large, a modest ground-air temperature difference leads to a very large increase in latent heat flux. This tends to make it hard for the ground temperature to differ much from the free air temperature in such cases.

Table 6.1 gives some typical values of E_o and b_L for water, carbon dioxide and methane. In all three cases, we see that the latent heat flux rises very strongly with temperature. For water, latent heat flux is insignificant at temperatures of $230K$ or lower. The feeble latent flux of a Watt per square meter or so would be utterly dominated by infrared cooling of the surface, or by the sensible heat flux arising from a ground-air temperature difference of as little as $1K$. This corresponds to the situation in the Antarctic night of the present Earth, or to the daily average tropical temperatures on a Snowball Earth. However, even at the freezing point of water, the latent heat flux is quite substantial. With a $5K$ ground-air temperature difference, the flux would be nearly $100W/m^2$, which is almost half of the typical midlatitude absorbed solar radiation in the ocean, and roughly equal to the typical absorbed solar radiation in ice. The latent heat flux is also comparable to the typical infrared cooling of the surface at such temperatures (inferred from Figure 6.1). As temperature is increased further to values characteristic of the modern tropics, the flux increases dramatically; it would take about 90% of the supply of absorbed solar energy going into the ocean in order to sustain the evaporation arising from just a $2K$ ground-air temperature difference. At these temperatures, the latent flux is considerably in excess of the surface infrared cooling.

For the other gases in the table, the latent heat flux becomes substantial at much lower temperatures. At temperatures comparable to the Martian polar Spring, the latent heat flux due to CO_2 sublimation is comparable to the water vapor values for Earth's midlatitudes or tropics (assuming the same degree of boundary layer saturation). These fluxes are particularly consequential in light of weak supply of solar radiation on Mars, relative to Earth. Alternately one may compare the latent flux to the infrared cooling of the surface in the thin Martian atmosphere (σT_g^4, or $37W/m^2$ at $160K$). Either way, we conclude that latent heat flux plays a key role in determining surface temperature at places on Mars where seasonal CO_2 frost is sublimating or being deposited. At Titan temperatures, latent heat flux due to methane evaporation is enormous; the solar radiation reaching Titan's surface is well under $5W/m^2$, which is two orders of magnitude less than the Methane evaporation one gets under the conditions of Table 6.1. Somehow or other, conditions near Titan's surface must adjust until the evaporation is reduced to the point where it can be balanced by the supply of energy to the surface, but the numbers in the table tell us that methane latent heat flux is the dominant constraint on the adjusted state. Ironically, Titan, at $95K$ is like an extreme form of the Earth's tropics, in that evaporation dominates the surface energy budget to an even greater extent than it does in Earth's tropics. If the temperature of the Earth's tropics were raised to $320K$, as might happen in the high CO_2 world following deglaciation of a Snowball Earth, then E_o on Earth, too would greatly exceed the available solar energy, though not to such an extent as it does on Titan. The way the surface conditions adjust to accomodate this state of affairs will be taken up in the Section 6.5.

When the surface is sufficiently cold relative to the air, vapor from the air can be deposited on the surface in the form of dew or frost. In this case the latent heat flux is negative, and carries energy from the atmosphere to the ground. If the boundary layer is saturated ($h_{sa} = 1$) then frost or dew deposition occurs whenever $T_g < T_{sa}$. If the boundary layer is unsaturated deposition won't occur until the ground temperature is made sufficiently cold that the saturation vapor pressure there falls below the partial pressure of the condensible in the overlying atmosphere (a temperature known as the "dew point" or "frost point"). When latent heat is being carried to the surface – as

	H_2O	H_2O	H_2O	H_2O	CO_2	CO_2	CH_4	CH_4
T_{sa} (K)	230	273	300	320	150	160	80	95
E_o (W/m^2)	.72	40.8	193.3	557.8	52.5	182.1	93.2	640.0
b_L (W/m^2K)	.28	11.2	38.6	98.0	24.4	74.4	55.6	243.

Table 6.1: Latent heat flux coefficients for various gases at selected temperatures T_{sa}. Computed with $U = 10m/s$, $C_D = .001$ and boundary layer relative humidity $h_{sa} = 70\%$.

it is during he seasonal polar CO_2 frost formation on Mars – the rate of condensation is limited by the rate at which the surface can get rid of the deposited latent heat. Since the surface is colder than the atmosphere during deposition, sensible heat flux carries heat the wrong way to balance the budget, so it is only infrared cooling of the surface that can sustain frost or dew. Otherwise, the surface will simply warm in response to the deposited latent heat until it is no longer cold enough for frost or dew to form.

Over land, there are two further complications that must be considered. The first is that land, unlike a deep ocean or lake or a thick glacier, can dry out. If the land surface is a mix of condensible and (essentially) noncondensible substance, the latent heat flux can exhaust the supply of condensible, whereafter the boundary condition $r_g = r_{sat}(T_g)$ is no longer appropriate. In the absence of further supply of condensible at the ground, the latent heat flux must fall to zero. In such a case, one must keep track of the mass of the condensible reservoir at the ground, and zero out the latent heat flux when the reservoir is exhausted. This would be the case for thin snow cover, scattered puddles, or soil moisture on Earth, for CO_2 frost layers on Mars and for liquid methane swamps on Titan. For soil moisture, a common simple model is the *bucket model*, in which each square meter of soil surface is treated as a bucket whose capacity is determined by its porosity and depth. The bucket is filled by rainfall, and emptied by evaporation. Once the bucket is full, any additional rainfall is assumed to run off into rivers (which may or may not be tracked, according to the level of sophistication of the model). As long as the bucket has some water in it evaporation is sustained, but when the bucket is empty latent heat flux is zeroed out and only radiative and sensible heat transfers at the ground are allowed. The bucket model may serve also as a model of conditions at Titan's surface, which may consist not only of liquid methane puddles but also bogs consisting of beds of granular water ice sand or pebbles whose pores are saturated with liquid methane.

The second complication over land concerns the effect of land plants. At present, Earth's climate provides the only example where this must be taken into account. Plants actively pump water from deep storage, at rates determined by their own physiological requirements. This is known as *transpiration*, and given that moisture flux over vegetated land is always some mix of transpiration and evaporation, the joint process is called *evapotranspiration*. In this case, the moisture boundary condition at the ground may be more appropriately represented as a flux condition determined by plant physiology rather than setting the moisture mixing ratio at the ground. The moisture flux may be limited by rate at which trees pump moisture, and not by rate at which turbulence carries it away. The mixing ratio at the ground still cannot exceed saturation, so when the transpiration becomes strong enough to saturate the air in contact with the ground, one can revert to the previous model of conventional evaporation. Yet a further complication in vegetated terrain is the very notion of ground and ground temperature. Is "the ground" the forest surface or the elevated leaf canopy? Is the ground temperature that of the leaf surface or the soil? How do we take into acount the mix of illuminated hot leaves and relatively cool leaves in shade? A proper treatment of these factors requires a detailed model of the microclimate in the vegetation layer, which is beyond what we aspire to in this book. One need not abandon all hope

of estimating conditions over vegetated terrain, however. As a rule of thumb, dense forests that get enough rainfall to survive in the long term tend to act more or less like the ocean, save for an eleveated C_D caused by greater surface roughness. Grasslands, shrub, tundra and prairie can be crudely modeled using the bucket model.

When evaporation dominates the surface budget, equilibrium requires $F_L = 0$, or equivalently $p_{sat}(T_g) = h_{sa}p_{sat}(T_{sa})$. Since p_{sat} is monotonically increasing in temperature, this relation requires $T_g < T_{sa}$ if the boundary layer air is unsaturated $h_{sa} < 1$. Thus, evaporation or sublimation drives the ground temperature to be *colder* than the overlying air temperature. However, the ground and surface could also achieve equilibrium by transferring enough moisture to the surface layer that it becomes saturated ($h_{sa} = 1$), in which case $T_g = T_{sa}$ in equilibrium, as for the case of sensible heat flux. The extent to which equilibrium is attained by adjusting temperature vs. humidity depends on the competition between the rate at which moisture is supplied to the boundary layer and the rate at which dry air from aloft is entrained into the boundary layer. Observed boundary layers on Earth and Titan are significantly undersaturated, leading to the conclusion that the ground temperature would be considerably less than the air temperature, if other fluxes did not intervene. Using the linearized form of the latent heat, the equilibrium ground-air temperature difference is $T_g - T_{sa} \approx -E_o/b_L$. For the conditions of Table 6.1, this is $-2.6K$ for Titan at $95K$. For a hot Earth at $320K$, the difference is about $-5.7K$. There are currently no observations of the state of saturation over the sublimating Martian CO_2 frost cap, but given the saturation assumed in the table the equilibrium occurs with $T_g - T_{sa} \approx -2.4K$ when the air temperature is $260K$. Thus, even when evaporation dominates, the equilibrium ground temperature does not differ greatly from the overlying air temperature. This was also found to be the case when the surface budget is dominated by sensible heat flux. It is only the radiative terms that can drive the ground temperature to be substantially different from the overlying air temperature.

6.4 Similarity theory for the surface layer

The surface layer theory based on dimensional analysis tells us most of what we need to know, but it doesn't tell us how the drag coefficient depends on the height at which the top-of-layer conditions are applied, nor does it say precisely how the coefficient depends on stratification or the surface roughness. We will now re-do the surface layer theory using a more precise form of the similarity assumption. The most important thing we will get out of this is a quantification of the suppression of turbulent mixing in stable conditions. This exerts a very important control on the fluxes at night-time and over ice or snow, where surface layer conditions are often stable. In particular, when ice or snow is melting the temperature is pinned at the freezing point, so if the atmospheric temperature is significantly above freezing, the surface layer is very stable, and this limits the delivery of heat available for melting.

Let c be any quantity whose flux we wish to determine in the surface layer. It might be temperature, water vapor, methane or some other chemical tracer. We will also consider the flux of horizontal momentum (proportional to horizontal wind u) using the similarity theory. Though we are not attempting to do much dynamics in this book, we will nevertheless need to talk a bit about momentum flux since this is what will tell us how the mean wind varies with height within the surface layer.

Within the surface layer, the fluxes of tracer and momentum are constant, by definition. This allows us to define the following velocity and tracer scales:

$$u_*^2 \equiv \overline{w'u'}, c_* \equiv \overline{w'c'}/u_* \tag{6.16}$$

As a consequence of the second definition the tracer flux is just $u_* c_*$. The velocity scale u_* is called the *friction velocity*, and is taken to be positive by convention. When the flux of tracer is upward,then the tracer fluctuation scale c_* is positive.

Next we derive equations for the vertical gradient of mean tracer and mean wind. We'll first consider the neutrally stratified case, in which buoyancy forces are negligible. Strictly speaking, this case only applies when the density within the surface layer is constant. Any situation with heat transport would involve some temperature fluctuations, and hence some density fluctuations. In practice, though, when the temperature difference across the surface layer is sufficiently weak, the neutrally stratified calculation yields accurate results. The definition of "sufficiently weak" will be made precise later, when we come to incorporate buoyancy forces.

When buoyancy is insignificant in the surface layer, the only length scale appearing in the problem is the height z above the ground. Since the only tracer scale is c_* and the only velocity scale is u_*, dimensional analysis then tells us that the equations for the vertical gradients must be

$$\begin{aligned} K_{vk} \frac{d\overline{c}}{dz} &= -\frac{1}{z} c_* \\ K_{vk} \frac{d\overline{u}}{dz} &= \frac{1}{z} u_* \end{aligned} \tag{6.17}$$

where K_{vk} is a nondimensional constant called the *Von Karman constant*. In principle, the nondimensional constant appearing in the tracer equation could be different from that in the momentum equation, but laboratory experiments indicate that in fact the same constant applies to both. The Von Karman constant has been measured in a wide range of turbulent laboratory experiments, which indicate that $K_{vk} \approx .4$. The sign choice in the tracer equation is dictated by the physical requirement that the tracer flux be upward when the concentration is greater at the surface than it is aloft.

The similarity equations allow us to relate the flux of tracer to the difference in tracer concentration between the ground and the upper edge of the surface layer, and similarly for momentum. Let z_1 be the upper edge of the surface layer. Integrating from a smaller height z_* to z_1 and assuming that the winds vanish at the height z_*, we find

$$\begin{aligned} K_{vk}(\overline{c}(z_1) - \overline{c}(z_*)) &= -c_* \ln(\frac{z_1}{z_*}) \\ K_{vk}\overline{u}(z_1) &= u_* \ln(\frac{z_1}{z_*}) \end{aligned} \tag{6.18}$$

The height z_* at which we set the lower limit of integration cannot be sent to zero because of the logarithmic divergence in that limit. In fact, it has a physical meaning, and is called the *roughness height*. It corresponds to the height at which the airflow is so perturbed by the irregularities in the boundary that the mean flow is essentially zero. The roughness height corresponds loosely to the typical height of the bumps on the surface, but is generally smaller than one would intuit from the physical height of the bumps. In practice, it is determined by fitting the observed mean wind profile with the logarithmic form. Over open water, the roughness length is on the order of 0.0002 m, though at strong wind speeds the wind-driven waves increase the roughness significantly. Over ice or smooth land, the roughness length is more like 0.005 m, increasing to $0.03m$ if there is grass or low vegetation, 0.5 m for low forest and 2 m for large forests or urban areas.

The logarithmic profile of wind and concentration is called *the law of the wall*, and has been verified in a great variety of turbulent flows, ranging from wind and water tunnel experiments to atmospheric measurements to velocity profiles in tidal surges in the Bay of Fundy. Next, the tracer

and velocity equations can be combined to give the tracer flux

$$\overline{w'c'} = u_* c_* = \frac{K_{vk}^2}{(\ln(\frac{z_1}{z_*}))^2}\overline{u}(z_1)(\overline{c}(z_*) - \overline{c}(z_1)) \tag{6.19}$$

from which we identify the drag coefficient

$$C_D = \frac{K_{vk}^2}{(\ln(\frac{z_1}{z_*}))^2} \tag{6.20}$$

This formula allows us to explicitly compute the drag coefficient given the roughness height and the height at which one chooses to apply the boundary condition at the upper edge of the surface layer; one is free to choose z_1 as a matter of convenience, so long as it is low enough that the fluxes are constant within the surface layer. Though different roughness lengths are sometimes applied for moisture and momentum, it is generally adequate to use the same drag coefficent for all mixed quantities. Using the previous values for roughness length and assuming $z_1 = 10m$, we get $C_D = 0.0014, 0.0028, 0.0047, 0.018, 0.062$ for open water, ice or smooth land, grassland, low forest and large forest, respectively.

Now let's introduce the effects of buoyancy. To do this, we must first define buoyancy quantitatively. Let ρ_g be the mean density at the ground. Then the net force (per unit volume) on an air parcel with density ρ will be $g \cdot (\rho_g - \rho)$ while the parcel is near the ground. The acceleration of the air parcel is obtained by dividing by the force by the mass, and is thus $g \cdot (\rho_g/\rho - 1)$. The buoyancy acceleration is a form of *reduced gravity*, reflecting the fact that buoyancy forces cancel part of the gravitational forces, leading to a reduction in acceleration. We will refer to the buoyancy acceleration as simply "buoyancy" for short, and denote it by the symbol β. The buoyancy is affected both by the temperature and the composition of the atmosphere. For uniform composition, warm air will be positively buoyant when surrounded by colder air. However, air that is rich in a low molecular weight substance will be buoyant when surrounded by air that has lower concentration of that substance, even if the temperature is uniform. For example, since water vapor has lower molecular weight than dry Earth air, moistening an air parcel adds to its upward buoyancy and drying it tends to make it sink. The same can be said for adding methane to N_2 in Titan's atmosphere. Similarly, the Martian atmosphere contains a few percent of Ar on average, which has a higher molecular weight than CO_2. Thus, when pure CO_2 sublimates from the Martian CO_2 seasonal frost cap, it will be positively buoyant in the background mixture of CO_2 and Ar. One can imagine a variety of situations in which an atmospheric constituent is released from or absorbed into the surface, but the most common situation involves a condensible substance which condenses onto or sublimates/evaporates from a reservoir at the surface. That could be N_2 ice on Triton, liquid CH_4 on Titan, CO_2 ice on Mars, or solid or liquid H_2O on Earth. Using the ideal gas law, the density is

$$\rho = \frac{p}{R_a T}(1 - \eta_c) + \frac{p}{R_c T}\eta_c = \frac{p}{R_a T}(1 + (\frac{M_c}{M_a} - 1)\eta_c) \tag{6.21}$$

where M_a and M_c are the molecular weights of the noncondensible background gas and the condensible component, respectively. Since the surface layer is thin, p can be assumed nearly constant within the layer. The buoyancy is then

$$\beta = g \cdot (\frac{T}{T_g}\frac{1 + (\epsilon - 1)\eta_{c,g}}{1 + (\epsilon - 1)\eta_c} - 1)) \tag{6.22}$$

where $\epsilon \equiv M_c/M_a$ and $\eta_{c,g}$ is the molar concentration of the condensible substance at the ground. When $\eta_c = \eta_{c,g} = 0$, or when $\epsilon = 1$, buoyancy is simply $g \cdot (T_g - T)/T_g$. In the general case, if $\epsilon < 1$

then increasing η_c makes the parcel more positively buoyant, while if $\epsilon > 1$ increasing η_c makes the parcel more negatively buoyant. In general, the buoyancy is a nonlinear function of temperature and concentration, but in the special case where both $(T - T_g)/T_g$ and η_c are small, the buoyancy takes on the simple form

$$\beta \approx g\frac{T - T_g}{T_g} + g \cdot (\epsilon - 1)(\eta_{c,g} - \eta_c) \tag{6.23}$$

in which the buoyancy is the sum of a temperature contribution and a composition contribution. When using either form of the buoyancy, it is assumed that the the buoyancy-generating tracer is saturated at the ground, in the typical case where its flux is maintained by a condensible reservoir there. Thus, $\eta_{c,g} = p_{sat}(T_g)/p$, where p_{sat} is given by Clausius-Clapeyron.

We can treat the buoyancy flux and mean buoyancy profile much as we did the tracer flux and tracer profile in the neutral case, defining $\overline{w'\beta'} = u_*\beta_*$. However, the buoyancy scale β_* is now a dynamically signficant quantity with dimensions of acceleration, which can affect the profile. This has the important consequence that z is no longer the only length scale that enters into the problem – in addition we can define the *Monin-Obukhov length*

$$\ell \equiv \frac{1}{K_{vk}}\frac{u_*^2}{|\beta_*|} \tag{6.24}$$

The inclusion of the Von-Karman constant in the definition of the Monin-Obukhov length is purely a matter of convention, and has no particular physical significance. The Monin-Obukhov length is on the order of the height to which a negatively buoyant plume with velocity u_* would rise before exhausting its kinetic energy, or the height to which a positively buoyant plume initially at rest would rise before attaining velocity u_*. For distances much closer to the boundary than ℓ, the turbulence is dominated by the kinetic energy of the wind shear, as in the neutral case. For heights much greater than ℓ buoyancy suppresses or enhances the turbulence. Using the Monin-Obukhov length, we define the non-dimensional depth $\zeta \equiv z/\ell$. In contrast to the neutral case, the equations for the gradient of wind, tracer or buoyancy can each depend on some function of ζ; in principle, because the function has a dimensionless argument, one can take data in the field or laboratory, and determine the function once and for all, as if it were a sine or cosine. The test of the validity of this bold assumption is to evaluate the functions from a wide variety of different field and laboratory datasets, and see if one gets essentially the same result from each. This is the assumption upon which the similarity theory rests, and it seems to work out well in practice. In terms of the similarity functions, the equations for buoyancy and wind gradient become

$$\begin{aligned} K_{vk}\frac{d\overline{\beta}}{dz} &= -\frac{1}{z}\beta_* F_\beta(\zeta) \\ K_{vk}\frac{d\overline{u}}{dz} &= \frac{1}{z}u_* F_u(\zeta) \end{aligned} \tag{6.25}$$

In general, one should allow for different scaling functions for the buoyancy and momentum equations, and indeed this is sometimes necessary to provide a good fit to data. For the stably stratified (i.e. negatively buoyant) case, it has been found that using the same scaling function for both equations is adequate. Some remarks on the positively buoyant case will be given later. In any event, let's assume $F_u = F_\beta$. In nondimensional form, the equations become

$$\begin{aligned} K_{vk}\frac{d\overline{\beta}}{d\zeta} &= -\frac{1}{\zeta}\beta_* F(\zeta) \\ K_{vk}\frac{d\overline{u}}{d\zeta} &= \frac{1}{\zeta}u_* F(\zeta) \end{aligned} \tag{6.26}$$

which integrate out to yield the relations

$$\begin{aligned} K_{vk}(\overline{\beta}(\zeta_1) - \overline{\beta}(\zeta_*)) &= -\beta_* G(\zeta_1) \\ K_{vk}\overline{u} &= u_* G(\zeta_1) \end{aligned} \tag{6.27}$$

where $G(\zeta_1) \equiv \int_{\zeta_*}^{\zeta_1} (F(\zeta)/\zeta) d\zeta$. To eliminate the buoyancy and velocity scales, we divide the first equation by the square of the second and multiply by $z_1 - z_*$, which results in

$$-(z_1 - z_*)\frac{(\overline{\beta}(z_1) - \overline{\beta}(z_*))}{\overline{u}^2} = \frac{\zeta_1 - \zeta_*}{G(\zeta_1)} \tag{6.28}$$

in which we have rewritten the mean buoyancies as a function of the dimensional height. The left hand side is a nondimensional number called the *bulk Richardson number*, denoted henceforth by Ri. The Richardson number can be computed in terms of known quantities at the upper and lower edges of the surface layer, and gives the relative importance of potential and kinetic energy; when $|Ri| \ll 1$, buoyancy forces are negligible, and the surface layer can be treated as if it were neutral.

Given Ri, Eq. 6.28 can be solved either explicitly or iteratively for ζ_1, which allows β_* and u_* to be determined from Eq. 6.27. The buoyancy flux is then

$$\overline{w'\beta'} = u_*\beta_* = \frac{K_{vk}^2}{G(\zeta_1)^2}\overline{u}(\overline{\beta}(\zeta_*) - \overline{\beta}(\zeta_1)) \tag{6.29}$$

from which we identify the drag coefficient

$$C_D \equiv \frac{K_{vk}^2}{G(\zeta_1)^2} \tag{6.30}$$

This gives the drag coefficient for the buoyancy flux, but what we really want is the drag coefficient for sensible and latent heat flux. When the atmosphere has uniform composition, the buoyancy flux is proportional to the sensible heat flux, and so the above-derived drag coefficient can be unambigously used for sensible heat flux. When the atmosphere has nonuniform composition contributing to buoyancy, however, the drag coefficients for sensible and latent flux could in principle be different from that for buoyancy. This is sometimes handled by introducing separate empirically determined scaling functions for moisture and heat flux. Refinements of the theory are quite straightforward, and can be found in the references given in the Further Readings section for this chapter. There is a fair amount of data pertinent to similarity functions for moisture flux on Earth, where the concentration of moisture reaches a few percent of the total atmosphere. These can almost certainly be applied to other buoyancy-generating substances at similar concentrations. However, behavior of the similarity functions when the buoyancy generating component accounts for a substantial fraction of the atmosphere, as is the case for methane on Titan, is essentially unexplored. In our calculations, we will be content to use the same drag coefficient for all fluxes. In the stably stratified case the resulting errors are probably not too consequential, since we'll see shortly that the main effect of the stratification is to choke off essentially all turbulent fluxes when the Richardson number exceeds a critical value; the refinements to the theory only modify the fluxes in the rather narrow window between neutral conditions and nearly complete suppression.

To proceed further, we need to specify an explicit similarity function $F(\zeta)$. In the stably stratified case ($Ri > 0$) field and laboratory experiments can be adequately fit by functions of the form $F(\zeta) = 1 + \zeta/Ri_c$, with $Ri_c \approx 0.2$. With this definition,

$$G(\zeta_1) = \frac{1}{Ri_c}(\zeta_1 - \zeta_*) + \ln\frac{\zeta_1}{\zeta_*} = \frac{1}{Ri_c}(\zeta_1 - \zeta_*) + \ln\frac{z_1}{z_*} \tag{6.31}$$

With this simple form of the similarity function, Eq. 6.28 can be analytically solved for ζ_1 in terms of Ri, though for more complicated functions commonly in use a numerical iteration is generally required. With this form of G it is a straightforward matter to solve Eq. 6.28 for $\zeta_1 - \zeta_*$ in terms of R_i, evaluate $G(\zeta_1)$, and then compute C_D from Eq. 6.30. Note that with the assumed form of G the right hand side of Eq. 6.28 has a maximum value of Ri_c when the argument approaches infinity. Thus, there is no consistent solution when $Ri > Ri_c$. It is assumed that the turbulence is completely suppressed, and that the turbulent fluxes vanish, for more stable values of Ri. Complete suppression of turbulence is somewhat unrealistic, and some formulations use alternate forms of G so as to allow a bit of flux to persist into the very stable case. However, the applicability of Monin-Obukhov theory to very stable conditions is a matter of considerable dispute.

Carrying out the above procedure, we find that the drag coefficient is

$$C_D = \begin{cases} \frac{K_{vk}^2}{(\ln \frac{z_1}{z_*})^2}(1 - \frac{Ri}{Ri_c})^2, \text{for } 0 \leq Ri \leq Ri_c \\ 0, \text{for } Ri > Ri_c \end{cases} \tag{6.32}$$

Note that this reduces to the previously derived neutrally stratified result when $Ri = 0$. As the surface layer is made more stable, the drag coefficient goes down monotonically, and approaches zero as $Ri \rightarrow Ri_c$.

Now let's do a few examples illustrating the effects of stable surface layer physics on turbulent heat flux. First consider a melting slab of sea ice or glacier ice, in an environment where the air temperature is $280K$. Since the ice is melting, the ground temperature is pinned at the freezing point, namely $273.15K$. Since the air is warmer than the ice, there will be a flux of sensible heat from the air to the ice, which will help sustain melting. At these temperatures, it is safe to neglect the contribution of water vapor to buoyancy. Suppose that the wind is $5m/s$, the air temperature and wind have been specified at $10m$ above the ice surface, and the roughness height is .005 m. Then, for a neutrally stratified boundary layer $C_D = .0028$ and the sensible heat flux would be $125W/m^2$. With the specified parameters, the Richardson number is 0.1, and incorporation of buoyancy effects on the turbulence bring C_D down to .0007, and the sensible heat flux falls to 32 W/m^2. If we increase the air temperature to $285K$, the Richardson number increases to .17, and the increasing stability reduces the sensible heat flux to a mere $4.8W/m^2$, whereas neutral surface layer theory would have led us to expect a substantial increase in flux.

Exercise 6.4.1 How long would the sensible heat fluxes computed above take to melt through a 5 meter thick layer of water ice, if all the energy is used to melt ice?

Next, let's consider the effect on the night-time temperature inversions appearing in cold, dry climates such as Antarctica or the tropics on Snowball Earth. The radiative balance for such cases was discussed towards the end of Section 6.2.3. For example, in Earthlike conditions with $T_{sa} = 240K$, the equilibrium surface temperature is $177K$ if the only coupling of ground to atmosphere is via infrared. Using the same assumptions as in the previous example, except for the new temperatures, neutral surface layer theory would predict a sensible heat flux of over 1500 W/m^2, which of course would mean that the turbulent transfers would keep the inversion from getting nearly as strong as it would be in the purely radiative case. However, with such a large temperature difference, the Richardson number is 1.4, which leads to a complete suppression of turbulence and allows the extremely strong inversion to be realized, if there is enough time for the surface to cool down to equilibrium. An interesting aspect of this problem, however, is that the sensible heat flux is not a monotonic function of $T_{sa} - T_g$. As T_g is decreased from T_{sa}, the flux first increases, reaches a maximum, then decreases to zero as the critical Richardson number is approached. This means that there is the possibility of multiple equilibrium states – one with

turbulence and heat flux, and another with turbulence suppressed. Whether or not this happens depends on the slope of the radiative flux, but even when there aren't multiple equilibria, there tend to be abrupt transitions between turbulent and non-turbulent states as a control parameter such as solar absorption is continuously varied. This behavior is explored in Problem **??**.

The inclusion of a light, condensible substance like methane or water vapor has an important effect because it allows the surface layer to remain neutrally stratified even when the ground temperature is significantly lower than the air temperature at the upper edge of the surface layer – provided that the air there is appreciably undersaturated. To get a feel for the numbers, let's do an example involving water vapor in air. Suppose that the air temperature is $300K$ and the relative humidity is 70%. Then, using the formulae for the Richardson number and for buoyancy, we find that the surface layer is neutrally buoyant ($Ri = 0$) when $T_g = 299K$. Without the effect of water vapor on buoyancy, the Richardson number would be 0.013 assuming $\overline{u} = 5m/s$, so in this case the suppression of turbulence caused by neglect of the moisture contribution to buoyancy is small. The effect increases sharply at higher temperature, though. For $T_{sa} = 340K$, the surface layer remains neutral down to 335.5 K, and the Richardson number without the moisture contribution to buoyancy would be 0.05. The maintainence of buoyancy by light vapor will figure importantly in our estimates of precipitation rates on hot planets, in Section 6.8.

Finally, let's take a quick look at the unstable case, where the surface layer is positively buoyant. In this case, the buoyancy-driven turbulence adds to the mechanically driven turbulence due to wind shear across the surface layer. Buoyancy-driven turbulence is particularly important when the mean winds at the top of the surface layer are weak. When $\overline{u} = 0$ the neutral theory would predict that there are no turbulent fluxes, but if the surface layer has upward buoyancy, then convection should in fact be able to sustain turbulence. The case $\overline{u} = 0$ with upward buoyancy is the *free convection limit*. In this limit, there is no longer an intrinsic velocity scale separate from that defined by the buoyancy scale, and there is no longer any intrinsic length scale such as enters the Monin-Obukhov theory. Instead, we can define a velocity scale $\sqrt{\beta_* z_1}$, which is the order of magnitude of the upward velocity attained by a buoyant plume when it reaches the top of the surface layer. Since there is no longer a characteristic length scale, the buoyancy profile $\overline{\beta}(z)$ is logarithmic, just as for the neutral case, and this allows us to relate the buoyancy gradient to the difference in buoyancy between the upper and lower edge of the surface layer. Using the characteristic velocity scale, the buoyancy flux in the free convection limit can be written

$$\overline{w'\beta'} = \frac{K_{vk}^2}{(\ln(z_1/z_*))^2}(a \cdot \overline{\beta}(z_1)\ln(z_1/z_*))^{\frac{1}{2}}(\overline{\beta}(z_*) - \overline{\beta}(z_1)) \equiv C_{D,neut}U_{free}\Delta\overline{\beta} \tag{6.33}$$

where $C_{D,neut}$ is the usual neutral drag coefficient, U_{free} is a characteristic buoyancy velocity and a is a nondimensional constant whose value has been empirically determined to be about 15. To apply this result to the flux of other quantities, such as latent or sensible heat, we use the same drag coefficient and U_{free}, but replace $\Delta\overline{\beta}$ with the difference in the quantity whose flux we wish to obtain. Note that because of the definition of U_{free}, the buoyancy flux scales like the $\frac{3}{2}$ power of buoyancy. However, by casting the flux formula in the above form we can see that it is just like the neutral case, but with a buoyancy velocity replacing the mean wind. As an example, consider a dry case with ground temperature of $305K$, air temperature $300K$, Earth gravity, a surface layer thickness of $10m$ and a roughness length of $.001m$. With these parameters $U_{free} = 15m/s$, so buoyancy driven turbulence becomes a significant player when the mean wind is $15m/s$ or weaker. Based on this estimate, it is clear that convection will lead to large sensible and latent heat fluxes whenever the ground temperature tries to get much bigger than the overlying air temperature. The upshot is that it is quite easy for the ground to get a lot colder than the overlying air, because of the inhibition of turbulence in stable conditions, but it is harder for the ground to get much hotter than the overlying air.

The general unstable case with nonzero mean wind can be treated similarly to the way we treated the stable case, though it is necessary to adopt different scaling functions for wind and momentum and the form of the functions is sufficiently complicated that a Newton's method iteration is generally needed in order to solve for ζ_1. In addition, the scaling functions most commonly in use do not, in fact, reduce to the correct free-convection limit when the mean wind becomes weak. This point, together with a resolution of the problem, is discussed in the paper by Delage and Girard given in the Further Readings section for this chapter. A simple expedient for dealing with the general unstable case, however would be to compute the turbulent fluxes for both the free-convection and neutral limits, and then take whichever of the two is greater. This procedure by definition gives the correct free convection limit, and also eliminates the chief shortcoming of the neutral theory, namely the spurious vanishing of fluxes when mean winds become very weak. One can easily implement this formulation by computing fluxes using the usual neutral C_D, but replacing $\overline{u}$ with U_{free} when $\overline{u} < U_{free}$.

6.5 Joint effect of the fluxes on surface conditions

Including turbulent heat fluxes, the surface energy budget can be written

$$0 = F_{rad} - F_{sens} - F_L \tag{6.34}$$

where F_{rad} is the net radiative flux into the surface, given by

$$F_{rad} = (1 - \alpha_g)S_g + \sigma e_a e_g T_{sa}^4 - \sigma e_g T_g^4 \tag{6.35}$$

Without turbulent fluxes, the surface budget would be $F_{rad} = 0$. F_{rad} in isolation can drive the ground temperature to be either larger or smaller (and perhaps *much* larger or smaller) than the air temperature, according to the circumstances discussed in Section 6.2. Sensible heat flux always drives the ground temperature and air temperature to become identical, whereas latent heat flux drives the ground temperature to be colder than the air temperature, by an amount that depends on the boundary layer relative humidity. When all three fluxes act in concert the resulting behavior depends on the relative importance of the fluxes.

We'll begin our tour of the range of possible behaviors by discussing how the surface balance is accomplished for typical conditions in the Earth's tropical oceans. Take $T_{sa} = 300K$, $C_D = .0015$, $U = 5m/s$ and $h_{sa} = 80\%$. We'll assume the absorbed solar radiation $(1-\alpha_g)S_g$ is $320W/m^2$, which is typical of clear-sky conditions over the tropical ocean. To determine the back-radiation, we need e_a. At tropical temperatures in the moist case, this coefficient is not very sensitive to CO_2, and has a value of about .9. The terms making up the surface balance are shown in the left panel of Figure 6.2. As noted previously, the equilibrium ground temperature would be exceedingly large without turbulent heat flux. In the figure, the no-turbulence equilibrium occurs where F_{rad} crosses zero, at around $336K$. Adding sensible heat flux to the budget makes the slope of the flux curve more negative, and brings the equilibrium ground temperature down to $316K$. Adding in evaporation steepens the curve yet more, and brings the ground temperature down to $303K$, which is only slightly warmer than the $300K$ temperature of the overlying air. At the equilibrium point, the dominant balance is between the evaporation ($206W/m^2$) and the absorbed solar radiation ($320W/m^2$), leaving only $114W/m^2$ to be balanced by the other terms. The sensible heat flux is weak because the ground temperature and air temperature are nearly identical, which also makes the net infrared cooling of the surface weak given that $e_a \approx 1$.

Next we'll discuss a typical set of Earth polar or midlatitude winter conditions. We set the absorbed solar flux $(1 - \alpha_g)S_g$ to $100W/m^2$, taking a low value on account of the high albedo of

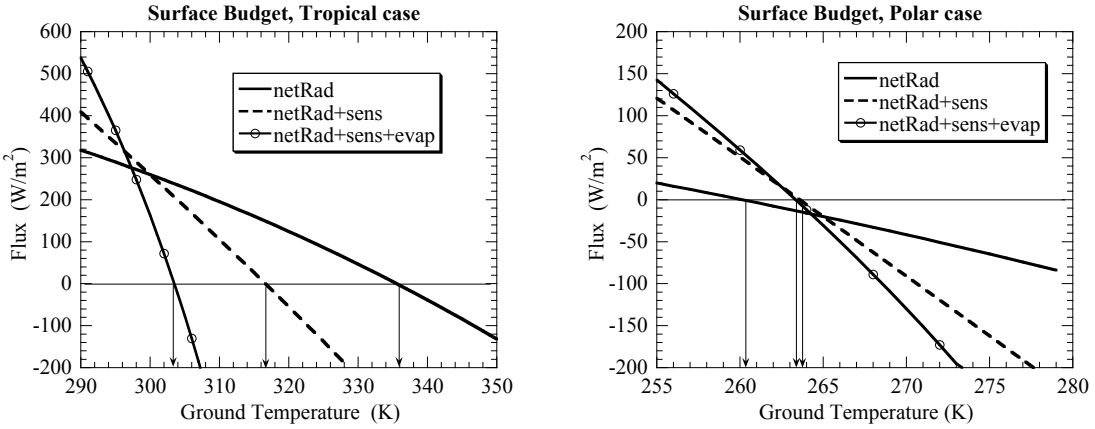

Figure 6.2: Terms in the surface balance for conditions representing the tropical oceans on Earth (left panel) and cold polar conditions on Earth (right panel). See text for the parameters of the calculation.

snow or ice and the reduced solar flux received at high latitudes. We'll set $T_{sa} = 265K$, in which case $e_a \approx .6$ with $300ppmv$ of CO_2 in the atmosphere. The remaining parameters are held at the same values used in the tropical case. The main differences from the tropical case are that in the cold case the latent heat flux and the infrared back-radiation are weaker – the latter doubly so because of the lower air temperature and the lower e_a. The right panel of Figure 6.2 shows that because of the weak solar radiation and the weak back radiation, the radiative equilibrium surface temperature is nearly $5K$ colder than T_{sa}, in contrast to the tropical case. The situation here is a less extreme version of the night-time radiative equilibrium temperature considered in Section 6.2.3. Since e_a is fairly small the temperature plummets at night when $(1 - \alpha_g)S_g = 0$. In the present case, S_g doesn't vanish, but its weak value is insufficient to warm up the ground temperature to the point where it exceeds the air temperature. This is the typical daytime condition in high latitude winter over ice and snow. Warm air imported from low latitudes helps to keep T_{sa} from getting too cold in the polar and midlatitude winter, but the weak sunlight and weak back-radiation leave the ground colder.

Since the radiative equilibrium ground temperature in the cold case is colder than the air temperature, adding in sensible heat flux conveys heat from the atmosphere to the ground, warming the ground up to just over $263K$. The sublimation is weak at such cold temperatures, and causes little additional change in the surface temperature. While the dominant balance in the tropical case was between solar heating and evaporative cooling, the dominant balance in the cold case is radiative, with slight modifications due to sensible heat flux. For any given air temperature, the amount by which the ground temperature departs from the air temperature depends on the absorbed solar radiation, but the sensible heat flux always pulls the ground temperature back towards equality with air temperature. For example, at higher latitudes or deeper in the winter or near sundown, we might take $(1 - \alpha_g)S_g = 50W/m^2$. In this case the radiation-only ground temperature is $246.6K$, which is substantially below the air temperature; however, addition of sensible heat flux brings the ground temperature back up to $260K$. Nearer to noon, or as summer approaches, we might have $(1 - \alpha_g)S_g = 150W/m^2$. In this case, the radiation-only ground temperature is $271.8K$; again addition of sensible heat flux brings the ground temperature closer to air temperature, in this case by cooling the ground to $267.1K$, rather than warming it. Note

that in these calculations of the effects of sensible heat transfer, the drag coefficient C_D was held constant. Incorporating the inhibition of turbulence in stably stratified layers has the potential to substantially reduce the warming effect of sensible heat fluxes, particularly when the absorbed solar radiation is weak, since the inversion is strongest in those cases. This is explored in Problem **??**.

Next let's estimate the maximum daytime temperature over a subtropical desert on Earth. Solid surfaces like sand or rock take little time to reach equilibrium, and so the maximum temperature can be estimated by computing the equilibrium temperature at local solar noon. Using the present Earth solar constant and a relatively high albedo of .35 (typical of Sahara desert sand), the absorbed flux is about $890 W/m^2$. Over the interior of a dry desert, there should be little moisture in the boundary layer, so set $e_a = .72$ corresponding to a boundary layer relative humidity of 20%. Finally, we take $T_{sa} = 300K$. In these circumstances the radiative equilibrium ground temperature is a torrid $383K$ – hot enough to boil water. When sensible heat flux is added into the budget, heat is transferred from the ground to the air, moderating the surface temperature. Taking a relatively high drag coefficient $C_D = .003$ on account of the roughness of land surfaces, the equilibrium ground temperature is brought down to $330K$ if the surface layer wind speed is $5m/s$. The temperature approaches the radiative temperature as the wind is made weaker; for example when the wind is reduced to $2.5m/s$ the temperature increases to $349K$. Consistent with these estimates, the hottest satellite-observed ground temperatures do indeed occur in subtropical deserts, and are near $340K$. With a wind of $5m/s$, making the ground moist and turning on evaporation brings the equilibrium temperature down from $330K$ to $306K$. The general lesson is that dry surfaces heat up greatly during the daytime. Their maximum temperature can greatly exceed the overlying air temperature, especially when the wind is light. This can contribute to the *urban heat island* effect, since constructed environments often replace moisture-holding surfaces with low albedo impermeable surfaces like asphalt, which hold little water and dry out quickly. The surface heating also leads to amplified climate change over land, in circumstances where a formerly moist soil becomes dry, or *vice versa*.

We'll conclude this section with a discussion of the linearized form of the surface balance, which enables simple, explicit solutions for the temperature jump across the surface layer. Using the linearized flux coefficients defined previously, the temperature jump is simply

$$\Delta T \equiv T_g - T_{sa} = \frac{(1-\alpha_g)S_g - e_g \cdot e^* \sigma T_{sa}^4 - E_o}{b_{ir} + b_{sens} + b_L} \tag{6.36}$$

The numerator in this expression is the energy imbalance the surface would have if the ground temperature were equal to the overlying air temperature. It can be either positive or negative and its sign determines the sign of ΔT, since all three terms in the denominator are positive. The denominator is a stiffness coefficient. For any given magnitude of the numerator, the denominator determines how much the ground and air temperature differ. In other words, when the denominator is large, the ground and air temperature are very tightly coupled, but when the denominator is small, they can vary independently. The coupling constants will prove useful in making simple models of the seasonal and diurnal cycle of temperature, as we shall do in Chapter 7.

Table 6.2 shows some typical coupling coefficients for Earth and Titan. Since these are derived by linearizing around $T_g = T_{sa}$, the effects of buoyancy on C_D to not affect b_{sens}. Moreover, since both methane and water vapor are positively buoyant in the background gas, the surface layer is in the unstably stratified regime, so that suppression of turbulence does not enter into the picture. The unstable buoyancy effects do cause C_D to increase slightly with T_g, and this would slightly increase b_L. Note that b_{sens} is nearly independent of temperature; the slight variation is due to the effect of the composition on mean specific heat and on surface layer density. The

	T	b_{ir}	b_{sens}	b_L
Water+Air	250.	3.54	21.00	2.76
Water+Air	280.	4.98	18.89	19.72
Water+Air	300.	6.12	17.99	57.95
Water+Air	320.	7.43	17.84	147.0
$CH_4 + N_2$	85.	0.14	95.07	161.36
$CH_4 + N_2$	90.	0.17	92.55	287.29
$CH_4 + N_2$	95.	0.19	91.88	365.75

Table 6.2: Some typical surface flux coupling coefficients. The "Water+Air" cases are done under Earthlike conditions, with a 1 *bar* Earth air noncondensible background. The $CH_4 + N_2$ cases are done under Titanlike conditions, with a 1.5 *bar* noncondensible N_2 background. Both cases were done with 70% relative humidity at the top of the surface layer, $U = 10m/s$ and C_D held fixed at 0.0015 . Units for all the coupling coefficients are W/m^2K

radiative coupling coefficient b_{ir} increases gently with temperature, but the latent heat coefficient b_L increases sharply, owing to the exponential behavior of Clausius-Clapeyron. For Earth, the sensible heat transfer dominates the coupling in cold conditions, which apply near the poles in climates like the present and globally for Snowball conditions. As temperature increases, latent heat fluxes come to increasingly dominate the coupling. In tropical conditions for the present Earth climate, evaporation accounts for 71% of the total coupling coefficient of 82.1 W/m^2K. To get an idea of how tightly coupled the ground temperature is to the overlying air temperature in this case, we note that an increase of 40 W/m^2 in absorbed solar radiation at the ground (arising perhas from a drastic decrease in cloudiness) could be accomodated by an increase of ground temperature by under a half a degree. In these circumstances, the most effective way to increase the surface temperature is not to alter the surface energy budget, but rather to increase the temperature of the atmosphere. This is the main way increases in greenhouse gases increase the ground temperature. As temperature increases beyond modern tropical values, evaporation becomes even more dominant, and coupling becomes even tighter.

In the Titan case, the infrared coupling is almost completely insignificant. It is interesting, however, that the sensible heat coupling coefficient is quite significant, owing to the high density of Titan's atmosphere. We have already noticed that on Titan evaporation easily dominates the weak absorbed solar radiation. Evaporation makes the numerator of Eq. 6.36 strongly negative if the relative humidity is appreciably less than 100%, which drives a temperature inversion at the surface. The dominant balance in this case is between evaporative cooling of the ground and transfer of sensible heat from the warmer atmosphere and the colder ground. In this situation, the suppression of turbulence by stable boundary layer effects, can play an important role in determining the strength of the inversion (see Problem **??**). The Earth's tropics is not hot enough to be in this regime, but on a much hotter Earth the increase of water vapor would lead to very similar effects.

6.6 Global warming and the surface budget fallacy

A common fallacy in thinking about the effect of doubled CO_2 on climate is to assume that the additional greenhouse gas warms the surface by leaving the atmospheric temperature unchanged, but increasing the downward radiation into the surface by making the atmosphere a better infrared emitter. A corrollary of this fallacy would be that increasing CO_2 would not increase temperature

if the lower atmosphere is already essentially opaque in the infrared, as is nearly the case in the Tropics today, owing to the high water vapor content of the boundary layer. This reasoning is faulty because increasing the CO_2 concentration while holding the atmospheric temperature fixed reduces the OLR. This throws the top-of-atmosphere budget out of balance, and the atmosphere must warm up in order to restor balance. The increased temperature of the whole troposphere increases all the energy fluxes into the surface, not just the radiative fluxes. Further, if one is in a regime where the surface fluxes tightly couple the surface temperature to the overlying air temperature, there is no need to explicitly consider the surface balance in determining how much the surface warms. Surface and overlying atmosphere simply warm in concert, and the top-of-atmosphere balance rules the roost.

Arrhenius properly took both the top-of-atmosphere and surface balances into account in his estimate of the effect of doubling CO_2, though he did so using a crude one-layer model of the atmosphere. Guy Stewart Callendar (1938) and Gilbert Plass (1959) employed more sophisticated multilevel models, but when it came to translating their radiation results into surface temperature change both got mired in the surface budget fallacy. The prime importance of the top-of-atmosphere balance was emphasized with crystal clarity in Manabe's work of the early 1960's, but one still encounters the surface budget fallacy in discussions of global warming from time to time even today.

Figure 6.3 shows how the budgets change when CO_2 is doubled from 300 $ppmv$. The case shown is typical fo the present Earth's tropics, for which water vapor makes the boundary layer optically thick. The system starts off in balance, at a surface temperature of $300K$. If CO_2 is immediately doubled, the downward radiation into the surface increases by a mere 1.2 W/m^2. However the OLR goes down by over 4 W/m^2. The atmosphere-ocean system is receiving more solar energy than it is losing, and so it warms up. The top-of-atmosphere balance is restored when the surface air temperature has warmed to $302K$. This increases the radiation into the ground by an additional 7.3 W/m^2. Part of this increase comes from the fact that the warmer boundary layer contains more water vapor, and therefore is closer to an ideal blackbody. Most of the increase, however, comes about simply because the low level air temperature T_{sa} increases, and hence σT_{sa}^4 increases along with it. This increase occurs even if the boundary layer is an ideal blackbody – i.e. completely opaque to infrared. In addition, the increase of T_{sa} would increase the latent and sensible heat fluxes into the surface if the surface temperature were to stay fixed, and this increase also contributes to the warming of the surface.

There are a few situations in which the detailed surface balance could have a significant effect on surface warming. This can only happen in the weakly-coupled regime. In that regime ΔT can be fairly large, and changes in ΔT can add to whatever warming is directly caused by the atmospheric warming that comes from satisfying the top-of-atmosphere budget. For example, if land dries out, the loss of evaporation will cause ΔT to increase. Conversely, if a formerly dry area becomes moist, ΔT would decrease, moderating the surface warming. The weakening of the low-level inversion in Antarctica can play a crucial role in Antarctic surface climate change. Finally, it should be noted that when the atmosphere is optically thick, ΔT does not affect the OLR. However, when the atmosphere is somewhat transparent to infrared from the surface, an increase in ΔT increases the OLR a bit, so that the atmosphere doesn't have to warm up quite as much as one thought in order to bring the top-of-atmosphere budget into balance.

The relative roles of the surface budget and the top-of-atmosphere budget in determining surface temperature change upon doubling of CO_2 are further explored in Problems **??** and **??**.

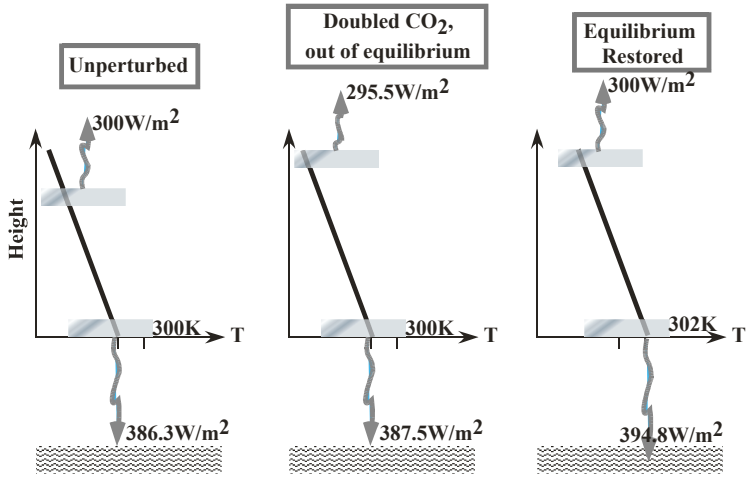

Figure 6.3: Changes in top-of-atmosphere and surface radiative fluxes upon doubling CO_2. Calculations were carried out with the `ccm` radiation model assuming the atmosphere to be on a moist adiabat patched to an isothermal 180 K stratosphere. The low level relative humidity is fixed at 80%, while the relative humidity in the free troposphere is 50%.

6.7 Mass balance and melting

When the surface consists of a solid ice which can undergo melting, the surface balance works in a somewhat different way once the ground was warmed to the melting temperature. We can no longer solve Eq. 6.34 for T_g, since T_g cannot rise above the melting temperature so long as there is any ice left to melt. Instead, we compute the surface residual at the melting temperature T_f, i.e.

$$F_{net}(T_{sa}, T_f) = F_{rad}(T_{sa}, T_f) - F_{sens}(T_{sa}, T_f) - F_L(T_{sa}, T_f) \tag{6.37}$$

If $F_{net}(T_{sa}, T_f) > 0$, then the energy flux F_{net} is available for melting. In that case, the mass melted per unit time per unit area is given by F_{net}/L_f, where L_f is the latent heat of fusion. Often this is converted to liquid equivalent depth per unit time by dividing by the density of the liquid phase. Expressed that way, the rate corresponds to the rate of growth of liquid layer that would be caused by the melting, if the liquid did not run off to some other place. Melting is a very powerful means of ablation of ice, be it mountain glacier, ice sheet or sea ice. The latent heat of fusion is much smaller than the latent heat of sublimation, so a given amount of energy can ice into rapidly movable form much more rapidly by melting than by sublimation. For example, the ratio of latent heats is 0.118, so a given amount of energy can get rid of 8.5 times as much ice by melting as it can by sublimation. Sublimation always carries the vapor away from the ice surface, but for melt to actually become realized ablation, the meltwater needs to go away somewhere. It may flow to the base of a glacier through an abyss called a *moulin*, or it may run away to a melt pond and form a temporary lake. It may also percolate into the snow and re-freeze, releasing latent heat int the process. In that case the melting actually constitutes an energy transport mechanism rather than a true ablation.

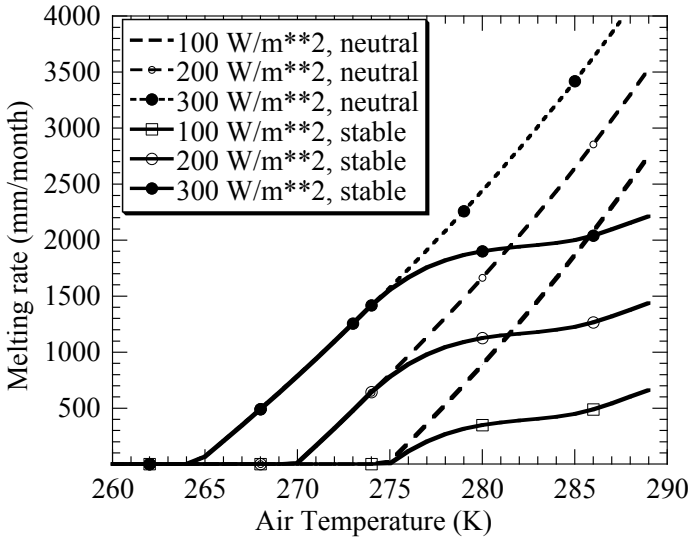

Figure 6.4: Melting rate in liquid water equivalent as a function of air temperature. Results are given for three different values of absorbed solar radiation. The calculations were performed with $U = 5m/s$, $z_* = .00033m$, and $h_{sa} = .8$.

Figure 6.4 shows the melting rate as a function of air temperature for three different values of surface absorbed solar radiation (see the caption for the rest of the conditions). To illustrate the importance of stable surface layer physics, each calculation is done in two ways: using Monin-Obukhov theory and using neutral surface layer theory. Note that melting can begin even when the air temperature itself is below freezing; this is simply because the energy balance allows the ground to be warmer than the air, if there is a sufficient supply of solar absorption. Note also that even for a fixed air temperature, the melting increases with solar absorption. This shows that the ablation of a glacier can be affected by the solar radiation, even if air temperature does not change. Various processes can change the absorbed solar radiation, among them clouds, changes in the Earth's orbital parameters (see Section 7), and the fact that fresh snow is more reflective than old ice. For any given amount of absorbed solar radiation, the melting rate increases dramatically as air temperature is increased, because this increases the delivery of heat to the surface by sensible heat flux and infrared heat flux. The development of a stable surface layer when the air temperature gets large sharply limits the increase of melting. In this regime, increasing the wind speed very strongly increases melting, since higher winds favor lower Richardson numbers and higher drag coefficients.

If some geological indicator of past glacial behavior tells us that a mountain glacier was more extensive at some particular time in the past, should we conclude that it was colder then or that it simply snowed more at that time? Similarly, if we see mountain glaciers retreating worldwide at present, should we take that as an indication of a warming climate, or of a reduction in snowfall? The sensitivity of melting rate to temperature has a great bearing on this question. Consider the case with 200 W/m^2 of absorbed solar radiation in Figure 6.4. Increasing air temperature from $270K$ to $274K$ increases the melting rate from zero to $650mm/mo$. To offset such an increase in melting, the snow accumulation rate would thus have to increase by $650mm/mo$. This is a very substantial increase. To put it in perspective, the required *increase* in precipitation is over

three times the *maximum* monthly precipitation rate observed in the past few years in central Iceland – which is a very snowy place. Recall, too, that the melting rate we have stated is in liquid water equivalent depth. Snow has low density, and the actual thickness of the corresponding snow accumulation would be 6.5 *meters* per month or more. It is not impossible for glacier extent to be affected by precipitation, but in situations where melting occurs, it takes a truly enormous change in precipitation to have the same impact as a rather modest change in temperature.

Melting is a very nonlinear process, which acts as a kind of rectifier capable of turning a fluctuating temperature signal into a secular growth or decay of a mountain glacier or continental ice sheet. Melting turns on when the air temperature approaches freezing, and greatly accelerates as temperature becomes warmer. On the other hand, the melting turns off sharply when the air temperature falls much below freezing. In consequence, ablation of ice cares little about just how cold it gets during the depth of winter, but a great deal about the length and warmth of the summertime melt season. This makes the growth and decay of midlatitude and polar mountain glaciers or continental ice sheets very sensitive to what is going on during the melt season.

6.8 Precipitation-temperature relations

There is more to climate than temperature, and for atmospheres that contain a condensible substance – e.g. water on Earth or methane on Titan – the precipitation rate is of as much interest as the temperature. Aside from the role of rainfall in making land habitable on Earth, there are many reasons for being interested in precipitation. For example, it is the precipitation of snow that feeds the growth and flow of glaciers. Precipitation of water on Earthlike planets exerts a controlling influence on the chemical weathering processes that ultimately control atmospheric CO_2, as will be discussed in Chapter 8. In the long term, the rate at which a condensible substance precipitates from an atmosphere must be balanced by the rate at which that substance evaporates or sublimates from a reservoir at the planet's surface – an ocean or glacier. Since latent heat flux can be turned into mass flux upon dividing by the appropriate latent heat, much can be learned about the evaporation or sublimation rate by a careful examination of the surface energy budget. To streamline the prose, we'll generally use the term "evaporation" to refer to evaporation or sublimation in the following, with the understanding that when the surface in question is ice the phase change is actually sublimation.

We begin by writing the surface balance in the form

$$F_L(T_{sa}, T_g) = [(1-\alpha_g)S_g - \sigma e^* T_{sa}^4] + c_p \rho_s C_D U (T_{sa} - T_g) + \sigma(T_{sa}^4 - T_g^4) \qquad (6.38)$$

In this equation we have assumed $e_g = 1$ for simplicity. The functional form of F_L given by Eq. 6.13. As usual, this equation must be solved for T_g given T_{sa} and the other parameters affecting the surface budget. We can distinguish two regimes: the weak evaporation regime and the strong evaporation regime. The weak evaporation limit is defined by the condition $F_L(T_{sa}, T_{sa}) \ll (1-\alpha_g)S_g$. This condition guarantees that F_L will be negligible compare to the surface solar flux as long as the solution does not require that T_g be enormously greater than T_{sa}. Given the form of the Clausius-Clapeyron relation, the weak evaporation regime applies at sufficiently low temperatures, since in that case the atmosphere can carry little vapor even when it is saturated. The notion of "low" vs "high" temperature must be understood with reference to the volatility of the substance undergoing the phase change. For methane, $95K$ is a "high" temperature in this sense, but for water vapor even $250K$ is a "low" temperature.

In the weak evaporation limit, we can set the left hand side of Eq. 6.38 to zero when solving for T_g, and then use the resulting ground temperature to evaluate the evaporation by plugging

it into the formula F_L. Since the latent heat flux is small, leaving it out of the surface balance causes only small errors in T_g. In this limit, the evaporation is not significantly constrained by the energy supply. In the regime where $T_g \geq T_{sa}$ – i.e. when the term in square brackets in Eq. 6.38 is positive – the exponential increase of saturation vapor pressure with temperature yielded by Clausius-Clapeyron leads to a roughly exponential increase of the latent heat flux with temperature, as T_{sa} is increased. Even when the surface absorbed solar radiation is so weak that an inversion forms in the surface layer, the control exerted by Clausius-Clapeyron is so strong that one still tends to get an exponential increase with temperature, unless the inversion gets very strong. The behavior is explored quantitatively in Problems **??** and **??**.

At high temperatures, when the energy carried away by latent heat flux has a strong effect on the ground temperature, the behavior is very different. Intuitively, one expects that the evaporation can't increase beyond the point where the entirety of the absorbed solar radiation goes into evaporating material from the surface. It's not quite as simple as that, owing to the effect of sensible and radiative heat fluxes, but a constraint very similar to this does come into play. Let's assume that the condensible substance is like water vapor and makes $e^* \approx 0$ in warm conditions. Then the surface balance becomes

$$F_L(T_{sa}, T_g) = (1 - \alpha_g)S_g + c_p \rho_s C_D U(T_{sa} - T_g) + \sigma(T_{sa}^4 - T_g^4) \tag{6.39}$$

If the second two terms on the right hand side were not there, we'd have the result that the latent heat flux is equal to the surface absorbed solar radiation. The two additional terms are positive when $T_g < T_{sa}$, and can thus allow the latent heat flux to somewhat exceed the available solar forcing under circumstances when an inversion can form at the surface. The strength of this inversion determines the amount of "excess evaporation" that can be sustained.

We can show that an inversion must always form in the strong evaporation limit, and also put a strict bound on the temperature difference across the inversion. First, note that when $T_g = T_{sa}$ the last two terms in Eq. 6.39 vanish, while by definition of the strong evaporation limit F_L is much greater than the remaining solar term. Thus, when $T_g = T_{sa}$, the left hand side exceeds the right hand side. Next, recall that latent heat flux vanishes when the saturation vapor pressure corresponding to the ground temperature equals the vapor pressure at the top of the surface layer. Thus, if the relative humidity is below 100%, $F_L(T_{sa}, T)$ will vanish at some temperature $T < T_{sa}$ which we will call T_o. Actually, it is possible that stable boundary layer physics will extinguish the turbulence at some temperature that is somewhat warmer than the temperature at which the gradient of vapor pressure vanishes. In any event, there is a $T_o < T_{sa}$ where $F_L(T_{sa}, T_o) = 0$. This temperature will be a function of T_{sa} and the other surface layer parameters, as well as the the thermodynamic properties of the atmosphere. Because the left hand side vanishes at $T_g = T_o$, while the right hand side is positive, it follows that the left hand side is less than the right hand side. Together with the previous result, we now know that there is a solution to the surface balance equation with $T_o < T_g < T_{sa}$. The maximun strength of the inversion is $T_{sa} - T_o$, and this determines the maximum excess evaporation, through determining the maximum possible size of the second two terms on the right hand side of Eq. 6.39. It can be shown that $T_{sa} - T_o$ increases very slowly with T_{sa} (Problem **??**), whence we conclude that the excess evaporation must increase only slowly with temperature.

In Figure 6.5 we show the latent heat flux as a function of T_{sa}, obtained by solving Eq. 6.38 for T_g using a simple Newton's method iteration. The calculation was carried out for a water/air atmosphere on Earth with $(1 - \alpha_g)S_g = 200 W/m^2$, 80% relative humidity in the free atmosphere $U = 5 m/s$ and a constant $C_D = .0015$. As expected from our analysis of the weak and strong evaporation limits, the flux grows approximately exponentially at low temperatures, but the growth levels off and becomes much weaker once the latent heat flux exceeds the absorbed

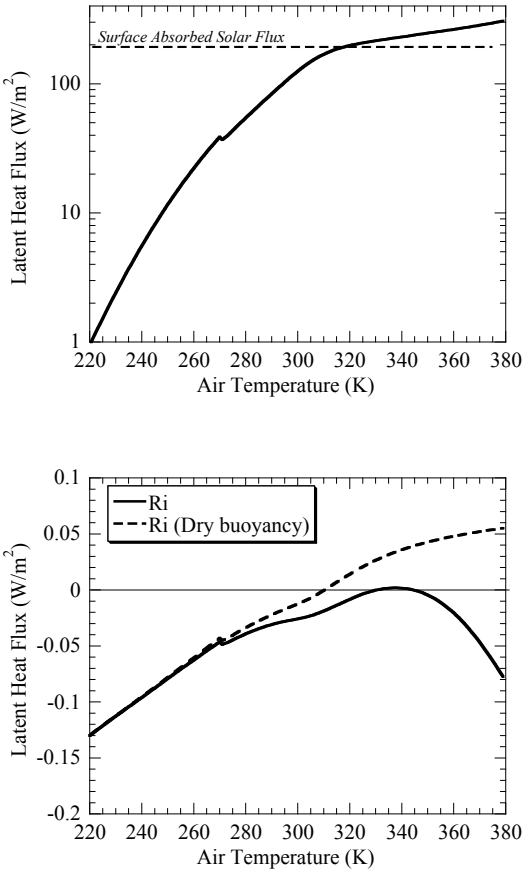

Figure 6.5: Left panel: Latent heat flux computed from the surface energy balance assuming surface absorbed solar flux = $200W/m^2$, relative humidity of 80%, $C_D = .0015$, and $U = 5m/s$. Right panel: Corresponding Richardson number computed with and without the contribution of water vapor to buoyancy.

solar radiation. Notably, the latent heat flux grows by more than two orders of magnitude as the air temperature is increased from $280K$ to $300K$, but hardly doubles as it is increased an equal amount to $380K$. An examination of the dry Richardson number (proportional to $T_{sa} - T_g$) shows that an inversion develops in this circumstance, and becomes stronger as the temperature increases. If water vapor were not positively buoyant, the inversion would become stable, limiting the turbulence and reducing the evaporation even more. However, the buoyancy of water vapor allows the surface layer to remain unstable despite the inversion, especially at high temperatures. In other circumstances permitting a stronger inversion, the surface layer can become stable despite water vapor buoyancy (see Problem **??**)

The preceding results shed some light on precipitation rates of water in both cold and warm climates. To turn the latent heat fluxes into precipitation rates, note that $1W/m^2$ of latent heat flux is equivalent to 1.21 cm/yr of liquid water equivalent precipitation if the flux is due to sublimation, or 1.26 cm/yr if the flux is due to evaporation. It is sometimes erroneously supposed

that in the cold conditions of a Snowball Earth, the hydrological cycle shuts down. Let's estimate the precipitation rate for the Snowball tropics. Taking the mean surface absorbed solar flux to be $130 W/m^2$ as is reasonable for ice subject to tropical insolation, and $T_{sa} = 240K$, in equilibrium we find that the ice surface temperature is $241.28K$. With these temperatures, the latent heat flux is $1.81 W/m^2$, which translates into a precipitation rate of $2.2 cm/yr$ liquid water equivalent. This may not sound like much, but the Snowball can last a very long time. Given 100,000 years to accumulate, this trickle of snowfall can build a glacier $2.2 km$ high, which is high enough to flow significantly. Thus, there is no essential incompatibility with cold Snowball conditions and geological evidence for active glacier flow. The instantaneous noontime precipitation can be much higher, because it is driven by greater solar flux. Only the mean is relevant for building glaciers, but the relatively heavy noontime sublimation, followed by snowfall as night approaches, can be important in modifying the surface conditions and covering dusty, dark ice with fresh, reflective snow during part of the day.

Turning attention next to the warmer conditions of the Earth's present tropical oceans, we take the absorbed solar radiation to be 200 W/m^2 and assume $T_{sa} = 300K$. Under these circumstances $T_g = 301K$ and the latent heat flux is 125 W/m^2, which translates into a precipitation rate of 156 cm/yr. This is reasonably close to the observed tropical precipitation rate. Now suppose that we introduce a high cloud which reflects a lot of sunlight back to space, but which has such a strong greenhouse effect that the change in OLR compensates, leaving the top-of-atmosphere radiation budget unchanged; as we saw in Chapter 5, Earth's actual high tropical clouds do something approximating this idealization. Since the air temperature is determined primarily by the top-of-atmosphere balance in the optically thick limit, we can keep T_{sa} fixed at $300K$ as in the previous case. However, the surface absorbed solar will be reduced, say to $100 W/m^2$, while the downwelling infrared is essentially unaffected by the cloud because the tropical atmosphere is optically thick. Under these circumstances, with reduced surface solar flux the surface temperature falls only modestly, to $T_g = 299K$. However, the latent heat flux falls dramatically, to $57 W/m^2$, by about the same proportion as the reduction in surface solar flux. This example shows that in the optically thick warm regime where the surface is tightly coupled to the air by latent heat exchange, the surface energy budget has little influence on temperature. For the purposes of estimating temperature, we could do pretty well by simply assuming that the ground temperature equals that of the overlying air. However, changing terms in the surface budget – as we did here by reducing the surface solar flux – has a profound effect on precipitation. In brief, in warm tropical conditions, the surface energy budget tells us about precipitation, while the top-of-atmosphere energy budget determines the temperature.

6.9 Simple models of sea ice in equilibrium

There are many circumstances in which one would like to know the thickness attained by a layer of ice floating on an ocean once it comes into a state of thermal equilibrium. This problem is relevant to the state of the sea ice cover which forms in the polar regions on Earth today and in other icy climates. It is also relevant to the thick-ice regimes prevailing on a globally glaciated Snowball Earth. Another application is the determination of the ice-crust thickness for icy moons such as Europa, which consist of a crust of water ice floating on a deep brine ocean. In this section, we'll lay out some elementary models which shed light on the determination of ice thickness. While the physics set forth here would apply equally well to the freezing of any liquid whose solid form has lower density than the liquid form, in practice, the condition that the ice floats for the most part restricts the applicability to water. An important exception to this is the determination of the thickness of the crust of rocky planets, which can be viewed a a form of "rock-ice" supported by

a more fluid interior.

In order to proceed we need to know a bit about the flux of heat through an immobile solid. Heat is conducted through such a substance through collisions of molecules with one another, which propagate information about changes in temperature in one part of the body to the remainder of the body. Both experiment and theory show that in most circumstances, the flux of heat is proportional to the temperature gradient, and the flux is in the direction opposite to the gradient. In other words, the heat flows in such a way a to try to wipe out temperature gradients and make the body isothermal. The constant of proportionality is called the *thermal conductivity*, and we shall call it κ_T. The thermal conductivity is determined by molecular properties of the solid, the specific heat, and the density. Low density substances generally have lower thermal conductivity, since there is less mass available to transmit heat. The thermal conductivity of various substances will be discussed at greater length in Chapter 7. For now, it will suffice to know that for water ice $\kappa_T \approx 2.24 W/m \cdot K$, and that of new fluffy snow is $\kappa_T \approx .08 W/m \cdot K$.

If there are no internal sources or sinks of heat within the ice layer, then the heat flux must be constant once the system reaches equilibrium. The value of the flux is set by the heat flux delivered to the bottom of the ice. In some cases, for example the heat flux through a quiescent ocean for a globally glaciated Snowball Earth, the flux would be just the geothermal heat flux escaping the interior of the planet. In other cases, for example in the case of sea ice or shelf ice abutting an open ocean, the flux would be the much larger value delivered by dynamic ocean heat transport – the delivery of heat in the form of relatively warm water by ocean currents. Regardless of the source, we will call this flux F_i. In this section, we'll denote the temperature profile within the ice as $T(z)$, with $T(0)$ being the temperature of the ice or snow surface (also called T_g) and $T(-h)$ being the temperature at the base of the ice, where h is the ice thickness. Then, the constant flux requirement yields the following differential equation for the temperature profile:

$$-\kappa_T \frac{dT}{dz} = F_i \tag{6.40}$$

Note that this equation remains valid even if the thermal conductivity varies with depth. For example, snow has much smaller conductivity than ice, owing to it's low density ahd the immobilization of air in the pore space. Hence,if a layer of ice were blanketed with snow, the temperature gradient within the snow layer would be much steeper than the temperature gradient within the rest of the ice. Suppose that the ice layer has not frozen all the way to the rock, so that the ice is floating on a layer of the same liquid (generally water) which freezes to make the ice. Where the ice is in contact with the liquid, the temperature of the ice must equal the freezing point of the liquid, which we will call T_f. Thus, the bottom boundary condition on temperature is $T(-h) = T_f$. Note that the freezing point of sea water or any other brine is lower than the freezing point of pure water.

The energy balance at the ice surface impose another condition on the temperature profile. This energy balance is identical to the surface energy budget given in Eq. 6.34, except that one must add in the contribution from the heat flux through the ice. Thus, at $z = 0$ we require

$$0 = F_i + F_{rad} - F_{sens} - F_L \tag{6.41}$$

The internal flux is usually small, and makes a negligible contribution to the surface balance. In most cases, we can drop the term and compute the ice surface temperature as if it were not there. This equation determines T_g (which is now the ice surface temperature, called $T(0)$) as before, and the inclusion of F_i would only make the ice surface temperature ever so slightly warmer than it otherwise would have been without heat diffusion through the ice.

Sublimation takes away mass as well as heat, and in general this mass loss must be taken into account when formulating the conditions for equilibrium thickness. Let's assume first that there is no net mass loss or gain at the surface. This could be because the temperature is so low that sublimation is negligible, or it could be because all the sublimated mass precipitates back out onto the surface locally. In this case, since there is no mass loss at the surface, there is no freezing at the base of the ice once equilibrium is attained, and hence no latent heat release there, and the only heat flux that needs to escape through the ice is F_i. If we divide Eq. 6.40 by κ_T and integrate over the ice layer, we find

$$T_f - T_g = F_i \int_{-h}^{0} \frac{dz}{\kappa_T} = \frac{F_i h}{\overline{\kappa_T}} \tag{6.42}$$

where $\overline{\kappa_T}$ is the harmonic mean of the thermal conductivity, that is,

$$\overline{\kappa_T} \equiv (\frac{1}{h} \int_{-h}^{0} \frac{dz}{\kappa_T})^{-1} \tag{6.43}$$

Solving for the ice thickness, we find

$$h = \frac{\overline{\kappa_T}(T_f - T_g)}{F_i} \tag{6.44}$$

This determines the ice thickness once T_g is known. The physical content of this statement is that the ice thickness grows until it is just thick enough to let through the amount of heat delivered to the bottom of the ice. Ice can exist in equilibrium if $T_g < T_f$, and the ice thickness approaches zero as T_g warms to the freezing point. Increasing the heat flux or decreasing the thermal conductivity would also thin the ice. The fact that it is the *harmonic* mean of the thermal conductivity that appears in thiss equation has important consequences. The harmonic mean gives greatest weight to regions of small conductivity, with the consequence that even a thin layer of very small conductivity can drive the harmonic mean to small values, and hence require thin ice. In particular, a relatively thin blanket of snow can hold in the heat diffusing through the ice, and cause the ice to thin dramatically – all other things being equal. The following exercise gives some feel for the numbers.

Exercise 6.9.1 (a) The mean geothermal heat flux on Earth is about 0.03 W/m^2, and it was not much greater back in the Neoproterozoic. Snowball Earth simulations indicate tropical ice surface temperatures on the order of $230K$ for a globally glaciated planet. How thick do you expect the tropical ice to be if the entire layer has the thermal conductivity of ice? How would the thickness change if you added a one meter layer of snow at the top?

(b) In a situation like the present Earth with a great deal of open water, ocean currents can deliver heat to the bottom of an ice shelf or sea ice layer at rate much greater than the geothermal heat flux. Suppose that ocean currents transport a mean flux of $2W/m^2$ to the bottom of a polar ice layer which has a surface temperature of $250K$. How thick is the ice in equilibrium?

(c) The atmosphere of Europa is so tenuous that it has essentially no radiative effect, so the temperature is determined by a balance between absorbed solar radiation and blackbody emission. Suppose that near the equator the annual mean absorbed solar radiation is 5 W/m^2. The internal heat flux for Europa is not well constrained. Compute the equilibrium ice thickness assuming a heat flux of 0.01, 0.1 and 1 W/m^2.

The second example in the exercise shows that a rather small amount of heat delivered to ice by oceanic heat fluxes can be very effective in melting back sea ice. This is true because essentially all the heat so delivered can be used in melting. In contrast, if one delivered the equivalent of 1

W/m^2 to high latitude regions by heat transport in the atmosphere, a great deal of the heat would be lost by radiation to space, and only a small portion would actually be usable for melting ice.

Note that although the insulating properties of a snow layer thin the ice if the ice surface temeperature is held constant, this effect is offset by the fact that snow has a considerably higher albedo than ice, which reduces the surface temperature. Moreover, the low thermal conductivity and low density of snow allow it to cool very rapidly at night, particularly when a stable inversion forms. The daytime warming tends to be not so extreme, owing to stronger turbulent heat fluxes in a neutral or unstable surface layer. This process tends to reduce the daily-mean snow surface temperature, which again has a thickening effect on the ice.

Now let's bring in the effects of mass loss at the surface. In equilibrium, mass loss from the top must be balanced by freezing at the base. The latent heat of fusion adds to the flux delivered to the base of the ice. Hence, if *all* the mass sublimated from the surface is carried away and precipitated elsewhere, F_i is replaced by $F_i + (L_f/L_{sub})F_L$, where L_f is the latent heat of fusion and L_{sub} is the latent heat of sublimation. This makes the ice thinner than it was our previous estimate. In conditions cold enough to form ice we are generally in the weak evaporation limit, so F_L is small; moreover, for water $L_f/L_{sub} = .118$, which brings down the additional flux even more. The net flux delivered to the base is on the order of 1 W/m^2 or less in typical conditions, and does not significantly increase the ice surface temperature. Thus, we can get the ice thickness including sublimation by simply replacing F_i in Eq. 6.44 with the modified basal heat flux. The latent heat flux F_L can be estimated using the results of Section 6.8 in the weak evaporation limit. For $F_L = 1W/m^2$, which is typical of the cold Snowball Earth tropics, the basal flux increases from $.03W/m^2$ to $.148W/m^2$, which thins the ice by a factor of 5. Clearly the effect of sublimation on ice thickness is very significant, and can allow the tropical ice on Snowball Earth to be much thinner than it otherwise might have been, though even with sublimation taken into account the ice is over 300 m thick when the mean surface temperature is $250K$. With stronger sublimation, as would happen as CO_2 increases and the ice surface warms towards the freezing point, the effect is even more pronounced.

If some regions are experience net ablation of ice through sublimation, others must be experiencing net accumulation, since the vapor that sublimates must ultimate precipitate out somewhere else. What happens to ice thickness in regions of net accumulation? To a point, accumulation at the surface can be balanced by melting at the base. However, the only supply of heat available to melt ice at the base is the geothermal heat flux, and a flux of .03 W/m^2 can melt sustain a melt of only 3 mm of ice per year. If the accumulation exceeds this tiny rate, then if no other process intervenes the ice thickness increases until it freezes to the bottom. In reality, the generation of regions of thick ice would drive a flow of ice into regions where it can ablate by sublimation, thickening tropical ice and thinning polar ice. Things that flow are beyond the scope of this book.

We'll now consider one last variation on the theme of ice thickness. In the preceding calculations, it was assumed that all solar radiation that was not reflected was absorbed at the surface of the ice. In reality, some radiation will penetrate into the ice and be absorbed in the interior. If the penetration is significant, this can have a powerful effect on thinning the ice, because the low thermal conductivity of ice means that heat buried in the ice has a hard time getting out, and therefore accumulates until a considerable degree of warming has been achieved. To model this process, we modify the steady state thermal diffusion equation to allow for internal heating, which we represent as the vertical gradient of the downward solar flux $F_\circledast$. The equation becomes

$$\frac{d}{dz}\kappa_T \frac{dT}{dz} = -\frac{d}{dz}F_\circledast \tag{6.45}$$

or, upon integrating once,

$$\kappa_T \frac{dT}{dz} + F_{\circledast} = const. \tag{6.46}$$

This says simply that the sum of the diffusive and radiative heat flux must be constant. The constant of integration is determined by the requirement that the net flux out of the surface of the icce must equal the total solar absorption in the ice layer plus the heat flux delivered to the base of the ice. We'll simplify the problem by assuming that there is no applied flux, and that the ice is thick enough that all the solar radiation is absorbed with none penetrating through into the ocean. In this case, the heat flux out of the top of the ice, $-\kappa_T dT/dz$, must equal the solar flux $F_{\circledast}(0)$, whence the constant of integration is zero. In this case, the temperature profile within the ice is

$$T(z) = T_g + \int_z^0 \frac{F_{\circledast}}{\kappa_T} dz \tag{6.47}$$

The temperature increases with depth in the ice, and becomes uniform at depths where $F_{\circledast} \approx 0$, i.e. below the layer within which most of the solar radiation is absorbed. The deep ice temperature becomes larger as the solar flux penetrates deeper into the ice. To make this more clear, suppose that κ_T is constant and $F_{\circledast} = (1-\alpha_g)S_g \exp(z/H_{\circledast})$. Then, the deep ice temperature is

$$T_\infty = T_g + (1-\alpha_g)S_g H_{\circledast}/\kappa_T \tag{6.48}$$

This temperature increases without bound as the penetration depth $H_{\circledast}$ increases. For sufficiently large $H_{\circledast}$, the deep ice temperature increases to the melting point, which in that case essentially limits the ice thickness to the solar penetration depth. A precise calculation of the ice thickness is done in Problem **??**. Note that much of what we have said about the effect of internal absorption of solar radiation applies equally well to other internal heat sources, notably tidal heating arising from flexing of the ice crust. This heat source may be particularly important in determining the ice crust thickness on Europa.

Exercise 6.9.2 Assume that $T_g = 240K$ and $(1-\alpha_g)S_g = 100W/m^2$. If the decay of solar flux is exponential, how great does the penetration depth have to be in order to bring the deep ice temperature to the freezing point?

The interest in mechanisms for thinning tropical ice in Snowball conditions arises from two challenges facing the Snowball hypothesis. The first is the obvious need to find a way to exit from the globally glaciated state. Accumulation of CO_2 can warm the planet, but it is not clear that this process is actually sufficient to deglaciate thick ice, or that the necessary levels of CO_2 can be achieved. Thin ice can help make deglaciation easier, especially if solar radiation can penetrate the ice and warm the underlying ocean. The second challenge is that photosynthetic eukaryotes, who are rather fragile creatures in comparison to cyanobacteria, seem to have made it through the Neoproterozoic snowball without any evidence of a major crisis (such creatures were not yet around at the time of the putative Paleoproterozoic snowball, and so pose less of a problem then). Thin ice allows more fractures and leads, which can provide open water refugia. If the ice is thin and clear enough, enough solar radiation may even be able to penetrate the ice to support photosynthetic life beneath the ice layer.

Though we have mainly had floating water ice in mind in the preceding derivation, the heating due to solar pentration is generically applicable to planets with an icy crust, whatever the ice may be made of. For example, burial of solar heating is thought to lead to cryovolcanism in the nitrogen-ice crust of Neptune's moon Triton. The interior heating of ice is a form of solid greenhouse effect, in that the solar energy penetrates well, but the heat can only escape slowly. In

this case, the heat transfer through the ice is by molecular diffusion rather than infrared radiation, but the general principle is the same.

The problem of solar absorption within ice is a radiative transfer problem nearly as challenging as that confronted for clouds. It depends on scattering off of air bubbles and brine pockets, and therefore requires some understanding of the distribution of these. The absorption and scattering is wavenumber dependent,so spectrally resolved radiative transfer should ideally be used, and is is the case for the atmosphere, the spectral absorption features generally lead to non-exponential attenuation of the solar beam. Theses issues are all at the frontier of climate research.

6.10 For Further Reading

For a comprehensive reference to the planetary boundary layer and atmospheric turbulence, see

- Garrett JR 1994: *The Atmospheric Boundary Layer*. Cambridge, 334pp.

Meaurements of the hottest ground temperature on Earth are discussed in:

- Mildrexler, Zhao and Running 2006: "Where are the hottest spots on Earth" *EOS* **87**.

A discussion of Monin-Obukhov scaling functions for the unstable case, with particular attention to the free-convection limit, can be found in:

- Delage Y and Girard C 1992: Stability Functions Correct at the Free Convection Limit and Consistent for Both the Surface and Ekman Layers. *Boundary-Layer Meteorology*, **58**, 19-31.

This journal is the primary source for results about turbulent fluxes in the surface layer and the planetary boundary layer.

Chapter 7

Variation of temperature with season and latitude

7.1 Overview

Why is the Earth generally hotter near the Equator than at the poles? Why is it generally hotter in Summer than in Winter, especially outside the tropics? Would this be true on other planets as well? How would the pattern change over time, as features of the planet's orbit vary? Would a very slowly rotating planet lose its atmosphere to condensation on the nightside? Would a planet whose rotation axis was steeply inclined relative to the normal to the plane of the orbit, or a planet in a highly elliptical orbit, have such an extreme seasonal cycle that it would be uninhabitable? The answers to these questions are to be found in the way the geographic and temporal pattern of illumination of the planet plays off against the thermal response time of the atmosphere, ocean and solid surface of the planet. Generally speaking, in this section we seek to understand the features of a planet that determine the magnitude and pattern of geographic and seasonal variations in temperature.

Most of the discussion of temporal variability will focus on seasonal rather than diurnal variations, but much of the same considerations apply to both cycles, and so some remarks will be offered on the diurnal cycle as well. It should be kept in mind that the distinction between diurnal and seasonal cycle is meaningful only for bodies such as the Earth, Mars or Titan whose rotation period is short compared to the period of orbit about the Sun. For a planet whose length of day is a significant fraction of its year, one should think instead of a hybrid seasonal/diurnal cycle. The formulation developed below i sufficientlly general to handle that case.

7.2 A few observations of the Earth

First, let's take a look at how the Earth's surface temperature varies with the seasons. Figure 7.2 shows the zonal-mean air temperature near the surface for representative months in each of the four seasons. The first thing we note is that the temperature is fairly uniform in the tropics (30S-30N), but declines sharply as the poles are approached. The temperature difference between the Equator and 60N is 39K in the Winter but only 12K in the Summer. The Southern Hemisphere has a much

weaker seasonal cycle, except over the Antarctic continent: The temperature difference between the Equator and 60S is 26K in the Winter and 22K in the Summer. However, over Antarctica, poleward of 60S the seasonal cycle is extreme. Noting that the Northern Hemisphere has more land than the Summer, the data imply that the oceans have a strong moderating effect on the seasonal cycle. The temperature patterns in Figure 7.2 are what we seek to explain in terms of the response of climate to the geographically and seasonally varying Solar forcing.

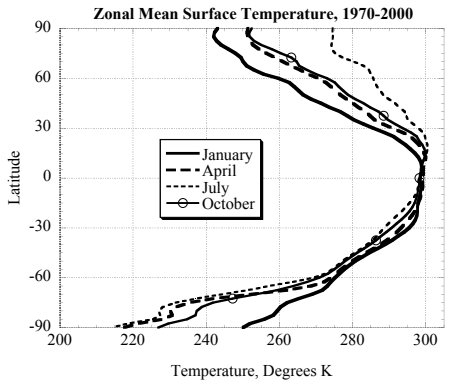

Figure 7.1: Observed zonal mean surface air temperatures for January, April, July and October. Computed from NCEP data for 1970-2000.

An even better appreciation of the effect of land masses on the seasonal cycle can be obtained by examining the map of July-January temperature differences, shown in Figure 7.2. This map shows that the strongest seasonal temperature contrast occurs in the interior of large continents, and that the ocean temperature varies by at most a few degrees over the year – and even less in the Tropics. The strong seasonal cycle of the Northern Hemisphere continents extends very little beyond the coastlines, and the seasonal cycle of the Northern oceans has similar magnitude to that of the more extensive Southern oceans.

7.3 Distribution of incident solar radiation

The geographical variations of temperature are driven by variations in the amount of sunlight falling on each square meter of surface, and also by variations in albedo. Seasonal variations are driven by changes in the geographical distribution of absorbed sunlight as the planet proceeds through its orbit. Therefore, the starting point for any treatment of seasonal and geographical variation must be the study of how the light of a planet's sun is distributed over the spherical surface of the planet. This section deals only with the distribution of incident sunlight, or *insolation*. The geographical distribution of the amount of sunlight absorbed is affected also by the distribution of the albedo. The albedo variations can also affect the seasonal distribution of solar forcing through seasonal variations in ice, snow, cloud and vegetation cover.

It will help to first consider an airless planet, so that we don't at once have to deal with the possible effects of scattering of the solar beam by the atmosphere. If our planet is far from its Sun, as compared to the radius of the Sun, the sunlight encountering the planet comes in as a beam of

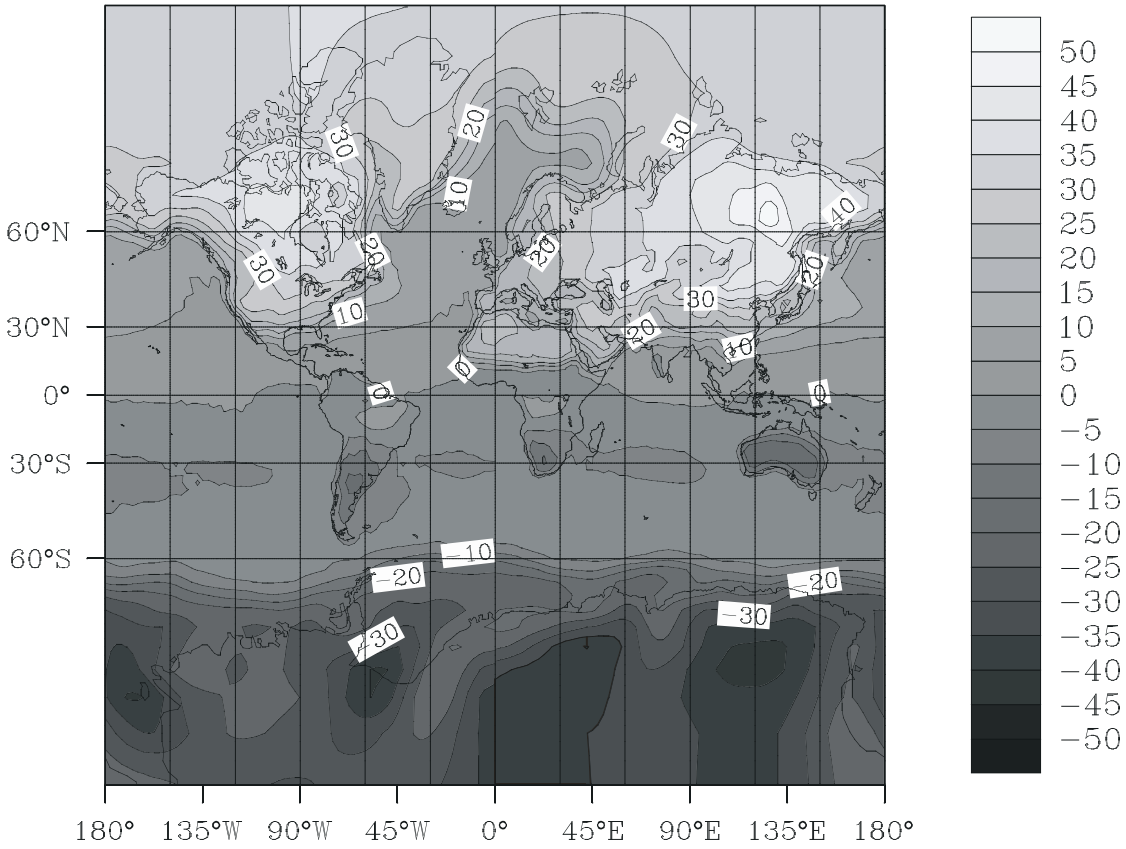

Figure 7.2: Map of July-January surface air temperature difference.

parallel rays with flux $L_{\circledast}$. Even if the surface of the planet is perfectly absorbing, the sunlight the planet intercepts is not spread uniformly over its surface; per unit area, parts of the planet where the sun is directly overhead receive a great deal of energy, whereas parts where the Sun grazes the surface at a shallow angle receive little, because the small amount of sunlight intercepted is spread over a comparatively large area, as shown in Figure 7.3. The night side of the planet, of course, receives no solar energy at all.

To obtain a general expression for the distribution of incident solar radiation per unit of surface area, we may divide up the surface of the planet into a great many small triangles, and consider each one individually. The solar energy intercepted by a triangle is determined by the area of the shadow that would be cast by the triangle on a screen oriented perpendicular to the solar beam. To compute this area, suppose that one of the vertices of the triangle is located at the origin, and that the two sides coming from this vertex are given by the vectors $\vec{r_1}$ and $\vec{r_2}$ By the definition of the cross product, the area of the triangle is given by $2A\hat{n} = (\vec{r_1} \times \vec{r_2})$ where $\hat{n}$ is the unit normal to the plane containing the triangle. To obtain the area of the shadow cast by the triangle, we apply the cross product to the projection of the vectors $\vec{r_1}$ and $\vec{r_2}$ onto the plane. These projections are given by $\vec{r_1} - \hat{z}\vec{r_1} \cdot \hat{z}$ and $\vec{r_2} - \hat{z}\vec{r_2} \cdot \hat{z}$, where $\hat{z}$ is the unit vector pointing in

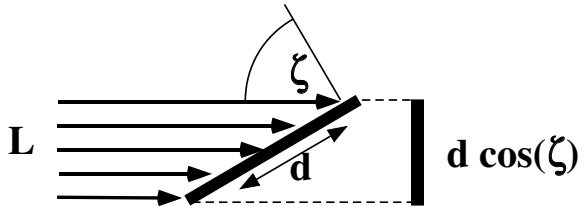

Figure 7.3:

the direction of the Sun. The cross product of these two vectors is

$$(\vec{r_1} - \hat{z}\vec{r_1} \cdot \hat{z}) \times (\vec{r_2} - \hat{z}\vec{r_2} \cdot \hat{z}) = \vec{r_1} \times \vec{r_2} - (\hat{z} \times \vec{r_2})(\vec{r_1} \cdot \hat{z}) - (\vec{r_1} \times \hat{z})(\vec{r_2} \cdot \hat{z}) \tag{7.1}$$

Now, the cross product of two vectors in the xy plane must point in the direction of the z axis. Hence, we can obtain the magnitude of the above vector by taking its dot product with $\hat{z}$. This is very convenient, since the dot product of $\hat{z}$ with the second two terms vanishes, leaving us with

$$2A_\perp = \hat{z} \cdot (\vec{r_1} \times \vec{r_2}) = 2A\hat{z} \cdot \hat{n} = 2A\cos(\zeta) \tag{7.2}$$

where $A_\perp$ is the area of the shadow and ζ is the angle between the normal to the patch of surface and the direction of the sun. This is known as the *zenith angle*. When the the zenith angle is zero, the Sun is directly overhead, and when it is 90^o the sunlight comes in parallel to the surface and leaves no energy behind. Zenith angles greater than 90^o are unphysical, since they represent light that would have to pass through the solid body of the planet in order to illuminate the underside of the surface; these are on the night side of the planet. If one draws a line from the center of the planet to the center of the Sun, the zenith angle will be zero where the line intersects the surface of the planet; this is the *subsolar point*. At any given instant, the curves of constant zenith angle make a set of concentric circles centered on the subsolar point, with a zenith angle of 90^o along the great circle which at the given instant separates the dayside of the planet from the nightside. If the surface were in equilibrium with the instantaneous incident solar flux, the subsolar point would be the hottest spot on the planet, with temperature falling to zero with distance away from the hot spot. As the planet rotates through its day/night cycle, a given point of the surface is swept through a range of distances from the hot spot, leading to a diurnal temperature variation. As the planet proceeds through its orbit in the course of the year, the diurnal cycle will change as the orientation of the planet's rotation axis changes relative to the Sun. Insofar as the surface actually takes a finite amount of time to heat up or cool down, the diurnal cycle will be attenuated to one extent or another.

As the next step toward realism, let's now consider a rapidly rotating planet whose axis of rotation is perpendicular to the line connecting the center of the planet to the center of its Sun. If the axis of rotation is in fact perpendicular to the plane of the orbit, this situation prevails all year round; otherwise, the condition is met only at the equinoxes, and indeed the condition defines the equinoxes. We assume that the planet is rotating rapidly enough that the day-night difference in solar radiation is averaged out and the corresponding temperature fluctuations are small. In other words, the length of the day is assumed to be short compared to the characteristic thermal response time of the planet's surface, a concept which will be explored quantitatively in Section 7.4. Consider a small strip of the planet's surface near a latitude ϕ, of angular width $d\phi$. If a is the planet's radius, then the area of this strip is $2\pi a^2 \cos(\phi) d\phi$, if angles are measured in radians. The cross section area of the strip seen edge-on looking from the Sun determines the amount of solar flux intercepted by the strip. This area is $2a^2 \cos^2(\phi) d\phi$ when $d\phi$ is small. In consequence,

the incident solar radiation per unit area at latitude ϕ is $L_\circledast \cos(\phi)/\pi$. At the Equator, the solar radiation per unit area is $L_\circledast/\pi$, which is somewhat greater than the value $L_\circledast/4$ which we obtained in Chapter 3 by averaging solar radiation over the *entire* surface of the planet. If the planet has no atmosphere to transport heat or create a greenhouse effect, the equilibrium temperature as a function of latitude is

$$T = (\frac{L_\circledast \cos(\phi)}{\pi\sigma})^{\frac{1}{4}} \tag{7.3}$$

The temperature has its maximum at the Equator, and falls to zero at the poles.

Exercise 7.3.1 For the geometric situation described above, derive an expression for the cosine of the zenith angle as a function of latitude and longitude. Re-derive the expression for the daily-average distribution of solar absorption by averaging the cosine of the zenith angle along latitude circles.

Now we turn to the general case, in which the axis of rotation of the planet is not perpendicular to the plane containing the orbit. The angle between the perpendicular to the orbital plane, and the planet's axis of rotation, is known as the *obliquity*, and we shall call it γ. It can be regarded as constant over the course of a planet's year, though there are longer term variations which will be of interest to us later. The near-constancy of γ arises from the conservation of angular momentum in the absence of torques; small torques on the planet give rise to the long-term variations.

For a moon, it is the obliquity of the rotation axis relative to the plane of the orbit of the host planet that counts. The typical behavior of the major Solar System moons is to be orbiting very nearly in the plane of the host planet's equator, and to be tide-locked to the planet. This is the case for Titan, for example. For such moons, the obliquity of the moon (for the purposes of determining the seasonal cycle) is essentially the same as the obliquity of the host planet. The obliquity of the Neptune's moon Triton is the principal known exception to this simple behavior. Its orbit is highly inclined to Neptune's equator, and this leads to a very complex seasonal cycle. In any event, a moon's "year" for the purposes of computing illumination is essentially the same as the host planet's year, since the radius of a moon's orbit is inevitably small compared to the distance of its planet to the host star.

The task now is to determine the solar zenith angle as a function of latitude, position along the latitude circle, and time of year. Let the point P be the center of the planet, and S be the center of the sun. If we draw a line from P to S, it will intersect the surface of the planet at a latitude δ, which is called the *latitude of the sun*, or sometimes the *subsolar latitude*. It is a function of the orientation of the planet's axis alone, and serves as a characterization of where we are in the march of the seasons. If the obliquity of the planet is γ, then δ ranges from γ at the Northern Hemisphere summer solstice to $-\gamma$ at the Southern Hemisphere summer solstice. Let Q be a point on the planet's surface, characterized by its latitude ϕ and its "hour angle" h, which is the longitude relative to the longitude at which local noon (the highest sun position) is occurring throughout the globe. For radiative purposes, we just need to compute the zenith angle ζ, defined previously. To get the zenith angle, we only need to take the vector dot product of the vector $\vec{QS}$ and the vector $\vec{PQ}$. To do this, it is convenient to introduce a local Cartesian coordinate system centered at P, with the z-axis coincident with the axis of rotation, the x-axis lying in the plane containing the rotation axis and $\vec{PS}$, and the y-axis orthogonal to the other two, chosen to complete a right-handed coordinate system.

First, note that by the definition of the dot product,

$$\cos(\zeta) = \frac{\vec{PQ} \cdot \vec{QS}}{|PQ||QS|} \tag{7.4}$$

Further, $\vec{PQ} + \vec{QS} = \vec{PS}$, so

$$\cos(\zeta) = \frac{\vec{PQ} \cdot \vec{PQ}}{|PQ||QS|} + \frac{\vec{PQ} \cdot \vec{PS}}{|PQ||QS|} \approx \frac{\vec{PQ} \cdot \vec{PS}}{|PQ||QS|} \tag{7.5}$$

where we drop the first term based on the assumption that the radius of the planet is a small fraction of its distance from the Sun. For the same reason, $|QS|$ in the denominator can with good approximation be replaced by $|PS|$, leaving the expression in the form of a dot product between two unit vectors. Letting $\hat{n}_1 = \vec{PQ}/|PQ|$ and $\hat{n}_2 = \vec{PS}/|PS|$, the unit vectors have the following components in the local Cartesian coordinate system.

$$\hat{n}_1 = (\cos(\phi)\cos(h), \cos(\phi)\sin(h), \sin(\phi)), \hat{n}_2 = (\cos(\delta), 0, \sin(\delta)) \tag{7.6}$$

whence

$$\cos(\zeta) = \cos(\phi)\cos(\delta)\cos(h) + \sin(\phi)\sin(\delta) \tag{7.7}$$

When $\cos\zeta < 0$ the sun is below the horizon. Note that this formula just amounts to a transformation from a latitude-longitude coordinate system with the pole placed at the subsolar point (in which case the zenith angle formula becomes very simple) to a coordinate system with the pole at some other point. For a rotating planet it is convenient to use a coordinate system with the poles lined up along the rotation axis, but the formula would be equally valid if one for some reason wanted to adopt a different coordinate system. The longitude in Eq. 7.7 is not geographically fixed, since the hour angle is defined relative to the longitude of local solar noon, but one can introduce a *longitude of the sun*, which we'll call $\lambda_\circledast$, which is the longitude of the subsolar point measured in a geographically fixed coordinate system such as the conventional latitude-longitude system in use at present on Earth. If λ is the longitude measured in this coordinate system, then the hour angle is given by $h = \lambda - \lambda_\circledast$. Both δ and $\lambda_\circledast$ are functions of time, but $\lambda_\circledast$ depends on the rotation of the planet as well as its position in its orbit [1].

The cosine of the zenith angle attains a maximum value $\cos(\phi - \delta)$ when $h = 0$, and a minimum value $-\cos(\phi + \delta)$ when $h = \pm\pi$. Both values are above the horizon when $|\phi| > |\pi/2 - \delta|$, corresponding to the perpetual polar summer day. Both values are below the horizon when $|\phi| > |\pi/2 + \delta|$, corresponding to the perpetual polar winter night. At the solstices, δ takes on its extreme values of $\pm\gamma$. Therefore, perpetual day or night are experienced at some time of year for latitudes poleward of $\pi/2 - \gamma$. These circles are known as the *Arctic* and *Antarctic* circles on Earth. Apart from the case of perpetual day or night, there is a *terminator* which separates the illuminated from the dark side of the planet. The position of the terminator is given by

$$\cos h_t = -\tan(\phi)\tan(\delta) \tag{7.8}$$

We shall adopt the convention $h_t = 0$ in the case of perpetual night, and $h_t = \pm\pi$ for perpetual day.

If the planet's day is short compared to the time required to orbit the star, the position and axis orientation of the planet can be considered nearly fixed as the planet rotates about its axis. In this case, h_t translates directly into the length of the daylight period. Let Ω be the angular velocity of rotation of the planet relative to the fixed stars, so that the duration of the planet's *stellar day* is $T_{day} = 2\pi/\Omega$. Then the duration of daylight in the rapidly rotating case is close to $2h_t/\Omega = (h_t/\pi)T_{day}$ [2].

[1] The reader should be cautioned that the "longitude of the subsolar point" defined here differs from the "ecliptic longitude of the Sun" in common use in astronomy. The latter uses a celestial reference frame for longitude.

[2] The stellar day is slightly different from what is misleadingly called the *sidereal day*, since the latter takes into

Exercise 7.3.2 For a given latitude ϕ, what δ yields the least hours of daylight? What δ yields the most hours of daylight? The Earth's present obliquity is 23.5 degrees, and it's length of day is 23.94hours. Sketch a plot of the maximum and minimum hours of daylight vs. latitude for the Earth. In all these calculations, you should make use of the rapidly-rotating approximation.

When the planet's stellar day is an appreciable fraction of its year, then the distinction between the stellar day and solar day becomes important. In such cases, the time variations of δ and of $\lambda_\circledast$ over the course of the day must be taken into account when computing the duration of daylight and the diurnal variation of insolation. Some calculations of this sort are carried out in Problem **??**.

The diurnal variations of the zenith angle lead to hot days and cold nights. Where the thermal response time is long enough to average out an appreciable portion of the diurnal temperature variation, the daily mean incident solar flux is an informative statistic. Since the incident solar flux per square meter of surface is $L_\circledast \cos\zeta$, where $L_\circledast$ is the solar constant in W/m^2, one can obtain the daily mean flux by averaging $\cos\zeta$ over a rotation period of the planet. This results in a nondimensional flux factor f, by which one multiplies the solar constant in order to obtain the daily mean solar radiation incident on each square meter of the planet's spherical surface. The daily average can be performed analytically, resulting in

$$\begin{aligned} f(\phi,\delta) &= \frac{1}{2\pi}\int_{-h_t}^{h_t} cos(\zeta)dh \\ &= \frac{1}{\pi}[\cos(\phi)\cos(\delta)\sin(h_t) + \sin(\phi)\sin(\delta)h_t] \end{aligned} \tag{7.9}$$

where h_t is determined by Eq. 7.8. This derivation of the daily average assumes that the length of the day is much less than the length of the year, so that δ and $\lambda_\circledast$ may be regarded as constant over the course of the day. If the length of the day is a significant fraction of the length of the year, as is the case for slowly rotating planets like Mercury or Venus, the expression still gives the correct instantaneous average along the latitude circle, but this average is no longer identical to the time average over a day. In that case, however, the distinction between the diurnal and seasonal cycle is no longer meaningful, and the time variations of insolation from Eq. 7.7 should be used directly, without diurnal averaging.

During the equinoxes, $\delta = 0$ and $f = \cos(\phi)/\pi$, independant of the obliquity. This agrees with the result we obtained earlier by direct geometrical reasoning. At other times of year, the daily mean flux is governed by two competing factors: the varying length of day, which tends to produce higher fluxes near the summer pole, and the average zenith angle, which tends to produce high fluxes near the subsolar latitude (which remains near the Equator if the obliquity is not too large). The latitude where the maximum daily mean insolation occurs is always between the subsolar latitude δ and the summer pole. For $\delta = 0$ the maximum occurs at the Equator, and a little numerical experimentation shows that the latitude of the maximum increases to about 43.4^o when $\delta = 23.4^o$ (and similarly, with reversal of signs, in the Southern hemisphere). For larger δ, the length-of-day effect wins out over the slant angle effect at the pole, and the maximum occurs at the summer pole itself. This state of affairs just barely happens at the solstice for the present obliquity of Earth and Mars; as a result, the summer hemisphere solstice insolation is fairly uniform

account the effect of the precession of a planet's rotation axis. Precession will be discussed in Section 7.6. It is usually a very slow process, so for the purposes of computing a planet's climate, the distinction between sidereal and stellar day is almost always immaterial. In the context of extrasolar planets the approved term "stellar day" becomes very confusing, since it refers to the distant stars and not the planet's host star;pending some agreement on better terminology, we'll use the term "solar day" to refer to the diurnal variation of incident radiation regardless of what star the planet is orbiting.

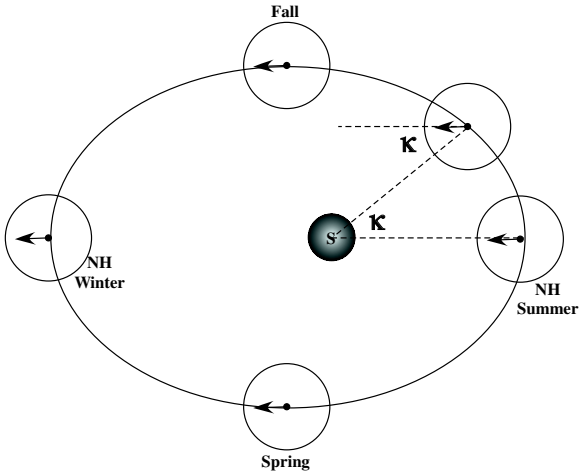

Figure 7.4:

in these two cases. It is also useful to note that the daily mean insolation at the summer pole exceeds the daily mean insolation at the Equator when $|\delta| > 17.86^o$.

To obtain a general appreciation of the seasonal cycle, recall that δ varies from $-\gamma$ during the southern hemisphere summer solstice to γ at the northern hemisphere summer solstice, taking on a value of zero at the equinox which lies between the two solstices. Consider a planet with uniform albedo, so that the absorbed solar radiation is determined by the distribution of incident solar radiation. Suppose further that the thermal response time is long enough to average out the diurnal cycle, but short compared to the length of the year. If the obliquity is below 23.4^o, the "hot spot" starts some distance poleward of the Equator in the Southern hemisphere, moves to the Equator as the equinox is approached, and then migrates a similar distance into the Northern hemisphere as the Northern summer solstice is aproached. If the obliquity is greater than 23.4^o, the hot spot starts at the South Pole, discontinously jumps to -43.4^o at the point in the season where the subsolar latitude crosses -23.4^o, smoothly migrates throught the Equator and on to 43.4^o when the subsolar latitude approaches 23.4^o, and then discontinuously jumps to the North Pole. Note that in either case, the hot spot crosses the Equator twice per year, at the equinoxes; the two solstices are the coldest times at the Equator. The climate at the Equator has a periodicity that is *half* the planet's year.

It only remains to express δ as a function of the position of the planet in its orbit. The planet is spinning like a top, and if there are no torques acting on the planet (an assumption we will relax later) its angular momentum is conserved. Hence the rotation axis keeps a fixed orientation relative to the distant stars throughout the year. This is why Polaris is the Northern Hemisphere pole star all year around. Let κ be the angle describing the position of the planet, as shown in Figure 7.4. We shall adopt the convention that $\kappa = 0$ occurs at the Northern Hemisphere summer solstice. We shall refer to κ as the *season angle*, though it is closely related to what astronomers call the *ecliptic longitude of the sun*. In our case, we have defined the season angle relative to the Northern Hemisphere summer solstice, but other choices are also common, for example defining it relative to the Northern winter solstice or the Spring equinox. When discussing the progression through the seasonal cycle on planets other than Earth, a season angle is almost universally used to describe where the planet is in its annual cycle, since this description obviates the need to make up names for months for each planet. If we project the rotation axis onto the plane of the ecliptic

(i.e. the plane containing the planet's orbit), then the angle made by this vector with $\vec{PS}$ is equal to κ. The rotation axis projected onto the plane of the ecliptic acts like the hand of a clock, which rotates around the clock face once per year, though at a non-uniform rate if the orbit is not perfectly circular.

Let $\hat{n}$ be the unit normal vector to the plane of the ecliptic, and $\hat{n}_a$ be the unit vector in the direction of the rotation axis. Introduce a new cartesian coordinate system with x pointing along $\vec{PS}$, z pointing along $\hat{n}$, and y perpendicular to the two in a right-handed way. Then $\hat{n}_a = (\cos(\kappa)\sin(\gamma), \sin(\kappa)\sin(\gamma), \cos(\gamma))$ and the latitude of the sun is the complement of the angle between $\hat{n}_a$ and the x axis, whence

$$\sin(\delta) = \cos(\frac{\pi}{2} - \delta) = \cos(\kappa)\sin(\gamma) \tag{7.10}$$

In the limit of small obliquity, this equation reduces to $\delta = \gamma\cos\kappa(t)$. For a circular orbit, $\kappa(t) = \Omega t$, where Ω is the orbital angular velocity (2π divided by the orbital period). In this special case, the subsolar latitude varies cosinusoidally over the year, with amplitude given by the obliquity. This is actually not a bad approximation even for the roughly 23^o current obliquity of Earth and Mars, agreeing with the true value to two decimal places. At the opposite extreme, when $\gamma = 90^o$, the subsolar latitude is given by $\delta = \pi/2 - \kappa$, which is not at all sinusoidal.

Exercise 7.3.3 Compute the length of day as a function of the time of year for the latitude at which you are currently located. Compare with data for the current day, either observed yourself or presented in the newspaper weather report. Compute the length of a shadow that would be cast by a tall, thin skyscraper of height 100m, as a function of the time of day and time of year at your latitude.

Contour plots of the diurnally averaged flux factor for various obliquities are shown in Figure 7.5. These plots assume the orbit to be perfectly circular, so that there is no variation in distance from the Sun in the course of the year. Over the course of the year, the hot spot moves from south of the equator to north of the equator, and back again, passing over the equator at the equinoxes. The amplitude of the excursion increases with obliquity, and goes all the way from pole to pole for sufficiently large obliquity. Earth, Mars, Saturn,Titan, and Neptune with present-day obliquities of 23.5^o, 24^o,26.7^o,26.7^o,and 29.6^o respectively, are qualitatively like the 20^o case. The pattern of variation of incident solar radiation which forces the seasonal cycle is similar in all these cases. However, the nature of the seasonal cycle will differ amongst these planets because the differing nature of the atmospheres and planetary surfaces will lead to different thermal response times. In the case of gas giant planets, another variable is the proportion of energy received from solar energy vs. the that received by transport from the interior of the planet. Insofar as the latter becomes dominant, the role of solar heating, and hence the prominence of the seasonal cycle, becomes less. Jupiter has a low obliquity (3.1^o), which, compounded by a fairly high proportion of internal heating (11 %) should lead to a minimal seasonal cycle. At the opposite extreme is Uranus, which has an obliquity of nearly 90^o, and apparently insignificant internal heating. Venus is so slowly rotating that its obliquity is of little interest. Obliquity is not constant in time; it varies gradually over many thousands of years. We will see in Section 7.6.1 that relatively slight variations in the Earth's obliquity are believed to contribute to the coming and going of the ice ages. The obliquity of Mars varies more dramatically, and perhaps with greater consequence; at various times in the past it could have reached values as high as 50^o and as low as 15^o.

If the thermal response time of the planet is a year or more, then a considerable part of the seasonal cycle is averaged out and the annual mean insolation becomes an informative statistic. It will be seen in the next section that this is the case for watery planets like the Earth. The

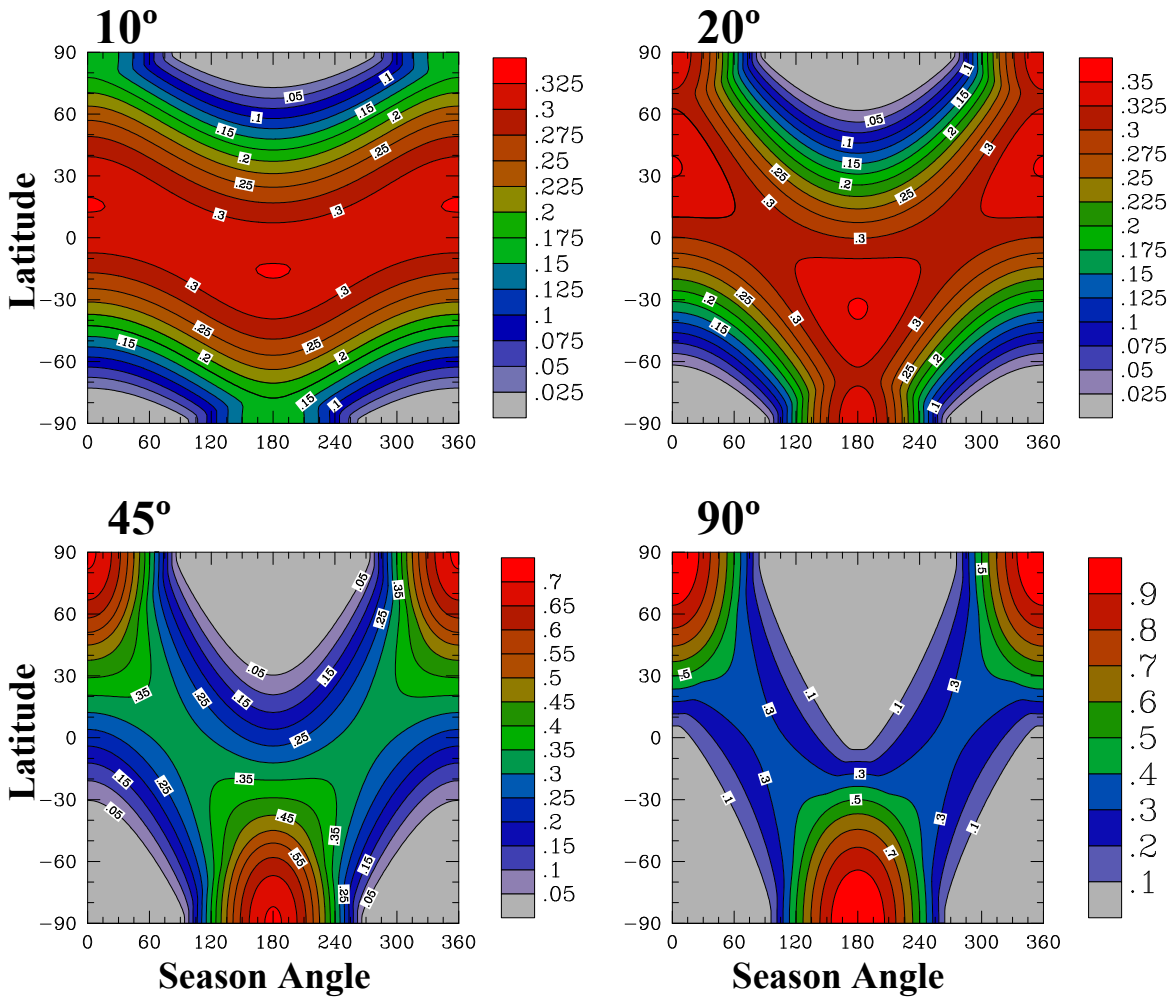

Figure 7.5: The seasonal and latitudinal distribution of daily-mean flux factor for four different values of the obliquity. In these plots, a circular orbit has been assumed. To obtain the daily mean energy flux incident on each square meter of the planet's surface, one multiplies the flux factor by the solar constant. For example, if the solar constant is $1000 W/m^2$, the incident solar flux at the pole during the Summer solstice is about $700 W/m^2$ if the obliquity is 45^o.

annual mean flux factor is shown in Figure 7.6. When obliquity is small, the poles receive hardly any radiation. As obliquity is increased, the polar regions receive more insolation, at the expense of the equatorial regions. For Earthlike obliquity, the maximum insolation occurs at the equator, which is why this region of Earth's surface tends to be warmest. When the obliquity exceeds 53.9^o, the annual mean polar insolation becomes greater than the annual mean equatorial insolation. For such a planet, the poles will be warmer than the tropics, provided that the thermal response time is long enough to average out most of the seasonal cycle. Consider a planet with 20^o obliquity, zero albedo, and a very long thermal response time. If the planet were put in Earth's orbit about the Sun, the Solar constant would be $1370 W/m^2$, yielding equatorial insolation of $422 W/m^2$ and polar insolation of $149 W/m^2$, based on the flux factors given in Figure 7.6. In the absence of any greenhouse effect or lateral energy transport by atmospheres or oceans, the equatorial temperature would be 294K and the polar temperature would be 226K. If one takes into account the clear-sky greenhouse effect of an Earthlike atmosphere with 300ppm CO_2 and 50% relative humidity using the OLR results given in Chapter 4, the polar temperature rises to 237K, but the equatorial temperature becomes problematic: The annual mean equatorial solar flux is near or above the runaway greenhouse threshold discussed in Chapter 4, leading to extremely high or even unbounded equatorial temperatures. Lowering the relative humidity to 20% to reflect the fact that much of the tropical troposphere is very dry (recall Chapter **??**) still leaves the tropics with temperatures in excess of 350K. Part of the problem lies in the neglect of albedo. Simply using the observed planetary albedo in the tropics gives the wrong answer, because almost all of the cloud albedo is offset by the cloud greenhouse effect in the present climate (Chapter **??**). Using an albedo of .15, based on the observed tropical clear-sky albedo, reduces the equatorial solar absorption to $360 W/m^2$, which is in balance with a tropical temperature of 318K assuming a relative humidity of 20%. This is still well in excess of the observed tropical temperature. In the real atmosphere, heat transports due to large scale atmospheric and oceanic motions remove some of the heat from the tropics and deposit it at high latitudes, reducing the tropical temperatures and increasing the polar temperatures. Since incorporation of ice-albedo effects would reduce the polar temperatures below the estimates given above, such transports are also needed to bring the polar temperatures up into the observed range. Some elementary models of heat transport will be discussed in Chapter 9, though a proper treatment of the subject requires a full understanding of the fluid dynamics governing atmospheric transport.

If we put the same planet at the orbit of Mars the temperatures become 238K and 183K without any greenhouse effect. Since there is little water vapor feedback at such low temperatures, the greenhouse effect is less dramatic in this case. Addition of an Earthlike atmosphere with 300ppmv CO_2 would increase the equatorial and polar temperatures to 256K and 192K, respectively, based on a linear OLR fit in the range 200K to 250K (specifically, $OLR = 80.6 + 1.83(T - 200)$). Thus, a waterworld placed at the orbit of Mars would require a much stronger greenhouse effect than the Earth's to avoid succumbing to a snowball state.

The effects of obliquity on the seasonal and latitudinal pattern of insolation may be summed up as follows. Increasing obliquity increases the intensity of the seasonal cycle at mid to high latitudes. The summer insolation gets steadily higher relative to the global mean, and a greater area of the winter hemisphere exposed to cold perpetual night or low insolation. Increasing obliquity also increases the annual mean polar insolation, though the way this affects polar climate depends on the thermal response time of the atmosphere-surface system. The increase might show up as very hot summers and bitterly cold winters, or as year-round warming, accordingly as the response time is fast or slow.

The preceding results on incident solar radiation have been derived in the absence of an atmosphere, but can still be used if there is in intervening atmosphere which may absorb or scatter

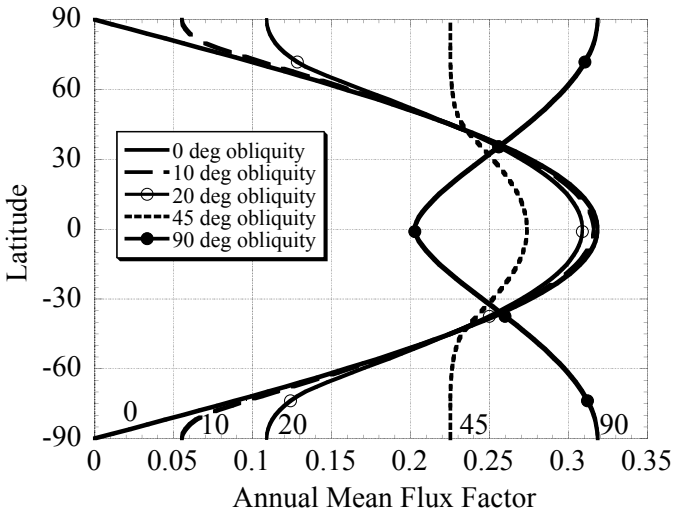

Figure 7.6: The annual mean flux factor for various obliquities, assuming a circular orbit.

solar radiation before it reaches the surface. The general geometry is illustrated in Figure 7.7. In this case, one suspends an imaginary sphere at an altitude above which the atmosphere is too thin to have a signficant effect on the solar radiation. The preceding results then give the solar flux entering each square meter of the surface of this sphere, and the angle at which the light enters the atmosphere. This is all that is needed as input to one-dimensional scattering models of the sort discussed in Chapter 5. One simply divides up the atmosphere into a series of patches near each latitude and longitude point, within which the properties are considered uniform, and applies a one-dimensional column model to each of these patches. As illustrated in 7.7, if the horizontal size of the patches is large compared to the depth of the atmosphere, the energy loss by horizontal scattering from one patch to another can be neglected, and each patch can be considered energetically closed, so far as radiation is concerned. The atmosphere has three effects, which can be inferred from the column radiation model: (1) Some of the incident solar radiation reaches the surface in the form of diffuse radiation at a continuous distribution of angles, rather than at the zenith angle, (2) Some of the solar radiation is absorbed in the atmosphere, rather than at the surface, and (3) some of the incident solar radiation is reflected back to space instead of entering the climate system. Of the three effects, it is the last – the effect of the atmosphere, including its clouds, on planetary albedo – that is most important for determining the climate. Diffuse radiation and atmospheric absorption do not change the amount of energy entering a column, but only the place and angle with which it enters. Often, this is of little consequence, so one can get a good estimate of the planet's temperature if one can obtain an estimate of the planetary albedo from one means or another.

The above reasoning can even to some extent apply to gas giant planets which have no surface. One can still define the imaginary sphere through which radiation enters the system, as before, but the problem comes in defining a characteristic depth scale. For the purposes of solar radiation, it suffices to consider the depth of atmosphere over which most of the solar radiation is absorbed, in effect a "photic" zone. This is typically shallow compared to the prodigious size of the gas giants. The full problem, including internal heat sources and dynamical motions, might

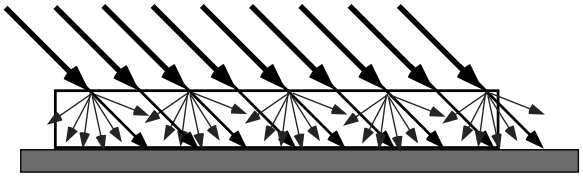

Figure 7.7: Schematic effect of atmospheric scattering on insolation

require consideration of a deeper layer. Whatever the depth of this "active layer," the preceding reasoning applies provided one can sensibly model the large scale aspects of the climate on the basis of averaging over patches of horizontal extent that is large compared to the characteristic depth scale. For giant planets, as for the Earth or any other planet, the essential difficulty is that clouds, temperature, water vapor and other climate variables are manifestly not uniform over length scales comparable to or longer than the depth of the atmosphere. One makes progress by boldly assuming that one can represent the effects of these fluctuating quantities by their large scale averages. It is an assumption that is difficult or impossible to justify mathematically, and in some cases may not even be true. With the present state of the art, one can only make progress by proceeding on the basis of the averaging assumption, and seeing how things work out.

7.4 Thermal Inertia

At several points in the preceding discussion, we have needed to make reference to the thermal response time of the system. In the present section, we shall make this notion precise. The heat storage in the planet's solid or liquid surface, and in its atmosphere means that it takes time for the system to heat up or cool down. The strength of this effect, known as *thermal inertia*, determines the extent to which the seasonal and diurnal fluctuations are averaged out in the climate response.

7.4.1 Thermal inertia for a mixed-layer ocean

The concept of thermal inertia is well illustrated by consideration of heat storage in the mixed layer of an ocean. Consider a layer of incompressible fluid with density ρ and specific heat c_p, which is well-mixed by turbulence to a depth H. The assumption of well-mixedness implies that any heating or cooling applied to the surface is distributed instantaneously throughout the depth of the mixed layer, whose temperature thus remains uniform. Let $(1-\alpha)S(t)$ be the absorbed solar flux heating the mixed layer. We have deliberately left off subscripts denoting whether this is a top-of-atmosphere or surface flux, since we will eventually see that there are circumstances in which the energy budget of the surface can be determined directly in terms of a top-of-atmosphere budget, without the necessity of considering in detail just how the flux is communicated to the surface. Assume further that the cooling of the mixed layer by infrared radiation or other means can be written as a function of temperature, which we shall call $F(T)$. For example, if the atmosphere above the ocean has no greenhouse effect and carries no heat away from the surface by turbulent transport, the cooling is the radiative cooling $F(T) = \sigma T^4$. The energy balance equation for the mixed layer is then

$$\frac{d}{dt}\rho c_p H T = (1-\alpha)S(t) - F(T) \tag{7.11}$$

Exercise 7.4.1 Consider a planet with a 50m deep mixed layer water ocean ($c_p = 4218 J/kg$, $\rho =$

$1000kg/m^3$). Suppose that the atmosphere for some reason has no effect whatsoever on the surface energy budget. (Why would this situation be hard to arrange, even for a pure N_2 atmosphere?) Hence $F(T) = \sigma T^4$. Suppose that the temperature of the polar ocean is $300K$ when the sun sets and the long polar night begins. Find a solution to Eq. 7.11 for this situation, and use it to determine how long it takes for the ocean to fall to the freezing point (about 271K for salt water)?

We may define a thermal inertia coefficient $\mu = \rho c_p H$ for the mixed layer ocean. If an amount of energy ΔE is added to or removed from a column of the ocean having a cross section of one square meter, the corresponding temperature change is $\Delta E/\mu$. For a 50m mixed layer water ocean, $\mu = 2.1 \cdot 10^8 J/(m^2 K)$, so that an energy flux of $100W/m^2$ out of the surface would lead to a cooling rate of $100/\mu = 4.74 \cdot 10^{-7} K/s = .04K/day$. Clearly, a rather shallow layer of well-mixed water can buffer a considerable surface flux imbalance. The Earth's ocean is several kilometers deep, but it is only the upper few tens of meters that are well mixed on short time scales; 50m is in fact a reasonable approximation to the overall mixed layer depth of Earth's ocean, though there are geographical variations. Most other liquids would do about as well as water at storing heat. It is primarily the mixing depth that determines the thermal buffering effect of a planet's ocean.

Atmospheres also have thermal inertia, which can be considered in a fashion analogous to a mixed layer. The entire mass of the troposphere is well-mixed, and this generally makes up most of the mass of an atmosphere. For a well-mixed atmosphere the temperature profile can be tied to the temperature at any convenient fixed level (usually the ground), and we need to determine how much energy it takes to change this index temeprature by one degree K. This is where the notion of moist or dry static energy, introduced in Chapter 2, comes into its own. For simplicity, let's consider a noncondensing atmosphere, for which case the dry static energy (per unit mass) $c_p T + gz$ is independent of height within the troposphere. Hence, by evaluating its value at $z = 0$, *the dry static energy per unit mass can be written* $c_p T_{sa}$, where T_{sa} is the near-surface air temperature. The mass per unit area of the atmosphere is p_s/g, so the energy required to change the surface air temprature by $1K$ while keeping the dry static energy well-mixed in the vertical is simply $\mu \equiv c_p p_s/g$. When the atmosphere has a condensible component, one needs to take into account latent heat storage as well. Exploration of that aspect of the problem will be relegated to the Workbook section of this chapter. The following discussion is most valid for noncondensing atmospheres, but will nonetheless apply approximately to condensible atmospheres undergoing fluctuations in which the latent heat storage doesn't change too much.

It is convenient to express μ in terms of the depth of a water mixed layer ocean, H_{eq} which would have the same thermal inertia coefficient. For the Earth atmosphere, $H_{eq} = 2.4m$, which is insignificant in comparison to the mixed layer depth of the ocean. Hence, one expects the Earth's atmosphere to come into equilibrium much more quickly than the ocean. The current 6mb CO_2 atmosphere of Mars has $H_{eq} = .03m$, while the massive atmosphere of Venus has H_{eq} in excess of $155m$. Neither Mars nor Venus has an ocean to buffer the seasonal cycle, but the Venus atmosphere alone can be expected to have a considerable moderating effect, whereas present Mars should behave more or less as if each point of the globe is in instantaneous equilibrium. Early Mars (circa 4 billion years ago) may have had a 2 bar CO_2 atmosphere, which would translate into a $10m$ equivalent mixed layer. This is considerably greater than that of Earth's atmosphere, but still not enough to have much moderating effect, in view of the fact that Mars' year is about twice as long as Earth's. Titan has a mostly N_2 atmosphere with a surface pressure only slightly in excess of Earth, but its weak gravitational acceleration of $1.35m/s^2$ means that this pressure translates into a much greater mass of atmosphere per square meter of planetary surface. Thus,the Titan atmosphere has an equivalent mixed layer depth of about $24m$. Given the low temperature of Titan, and consequent low rate of energy loss by infrared emission, this value is expected to

yield a very considerable buffering effect on Titan's seasonal cycle, regardless of whether there is a liquid ocean at the surface. For example, based on a typical surface temperature of 90K, blackbody emission would cool the planet only at a rate of about 1K per 300 Earth days, if insolation were completely shut off.

Exercise 7.4.2 The specific heat of liquid Methane is 3450 J/K. How deep would a well-mixed methane ocean on Titan have to be for it to have thermal inertia comparable to Titan's atmosphere?

We shall now consider some simple solutions to the mixed layer model, keeping in mind that this model applies to atmospheres as well as oceans, with a suitable choice of the equivalent mixed layer depth. At this point we assume that $\rho c_p H$ is constant, though models with a time-varying mixed layer depth are possible. Without any loss of generality we may write the insolation and temperature in the form

$$S = S_o + S'(t), T = T_o + T'(t) \tag{7.12}$$

where S_o and T_o are the time means of S and T and the deviations have zero time mean. Now, suppose that $T' << T_o$ for whatever reason; this need not require that $S' << S_o$, since the temperature fluctuations might be small by virtue of a slow response time of the system. Because the temperature fluctuations are small, the surface cooling can be expanded about T_o and approximated by a linear function:

$$F(T) = F(T_o) + bT'(t), b = \frac{dF}{dT}(T_o) \tag{7.13}$$

Now we choose T_o to be the equilibrium temperature corresponding to the mean insolation S_o, i.e. $F(T_o) = (1-\alpha)S_o$. With these assumptions the equation for the temperature fluctuation becomes

$$\frac{dT'}{dt} = \frac{1}{\rho c_p H}((1-\alpha)S'(t) - bT') \tag{7.14}$$

or equivalently, if we define the relaxation time $\tau = (\rho c_p H)/b$,

$$\frac{dT'}{dt} + \frac{T'}{\tau} = \frac{1}{\rho c_p H}(1-\alpha)S'(t) \tag{7.15}$$

We can distinguish two limiting cases for Equation 7.15. When the time scale over which S' varies is slow compared to τ, then the first term on the left hand side is negligible compared to the second, whence the solution becomes $T' = \tau \cdot (1-\alpha)S'(t)/(\rho c_p H)$, which is $(1-\alpha)S'(t)/b$. In other words, the system acts as if it's in equilibrium with the instantaneous solar radiation at each time. In the opposite limit, the time scale of the solar fluctuation is rapid compared to τ, in which case it is the second term on the left hand side that may be neglected. Thus,

$$T'(t) = \frac{1}{\rho c_p H}\int_0^t (1-\alpha)S'(t')dt' \tag{7.16}$$

In this case, the temperature is out of phase with the heating, and represents a time average of the fluctuating heating. The peak temperature occurs later than the peak solar heating, since it takes time for the mixed layer to respond to the accumulating heating. Further, in this case, the seasonal temperature fluctuation becomes small as the mixed layer depth is made large, since the mixed layer becomes more and more efficient at averaging out the seasonal fluctuations of solar flux.

The variations in solar radiation over the course of a year are not sinusoidal, but we can nonetheless gain some further insight into the seasonal cycle by writing $S' = S_1 \cos(\omega t)$. For this form of forcing, Eq. 7.15 can be solved most easily by using complex exponentials. Since $S' = S_1 Real(\exp(-i\omega t))$, the solution may be written $T' = Real(A \exp(-i\omega t))$. Substituting this form of solution into Eq. 7.15 we find

$$A = \frac{(1-\alpha)S_1}{\rho c_p H} \frac{1/\tau + i\omega}{1/\tau^2 + \omega^2} = |A| e^{i\Delta} \tag{7.17}$$

where the phase and amplitude are

$$\Delta = \arctan(\omega\tau), |A| = \frac{(1-\alpha)S_1}{\rho c_p H} \frac{1}{\sqrt{(1/\tau^2 + \omega^2)}} \tag{7.18}$$

With these definitions, the solution can be written

$$T'(t) = |A| \cos(\omega t - \Delta) \tag{7.19}$$

The character of the response depends on the period of the forcing relative to the characteristic response time of the system. This determines both the amplitude of the fluctuation and the phase shift relative to the forcing. For $\omega\tau << 1$ we have $\Delta = 0$ and $|A| = (1-\alpha)S_1/b$. For $\omega\tau >> 1$ we have $\Delta = \pi/2$ and $|A| = (1-\alpha)S_1/(\rho c_p H \omega)$. Note that in this case the temperature fluctuation becomes weak in inverse proportion to the frequency of the solar forcing fluctuation. These are special cases of the limits discussed previously, but we now have the further advantage of an explicit formula showing how the phase and amplitude of the seasonal cycle vary between the two extreme cases.

So far, we have not specified the flux which is to be used for the solar heating term ($(1-\alpha)S$) and the heat loss term $F(T)$ in Eq. 7.11. One possibility is to use the top-of-atmosphere fluxes, which are purely radiative. The other is to use surface fluxes, which include turbulent as well as radiative heat exchange between the planetary surface and the atmosphere, as discussed in Chapter 6. There are two thermal reservoirs at play – the atmosphere and the ocean – and a full treatment would require writing separate energy budgets for each, and accounting for the fluxes between them. There are several important cases, however, in which the thermal inertia of one or the other reservoir dominates, making it possible to make do with a one-layer model. The appropriate flux to use in such a one-layer model depends on which reservoir dominates. There are four cases to consider.

There are two circumstances in which the top-of-atmosphere fluxes are the appropriate ones to use. First, if the time scale under consideration is long enough that the surface budget stays near equilibrium, then the net solar flux transmitted by the atmosphere and absorbed at the surface is equal to the net turbulent and infrared flux passing from the surface into the atmosphere. In this case, the bottom boundary is energetically closed, and the energy budget of the atmosphere-ocean column can be determined from the top-of-atmosphere fluxes. In this case, the thermal inertia is provided by the atmosphere, and one uses the atmosphere's equivalent mixed layer depth in the mixed layer model equations. In this limit, the atmosphere is considered to be the mixed-layer "ocean," and the underlying surface simply acts to hand back to the atmosphere any energy it receives.

Alternately, if the response time of the atmosphere is short enough compared to the time scale under consideration, the energy budget of the atmosphere comes into equilibrium. In this case, by the definition of equilibrium, the net flux (solar plus infrared) entering the top of the atmosphere must equal the net flux (solar plus infrared plus turbulent) leaving the bottom of the

atmosphere. In this case, one can use the top-of-atmosphere absorbed solar radiation for the heat gain term and the OLR for the heat loss term in the surface energy budget, obviating the need to know the detailed physics behind the surface-to-atmosphere energy transfer. In this case, the thermal inertia is provided by the heat capacity of the mixed layer ocean.

In either case, one can compute $OLR(T)$ using a radiation model and some assumption linking the temperature and humidity profile to surface temperature, or one can use one of the linear or polynomial fits to the OLR curve discussed in Section 4. For example, with a linear fit to the OLR curve for a terrestrial atmosphere with 300ppmv CO_2 and 50% relative humidity, b is about $2(W/m^2)/K$ in the range 250K to 310K. The corresponding relaxation time τ is 1200 days for a 50 meter mixed layer, or 60 days for the 2.4m mixed layer which is equivalent to the thermal inertia of the Earth's atmosphere. In consequence, the seasonal cycle is expected to be strongly attenuated on the ocean- covered parts of the Earth (apart from coastal effects). The atmosphere alone does not have enough thermal inertia to damp out the seasonal cycle, but it does have enough thermal inertia to keep the atmospheric temperature roughly constant in the course of the diurnal cycle. Colder temperatures tend to make the relaxation time longer. For example, in an Earthlike atmosphere with 300ppm CO_2, the relaxation time roughly doubles at 160K. As noted earlier, Titan has a very long relaxation time owing to its thick atmosphere and low temperature; now we can make the statement more precise. Ignoring the greenhouse effect and setting $b = 4\sigma T^3, T = 90K$ we find a relaxation time of 20 Earth years, based on the equivalent 24m mixed layer depth of Titan's atmosphere. Since Titan's year (which is the same as Saturn's year) is about 30 Earth years, the seasonal cycle on Titan is expected to be considerably damped, though not so much so as the seasonal cycle over the Earth's oceans. The weak greenhouse effect from methane in Titan's atmosphere would somewhat enhance the damping. In contrast, a similar calculation for the thin atmosphere of present Mars gives a relaxation time of only .8 Earth days, based on $T = 200K$. Since a Mars day is approximately the same as an Earth day, the thermal inertia of the Martian atmosphere at present has relatively little damping effect on the diurnal cycle.

The thermal relaxation process is different if the time scale under consideration is short compared to the response time of the atmosphere, but long compared to the response time of the surface. In this case, which constitutes the third basic type of behavior, the atmospheric temperature remains approximately constant while the surface temperature fluctuates. This is the way the diurnal cycle works on Earth over ice or land. The relaxation time of surface temperature is then determined using the turbulent and radiative surface-atmosphere flux formulae discussed in Chapter 6, rather than the OLR. Because of the great thermal inertia of Titan's atmosphere and the low thermal inertia of the mostly solid surface, the seasonal cycle of Titan is somewhat in this regime as well. The situation of present Mars is not in this regime, since the atmosphere has little thermal inertia. There, the diurnal cycle affects the entire depth of the atmosphere, and the diurnal response is approximately governed by the OLR and the thermal inertia of the surface, much as for the Earth's seasonal cycle.

The fourth class of behavior in which a one-layer model suffices arises when the ocean mixed layer is so deep that the surface temperature can be regarded as nearly constant throughout the year. If there were no solar absorption in the atmosphere, the atmosphere would not have any seasonal temperature variation either, but solar absorption will drive a seasonal cycle in atmospheric temperature, for which the thermal inertia is provided by the atmosphere. Even if the solar absorption is small, the resulting temperature variations can be considerable if the atmosphere has fairly low thermal inertia, as is the case for Earth. This leads to a paradoxical situation in which the atmospheric temperature cycle is in phase with the insolation, even though the deep ocean mixed layer might lead one to think that the seasonal cycle should be a quarter year out

of phase. In a seasonal cycle of this type, neither the atmosphere nor ocean are near equilibrium. One fixes the ocean surface temperature and then computes the turbulent and radiative fluxes out of the bottom of the atmosphere using a surface budget model. One also needs to compute the net radiative fluxes into the top of the atmosphere. The difference, which is a function of the atmospheric temperature, give the heating that drives the atmospheric seasonal cycle. This case is explored quantitatively in Problem **??**. It is a case of real interest, since some aspects of the Southern Hemisphere seasonal cycle of atmospheric temperature appear to fall into this category.

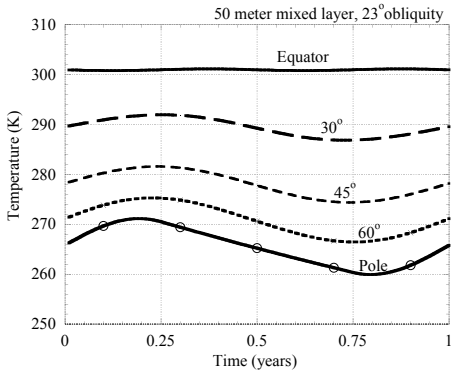

Figure 7.8: Numerically computed seasonal cycle for a $50m$ mixed layer ocean subject to a realistic seasonal cycle of insolation. The planet is in a circular orbit having an obliquity of 23^o. Infrared cooling was calculated from top-of-atmosphere balance with a CO_2 concentration of 300 $ppmv$ and fixed relative humidity of 50%. The time axis represents one Earth year. The Northern Hemisphere summer solstice occurs at time zero, and the Northern winter solstice is at time $0.5y$. See text regarding treatment of dynamical heat fluxes.

We are now equipped to compute the the seasonal cycle for an idealized Earthlike waterworld subject to realistic seasonal insolation. The object of this calculation is to provide a more quantitative feel for the extent to which oceans damp the seasonal cycle. The time series of insolation $S(t)$ is far from sinusoidal, particularly in the polar regions of planets with nonzero obliquity. The solution for this insolation cannot be obtained analytically, but it is an easy matter to numerically integrate the mixed layer equation subject to solar forcing obtained from the time dependence of the zenith angle given by Eq. 7.7 or its diurnally averaged version in Eq. 7.9. As we discussed in the context of the annual mean case, the temperatures obtained by doing a purely local balance are far too hot at the equator and far too cold at the poles, compared to Earth observations. This is because the real atmosphere takes some heat away from the tropical regions and deposits it in the polar regions. In order to represent this transport and get the model in a roughly realistic temperature range, one needs to put in the effect of the heat transport by *fiat*, in the absence of a fully dynamical model of heat transport. Similar considerations apply to any planet with an atmosphere that is dense enough to transport a significant amount of heat. In the present calculation, we will represent the effects of the heat transport by simply adding a latitude-dependent dynamic flux F_d to the column budget at each latitude. We'll assume this flux to be independent of time. F_d is negative in the tropics, where dynamics takes heat out of the column, and it is positive in the polar regions. For the present Earth, the magnitude of F_d can be determined by examining satellite-derived top-of-atmosphere energy budgets, or more simply it can be set at a value that gives approximately the right annual-mean temperature at each latitude. For the calculation to

follow, we set F_d to $-35W/m^2$ at the Equator, $-20W/m^2$ at $\pm 30^o$ latitude, zero at $\pm 45^o$ latitude, $40W/m^2$ at $\pm 60^o$ latitude and $80W/m^2$ at the poles. These values are somewhat weaker than the observed dynamical fluxes on Earth, but given the other simplifications we have made in this calculation, they suffice to keep the temperatures in a reasonably Earthlike range.

The calculations were carried out for a 50 m mixed layer ocean, which is roughly consistent with observed mixing in Earth's ocean, though the real mixed layer depth does vary somewhat both geographically and seasonally. For this depth of mixed layer, the ocean dominates the thermal inertia at seasonal time scales, and the atmosphere can be considered to be in equilibrium. Hence, it is appropriate to use top-of-atmosphere fluxes to drive the model. We'll neglect the albedo variations that would be caused by the cloud distribution and the formation of sea ice, and adopt a uniform top-of-atmosphere value of 0.2. The infrared cooling was computed based on the clear-sky polynomial OLR fit in Table 4.2, for a CO_2 concentration of $300ppmv$. Results are shown for Earthlike obliquity, using a circular orbit of period one Earth year, in Figure 7.8. These results use the approximate diurnally averaged form of the flux function.

The seasonal cycle at the equator is so weak as to be barely detectable in the figure, but the amplitude grows steadily as one moves towards the poles. At 45^o the peak-to-trough amplitude is $6K$, which agrees quite well with the observed midlatitude amplitude within extensive ocean basins. The weak tropical seasonal cycle is also consistent with observations. Note that from 30^o through 60^o latitude, the peak and trough is shifted a quarter year relative to the solstices, as is expected in the case of strong thermal inertia. Also, at these latitudes the seasonal cycle in temperature looks quite sinusoidal; this is in part because the insolation is fairly sinusoidal at these latitudes, but also because the mixed layer acts as a low-pass filter, damping down the response to higher harmonics of the insolation time series.

Because ice albedo feedback is not included, the high-latitude results cannot be expected to yield a realistic seasonal temperature pattern once ice forms. The calculations do give an indication of how strong the high-latitude seasonal cycle would be if the climate were warmed up enough to suppress the formation of ice, and also provide a reasonable estimate of temperature variations during the ice-free part of the year. It is significant that despite the long polar night, the ocean is only able to cool by about 10 K from its summer peak. Therefore, with a moderate depth mixed layer, formation of midwinter polar ice is far from inevitable.

The formation of polar ice is very sensitive to the values chosen for the high latitude dynamical heat flux and the albedo of the ice-free ocean; the latter of these is strongly influenced by low cloud cover. For the conditions of the present case, the ocean gets cold enough to form ice after the winter solstice at 60^o latitude, while at the pole the ocean is below freezing the entire year. Once ice forms, not only does the albedo increase, but also the effective thermal inertia is drastically reduced, since the solid ice layer insulates the atmosphere from the heat storage of the mixed layer. A discussion of these effects is deferred to Section 7.7, where we take up the very challenging question of how to account for the absence of high-latitude ice in hothouse climate like the Cretaceous.

Although this calculation has been tuned to Earthlike conditions, it gives a pretty good picture of the general state of affairs for moderate-obliquity waterworld planets in nearly circular orbits. The picture also applies to planets with atmospheres so thick they have equivalent mixed-layer depths comparable to $50m$. The general picture to take away is that such planets have relatively weak seasonal cycles in circumstances where polar ice fails to form; when polar sea ice forms, the high latitude seasonal cycle can become much stronger. All other things being equal, planets with a deeper equivalent mixed layer will have a seasonal cycle that is even more moderate than the one shown here, while planets with a shallower equivalent mixed layer will

exhibit more extreme seasonal variations. The results can be scaled to planets with different atmospheric radiative damping constants and different year lengths using the fact that the mixed layer equations with linearized radiative flux depend on the mixed layer depth and length of year only through the combination τ_y/τ, where τ_y is the duration of the year and τ is the response time derived previously. Since τ is proportional to mixed layer depth, a planet with twice the length of Earth's year but a mixed layer of depth $100m$ would have a seasonal cycle of similar amplitude to Earth's (other parameters being equal), and so forth. Halving the radiative damping coefficient $dOLR/dT$ would have the same effect as doubling the mixed layer depth. A detailed exploration of the behavior of the mixed layer equations for various planetary configurations is carried out in Problem **??**. To determine the appropriate dynamical heat flux for an arbitrary planet, however, requires either doing a full dynamical calculation, or invoking observations to constrain the annual mean temperature or radiation balance.

7.4.2 Thermal inertia of a solid surface

Heat diffuses slowly through a non-metallic solid, so when the underlying surface is solid it is typically necessary to consider the continous distribution of temperature as a function of depth within the solid. To a good approximation, heat flux within a solid is proportional to the temperature gradient; the proportionality constant is called the *thermal conductivity*, which we shall call κ_T. Balancing the rate of change of heat content against the convergence of heat flux yields the *diffusion equation*

$$\partial_t \rho c_p T = \partial_z \kappa_T \partial_z T \tag{7.20}$$

In this equation it is assumed that there are no internal heat sources. The surface heat budget enters the problem through the boundary condition at the surface ($z = 0$), which states that the diffusive heat flux into the surface equals the net heating of the surface by insolation and radiative and turbulent heat transfers. Using the same notation as we employed for the mixed layer case, this boundary condition reads

$$\kappa_T \partial_z T|_{z=0} = (1-\alpha)S(t) - F(T) \tag{7.21}$$

When S is a constant S_o, the problem is solved with a constant temperature T_o satisfying $(1-\alpha)S_o = F(T_o)$, just as for the mixed layer case. Linearizing the boundary condition about T_o and substituting the complex exponential form for S' yields

$$\kappa_T \partial_z T'|_{z=0} = (1-\alpha)S_1 e^{-i\omega t} - bT' \tag{7.22}$$

If ρc_p is constant, this boundary condition can be satisfied by a solution of the diffusion equation of the form

$$T' = Ae^{i(kz-\omega t)}, k = \sqrt{\frac{\omega}{D}}\frac{1-i}{\sqrt{2}} \tag{7.23}$$

where A is a constant and D is the diffusivity $\kappa_T/(\rho c_p)$. The complex vertical wavenumber k has been determined by substitution of the exponential form of T' into the diffusion equation. A will be determined by substitution of the solution into the boundary condition, but before doing so it is worth pausing to make some remarks on the solution Eq. 7.23. This solution was first obtained by Fourier, in his study of diurnal and seasonal variations of temperatures in the interior of the Earth. Eq. 7.23 shows that the characteristic depth to which temperature fluctuations penetrate is $\sqrt{(D/\omega)}$. Low frequency fluctuations penetrate to a greater depth than high frequency fluctuations, because heat has a longer time to diffuse before the surface temperature reverses. Note also that the phase lag of the time of maximum temperature with depth also reflects the time required for

	$\rho c_p (J/m^3)$	Conductivity $(Wm^{-1}K^{-1})$	Diffusivity (m^2/s)
Water Ice	$1.93 \cdot 10^6$	2.24	$1.16 \cdot 10^{-6}$
Fresh Snow	$.21 \cdot 10^6$	.08	$.38 \cdot 10^{-6}$
Old Snow	$1.0 \cdot 10^6$	.42	$.05 \cdot 10^{-6}$
Sandy Soil	$1.28 \cdot 10^6$	.3	$.24 \cdot 10^{-6}$
Clay Soil	$1.42 \cdot 10^6$	.25	$.18 \cdot 10^{-6}$
Peat Soil	$.575 \cdot 10^6$	.06	$.1 \cdot 10^{-6}$
Rock	$2.02 \cdot 10^6$	2.9	$1.43 \cdot 10^{-6}$
Lunar Regolith	$1 \cdot 10^6$	.01	$.01 \cdot 10^{-6}$

Table 7.1: Thermal properties of some common surface materials

the surface conditions to penetrate to the interior. For the diffusivity of water ice (Table 7.1) the characteristic depth is 12 cm for the diurnal period, 2.4m for the annual period, 24m for a century and 76m for a millennium. Solid rock yields similar numbers. Hence, the temperature profile within ice or rock still contains information about temperatures centuries or even millennia in the past, albeit in a rather smoothed and degraded form. This fact has been exploited in reconstructions of past temperatures.

Exercise 7.4.3 You are designing a lunar colony to be placed at a Lunar latitude where the sun is directly overhead at noon. The moon has an albedo close to zero, and the response time of the surface is rapid, so that the noontime surface temperature is close to the instantaneous equilibrium temperature of 394K (re-derive this temperature yourself). At night, the equilibrium temperature would be absolute zero, but there is not enough time to reach equilibrium; still the night-time temperature plummets to 100K. Since the Moon is tide-locked to the Earth, the Lunar day is 28 Earth days. The diffusivity of the Lunar regolith ("soil") is about $10^{-8} m^2/s$.

Approximate the day-night temperature variation by a sinusoidal curve. What would be the constant temperature far below the surface (neglecting internal heat sources)? How deeply would the colony habitat have to be buried in order for the ambient diurnal temperature fluctuations to be less than 1K?

NB: Given the low diffusivity of the regolith, your main difficulty is likely to be getting rid of the heat generated by energy use (biological and otherwise) within the colony.

Now we substitute Eq. 7.23 into the boundary condition (7.22). The result is

$$\begin{aligned} A &= \frac{(1-\alpha)S_1}{b + \rho c_p \sqrt{\omega D}\frac{1+i}{\sqrt{2}}} \\ &= \frac{(1-\alpha)S_1}{b}\frac{1}{1+\sqrt{\omega\tau_D}\frac{1+i}{\sqrt{2}}} \\ &= \frac{(1-\alpha)S_1}{\rho c_p\sqrt{D/\omega}}\frac{1}{\frac{1}{\tau_1}+\frac{1+i}{\sqrt{2}}\omega} \end{aligned} \tag{7.24}$$

where $\tau_D = (\rho c_p)^2 D/b^2$ and $\tau_1 = \rho c_p \sqrt{D/\omega}/b$. Upon comparison of the third line of this equation with the solution for the mixed layer model, it is seen that the solid case acts somewhat like a mixed layer model with frequency dependent layer depth $\sqrt{D/\omega}$. For low frequency forcing, $\omega\tau_D \ll 1$, the surface temperature follows the instantaneous equilibrium, $A = (1-\alpha)S_1/b$, just

as for the mixed layer case. For high frequency forcing, the amplitude of the surface temperature fluctuation decays like $1/\sqrt{(\omega)}$. This is slower than was the case for the fixed-depth mixed layer, since the layer determining the thermal inertia now gets thinner as frequency is increased. Note also that the phase lag of surface temperature relative to insolation differs from the mixed layer case. For the diffusion equation, the surface temperature lags the insolation by $\pi/4$ radians in the high frequency limit rather than $\pi/2$. The thinning of the active thermal layer keeps the surface temperature closer to instantaneous equilibrium than it would be in the fixed-depth case.

Apart from some exceptional circumstances, the thermal inertia of a solid surface has little effect on the seasonal cycle, though it can substantially moderate the diurnal cycle. This can be seen easily through the evaluation of tau_D in a few typical cases. First we consider the case of Antarctic or Arctic ice-covered regions. The flux coefficient based on a linear OLR fit in the temperature range 240K to 270K is $b = 2.16W/(m^2K)$. Using the heat capacity and thermal diffusivity for water ice, given in Table 7.1, we find $\tau_D = 11d$. At latitudes somewhat away from the poles, the diurnal cycle of insolation becomes significant, particularly during the equinoxes. Since the time scale for the surface is shorter than that for the atmosphere, it would be more appropriate to use surface flux coefficients than OLR in analyzing the terrestrial diurnal cycle. As noted in Chapter 6, the turbulent heat transfer is strongly inhibited at night-time, when the boundary layer is statically stable. In this case, the flux coefficient is dominated by the radiative term $4\sigma T^3$ based on surface temperature. For temperatures around $255K$ this yields an even shorter response time $\tau = 4d$. In the midlatitudes and Tropics, the estimate differs only in the use of the slightly larger values of b appropriate to the warmer temperatures, and the somewhat different thermal properties of rock or soil, but the result remains that τ_D is on the order of a few days or less. For Mars, one may use $b = 4\sigma T^3$ based on $T = 200K$, given the thin atmosphere. This yields $\tau_D = 15d$, which is still not sufficient to appreciably affect the seasonal cycle. It is only at the extremely cold temperatures of Titan that the response time of a solid ice surface becomes significantly longer (roughly 1300 Earth days), but even there the effect is of little interest, owing to the much longer response time of Titan's atmosphere. In sum, a solid surface can generally be considered to be in equilibrium for the purpose of computing temperature fluctuations on the seasonal time scale.

It should not be concluded from the above estimates that the thermal inertia of solid surfaces is sufficient to eliminate the diurnal cycle. The variation of insolation between noon and night-time is huge; On Earth, at a latitude where the Sun is overhead at noon, the amplitude of the variation is $1370W/m^2$, which leads to an undamped temperature fluctuation of 685K based on a flux coefficient $b = 2W/(m^2K)$. Even damped by a factor of 20, this amounts to a very considerable diurnal fluctuation. Similar considerations apply to the Martian diurnal cycle.

It should further be noted that while solids don't provide much thermal inertia for the surface temperature, they do provide substantial thermal inertia for the subsurface temperatures. Consider a nearly airless body,like the Moon or Mars or Triton. One sees a diurnal cycle corresponding nearly to instantaneous equilibrium at the surface, but at a sufficient depth the diurnal cycle is filtered out and one sees only the seasonal cycle. At still greater depths the seasonal cycle is filtered out and one sees only the annual mean. At yet greater depths, what remains are the effects of long-term climate variations. A calculation illustrating this behavior is carried out in Problem **??**.

As a complement to the periodically forced solution, Figures 7.9 and 7.10 show the solutions for the diffusion equation in water ice which is initialized at a uniform temperature of 300K and allowed to cool without solar heating subject to a flux upper boundary condition. The heat loss from the surface was computed using an Earthlike OLR fit $OLR(T_s) = 48.461 + 1.5866(T_s - 180) + .0029663(T_s - 180)^2$ A quadratic fit was used so that the the fit would remain accurate over a large temperature range. Except for the high initial temperature, which turns out to be inconsequential,

this problem can be thought of as representing the cooling of the Antarctic ice cap after permanent winter night closes in. Figure 7.9 illustrates the progressive penetration of the surface cooling into the depth of the ice; at time t, the cooling has penetrated to a depth on the order of $\sqrt{Dt}$, where D is the thermal diffusivity of the ice. Figure 7.10 shows that there is an extremely rapid initial cooling, owing to the thin layer of ice affected at short times. After a half day, the temperature has already fallen below freezing. Therafter, the temperature drop becomes slower, as the depth of the ice layer involved becomes greater. The reduction in OLR as temperature drop also contributes to the reduction in cooling rate. Nonetheless, after two months, the temperature has fallen to 190K, which is well below the 235K minimum temperature observed at the South Pole. Incorporation of the atmosphere's thermal inertia reduces the cooling rate somewhat, but does much increase the extremely cold temperature encountered at the end of the winter. Clearly, the Antarctic interior relies on heat transport from warmer latitudes to limit its winter temperature drop.

We conclude this section with a few remarks on the special effects of snow and ice (whether from water, CO_2 or some other substance) on the seasonal and diurnal cycle. Snow has a profound effect on the diurnal cycle, because of its very low thermal conductivity, which is nearly an order of magnitude lower than that of ice (see Table 7.1 for the case of water snow). The low thermal conductivity arises from the high proportion of the snow's volume which consists of air trapped in pores which are too small to allow the air to flow; since air itself has extremely low thermal conductivity, heat must primarily make its way through the contorted pathways of snow crystals in contact with each other. Other gases, trapped in snows made of other substance, have a similar effect. The low conductivity dramatically reduces the characteristic response time of the surface, even for a snow layer of modest thickness. In the Antarctic case discussed above, τ_D drops to a mere 60 minutes for old snow, and 20 minutes for fresh snow. At night, the temperature of the snow surface plunges almost instantaneously to its equilibrium value. In the case of the Earth, the atmosphere has sufficient thermal inertia that it doesn't cool much at night, above the boundary layer. Given the suppression of turbulent flux in the stable nocturnal boundary layer, the night-time equilibrium temperature is maintained mainly by the downwelling infrared flux from the atmosphere, as discussed in Chapter 6. When the low level air temperature is 255K, the downwelling infrared flux is about $120 W/m^2$, maintaining a snow surface temperature of 214K. On present Mars, the atmosphere cools down markedly at night, and in any event is too thin to provide much downwelling flux, so it is less obvious what limits the night-time temperature drop over the CO_2 snow fields that form in the winter hemisphere. One relevant consideration is that the flux coefficient b drops dramatically at very cold temperatures, leading to an increase of the relaxation time; when the surface temperature falls to 150K, τ_D increases to 23 hours even over snow. However, at such low temperatures the saturation vapor pressure of CO_2 is only 1.26mb, well below the ambient surface pressure. Hence, the night-time temperature minimum is likely to be governed by the latent heat release due to CO_2 condensation, which sets in at surface temperatures near 160K.

Snow cover on any planet can change rapidly in the course of the seasons, and on Earth, sea ice cover similarly expands and retreats. Since snow and ice have higher albedo than the surfaces they generally cover, this has an important feedback effect on the seasonal cycle. It enhances the winter-time cooling once ice or snow begin to accumulate, delays the springtime warming, but then accelerates the warming once ice or snow begin to retreat. The albedo feedback of snow is especially pronounced, since snow has a much higher albedo than ice. For water snow, for example, the albedo of fresh snow averaged over the solar spectrum can exceed .85, whereas a typical albedo for sea ice is on the order of .6. The high albedo of snow, like its low diffusivity, arises from its highly porous nature which offers many opportunities for light to encounter discontinuities in index of refraction, leading to scattering. It is a generic property of the snow of any weakly-absorbing substance. Note that the concept of "sea ice" is peculiar to planets with water oceans. On a planet

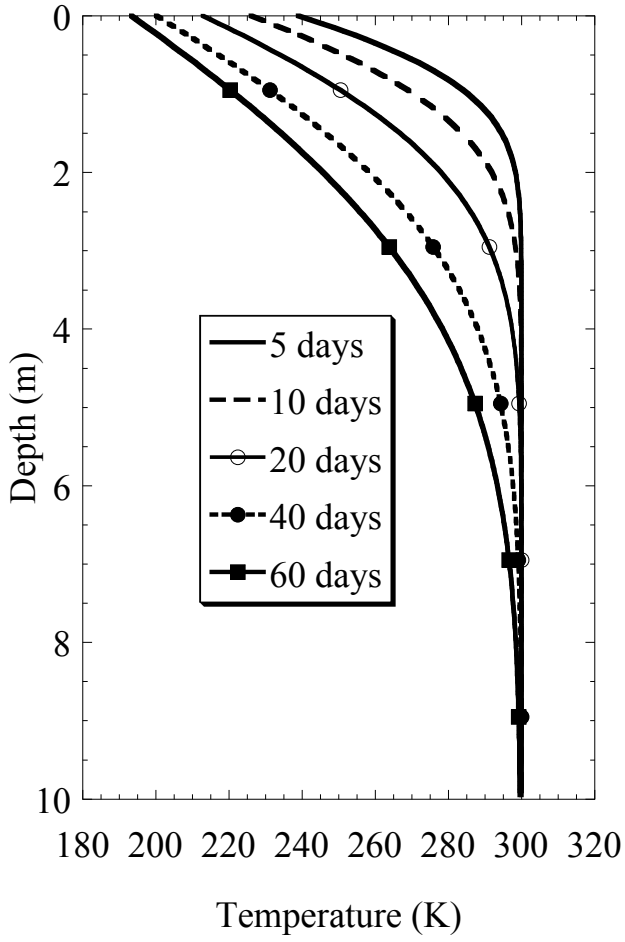

Figure 7.9: Temperature vs. depth at various times, for an ice layer subject to temperature-dependent heat loss at the surface. See text for specification of the heat loss rate.

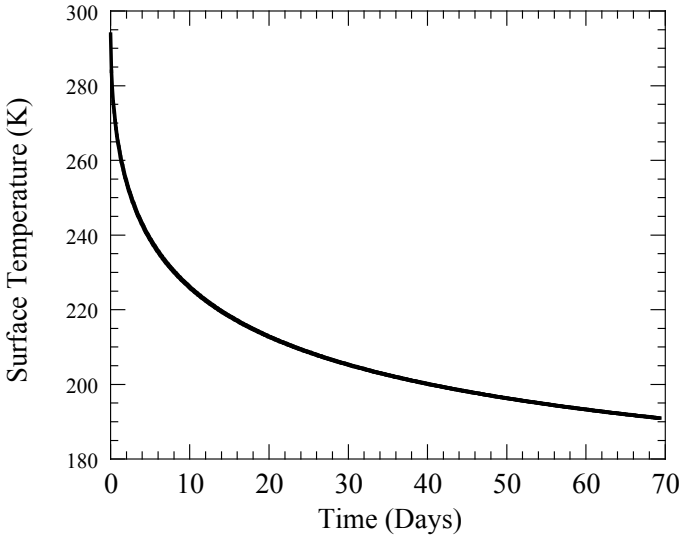

Figure 7.10: Time evolution of surface temperature for the solution shown in Figure 7.9

with a liquid methane or CO_2 ocean "sea ice" would sink, and not have any chance to affect the surface albedo until the ocean were frozen to the bottom.

The presence of a solid phase on the surface of the planet also introduces a new form of thermal inertia, associated with the latent heat of phase change from the solid to liquid form. Where there is ice, whether it be in the form of sea ice or land glaciers, the surface temperature cannot rise above the triple point (the "melting point") until all the ice has been melted. The phenomenon is familiar from an experiment commonly performed in elementary school science classes, in which one tries to boil a pot of water containing ice cubes,and finds that the temperature doesn't start to rise above freezing until all the ice is gone. Thermal inertia effects due to growth and decay of ice are discussed further in Section 7.7 and Problem **??**.

7.4.3 Summary of thermal inertia effects

The preceding discussion has revealed two limiting forms of behavior a planet can exhibit in the course of its seasonal cycle. A "waterworld," having high thermal inertia in the ocean-atmosphere system, responds primarily to the annual average insolation. Such a world will be coldest at the poles and warmest at the equator, unless the obliquity exceeds about 54^o, in which case the warmest climates will be found near the two poles. A "desertworld," having little thermal inertia in either the surface or the atmosphere, responds to the instantaneous insolation at each time of the year. The location of the highest temperature moves from some latitude north of the equator to the same latituded south of the equator, and back again, in the course of the year. The hot-spot crosses the equator twice per year, leading to a basic rhythm for tropical climate which is twice as fast as higher latitude regions. For small obliquity, the poles are frigid throughout the year, and the hot spot executes modest excursions about the equator. For obliquities greater than about 18^o, the excursion goes all the way from pole to pole, assuming a uniform albedo. Geographical and temporal albedo variations alter this picture. Formation of permanent ice or snow cover near

the poles will tend to keep the polar regions cold throughout the year; this effect is assisted by the thermal inertia implied by the latent heat required to melt or sublimate ice, which limits the summertime temperature increase. The desertworld case is explored in Problem **??**.

Thermal inertia sufficient to moderate the seasonal cycle can be provided either by a thick atmosphere or a well mixed liquid layer at the surface. A liquid layer need only have a depth of some tens of meters to have a significant moderating effect, though the precise amount of thermal inertia needed to moderate the seasonal cycle must always be considered relative to the length of the planet's year. Heat storage provided by non-melting solid surfaces is almost never sufficient to have a significant affect on the seasonal cycle, though it can substantially moderate the diurnal cycle for planets with rotation periods on the order of a few Earth days or less.

The Earth shows some characteristics of both limiting cases, with extreme continental climates and equable maritime climates. In the deep Tropics, the twice-per-year rhythm of land temperature contrasts with the nearly constant ocean temperature,leading to monsoonal circulations driven by the strong land-sea temperature gradient.

7.5 Some elementary orbital mechanics

Sir Isaac Newton showed that the orbit of a single planet revolving about its star takes the form of an ellipse, with a focus of the ellipse at the center of mass of the system. Since stars are typically much more massive than their planets, the center of mass for most purposes is identical to the center of the star. The elliptical nature of orbits has an important effect on the seasonal cycle, since the planet is farther from its sun at some parts of the year than it is at others. This makes the solar "constant" $L_\circledast$ a function of time of year. On Earth, we don't notice this effect too much because our orbit is nearly circular. Nonetheless, the effect has an important influence on the long-term evolution of climate. On other planets, it can be even more important.

The distance of closest aproach of a planet to its star is called the *perihelion*, which we shall call r_p. The greatest distance is called the *aphelion*, which we shall call r_{ap}, The *semi-major axis* is then $a = (r_p + r_{ap})/2$. Let κ_1 be the angle made by the line between the star and the planet, defined so that $\kappa_1 = 0$ at the perihelion. Then, in polar coordinates, the equation of the elliptical orbit is

$$r = a\frac{1-e^2}{1+e\cos(\kappa_1)} \tag{7.25}$$

where e is the *eccentricity* of the orbit, which lies in the interval $[0, 1]$. $e = 0$ yields a circular orbit, while the ellipse becomes progressively more elongated as $e \to 1$. Specifically, perihelion is $(1-e)a$, the aphelion is $(1+e)a$ and the ratio of the distance at aphelion to the distance at perihelion is $(1+e)/(1-e)$. To get the semi-minor axis, we maximize $r(\kappa_1)sin(\kappa_1)$, yielding $a\sqrt{1-e^2}$. Hence, the ratio of the minor to major axis is $\sqrt{1-e^2}$. The geometry of the orbit is summarized in Figure 7.11.

Exercise 7.5.1 What eccentricity would yield an ellipse with a 3:1 axis ratio? Sketch such an ellipse, indicating the correct location of the Sun relative to the orbit.

The variation of the solar constant is then given by

$$L_\circledast = \frac{1}{4\pi}\frac{I_o}{r^2} \tag{7.26}$$

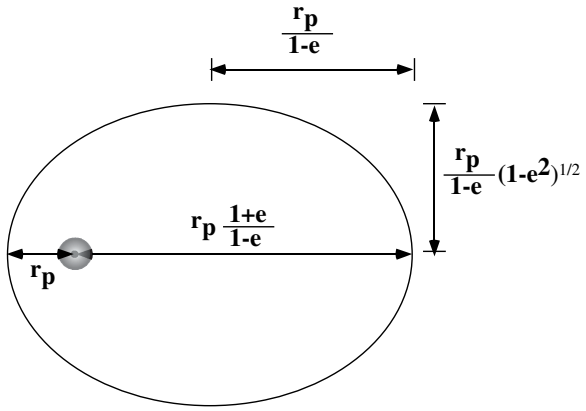

Figure 7.11: Geometry of an elliptical orbit with eccentricity $e = .66$

where I_o is the power output of the star (e.g about $3.8x10^{26}$ Watts for the Sun at present). The annual variation in distance from the Sun leads to "distance seasons," which are synchronous between the hemispheres. This contrasts with the "obliquity seasons" (dominant for Earth and Mars) which are out of phase between the hemispheres, one hemisphere enjoying winter while the other suffers under the torrid heat of summer. In the limit of small eccentricity, the ratio of solar constant at aphelion to that at perihelion is $1 + 4e$. This represents a very considerable variation, even for modest eccentricity. For the present eccentricity of the Earth (.017), it amounts to 6.8%, or $93W/m^2$ difference in the solar constant between perihelion and aphelion. To turn this flux into a crude temperature estimate, we divide 4 to account for the averaging over the Earth's surface, and apply a typical terrestrial $OLR(T)$ slope of $2W/(m^2K)$, yielding a temperature difference of more than 11K between perihelion and aphelion. This represents the amplitude of the distance seasons. For Mars, with its present eccentricity of .093, the effect is even greater. The perihelion to aphelion flux variation is 37%, or $219W/m^2$. For Martian conditions, where the atmosphere has a weak greenhouse effect, this translates into an amplitude of 30K.

To determine the time dependence of r, we must know $\kappa_1(t)$. Because the orbit is no longer circular, the angular velocity is no longer constant; the planet moves faster when is close to the sun than when it is farther away. There is no analytic expression for the time variation of the orbital position. However, it can be easily computed by numerically solving a first order differential equation, which can be derived either from Kepler's equal-area law, or directly from angular momentum conservation. We shall take the latter route. Let $v_\perp$ be the component of velocity perpendicular to the line joining the planet to its star. Then, by conservation of angular momemtum, $rv_\perp = J$ is independant of time. However, the angular velocity of the orbit is simply $v_\perp/r$, so the angle satisfies the equation

$$\frac{d\kappa_1}{dt} = \frac{J}{r^2} = \frac{J}{a^2}\frac{(1 + e\cos(\kappa_1))^2}{(1 - e^2)^2} \tag{7.27}$$

This equation shows that the angular velocity of the planet speeds up as it approaches perihelion, and slows down as it approaches aphelion. In consequence, the planet spends less time near the sun than it does at greater distances, and "distance summer" is shorter than "distance winter."

The average of the solar constant over the course of the year can be written

$$< L_\circledast >= \frac{1}{4\pi}\frac{I_o}{a^2} < \frac{a^2}{r^2} >= L_a < \frac{a^2}{r^2} > \tag{7.28}$$

where angle brackets denote the average over the planet's year and L_a is the solar constant evaluated at a distance equal to the semi-major axis of the orbit. We can take advantage of the fact that the same $1/r^2$ factor appears in Eqn. 7.27 to relate the mean solar constant to the nondimensionalized duration of the planet's year. Specifically, integrating Eqn. 7.27 over one year and dividing by the length of the year yields

$$< L_{\circledast} >= \frac{1}{\tau_y^*} L_a \tag{7.29}$$

where $\tau_y^* = \tau_y/(2\pi a^2/J)$, τ_y being the length of the year in dimensional terms. The quantity $2\pi a^2/J$ is the length of year for a circular orbit with radius a. Numerical integration of Eqn. 7.27 shows that the nondimensional year defined in this way decreases as the orbit becomes more eccentric. For $e = .1$, τ_y^* is .995, for $e = .25$, τ_y^* is .968, and $e = .5$, τ_y^* is .866.

Most Solar System planets have nearly circular orbits; leaving out Mercury and Pluto, the planets have current eccentricities ranging from .007 to .093. Even Pluto only has a value of .244, though other large Kuiper Belt Objects have higher eccentricity. The most nearly habitable of the close-orbiting Super Earths in the Gliese 581 system have orbits that are quite eccentric by Solar System standards: 0.16 for Gliese 581c and 0.38 for Gliese 581d, based on the best estimates availailable in the year 2009. Planetary systems with more highly eccentric orbits are common,and appear at present to represent the rule rather than the exception (See 1.3). This makes it important to understand the seasonal cycle of planets with highly eccentric orbits, though most of the detailed exploration of that regime will be left to the Workbook problems.

Note that the difference between perihelion and aphelion distance is $O(e)$, whereas the ratio of major to minor axes deviates from unity by only $O(e^2)$. Hence, for small e, the orbit still looks like a circle, but with the Sun displaced from the circle's center by $O(e)$. For small e, Eqn. 7.27 can be solved approximately by a straightforward expansion in e. Substituting

$$\kappa_1(t) = \frac{J}{a^2}[t + eF(t) + e^2 G(t)] \tag{7.30}$$

into the equation and matching like terms in e yields the solution

$$\kappa_1 = 2\pi t^* + 2e\cos 2\pi t^* + e^2[\pi t^* + \frac{5\pi}{2}\sin 4\pi t^*] + O(e^3) \tag{7.31}$$

where $t^* = tJ/(2\pi a^2)$. The first order term causes an $O(e)$ variation in the orbital angular velocity over the course of the year, but it this term by itself does not alter the length of the year. Taking into account the second order term, it may be inferred that the nondimensional length of the year is approximately $\tau_y^* = 1 - e^2/2$. In consequence, the annual mean insolation varies very little from what it would be for a circular orbit with radius equal to the semi-major axis. For $e = .1$, close to the present value for Mars, the eccentricity increases mean insolation by only .5%. For $e = .02$, similar to Earth at present, the increase is a meager .02%, or $.274 W/m^2$. Except in very unusual cases, orbital eccentricity affects the climate through the intermediary of the seasonal cycle, and not through any effect on the annual mean radiation budget.

The consequences of orbital eccentricity for a planet's climate derive from the way the distance seasons interact with the tilt seasons. Each of these types of seasons has a period of one planetary year, so the nature of the interaction is governed by the position in the orbit at which the Northern Hemisphere summer solstice occurs, measured relative to the position of the perihelion. This can be measured by an angle, called the *precession angle* or *precession phase*. We will define the phase such that when it is zero, the Northern Hemisphere solstice occurs at

the perihelion. It is also common to define the phase as the angle between the perihelion and the Northern Hemisphere spring ("vernal") equinox. When the precession angle is zero, the distance seasons make the Northern Hemisphere seasonal cycle stronger, since "Northern tilt summer" happens when the planet is closest to the Sun and ""Northern tilt winter" happens when the planet is farthers from the Sun. Conversely, the Southern Hemisphere seasonal cycle is attenuated when the precession angle is zero. When the precession angle is 180^o, the situation is reversed between the hemispheres, with the Southern Hemisphere getting very hot summers and very cold winters, and the Northern Hemisphere experiencing more moderate seasons. When the precession angle is 90^o or 270^o, the solstices conditions are no longer modulated by the distance seasons, but instead the vernal equinox becomes warmer than the autumnal equinox, or *vice versa*.

Figure 7.12 illustrates the effect of eccentricity and precession on the seasonal cycle of insolation. These results were computed by numerically solving Eq. 7.27, and substituting $\kappa_1(t)$ into the flux distribution function given by Eq 7.10 and Eq 7.9, after shifting its phase to account for the precession angle. Given $\kappa_1(t)$, we also know $r(t)$. Using this, we multiply the flux factor by $(a/r(t))^2$ to account for the variations in orbital distance. This is the quantity plotted, at selected latitudes, in Figure 7.12. One multiplies this flux factor by the solar constant at a distance equal to the semi-major axis, in order to obtain the actual insolation in W/m^2. Using the symmetries of Eqn. 7.9, the results for precession angles of 180^o and 270^o can be obtained from those shown in Figure 7.12 by simply shifting the curves shown by a half year, and interchanging the two hemispheres, so these cases do not require separate discussion.

For both eccentricities, we see that the Northern Hemisphere extratropical seasonal cycle is made more extreme when the precession angle is 0^o, while that in the Southern Hemisphere is moderated. At the Equator, the two equinoxes have identical insolation, but the time of maximum equatorial insolation is shifted towards the Northern summer solstice, which is also the time of perihelion in this case. For the larger, Marslike, eccentricity ($e = .1$), the maximum equatorial insolation in fact occurs at the solstice. For the case of 90^o obliquity, the extratropical seasonal cycle has identical strength in both hemispheres, but the equinox conditions now differ from each other, the Autumnal equinox receiving less insolation than the Vernal (Spring) equinox. Also, the time of maximum and minimum extratropical insolation is also significantly displaced from what it would be for a circular orbit. The effect of orbital velocity variations on the seasonal cycle is just barely visible for the lower, Earthlike, eccentricity, but it is prominent for the higher eccentricity case. For 0^o precession, Summer is longer than Winter in the Southern hemisphere, while Winter is longer than Summer in the Northern hemisphere; for 90^o precession, there is a marked asymmetry between the rate of increase of insolation going into each season, and the rate of decrease coming out of it. For example, in the Northern Hemisphere, Summer sets in rapidly, but the transition to Winter takes a long time. In fact the Northern hemisphere, Southern hemisphere and equatorial insolation maxima are all bunched up within a period of about a quarter of a year, indicating that the distance seasons are beginning to dominate the tilt seasons even at this modest eccentricity. The effect of precession phase on the annual average insolation at each latitude is insignificant; for both the high and low eccentricity cases shown in Figure 7.12, changing the precession phase leaves the annual mean flux factor unchanged to at least four decimal places.

Note that the precession angle has a big effect on climate when the eccentricity is large, but has no effect when the eccentricity is zero. The effects of precession angle and orbital eccentricity work in conjunction with each other, and cannot be disentangled.

At present, Earth's precession angle is close to 180^o, so that the Southern hemisphere is driven towards hotter summers and colder winters, while the Northern hemisphere is driven towards a weaker seasonal cycle. This pattern is not manifest in the observations (Figure 7.2) because the Northern Hemisphere has more land than the Southern Hemisphere, giving it a stronger seasonal

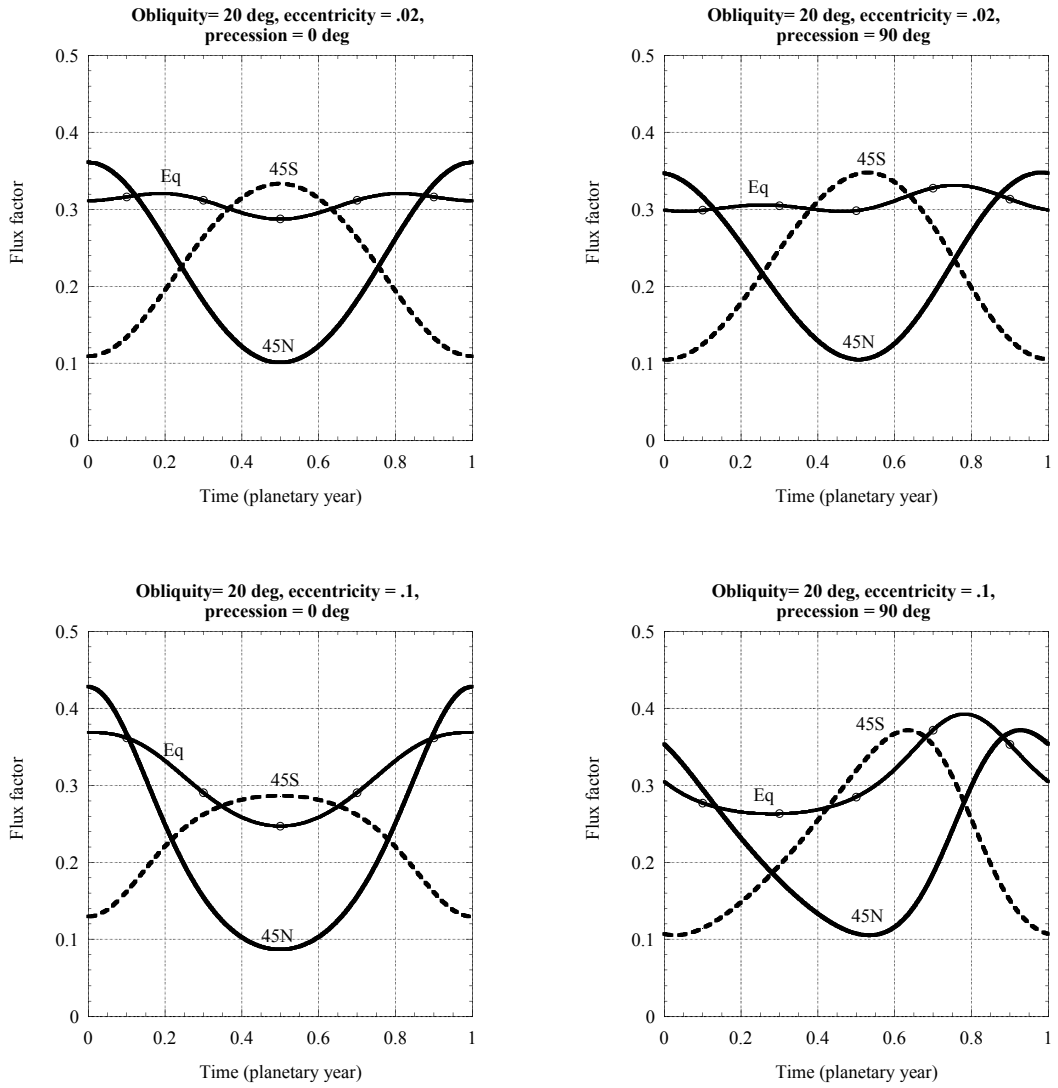

Figure 7.12: The seasonal cycle of solar flux factor for a planet with 20^o obliquity, at the Equator, 45N and 45S. To obtain the insolation at any given time of year, this flux factor is multiplied by the solar constant at the time of perihelion. Results are shown for an Earthlike eccentricity of .02 (top row), and a Marslike eccentricity of .1 (bottom row). The left column gives results for a precessional phase of zero degrees, while the right gives results for 90 degrees, both measured relative to the Northern Hemisphere summer solstice.

cycle, owing to its lower thermal inertia. Relatively speaking, though, the Northern Hemisphere seasonal cycle is weaker than it would be if the precession angle were 90^o or 0^o. Coincidentally, the precession angle of Mars is also about 180^o at present, so that the Southern Hemisphere Martian winters are expected to be considerably colder than those in the North. Evidence that this indeed occurs, and its broader implications for Martian climate, will be taken up in Section 7.7.

The precession angles and orbital eccentricities of Earth and Mars have been different in the past, and will be different in the future. This has some extremely important implications for the evolution of climate, to which we now turn our attention.

7.6 Effect of long term variation of orbital parameters

The three *orbital parameters* that govern the seasonal and geographical distribution of insolation are the precession angle, obliquity, and eccentricity. All three change gradually on a scale of many thousands of years, owing basic laws of mechanics which apply to any planet in any solar system.

The evolution of the precession angle derives from a fairly elementary property of the mechanics of rigid-body rotation. The rotation axis of a rotating body subject to a net torque executes a rotation at constant rate about a second axis whose orientation is determined by the torque. The *precession rate* is determined by the magnitude of the torque and the angular momentum of the rotating body. The phenomenon of precession can be easily observed on a tabletop, by setting down a toy gyroscope with its axis inclined from the vertical. The top will precess, because there is a torque caused by the Earth's gravity and the force of the tabletop pushing up on the point of the top. For planets, the torque instead is provided by the slight deviations of the mass distribution from spherical symmetry. The equatorial bulge caused by rotation is a major player, but other asymmetries, including those due to the distribution of ice, and of major geographic features, are also of consequence.

Obliquity variations also stem from the basic properties of rigid-body rotation, but these variations arise from fluctuations in the torque on the planet, rather than the mean torque. The obliquity cycle is inextricably linked with the precessional cycle, which modulates the orientation of the aspherical planet with respect to the non-uniform gravitational field caused by the Sun, the planet's moon(s) (if sufficiently massive), and all the other planets.

Eccentricity evolves because the periodic elliptical orbit is a solution only of the two-body problem, consisting of a planet and its star in isolation. Although the gravity of the Sun greatly dominates that of the other planets in our Solar System (and most likely in other planetary systems as well) the relatively small tugs of the planets on each other causes eccentricity to change gradually. Early in the history of this subject, it was shown by Laplace and Lagrange that the semi-major axis remains very nearly constant in the course of such eccentricity changes. The results of the preceding section therefore imply that eccentricity cycles have only a weak effect on annual mean insolation, since the mean insolation changes little if the semi-major axis is held fixed, except for extremely non-circular orbits.

Tiny deviations of the stellar gravity field from the ideal $1/r^2$ law add up to significant effects on obliquity and eccentricity over sufficiently long periods of time. The fact that the Sun is not perfectly spherical enters the problem, and even general relativistic deviations from Newtonian gravity have major effects.

Eccentricity modulates the distance seasons, and precession determines whether they constructively or destructively interfere with the tilt seasons. Meanwhile, obliquity variations modu-

late the strength of the tilt seasons. The net result is a rich variety of rhythms and patterns in insolation, which may lead to dramatic cycles in the state of a planet's climate.

In the following we discuss Milankovic variations for Earth and Mars, but similar Milankovic cycles should be a generic feature of planetary systems.

7.6.1 Milankovic cycles on Earth

Earth's precessional cycle is shown in Figure 7.13. The precession angle increases at a nearly constant rate, completing a cycle every 22,000 years. Though the variation in rate is not evident over any one cycle, the rate is not exactly constant, and therefore the phase drifts over the course of hundreds of thousands of years.

The precessional cycle is very rapid, and the precession angle has changed markedly even over historical times. Eight thousand years ago, when the first Sumerians poured into the valleys of the Tigris and Euphrates, the star we now call Polaris (the "Pole Star", in the tail of the Little Bear) was about 40^o of arc away from the star that the the North Polar axis then pointed to,and about which the constellations rotated at the time. The consequences of precession for change in seasonality are potentially highly consequential. In Figure 7.13, the July insolation at 65N is shown as a general indication of the magnitude of the seasonality effect; high northern July insolation in the precessional cycle goes with low January northern insolation, weak southern January (summer) insolation, and relatively strong southern July (winter) insolation. Ten thousand years ago, the Northern Hemisphere summer insolation was fully $40W/m^2$ greater than at present, and so the northern summers should have been considerably warmer than today, while the northern winters should have been considerably colder. The effect should show up especially over land, which is dark enough to absorb most of the solar radiation and has low enough thermal inertia to respond nearly instantaneously to seasonal changes. The climate system in its full glory is nonlinear and complex, so the response of climate to this change in seasonality could show up in any number of unexpected ways, and not simply as an enhancement of the Northern Hemisphere seasonal cycle over land.

The event which is most likely to be a recent manifestation of the precessional cycle is the "Climatic Optimum," covering the period of about 5000 to 7000 years ago (see Chapter 1). The term is most often used to refer to a period of generally warmer Eurasian temperatures. The "optimum" is sometimes said to be about 1-2K warmer than present, but it is difficult to get reliable estimates of global mean temperatures, or even annual means. What is certain is that some regions during some seasons were warmer than they were at recent pre-industrial times. At about the same time, the Sahara, which is now a torrid desert, experienced a period of greening, with currently dry riverbeds ("wadis") filled with water, and a teeming variety of animal life and flora not known at present. The greening of the Sahara is thought to be associated with atmospheric circulation systems known as "monsoons," forced to a greater extent by the enhanced heating of Northern Hemisphere subtropical land. A central question, though, is why the greening of the Sahara, and the Climatic Optimum occurred several thousand years after the precessional peak in Northern Hemisphere insolation. There are some indications that the warming may have *begun* as much as 10,000 years ago, but the question of the physics accounting for the time delay in response remains unsettled. Candidates for the necessary inertia in climate response include vegetation adaptation, land ice, and deep ocean heat storage.

Looking further back in time, the obliquity and eccentricity variations become significant, though of course, the precession cycle also continues to have a large effect. The Earth's obliquity and eccentricity cycle is shown in Figure 7.14. The amplitude of the obliquity cycle varies consider-

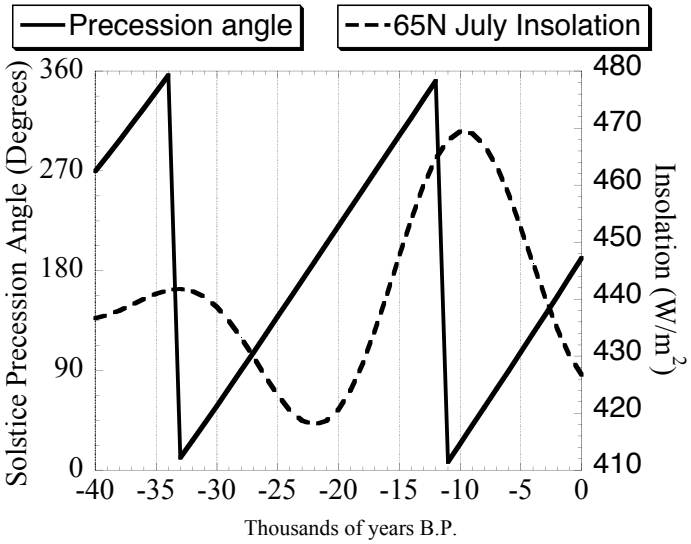

Figure 7.13: Evolution of precession angle relative to the Northern Hemisphere Summer Solstice, and the associated July insolation at 65N. Data taken from Berger and Loutre (1991).

ably over time, but its dominant period is on the order of 40,000 years. The Earth's obliquity varies narrowly in a range from about 22^o to 24.5^o. At present, the Earth is in the middle of its obliquity range. Eccentricity varies on a longer time scale of approximately 100,000 years. However, in Figure 7.14 there are also hints of 400,000 year cycle of eccentricity, whose fingerprint consists of two high eccentricity cycles followed by two low eccentricity cycles. This visual impression is borne out by spectral analysis. Currently, the Earth is near the low end of its eccentricity range, though it has gotten quite close to zero during the past two million years. At the other extreme, Earth's eccentricity has gotten as high as .055, or more than half that of Mars.

The idea that ice ages are due to changes in Earth's orbital parameters is nearly as old as the discovery of ice ages themselves. The idea has gained currency, but it is nearly as hard to justify today on basic physical principles as it was when first proposed. The main reason for its acceptance is circumstantial, in that increasingly detailed data on the observed rhythm of the ice ages shows the unmistakable imprint of the calculated rhythm of the orbital forcing. James Croll first proposed in the 1870's that changes in the Earth's eccentricity led to ice ages, and his idea was refined a half century later by Milutin Milankovic, whose name is now generally attached to the theory. The centerpiece of Milankovic's idea is that ice ages require the accumulation of snow on land, and that this in turn is favored by mild summers (limiting melting of old snow and ice) and warmer, but still sub-freezing, winters (favoring snow accumulation, since warmer air contains more water). The gaping hole in Milankovic's theory is that it predicts that ice ages should follow the precessional cycle. In particular, the Northern Hemisphere and Southern Hemisphere should have ice ages in alternation every 10,000 years, with the severity of the ice ages modulated by the eccentricity cycle. This is not at all what is observed. Figure 7.15 shows the Antarctic temperature record for the past 400,000 years, together with eccentricity and the July insolation at 65N. Numerous other temperature proxies worldwide show that the Northern Hemisphere temperature, and global glacier ice volume, is nearly in phase with the Antarctic temperature record, so that the Antarctic temperature can be taken as an index of when the world is in an ice age. The dominant signal

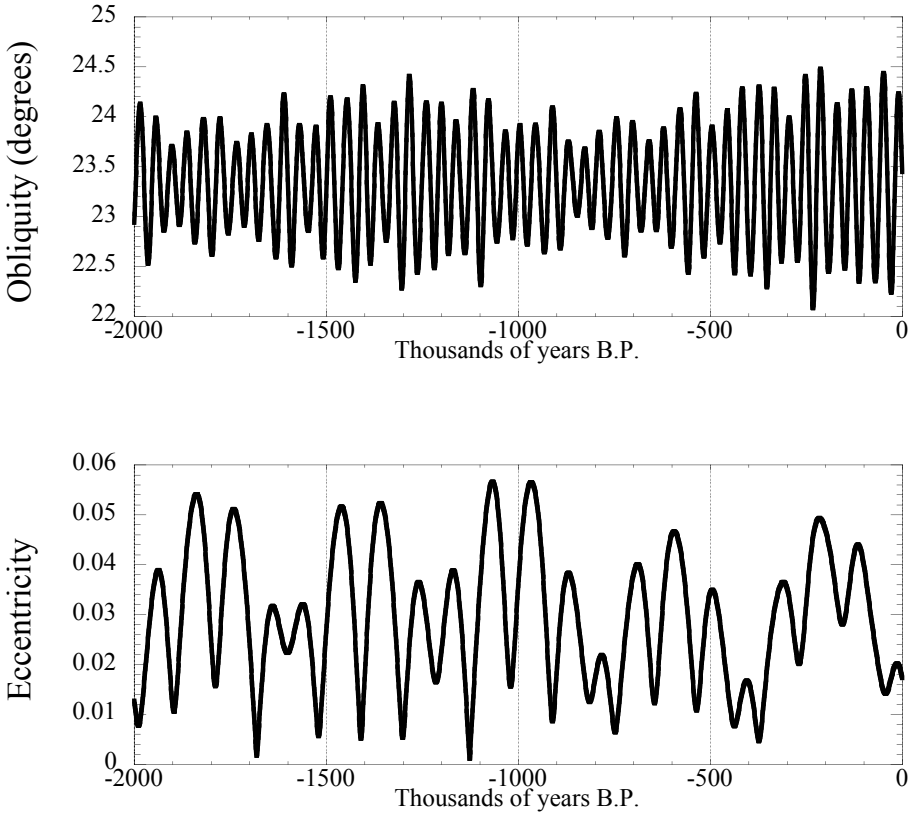

Figure 7.14: Evolution of the Earth's obliquity and eccentricity. Data taken from Berger and Loutre (1991).

in the climate response is an approximately 100,000 year spacing in the major interglacial warm periods, and a similar spacing in the coldest glacial periods. Crudely speaking, each interglacial corresponds to a peak in eccentricity, and a time within which (during parts of the precessional cycle) the Northern Hemisphere seasonality is unusually strong. This is somewhat reminiscent of the Milankovic mechanism, but what filters out the high frequency precessional cycle? Why does the entire Earth fall into an ice age at the same time, rather than alternating between hemispheres? A closer examination of the 65N July insolation strongly suggests that major global deglaciations occur when the Northern Hemisphere seasonality is weak, suggesting that the Earth listens to the Northern hemisphere forcing more than the Southern, in deciding when to have an ice age. This probably has something to do with the fact that the Northern hemisphere has more land, and hence more seasonality, than the Summer, but the precise way this asymmetry influences global glaciation remains largely obscure.

The problem is not that the amplitude of radiative forcing associated with Milankovic cycles is small: it amounts to an enormous $100W/m^2$, with the amplitude determined by the eccentricity cycle. The problem is that the forcing occurs on the fast precessional time scale, whereas the climate response is predominately on a much slower 100,000 year time scale. One does not so much need an amplifier of Milankovic forcing, as a "rectifier," which is sensitive to the *amplitude* of the precessional variation, rather than to its mean. Recall that atmospheric CO_2 is observed to vary on the glacial-interglacial time scale. Certainly, this is a major piece of the puzzle, since the drop in CO_2 during glacial times is sufficient to account for a major portion of the cooling of the climate, particularly in the Southern Hemisphere (see Chapter 4). CO_2 is a globalizing effect, and (insofar as it is linked to the glacial-interglacial physical climate changes) an amplifying feedback. The circumstantial role of CO_2 in ice ages is also a reprise of an old idea. The 19th century physicist Tyndall, whose work on infrared spectroscopy is at the foundations of our current understanding of the greenhouse effect, was primarily interested in explaining the ice ages, and the association reappeared later in the work of Chamberlain. The mechanism of the CO_2 cycle not known, but almost certainly involves CO_2 storage in the deep ocean. The lack of a theory for the glacial-interglacial CO_2 cycle is the central impediment to a theory of the ice ages. The presence of ice does seem to be a prerequisite for a strong climate response to orbital forcing. Before the onset of permanent polar ice at the beginning of the Pleistocene, response to orbital forcing was weak (see Chapter 1). Besides CO_2, ocean circulations can potentially play a major role in globalizing and rectifying the Northern Hemisphere signal, through direct heat transport as well as indirect effects on CO_2. The answer to the mystery of the ice ages lies somewhere in the space: *ice,ocean,* CO_2, but how the system works its miracles to yield a 100,000 year cycle is still unknown.

In recent years, an alternate picture of the way the joint precesssional/eccentricity cycles affect glaciation and deglaciation has emerged. This picture proceeds from the observation that when the tilt seasons line up with the distance seasons, Kepler's law implies that the intense summers in the in-phase hemisphere are also short, while the moderate summers in the opposite hemisphere are long. Then, if one assumes that deglaciation is primarily responsive to some measure of integrated summer melt energy – to be thought of as the net energy input for that part of the season where the glaciers are brought to the melting point – the variations of duration of the seasons for an Earthlike obliquity and eccentricity range largely cancel out the variations of the peak-intensity. This makes the precessional and eccentricity cycles mostly drop out of the picture, leaving obliquity variations as the main player. In this view, the obliquity cycle is the natural rhythm for glacial/interglacial cycles, and the early Pleistocene, in being dominated by the obliquity cycle, is responding in the simplest and most readily understood way. The longer cycles that emerge in the later Pleistocene are seen not as a consequence of the eccentricity variations, but rather as a matter of skipping of obliquity cycles due to some as-yet unknown nonlinearity in glacial dynamics. What we see as a 100,000 year cycle is thought of instead as an irregular blend

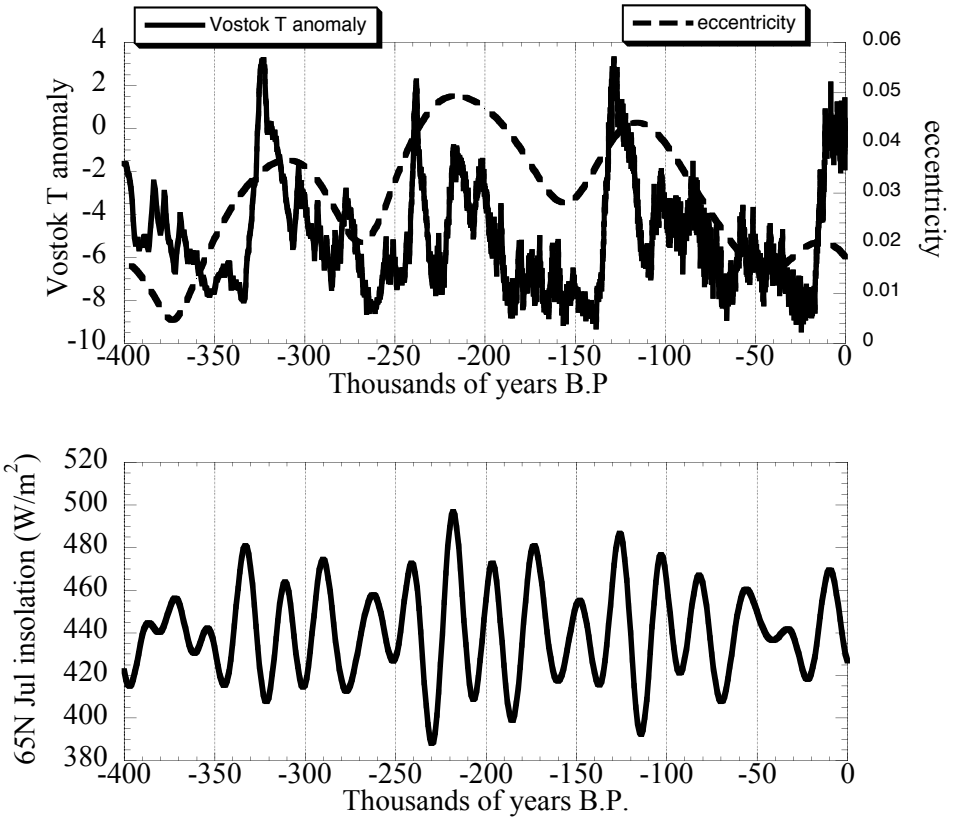

Figure 7.15: Comparison of Antarctic temperature reconstructed from Vostok ice core deuterium measurements, with the Earth's eccentricity cycle. The bottom panel shows the corresponding July insolation at 65N. Temperature is given as deviation from the mean modern value. Vostok temperature data was taken from Peteet *et al.* (1999) (See Chapter 1).

of 80,000 year cycles coming from a single skip, and 120,000 year cycles coming from a double skip. The glacial/interglacial CO_2 variations are still an important amplifying feedback for this mechanism, and indeed could play a role in the mechanism accounting for skipping of obliquity cycles.

7.6.2 Milankovic cycles on Mars

As expected from general mechanical considerations, Mars has Milankovic cycles analogous to those of Earth. Mars' cycles differ in some key respects, because of the lack of a massive moon, and because of the proximity of Jupiter.

As for Earth, the precession angle of Mars increases at a nearly constant rate. However, because Mars does not have a moon as massive as Earth's, the precession is dominated by Solar gravity, and is slower. The Mars precessional cycle has a period of approximately 50,000 Earth years. The current precession phase is 145^o, and will reach 180^o in about 5000 years.

The obliquity and eccentricity variations are shown in Figure 7.16. Obliquity has short term variations with amplitude on the order of 20^o. The period is not visible in the figure, but a finer scale examination of the data shows that the period is about 125,000 Earth years in recent times. The amplitude is markedly larger than that of Earth's obliquity cycle, but what is even more remarkable is that the obliquity drifts to values as large as 47^o over 10 million years. The extreme obliquity variations are directly linked to the absence of Earth's massive moon, which can be shown to provide a considerable damping effect on obliquity. This raises the intriguing possibility that a massive moon may be a necessary condition for a planet to avoid extreme climate fluctuations that could compromise its habitability. Calculations of the Earth's obliquity have also been carried out for tens of millions of years, and do not yield any greater variations than have been encountered in the past million years.

Mars is close to its maximum eccentricity at present, though it can get somewhat larger. The eccentricity of Mars undergoes quasiperiodic large amplitude cycles with a period on the order of 3 million years. In addition, there are short period, lower amplitude eccentricity variations with a period on the order of 100,000 years, rather similar to Earth's. In contrast, the very long period variations are not found in Earth's eccentricity.

Mars has no ocean, little thermal inertia, and a thin atmosphere that has a relatively modest effect on the planet's surface temperature. These features should lead to a different, and perhaps simpler, response to orbital forcing on Mars as compared to Earth. The predicted climate changes have been simulated in detail using comprehensive climate models, but we will confine ourselves here to some general remarks. The main effect of Martian Milankovic cycles is likely to be the redistribution of water deposits, in the form of either glaciers or permafrost. There are two aspects to this redistribution. On the short precessional time scale, the asymmetry between the Northern and Southern polar ice caps should reverse. For example, about 25000 years ago, the Southern hemisphere should have had milder summers and winters, while the Northern had cold winters and hot summers; the default reasoning would imply that at such times, the Southern ice cap should be large and be composed mainly of water ice, whereas the Northern ice cap becomes smaller and experiences massive seasonal CO_2 snow deposition. On the time scale of millions of years, the obliquity of Mars becomes much greater, leading potentially to a situation where water my migrate from poles that are seasonally very hot, and re-deposit in the tropics. At times of much lower obliquity, permafrost ice may migrate to both poles. The migration of water deposits and changes in patterns of deposition of CO_2 snow probably leaves some imprint on the surface geology of Mars, and the growth and decay of glaciers certainly does. These offer some prospects

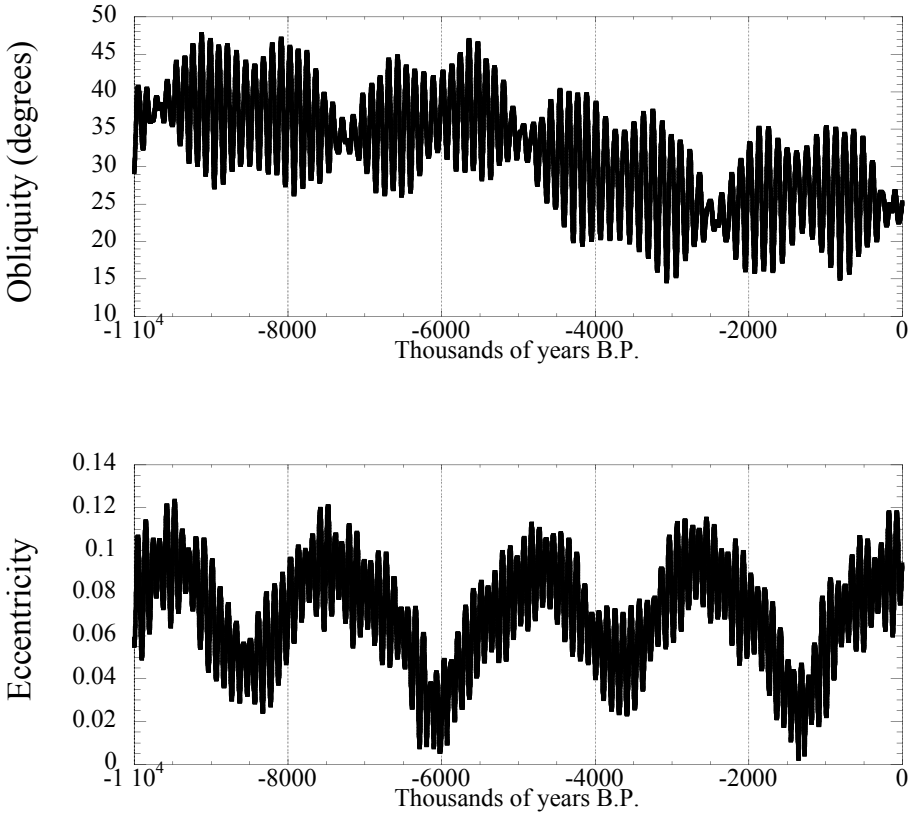

Figure 7.16: Evolution of Mars' obliquity and eccentricity. Data taken from Laskar *et al.* (2004).

for reconstructing the consequences of Milankovic cycles on Mars. Even better information would be obtained through analyzing cores of the polar ice caps, much as is done in Antarctica and Greenland. It is a very exciting development that the technology for doing this robotically on Mars is already under development. With respect to Mars, we are more or less at the stage of Croll or Milankovic, who thought they found the key to Earth's ice ages. Data showed they were on the right track, but that the climate system is much more intricate than they imagined. Given that we do not yet have a satisfactory theory leading from orbital variations to climate response on Earth, one can look forward to many surprises, once data on the Martian climate response becomes available.

7.7 A palette of planetary seasonal cycles

This section provides a sampler of the many ways in which thermal inertia and orbital characteristics can combined to yield seasonal cycles with interesting consequences. These are presented as a series of vignettes, which are primarily meant to serve as impetus to further inquiry.

7.7.1 Formation and inhibition of polar sea ice

For continental configurations similar to Earth's present state, the processes governing high-latitude climate variations differ greatly between the North and South polar regions. At present there is an ocean at the North Pole, so ice can form there through freezing of ocean water, but in the absence of ice the seasonal cycle would be moderated by the thermal inertia of the ocean. The South Pole is surrounded by the Antarctic continent, so that ice can accumulate only through snowfall; moreover, the seasonal cycle there will be strong even if there is no glacier. The conditions for formation of ice in a polar ocean are germane to Earth's climate variations at times going back at least to the Cretaceous, and there are numerous other times in Earth's history when one or the other pole was surrounded by ocean. Since most planets with an ocean will undergo some periods when there is no continent near one of the poles the problem is of general interest for exoplanet climate as well.

At the time of writing, the high northern latitudes are covered by sea ice for most or all of the year, though that appears to be changing rapidly as a result of anthropogenic global warming. In the hothouse climates of the Eocene and Cretaceous, these regions were nearly or completely ice free, and deep ocean temperatures as well as other proxies indicate annual mean temperatures of at least 10C, and perhaps at times as high as 20C. What conditions could lead to such warm polar seas, given that we know the Earth can also support icy polar conditions?

There are two points of reference to keep in mind when thinking about the formation of polar ice. Both depend on the dynamic heat flux and ice-free ocean albedo, as well as the insolation parameters and atmospheric greenhouse effect. The first point is whether the temperature that would be in equilibrium with the annual mean absorbed solar radiation for an ice free ocean is above the freezing point of sea water. This temperature is the temperature that would be attained for an infinitely deep mixed layer. If the mean equilibrium temperature is below freezing, then it is inevitable that sea ice will form at least seasonally, since the seasonal cycle can only make the winter temperature fall below the mean. If the mean equilibrium temperature is above freezing, then it is possible that sea ice can be entirely suppressed. That would happen, for example, if the seasonal cycle were weak enough that the winter temperature never hit freezing. However, even if the winter temperature falls to freezing, sea ice might still be suppressed if the ocean stratification

were such that the dense, cold water at the surface sank deep into the ocean. This case is equivalent to having a very deep mixed layer, and since there would not be time in a single winter to freeze the deep ocean, ice would never form. Salt stratification plays a big role in the formation of sea ice on Earth, since relatively fresh Arctic surface water (coming from ice melt, river runoff, and net rainwater input) can hit the freezing point without triggering oceanic deep convection.

The second point of reference is whether the midsummer instantaneous equilibrium temperature attained by an ice surface is above the melting point. If it is, there is the possibility (but not the certainty) that the ice can be melted away during the summer. If it is not, then it is inevitable that ice will survive from one winter to the next, leading to perpetual ice cover.

The ice evolution occurring in modern times, as well as the transition to ice-free states in the Eocene, suggests that the Earth is not deeply in the perpetual-ice regime, but rather hovering at the boundary between where perpetual ice is conditionally possible, and the regime where the annual mean conditions rule it out. With simple models– in fact even with very complex models – it is only possible to make the case that the radiative conditions can be made compatible with such a state. Let's take conditions at $80N$ as an example. The satellite observed annual mean dynamical heat flux at that latitude is about $110W/m^2$ at present. Let's assume that the albedo of an ice-free ocean would be 0.33, which allows for partial cover by low clouds. In these circumstances, for Earth's present obliquity the annual mean temperature for a deep mixed layer would be $270K$ at a CO_2 concentration of 300 $ppmv$ – just below freezing. At 1200 $ppmv$, it rises to $273.5K$ – just above freezing. On the other hand, even at $300ppmv$ the equilibrium ice temperature at the summer solstice for an ice albedo of 0.6 is over $300K$ including the effects of dynamic heat flux, far in excess of what is needed to sustain summer melt. In fact, it would only take about 20 W/m^2 of dynamic heat flux to bring the summer ice surface to the melting point. This is consistent with the observation that high latitude Northern Hemisphere temperatures hover around freezing in the summer, but it also says that apart from extreme changes in obliquity, summer melt is inevitable. The real question is whether ice starts to form at all, and whether it becomes thick enough to last through the summer.

To give these ideas a more quantitative expression, let's extend the basic mixed layer model to include some idealized sea-ice feedbacks. We'll assume that whenever the ocean temperature drops below the freezing point of sea water (say, $271K$), ice forms and the albedo is increased to the albedo of sea ice (say, 0.6). Further, in order to crudely represent the low thermal inertia of ice, we'll decrease the effective mixed layer depth to $1m$ when ice is present. Finally, we'll ignore the time it takes to melt ice, and instantaneously remove the ice when its surface temperature reaches the melting point. Results for this model, adopting the dynamic heat flux and ocean albedo values stated previously, are shown in Fig. 7.17. The results show that there is a quite extensive and very cold winter ice season when the CO_2 is at $300ppmv$ (somewhat above pre-industrial levels) or at $150ppmv$ (somewhat below ice-age values). In neither case, though, does the ice cover become perpetual. This failing arises from the unrealistic assumption that ice melts instantaneously. In reality, the longer and colder the winter icy period, the thicker the ice will become, and the longer it will take to melt (see Problem **??**). The following exercise shows that the time required for melting could easily extend the icy period to the rest of the year.

Exercise 7.7.1 Suppose that partway through winter, when the sea ice is at its thickest, the ice has a thickness of 5 m. Once the ice surface reaches the melting point, the temperature will stay fixed until all the ice has melted, and this delays the disappearance of the sea ice. Assuming that the albedo of melting ice is 0.5 and that the incident solar radiation at the surface is $200W/m^2$, estimate how long it takes to melt the ice.

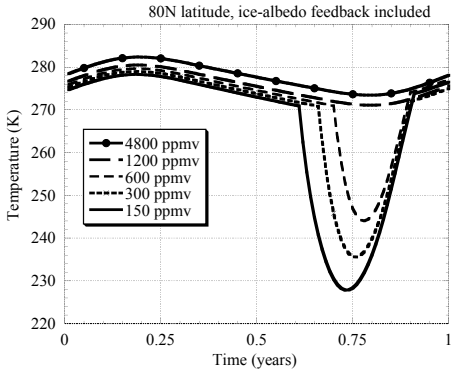

Figure 7.17: Numerically computed seasonal cycle for a mixed layer ocean subject to a realistic seasonal cycle of insolation at $80N$ latitude. The albedo for ice-free conditions is 0.33, which allows for partial low-cloud cover. The dynamic heat flux is 110 W/m^2, which is close to the observed annual mean value at this latitude. This calculation includes an idealized representation of the feedback of sea ice on albedo and thermal inertia (see text for details). Results are shown for a range of values of the CO_2 concentration.

With the assumed parameters, the ice season shortens when CO_2 is increased to $600ppmv$ and disappears completely when CO_2 exceeds $1200ppmv$. Even at $4800ppmv$, the mean polar temperatures are not as great as the proxies suggest, though, indicating that some additional amplifying factor besides ice-albedo feedback must come into play. The winter warming effect of high clouds is a likely candidate.

The calculation we have presented is only a plausibility argument. It only shows that, without going outside a defensible range of parameters, the basic model of the seasonal cycle can be put in a state where a CO_2 increase comparable to what is plausible for the Eocene or Cretaceous can lead to the elimination of sea ice. The actual proof that the Earth system is near this transition depends on myriad details governing ocean stratification, clouds, realistic sea ice physics and dynamical heat flux. Indeed comprehensive general circulation models vary considerably in the threshold CO_2 at which sea ice is suppressed.

7.7.2 Continental climates on Hothouse Earth

One of the toughest problems regarding the nature of hothouse climates like the Cretaceous is accounting for the absence of cold winters in the interiors of continents. There is abundant evidence for mild high latitude winters, in the form of fossils of creatures (notably alligators) and plants (notably palms) that cannot tolerate freezing conditions. The problem is that there is little local solar flux in winter, while the thermal inertia of land is too weak to keep tempetures from plummeting. For example, with 23^o obliquity, the flux factor at $50N$ latitude at the winter solstice is only 0.063. Even ignoring snow cover, the absorbed solar radiation for Earth is only 61 W/m^2, based on an albedo of 0.3. With $300ppmv$ of CO_2, the local equilibrium temperaature would be well below $220K$. In fact, even increasing the CO_2 concentration all the way to $10^5 ppmv$ is insufficient to bring the local equilibrium temperature up to 220 K, let alone the freezing point. Including the albedo of typical winter snow cover at this latitude, the absorbed solar flux drops further

to $34W/m^2$, leading to still more frigid conditions. A more complete exploration of continental temperature is carried out in Problem **??**.

Actual continental winters never actually get nearly as cold as the equilibrium temperature. At $50N$ latitude in the coldest part of Eurasia and North America, monthly mean winter temperatures hover around $250K$. Continental interior temperatures are warmed by lateral heat transports from the warm upstream ocean and from the warm tropics. Under clear sky conditions, the OLR with a surface temperature of $250K$, for the conditions of Table 4.2, is 178 W/m^2 with $300ppmv$ of CO_2; it falls modestly to 169 W/m^2 at $2400ppmv$. The difference between either of theses numbers and the absorbed solar radiation gives the rate at which heat needs to be imported. Given the slight absorbed solar flux in winter for a snow-covered surface, it is clear that dynamic heat transports are far and away the *dominant* factor in determining winter high-latitude temperatures in the continental interiors. A similar calculation tells us that the total flux needs to be brought up to 222 W/m^2 in order to bring the surface temperature up to $273K$ when the CO_2 is increased to $2400ppmv$, requiring in increased flux of $53W/m^2$ over and above the help provided by increasing CO_2. You could get about $27W/m^2$ of this from the albedo reduction that occurs when the snow disappears, but where does the other $26W/m^2$ come from? The problem is inherently dynamical, and requires detailed consideration of fluid dynamical heat transports. Current general circulation models do not generate enough heat transport to resolve the paradox, even if CO_2 is increased to the point where the upstream oceans have temperatures similar to paleoclimatic reconstructions. It is hard to see that there could be any major effect missing from the large scale dynamics, but low-level inversions form in the winter, as discussed in Chapter 6, and it is possible that some flaw in the representation of the boundary layer spuriously weakens the heat transfer from the atmosphere to the underlying surface.

In the wintertime there is little solar flux to reflect and a fairly high surface albedo. In these circumstances mid and high-level clouds could exert a substantial warming effect. Deep convection does not occur in the winter extratropics, and cloud formation in this regime is inextricably linked to large scale dynamics. Current general circulation models do not form enough high clouds to solve the problem of continental hothouse winter, but given uncertainties in the representation of convection and clouds, there is plenty of room for creative thinking in this area.

7.7.3 Snowball Earth

The high albedo of a Snowball state leads to very low temperatures in comparison to an unglaciated state for the same orbital position, but from the standpoint of the seasonal cycle, the most important effect is that replacing the liquid ocean surface with a solid ice/snow surface practically eliminates surface thermal inertia from the system. For a planet with a very thick atmosphere, the atmospheric thermal inertia could still moderate the seasonal cycle, but for an Earthlike atmosphere, the surface temperature would be nearly in equilibrium with the instantaneous diurnally-averaged insolation in the course of the seasonal cycle, in the absence of lateral atmospheric heat transport. The equilibrium surface temperatures for various obliquities are computed in Problem **??**.

For Earthlike obliquity, the insolation gradients are weak in the summer hemisphere, so the temperature will be fairly uniform there and one doesn't expect a big role for dynamical heat transport. In the winter hemisphere, however, dynamical heat transports will moderate the temperature contrast implied by the extreme gradients in insolation found there. These expectations are borne out by simulations of the Snowball Earth state using full general circulation models (see the Further Readings). For much higher obliquities the summer pole begins to become substantially hotter

than the summer tropics, and for some orbits and surface albedos could even seasonally reach the melting point.

Overall, the Snowball Earth climate has more in common with the climate of Mars than it does with the present Earth. The high albedo reduces the solar forcing to values similar to those of rocky Mars, while the frozen surface yields a low thermal inertia similar to that for the dry, rocky surface of Mars. The cold conditions also reduce the role of water vapor and water clouds in climate, while making CO_2 condensation possible at the winter pole. The principal difference between the climate states of Snowball Earth and Mars arise from the surface pressure of the atmosphere and the size of the planet; the latter of these comes in exclusively through dynamical effects.

Note that while the surface temperature undergoes extreme seasonal variations, the results of Section 7.4.2 show that the deep ice temperature will be equal to the annual mean surface temperature. This temperature will be similar to, though not exactly equal to, the surface temperature that would be in equilibrium with the annual mean flux; for Earthlike obliquity, this temperature has its maximum at the equator. For thick ice, deglaciation requires that the deep ice temperature reach the melting point, so that it is the annual mean conditions that primarily determine when deglaciation can occur. The formation of seasonal summer melt ponds at the surface can indirectly affect the deep ice temperature, however, through reducing the surface albedo and delaying the onset of autumn freezing conditions.

7.7.4 Venus

The deep atmosphere of Venus has an enormous thermal inertia, both because of its great mass and because it is so optically thick that heat escapes only slowly via sluggish convection and slow radiative diffusion. The slow rotation of Venus means that the nightside atmosphere has a long time to cool down, but it also makes it easier for the atmosphere to transport heat effectively from the dayside to the nightside. Observations confirm that there are essentially no geographic or seasonal variations of surface temperature apart from those that can be attributed to surface height variations.

This does not mean that the seasonal cycle of Venus is without interest, for the top bar of its atmosphere has as much mass as Earth's entire atmosphere, and this portion of Venus exhibits a lively mixed seasonal/diurnal cycle. The cycle here is driven by *in situ* atmospheric solar absorption, rather than by heating from below – a situation rather similar to that which arises due to ozone heating in the Earth's stratosphere. The dayside is considerably hotter than the nightside at these altitudes, and this drives a circulation carrying heat and constituents from one side to the other. Moreover, for rather subtle dynamical reasons, the upper atmosphere takes on a rotation of its own even though the surface of the planet is hardly rotating. The upper atmosphere rotation is not a rigid-body rotation, but it carries atmosphere along latitude circles with a time scale on the order of 5 Earth days. Because the sense of this circulation is in the same sense as the equatorial surface motion, but faster, it is called a *super-rotation*. The super-rotation of the upper atmosphere of Venus gives the seasonal cycle there many of the characteristics familiar from more rapidly rotating planets.

Venus can be considered to be the archetype for a probably common class of planets with thick CO_2 atmospheres containing effective solar absorbers. Understanding the seasonal cycle in this regime has important ramifications for the interpretation of exoplanet observations, since most observations of Venus class planets would only be able to see fluctuations in the infrared and reflected solar radiation from the upper portions of the atmosphere. Determination of the seasonal

cycle of cloud albedo is a key part of this story, and while that certainly involves both dynamics and atmospheric chemistry, the seasonal temperature variations provide the context against which both effects take place.

7.7.5 Mars, present and past

The seasonal cycle of Present Mars typifies a generic class of seasonal cycle involving condensation of a major constituent of the atmosphere. This kind of seasonal cycle is likely to be common throughout the Universe, and over time for a broad class of planets. Figure 7.18 shows the seasonal cycle of surface pressure for the Viking 2 lander, which sat in the Northern Hemisphere midlatitude of Mars. The surface pressure varies from $7.3mb$ to $11mb$ over the course of a year, amounting to a 20% variation around the mean. Earth's surface pressure typically varies by no more than 1% on large scales, or perhaps as much as 5% if one includes the minimum pressure at the center of strong hurricanes. The difference between Earth and Mars is that only a small portion of the Earth's atmosphere is condensible, whereas the primary constituent of the atmosphere of Mars can condense at the winter poles. If the major pressure variations on Mars are similar over then globe, then the pressure fluctuation implies that up to 16% of the maximum mass of the atmosphere can be sequestered in condensed form at one of the winter poles. A significant feature of the Martian seasonal pressure cycle is that the minimum pressure occurs somewhat after the Southern Hemisphere winter solstice, and that the secondary minimum at the Northern Hemisphere winter solstice is less pronounced. This is due to the high eccentricity of the present Martian orbit, and the current phase of its precessional cycle, which conspire to create Southern winter conditions that are far colder than Northern winter conditions.

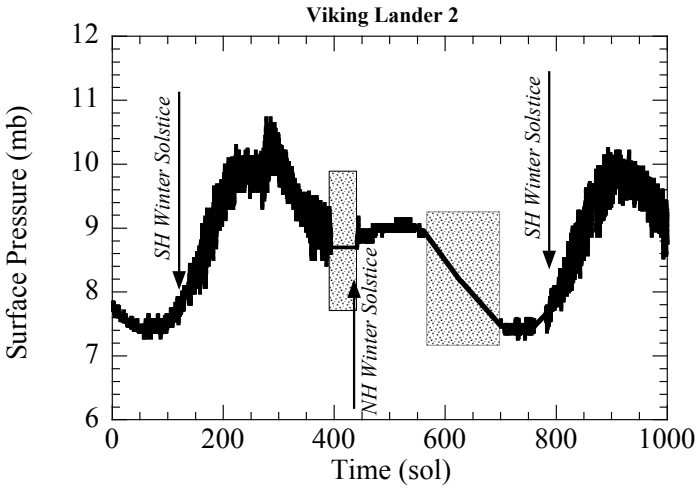

Figure 7.18: Surface pressure time series for the Viking 2 lander, which was located at 47.97^oN latitude. The time scale is in Mars days (*sol*) since landing. Solstices are marked. The shaded areas indicate missing data.

The simplest model of the seasonal cycle of pressure would be to posit that the global pressure is equal to the saturation vapor pressure corresponding to the tempererature at the winter pole. But how does one determine that temperature? Thermal inertia, and perhaps also dynamical heat fluxes, must still be involved since the polar night equilibrium temperature would be very nearly

absolute zero. We have seen that at temperatures of $150K$ or less, even a solid surface can provide significant thermal inertia, since the radiative relaxation times become so long at low temperatures. Another factor is that the latent heat of condensation must be radiated away to space in the polar regions, and this provides a form of thermal inertia limiting the condensation rate. For example, at a surface temperature of $150K$, one radiates only 28 W/m^2 to space, and using the latent heat of sublimation of CO_2 one could only condense at a rate of $4.8 \cdot 10^{-5} kg/m^2 s$. Carried out over a tenth of the Martian surface, this would condense out about a quarter of the Martian atmosphere in the course of one winter. If the polar temperature were to get much colder, however, polar condensation would practically halt, even though the surface saturation vapor pressure might be very low compared to the global mean pressure. If one tried to make the condensation rate much higher, however, the latent heat release would raise the surface temperature to the point where the surface pressure became subsaturated, and condensation would have to halt. The thermal inertia due to latent heat release is therefore a very significant player in Marslike seasonal condensation cycles. Similar considerations would come into play for the rate of condensation of atmosphere on the nightside of a tide-locked planet. In both cases, the condensation rate could also be limited by the dynamics governing mass transport into the condensing region.

The greenhouse effect of the thin atmosphere of Present Mars is slight, but in cases with a more substantial atmosphere, massive seasonal pressure fluctuations due to polar condensation have even more interesting consequences. In that case, the polar condensation amounts to an amplifying feedback on the seasonal cycle. During each solstice, sequestration of CO_2 in polar regions reduces the greenhouse effect *globally*, leading to very cold conditions in the winter hemisphere and moderated temperatures in the summer hemisphere. The situation is somewhat akin to water vapor feedback, in that Clausius Clapeyron controls the mass of a greenhouse gas present in the atmosphere. It is less local than water vapor feedback, however, since the reservoir of condensate is localized in polar glaciers, rather than spread out over a nearly ubiquitous ocean.

With a more massive atmosphere, the polar condensation engages interesting glaciological questions as well. A massive CO_2 glacier would flow, particularly in view if the fact that it would be rather easy to liquefy CO_2 at the base. The glacier flow would bring condensed CO_2 to warmer parts of the planet, where it could be recycled to the atmosphere.

Returning to Present Mars, let's take a look at the seasonal and diurnal cycle of temperature. Some typical midlatitude summer and winter diurnal cycles are shown in Fig. 7.19. As expected from the low thermal inertia of the system, the seasonal and diurnal cycles are large. The daytime peak summer temperature is $63K$ warmer in summer than in winter, while the seasonal cycle in night-time temperature is a more moderate $32K$. Because of the low thermal inertia, night-time temperatures are almost as low in summer as they are in winter. The diurnal cycle in summer, at about $50K$, is nearly as large as the seasonal cycle. Note also that the diurnal cycle in winter is weaker than it is in summer. This is because the low temperatures lead to longer radiative relaxation time, and also because the daytime peak insolation is weaker. A more quantitative analysis of the seasonal cycle in this dataset is carried out in Problem **??**.

The reader has no doubt noticed that the winter diurnal cycle is much less regular than the summer cycle. This is a general feature, and is not peculiar to the snippets of data shown in Fig. 7.19. The irregularity of the winter diurnal cycle is due to the presence of wavy instabilities of the Martian winter jet stream – an analog of Earthlike midlatitude storms. These are absent in the summer, for reasons that are inherently dynamical.

For Early Mars, an additional set of questions present themselves. Let's suppose that the planet has a thick CO_2 atmosphere of perhaps $2bar$, but that the radiative effect of this atmosphere is not sufficient to bring the annual mean temperature above freezing at any point. Suppose further

that Early Mars does not have extensive deep oceans, so that the high albedo of a Snowball state does not come into play. Under these circumstances, the surface would have low albedo and little thermal inertia. Could the seasonal cycle lead to seasonal summer glacier melt sufficiently strong to account at least in part for the ancient river beds seen on Mars? The extreme obliquity cycle of Mars accentuates this picture, since the summer hemisphere would become particularly hot during high obliquity phases, especially when they line up with perihelion during a high eccentricity phase.

Earlier, we estimated that the heat capacity of a $2bar$ Early Mars atmosphere was equivalent to a 10 m water mixed layer. This is far less than Earth's mixed layer depth, but to complete the calculation of thermal response time we also need to know the radiative damping coefficient for the atmosphere. Using the homebrew radiation model for the $2bar$ pure CO_2 atmosphere, for the conditions of Fig. 4.30, we find that $dOLR/dT$ is about 1 W/m^2K for surface temperatures near $273K$. With the effective mixed layer depth, this yields a thermal response time of 484 Earth days, which is notably less than the Martian year of 687 Earth days. Therefore, the thermal inertia of the Martian atmosphere would have some moderating effect on the seasonal cycle, but the seasonal cycle would be far less strongly attenuated than Earth's. Therefore, one can expect quite hot summers, especially during high obliquity stages, perhaps even leading to extensive seasonal melt. The relatively weak thermal inertia is a two-edged sword, however. During periods of high obliquity, the long dark winter will be exceptionally cold, and if enough CO_2 condenses at the winter pole, the drawdown of atmospheric pressure could limit the greenhouse effect. Precisely how much CO_2 condenses in these circumstances depends on the moderating effect of CO_2 ice clouds discussed in Section 5.9, and on the rate of heat and mass transport from the summer to the winter hemisphere. These, together with the volume of summer melt, constitute outstanding Big Questions.

7.7.6 Nearly airless bodies

At first glance, it might be thought that the seasonal cycle of nearly airless solid bodies like the Moon or Europa or Triton would be fairly dull. WIth little thermal inertia at the surface, and no significant atmospheric effect on surface temperature via heat transport or greenhouse effect, the seasonal cycle would seem to be a simple matter of the hot spot moving along with the subsolar point. On reflection, though, it is easily seen that the problem offers a rich variety of novel features worthy of the attention of the most discriminating planetary climatologist.

First of all, the thermal inertia of even a solid surface can become significant when temperatures become cold enough, since the radiative cooling decreases so rapidly with temperature. Thus, the temperature of the unilluminated part of the Moon is determined mainly by heat diffusing up from the subsurface and by the amount of time that has passed since the most recent illumination. For very cold bodies such as Europa or Triton, the thermal inertia becomes significant even under the subsolar point, so that rather than a simple hot spot, one should see a warm tail trailing the hot spot along the past path of the subsolar point. There can be interesting patterns arising from variations in surface thermal inertia and surface albedo, and moreover, the transport of trace amounts of volatile substances sublimated from the crust can over long time periods modulate both the albedo and thermal inertia. Mars is practically in the nearly-airless class, and the effect of the Martian glaciers provide a prime example of the feedback from slow transports in a thin atmosphere. In the course of the Martian obliquity cycle, the glaciers slowly migrate from the poles to the low latitudes, and this is mediated by transport in the thin atmosphere. Even the Moon and Mercury have interesting volatile migration patterns; there appears to be water ice near the Lunar poles, and an interesting seasonal redistribution of sodium on the surface of Mercury. Triton has a complex and young surface, which is probably due at least in part to sublimation

and redistribution of N_2 and CH_4 ices, and it is even possible that the redistribution of water ice on Europa affects that body's albedo. Europa has a $3.5d$ diurnal cycle but little obliquity, so the main redistribution of water should take the form of sublimation and redeposition in the equatorial belt, plus some leakage of snow into the midlatitude and polar regions. This would be a very slow process, but over the course of hundreds of millions of years could result in important evolution in the surface.

Moreover, as one goes deeper below the surface, the temperature is determined by the surface temperature averaged over longer and longer time scales. Thus, at a moderate depth, the temperature will be given by the diurnal average, while still deeper layers will have a temperature determined by the annual mean. To some extent, subsurface temperature can be probed via radio frequency emissions, and the variations can be used to recover information about composition and structure.

For icy bodies composed of substances that have sufficiently elevated vapor pressures when exposed to temperatures encountered over the course of the seasonal cycle, things get even more interesting. This is the case for N_2 and CH_4 ice on Triton and Pluto. In the former case, subsurface heating even leads to spectacular cryovolcanism. The "solid greenhouse effect" arising from penetration of solar radiation into ice, as discussed in 6.9, can lead to even more elevated subsurface temperatures and correspondingly more dramatic phenomena.

Finally, even if the atmosphere is so thin that it feeds back little on surface temperature, that doesn't mean that its dynamics is uninteresting or unimportant. It only means that the problem of atmospheric dynamics becomes simpler to think about. On Earth one needs to determine the surface temperature simultaneously with atmospheric temperature, both of which are then strongly affected by dynamical heat transports. On Mars, to a good approximation one can determine the surface temperature evolution as if the atmosphere weren't there at all, and then use the resulting time-space temperature pattern as the lower boundary condition to drive the atmospheric circulation. The feedback of the Martian atmosphere is not completely negligible, but as the atmosphere gets thinner, this approach becomes more and more accurate. The transient N_2 atmosphere of Triton is a case in point, and is all the more novel in that its dynamics involves supersonic expansion of sublimated gas into a near-vacuum.

7.7.7 Titan

Titan is one of the few bodies in the Solar System having both a thick atmosphere and a solid surface, and in this regard is more Earthlike than Present Mars, whose thin atmosphere makes it a candidate for inclusion in the "nearly airless" category. Like most moons, Titan is tide-locked to its primary, orbits in the plane of the primary's equator, and therefore shares the obliquity and the year of the primary. Titan's day is its orbital period about Saturn, or $16d$.

As already mentioned in Section 7.4, the distinguishing feature of Titan's seasonal cycle is provided by the great thermal inertia of its cold,thick N_2 atmosphere. One expects very little seasonal fluctuation in the interior of the atmosphere, but there can be a considerable seasonal cycle in surface temperature, owing to the low thermal inertia of the mostly solid surface. As for Venus, one expects a considerable seasonal cycle in the thin outer portions of the atmosphere. Even though the interior atmosphere temperature may not change much in the course of the season, modelling studies show that the variations in surface temperature modulate convection, which affects the circulation, methane precipitation, and methane cloud patterns of Titan's troposphere. Since Titan's year, like that of Saturn, is 29 Earth years, one will have to wait a decade or so to see whether the various predictions that have been made regarding the evolution of these features

are borne out.

7.7.8 Gas and Ice Giants

Jupiter has very low obliquity and a nearly circular orbit, so there is not much to drive a seasonal cycle there. However, Saturn and Neptune have obliquities in excess of 26^o and Uranus has the highest obliquity of all, at 97^o. This demonstrates that high obliquity can be imparted to a gas or ice giant either during the process of formation or at some point thereafter. The gas and ice giants in the Solar System all have quite circular orbits, but high eccentricity extrasolar gas/ice giant planets are very common. In addition. some of the extrasolar giants are in sufficiently close orbits that the mean rotation may be slow or tide-locked, leading to interesting hybrid seasonal/diurnal cycles. The main issue concerning the expression of the seasonal cycle on giant planets stems from the thermal inertia of the atmosphere and the depth of penetration of solar radiation. The very deep atmosphere receives little illumination, and is massive enough and opaque enough in the infrared that little seasonal variation should be expected; this situation is much the same as that for the deep layers of the atmosphere of Venus. However, there will always be some layer near the top of the atmosphere which will have low enough thermal inertia to respond seasonally. The depth of this layer will depend on the composition of the atmosphere, and the observational expression will depend on the nature of cloud-forming substances, if any. For a fairly clear atmosphere, the seasonal variations would be most apparent through variations in infrared emissions tied to temperature. With clouds, modulations in temperature could also be expressed as variations in cloud albedo and cloud pattern. For the case of extrasolar planets, understanding such seasonal variations is particularly important, since the observable quantities are tied to emission and reflection from the outer portions of the atmosphere, and one would like to know how deep a layer of the atmosphere these observables are probing. Extrasolar giant planets abound in orbits where they receive solar flux at rates similar to or even greatly exceeding Earth's solar constant. Higher fluxes and higher temperatures generally lead to larger seasonal fluctuations. Opportunities abound for exploring the novel seasonal cycles of giant planets operating in a thermal range far hotter than any seen for giants in the Solar System.

Gas giants are primarily made of H_2, which doesn't become a good infrared absorber until quite high pressures are reached. The thermal emission contribution to thermal inertia is therefore likely to be dominated by minor constituents, both in the form of gases and cloud particles. Ice giants have a more complex composition, and a correspondingly greater range of possible optically active constituents. A detailed consideration of the depth of the seasonally active layer of known and hypothetical planets is beyond the scope of what we wish to attempt here, particularly in view of considerable uncertainties regarding the vertical distribution of solar-absorbing constituents.

The study of seasonal cycles of giant planets is in its infancy, perhaps in part because one has to be very patient indeed to observe the seasonal changes. Saturn, with its orbit of 29 Earth years, presents the best near-term observational possibility in our own solar system. Some hints of a seasonal cycle on Saturn have already been observed. Between 1980 and 2002 the equatorial winds in the visible cloud layer of Saturn have decreased by about 40%, and there have also been substantial changes in the cloud patterns. It has been suggested that in the case of Saturn, the shading provided by the ring system could be playing a significant role in the seasonal insolation pattern. At the time of writing, the southern hemisphere of Saturn is just coming into midsummer. It will be interesting in the coming years to track the evolution of Saturn's atmosphere through its full seasonal cycle. Along with Titan, Saturn provides a rare opportunity to test our understanding of seasonal cycles by attempting to predict the evolution in the coming decade, with little guidance from past observations.

The case of Uranus should be even more dramatic, but given an orbital period of 84 Earth years, one will have to wait two decades or so in order to see a hint of what is going on. Uranus is a fairly featureless planet, so it is difficult to observe changes in the atmospheric circulation. The observing systems in place for the past two decades do not provide a suitable basis for probing the seasonal changes of Uranus so far, but improvements in orbital telescopes should make it possible to do better in the coming decades.

7.7.9 Habitability of planets with extreme orbital configurations

Outside the Solar System, planets with highly eccentric orbits are common. If the Earth did not have the benefit of its large Moon, its obliquity would probably undergo the same extreme fluctuations as that of Mars. Either of these situations could potentially lead to such extreme seasonal cycles as to pose a threat to habitability, even if the planet had an Earthlike composition and annual mean insolation. The high-eccentricity cases would seem particularly inhospitable, as there would be short periods at periastron with intense solar radiation which could lead to perilously hot conditions, followed by long periods far from the star when the oceans could well freeze over, and perhaps even the atmosphere could snow out.

If the planet had a mostly solid surface, these would indeed be severe habitability barriers, but even a moderately deep mixed layer ocean can average out a quite extreme seasonal variation of insolation. The essential issue, as always for seasonal cycle problems, is the length of the planet's year relative to the thermal response time of the ocean. A 50 m mixed layer can even out most of the seasonal cycle for a planet with high obliquity or a fairly high eccentricity even if the planet's year is as long as 5 Earth years. The critical point to consider is whether there is enough time at apastron for the ocean to freeze over, since if that happens one loses the benefit of the ocean's thermal inertia and extremely cold conditions can ensue. This problem has much in common with the problem of formation of polar sea ice on Earth. In particular, ocean stratification comes into play, since the mixed layer will deepen and prevent freezing even for long winters, in cases where the ocean stratification permits deep oceanic convection in near-freezing conditions. That situation could preserve habitability even for planets with quite long years and very high eccentricity, provided the ocean is reasonably deep.

The question of ocean stratification and sea ice suppression is inherently dynamical, but one can at least get an idea of the depth of mixed layer needed to preserve habitable conditions for any given orbital configuration. Another thing to keep in mind in this connection is that planets with a close-in periastron tend to adopt a variety of slowly-rotating spin states, the simplest of which is quasi-tidelocked in the sense that the planet always presents the same face to the host star at periastron. Such planets will be in the regime where there is a hybrid seasonal/diurnal cycle, rather than two distinct cycles.

Some simple calculations pertaining to habitability of extreme orbital configurations are developed in Problem **??**.

7.8 For Further Reading

The Milankovic cycle of Earth's orbital parameters is discussed in

- Berger A and Loutre MF 1991: Insolation values for the climate of the last 10 million of years. *Quaternary Sciences Review*, **10**, 297-317.

The "summer melt energy" theory of response to Milankovic forcing was introduced in

- Huybers,P 2006: Early Pleistocene glacial cycles and the integrated summer insolation forcing, *Science*, **313**, 508-511.

The Martian Milankovic cycle, and its climatic expression, are discussed in

- Laskar J, Correia ACM *et al* 2004: Long term evolution and chaotic diffusion of the insolation quantities of Mars, *Icarus*, **170**, pp 343-364.
- Levrard B, Forget F *et al* 2007: Recent formation and evolution of northern Martian polar layered deposits as inferred from a Global Climate Model. *J. Geophys. Res. Planets* **112**, E06012.

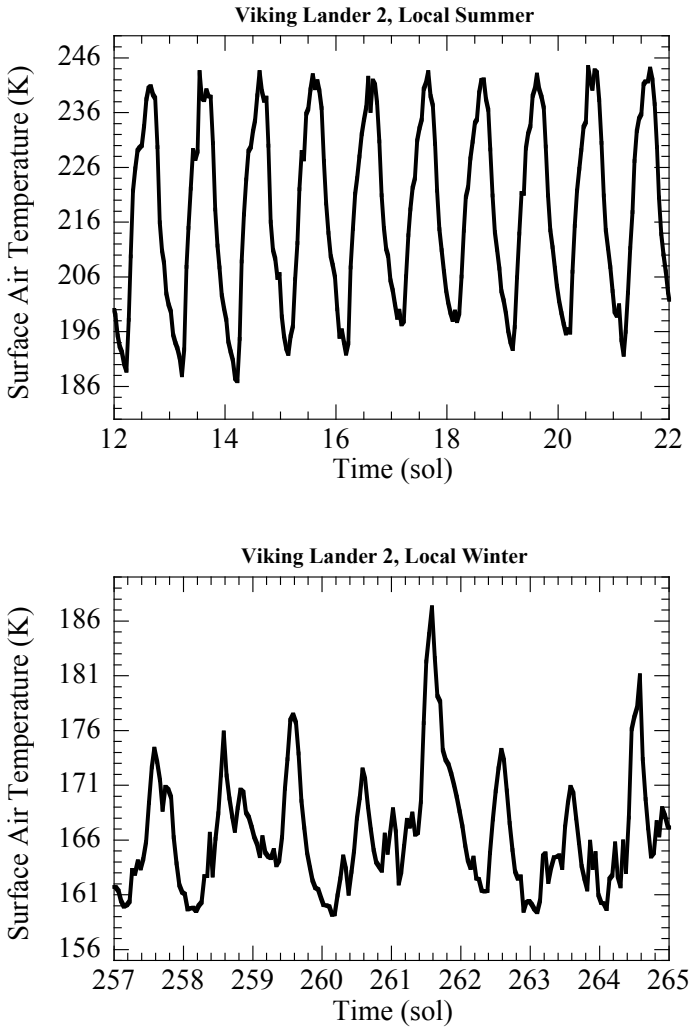

Figure 7.19: A segment of the near-surface air temperature time serie for the Viking 2 lander showing the diurnal cycle in summer (upper panel) and winter (lower panel).

Chapter 8

Evolution of the atmosphere

8.1 Overview

As we emphasized back in Chapter 1, atmospheres are not static. The mass and composition of an atmosphere evolves over time, as a result of a great variety of chemical, physical and biological processes. Now it is time to survey those processes in greater detail, and to put numbers on them to the extent possible in the limited space available in this chapter.

Throughout the following we will need to refer to some constituents of a planet as *volatiles*. These are "not rocks" – things that can become gases to a significant extent. The concept of a volatile is relative to the temperature of a planet. On Earth water is a volatile but on Titan it is basically a rock, as is CO_2, though N_2 and CH_4 remain as volatiles even at the low temperatures of Titan. On Earth, sand (SiO_2) is a rock, but on a roaster – a hot extrasolar Jupiter in a close orbit – it could be a volatile.

For planets in which some atmospheric volatiles exchange with a condensed reservoir, as in the case of Earth's ocean and glaciers, the whole atmosphere-ocean-cryosphere system is best treated as a unit for many purposes, and we will refer to this as the *volatile envelope*. In other cases, the portion of the volatile envelope which resides in the atmosphere plays a distinguished role. Only the atmospheric portion provides a greenhouse effect, and volatiles must first enter the atmosphere before they can escape to space as gases.

The main factors that govern atmospheric evolution of rocky terrestrial type planets or icy bodies with a thick solid crust are as follows. First, there is the matter of what, if anything, outgasses from the planetary interior, and at what rate. The composition of the outgassing depends on the chemistry and physics of the planet's interior; for example the early segregation of the Earth's core took iron out of the rest of the planet which allows oxygen to react with other elements. This favors the outgassing of oxidized gases such as CO_2, SO_2 and water vapor, though limited amounts of H_2 and CH_4 do exit the interior. On Titan it is speculated that ice-volcanism can release NH_3 and possibly CH_4 to help maintain the atmosphere, and on Venus it is speculated that the sulfuric acid clouds are maintained by outgassing of SO_2 (though no active volcanism has yet been observed there.)

Next, whatever enters the volatile envelope is subject to a number of further alterations. Atmospheric constituents are lost to space, either by gradual mechanisms or in catastrophic events such as giant impacts. Different constituents are generally lost at different rates, leading to evolu-

tion of the composition. Then, too, chemical reactions between the volatile envelope and the solid crust can bind up constituents in mineral form, selectively removing material from the volatile envelope. If the planet has no way to engulf bits of the crust and close the cycle by cooking out the volatiles, then the volatile envelope will eventually come into equilibrium with the static upper part of the crust, and this means of evolution will mostly cease, though changes in climate could still lead to changes in partitioning between the volatile envelope and the crust. This is the situation on Mars at present. If the planet is tectonically active, as is the case for Earth, then crustal material is mixed down into the interior, and there is instead a dynamic geochemical equilibrium involving a much greater proportion of the mass of the planet. The question of whether crustal recycling occurs on Venus and Titan is one of the current Big Questions of planetary science. The discovery of extrasolar planets of the Super-Earth class raises the Big Question of whether plate-tectonics or some other form of resurfacing becomes more or less likely on rocky planets larger than the Earth or Venus.

The main energy sources that sustain plate tectonics or other forms of resurfacing are fossil heat left over from the formation of the planet and heat release by radioactive decay in the interior. These heat sources are small, but they can have a big effect on interior temperature because the diffusivity of heat through solids is so low. The importance of radioactive decay introduces another dependence of climate evolution on planetary composition, since planets can be formed with a greater or lesser endowment of radioactive materials than the Earth. Further, for planets in close orbits about their stars, or satellites in close orbits about giant planets, interior heating by tidal stresses could be important. Such processes drive spectacular volcanism on Jupiter's satellite Io, and could well play a role on Super-Earths in the habitable zone of M-dwarfs, which is quite near in to such stars.

Finally, chemical reactions within the atmosphere determine how the elemental composition is arranged into molecules. These are usually fast processes on geological time scales, requiring seconds to a few million years to operate. They nevertheless affect the long term evolution by affecting what can escape to space, what can react with the crust, and whether the elements are arranged as greenhouse gases (e.g. oxygen in the form of CO_2) or not (oxygen in the form of O_2). Some examples include the breakup of water vapor by ultraviolet light, the oxidation of methane or hydrogen which limits their atmospheric concentration, and the breakup of methane on Titan followed by resynthesis into ethane.

Life profoundly alters virtually every aspect of atmospheric evolution. Through the use of complex enzymes, life can break stable bonds and synthesize componds in low-temperature low-energy environments where inorganic processes do very little. Nitrogen fixation and oxygenic photosynthesis are two prime examples. We'll see in Section 8.7 that the oxygen generated in the latter process actually raises the temperature of Earth's outermost atmosphere from about $250K$ to over $1000K$, affecting the escape of gases to space. Life can synthesize methane at rates far greater than the methane flux produced by volcanic outgassing. Life also alters the chemical environment at a planet's surface, altering the rate of reaction of atmospheric components with the crust.

8.2 About chemical reactions

The next few topics will require some basic knowledge about how chemical reactions work, so we will pause here to provide some elementary background. Consider, for example two hypothetical

substances A and B which can react to form a product C. The reaction is written as

$$A + B \rightleftharpoons C \tag{8.1}$$

The double arrow symbol, $\rightleftharpoons$, is there for an important reason: it reminds us that chemical reactions proceed in both directions. When a molecule of A encounters one of B, with a certain probability it will react to form C. However, from time to time, a molecule C will also break up into its components A and B. The net reaction depends on the competition between the forward reaction and the back reaction, and when the two rates are equal the system is in *chemical equilibrium* and concentrations of the substances do not evolve. As an example, in any glass of common liquid water, the following reaction is taking place:

$$H^+ + OH^- \rightleftharpoons H_2O \tag{8.2}$$

The superscripts indicate that in this particular reaction, the reactants are *ions*, having a positive or negative charge (in this case, a single charge equal to that of an electron or proton). Note that both charge and the count of atoms must balance between the left hand and the right hand side of the reaction. In water, a certain proportion of the H_2O molecules will decompose into the indicated ions, until the rate of recombination equals the rate of production.

The rate of a chemical reaction depends on the probability with which molecules of the reactants collide multiplied by the probability that they react upon collision. In simple cases, such as reactions between molecules in a gas or a dilute solution of molecules dissolved in a liquid, the chance of encounter of reactants is proportional to the product of the concentrations of the reactants. In chemical calculations, it is almost invariably most convenient to measure concentrations as molar concentrations, rather than mass concentrations. For liquid phase reactions between solutes, molar concentration and molar density (e.g Moles of solute per liter of solvent) are practically the same thing, since the density of the liquid varies little. For gases, it is often more convenient to represent the quantity of a substance in terms of its partial pressure instead of its mass or molar density.

It is commonly the case that the availability of a molecule to participate in reactions is not simply proportional to its concentration. This might happen, for example, because other molecules in a solution cluster around a solute molecule, partially shielding it from reactions. To deal with this, chemists have introduced the notion of *activity*, as a generalization of concentration. The represention of activity, when it is something other than a simple concentration, is particular to the class of reactions under consideration. To distinguish the activity of a substance, which is a number, from the abstract symbol denoting the substance itself, the activity is written in square brackets: $[A]$ is the activity of substance A. However the activities are defined, the reaction rate is expressed as the product of all the activities multiplied by a rate coefficient. By convention, we'll call the rate coefficient k_+ for the forward reaction and k_- for the back reaction. Thus, the reaction rate for the forward reaction in Eq. 8.1 is $R_+ = k_+[A][B]$. The product of activities represents the probability of encounter of the two reactants, while the rate coefficient represents the probability of reaction given a collision. For the unary back-reaction involving decomposition of C, the rate at which the decomposition proceeds is proportional to the amount of substance C present, so we write $R_- = k_-[C]$.

By way of example, let's consider the net of forward and back reactions in Eq. 8.1, in the simple case where the activities are just concentrations, in which case the reaction rates are the time-derivatives of the concentrations. The evolution of the concentrations is then given by

$$\frac{d[A]}{dt} = \frac{d[B]}{dt} = -k_+ \cdot [A][B], \frac{d[C]}{dt} = -k_- \cdot [C] \tag{8.3}$$

The decay rates of $[A]$ and $[B]$ are equal because each reaction consumes one particle of A and one of B, producing a particle of C. The rate depends on the product of the activities of the two reactants, since the product gives the probability of particles of the reactants encountering each other. Now, the first equation given only determines the decay of A owing to the reaction with B. This will be the net reaction when there is no C present, but after the reaction proceeds a while in a closed vessel, some C will accumulate and the decomposition of C back to A and B needs to be taken into account. Thus, when the activity is simply a concentration or density, the full system is governed by

$$\frac{d[A]}{dt} = -k_+ \cdot [A][B] + k_- \cdot [C] \tag{8.4}$$

which will come into equilibrium when the left hand side is zero, namely when

$$\frac{[C]}{[A][B]} = \frac{k_+}{k_-} \equiv k_{eq} \tag{8.5}$$

The quantity on the right hand side is the *equilibrium constant* for the reaction. This equation constrains the relative proportion of the three substances once equilibrium has been achieved, but how one uses this information depends on what is specified in the setup of the problem. For example, if for some reason we know the activity $[A]$, then we immediately know the ratio $[B]/[C]$, though we don't know the absolute amounts unless something in addition is specified. As a slightly more complicated example, suppose we put $2Mole/m^3$ of A and 1 $Mole/m^3$ of B into a closed vessel which initially contains no C. Then, after equilibrium has been reached $xMole/m^3$ of C will have been produced, which depletes each of $[A]$ and $[B]$ by $xMole/m^3$, since it takes one of each to produce a particle of C. The equilibrium equation then tells us that

$$\frac{x}{(1-x)(2-x)} = k_{eq} \tag{8.6}$$

Given knowledge of the equilibrium constant, this allows x to be solved for using the quadratic equation.

Exercise 8.2.1 Carry out the algebra to determine x in the above example, and describe how it behaves as k_{eq} is varied from very small to very large values. Do you ever completely use up the reactants?

When the activity is something other than simply concentration or density, the reaction rate is no longer the time-derivative of activity, and the forward and back reaction rates should be written abstractly as R_+ and R_-. These are measured in various ways, depending on the nature of the reaction. For example, if the reaction is between a gas or dissolved substance and the surface of a solid, it would be typical to characterized the reaction rate in terms of $Moles$ per unit time, per unit surface area of the reacting solid. The characterization of reaction rate affects how one calculates the approach to equilibrium, but it does not affect the equilibrium itself, since the equilibrium is defined by $R_+ + R_- = 0$, which yields the same equilibrium conditions on activity as before.

As a concrete example of the use of equilibrium coefficients, let's consider the dissociation reaction that occurs in liquid water.

$$H_2O \rightleftharpoons H^+ + OH^- \tag{8.7}$$

This reaction is of ubiquitous importance in aqueous chemistry. Only a tiny fraction of the water molecules dissociate, so it is customary to set the activity of H_2O to unity in writing the equilibrium relation

$$[H^+][OH^-] = k_w \tag{8.8}$$

The activities of the ions are not identical with concentrations, but it is customary to express the activities as a nondimensional coefficient times the actual concentration, so that activities and concentrations have the same units. In the units of moles per liter (that's *gram* moles) that aquatic chemists like, the equilibrium constant k_w is very close to 10^{-14} at room temperatures. If we take the $\log_{10}$ of the equilibrium equation written in these units, then

$$(-\log_{10}[H^+]) + (-\log_{10} OH^-) = 14 \tag{8.9}$$

This leads to the definition $pH \equiv -\log_{10}[H^+]$, assuming the activity to be expressed as moles per liter. If water dissociates in the absence of any solute that changes the H^+ or OH^- ion concentrations, then equal amounts of the two ions are produced, and the equilibrium relation in the form Eq. 8.9 implies that $pH = 7$. This case, without an excess of protons, is referred to as *neutral*. If a substance like H_2SO_4 dissociates and adds H^+ ions without adding OH^- ions, then Eq. 8.8 implies that the concentration of OH^- must go down, whence pH falls below 7 resulting in an *acid*, which has an excess of protons. The opposite situation, with a deficit of protons, is a *base* and has $pH > 7$. If one knows the pH, then the activity of H^+ expressed in moles per liter is 10^{-pH}, and if the activity coefficients are near unity this is nearly the actual concentration. To convert to units of mole fraction (molecules of substance per total molecules) that we favor, one must divide this concentration by 55.56, since that is the number of gram-moles in a liter of water. Note also that, like all equilibrium constants, k_w depends on temperature, so that the the H_+ concentration corresponding to a neutral solution is somewhat temperature dependent; the definition of pH, however, is always tied to the H_+ concentration itself.

Rate and equilibrium equations generalize to reactions involving more reactants and more products in the obvious way. For example, a reaction of the form $A+2B+3C$ has a rate proportional to $[A][B]^2[C]^3$. The equations for equilibria are modified correspondingly. Sometimes one of the "reactants" is a molecule that participates in the reaction only to the extent of providing some extra energy by collision. A reaction like that is written, for example, as $A + B + M \rightleftharpoons C + M$, where M is a generic colliding molecule; in kinetics, its activity $[M]$ would generally be the number density of molecules of any sort available for collision with the other reactants. For example, an atmospheric reaction between substances A and B might proceed more rapidly in a background of N_2 even though N_2 does not itself react with either A or B. It is also common in atmospheric chemistry for one of the "reactants" to be a photon, usually one within a designated range of frequencies. A photon is designated in reactions by the symbol $h\nu$, as in the dissociation reaction $AB+h\nu \rightleftharpoons A+B$. The symbol used for photons is meant to be reminiscent of the energy carried by a photon of frequency ν. The activity of the photons is usually the number flux of photons having the correct energy range to react. This is obtained by dividing the energy flux at frequency ν by the energy of an individual photon $h\nu$, then summing up the result over all frequencies involved in the reaction.

When one or more of the reactants is in the form of a pure, solid body – for example a lump of solid substance A reacting with a gaseous substance B within a given volume – the quantity of the solid within the volume is not the limiting factor in determining the availability of molecules of the solid substance to react. Rather, in this case it is the area of solid exposed to other reactants that counts. In such a case, as a matter of convention the activity of the pure solid substance is set to unity,and the effect of available surface area for reaction is taken into account by expressing the reaction rate as Moles per unit area of contact per unit time. If the pure phase is reacting with a gas, the reaction rate expressed this way will depend on the partial pressure of the gas, but not on the amount of solid present (so long as there is some present and it is in contact with the gas). Equilibrium at any given temperature yields a unique value of partial pressure, just as Clausius-Clapeyron yields a unique vapor pressure in equilibrium with the solid or liquid form of a substance, regardless of how much of the solid or liquid form is present. The same behaviour

applies when a pure solid phase reacts with a substance B dissolved in a fluid in contact with the solid – the reaction rate per unit area will depend on the concentration of B, but not on the amount of solid present, and equilibrium at any given temperature will yield a unique value of $[B]$, regardless of how much solid phase is present, as long as there is some. We will encounter this situation soon, in our study of weathering reactions.

For all reactions considered in this chapter, the activity will be either a concentration (or its equivalent, such as a partial pressure), or it will be unity for pure condensed phase reactants. The reader will not need to be concerned with more exotic expressions for activity, though it is good to be aware that they exist.

The equilibrium and rate constants depend strongly on temperature. Most of the temperature dependence of the rate constants k_+ and k_- is captured by the Arrhenius law, which states that $k(T) = A \cdot \exp(-E/R^*T)$ where A is a constant and E is a quantity known as the *activation energy*, measured in $J/Mole$ when the Arrhenius law is written in this form. The temperature dependence can be fit a bit more accurately if the constant A is replaced by a power law in T, but for most of our purposes the unmodified Arrhenius law suffices. Since the equilibrium constant is k_+/k_-, it follows that the equilibrium constant has temperature dependence $A_{eq} \exp(-\Delta E/R^*T)$, where ΔE is the difference in the activation energy between the reactants and the products, and A_{eq} is the ratio of the prefactors of the forward and back reactions. While activation energy must be positive, and hence all reactions speed up with temperature, ΔE may be either positive or negative; therefore the equilibrium constant may either increase or decrease with temperature, accordingly as whether ΔE is positive or negative.

The activation energy in the Arrhenius law is a generalization of the concept of latent heat, which we have discussed in connection with phase transitions, and which appears in the Clausius-Clapeyron equation in a manner similar to the way activiation energy appears in the Arrhenius law. Indeed, a phase transition between, say, the gas and liquid forms of a substance can be considered as a binary reaction in which two molecules of "gas" substance collide and react to form one molecule of "liquid" reactant. To relate the latent heat L to the activation energy in the Arrhenius law, we need only observe that Clausius-Clapeyron was written using the gas constant specific to the gas in question, rather than the universal gas constant. Rewriting in terms of the universal gas constant, the temperature dependence of saturation vapor pressure becomes proportional to $\exp(-(L \cdot M)/R^*T)$, where M is the molecular weight of the gas. Thus, the activation energy for condensation of water vapor into liquid water is $4.49 \cdot 10^7 J/Mole$.

The constants also depend on the total pressure at which the reaction takes place, but the pressure dependence is usually weak over the range of pressures of interest in atmospheric problems.

8.3 Silicate weathering and atmospheric CO_2

Carbon dioxide is one among many greenhouse gases that can be present in a planetary atmosphere, but it plays a distinguished role in the evolution of atmospheres of rocky planets like Earth, Mars and Venus because of its participation in chemical reactions that allow it to be exchanged between the atmosphere and minerals in the crust and planetary interior. What's more, acting over sufficiently long time scales, the dependence of the interchange on temperature can potentially act as a thermostat and help to keep the planet in the habitable range despite considerable changes in solar luminosity and other factors. Understanding the precision with which this feedback mechanism controls planetary temperature, the limits within which it can operate, and the circumstances under which it can break down, is one of the most central of the Big Questions.

It is likely that there are other gases important to climate which undergo similar exchanges, but CO_2 exchanges have been far more extensively studied than any other case, and so this problem serves as a template for thinking about more exotic possibilities, some of which will me mentioned at the end of this section.

Carbon dioxide exchanges with the solid crust and interior of a rocky planet through reaction with silicate minerals to form carbonates. The class of reactions involved in this exchange is typified by the idealized *Ebelmen-Urey reactions*

$$\begin{aligned} MgSiO_3(s) + CO_2(g) &\rightleftharpoons MgCO_3(s) + SiO_2(s) \\ CaSiO_3(s) + CO_2(g) &\rightleftharpoons CaCO_3(s) + SiO_2(s) \\ FeSiO_3(s) + CO_2(g) &\rightleftharpoons FeCO_3(s) + SiO_2(s) \end{aligned} \tag{8.10}$$

In all of these reactions, the carbon in gaseous CO_2 exchanges with the silicon in solid silicate minerals, to form a solid carbonate mineral plus solid silica (SiO_2, of which quartz is one form, and which is also common in beach sand). The reader should keep in mind that the actual silicate minerals involved in the formation of carbonates can be considerably more complex than the simple chemical compounds referred to in the above reactions, and may have different equilibrium and kinetic properties. The feldspar family of minerals is one of the most important players in silicate weathering in Earth's present crust. These minerals are aluminum silicates involving varying amounts of sodium, potassium, and calcium; their weathering products include a broad variety of clay minerals, rather than just simple silica. Nonetheless, the Ebelmen-Urey reactions are often taken as indicative of what is going on in more realistic cases.

First, let's take a look at the equilibria that would be reached in the Ebelmen-Urey reactions after a sufficiently long time has passed. Imagine, for example, putting a pile of powdered $MgSiO_3$ in a closed chamber filled with a large quantity of CO_2 gas, and holding the entire apparatus at constant temperature T. The gas will react with the silicate to form carbonate and silica, drawing down the pressure until the products have built up to the point that the rate of recombination of carbonate plus silica is equal to the rate of reaction of CO_2 with silicate, at which point equilibrium has been reached. As long as the silicate has not been used up, the pressure will equilibrate at a value that depends only on T. Since the activities of the solid phases in Eq. 8.10 are unity, the only activity that can vary is that of the gaseous CO_2, and this activity can be characterized by the partial pressure of the gas. As long as none of the solid reactants has been exhausted, chemical equilibrium for any one of the reactions in Eq. 8.10 is described entirely in terms of the way the partial pressure of CO_2 gas (pCO_2 in chemical parlance) depends on temperature, when in the presence of the three solid reactants in the equation. In this situation, when there is only one activity which can vary, the equilibrium constant can be taken to be pCO_2 itself, and its temperature dependence follows the Arrhenius law

$$pCO_2 = p_1 \exp(-\frac{\Delta H}{R^* T}) \tag{8.11}$$

where p_1 is some constant, R^* is the universal gas constant and ΔH is a characteristic energy, specifically the difference in *enthalpy of formation* between the products and reactants. For the Ebelmen-Urey reactions, the reaction releases energy (i.e. is *exothermic*), so ΔH is positive. If the gas constant R^* is given in $J/Mole \cdot K$ then ΔH has units of $J/Mole$. The equation for partial pressure looks just like the Clausius-Clapeyron relation, and this is no accident since the thermodynamic formalism for the temperature dependence of equilibrium pressure is identical in the two cases: formation of a condensed phase from a gas is just a form of chemical reaction involving a single substance. The latent heat of fusion for water is 158 $kJ/Mole$, which compares with ΔH values of 79496, 88700 and 64852 $kJ/Mole$ for the Mg, Ca and Fe reactions measured in the laboratory at $298K$.

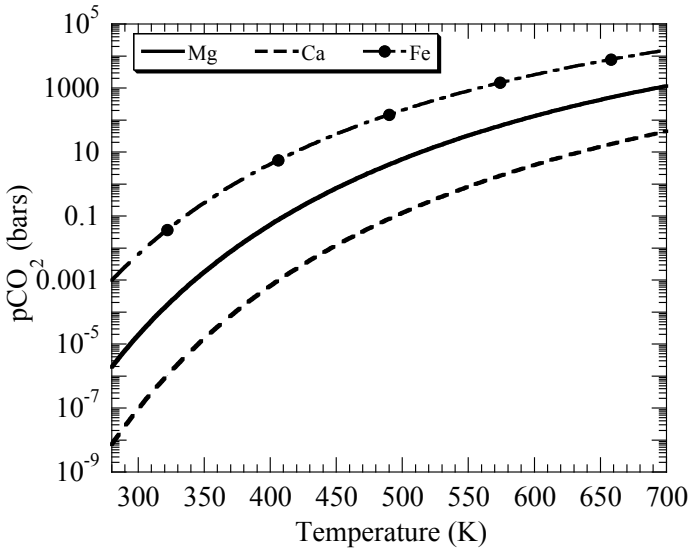

Figure 8.1: Equilibrium CO_2 partial pressure as a function of temperature for equilibrium with magnesium (Mg), iron (Fe) and calcium (Ca) silicates.

The equilibrium pressures for the three reactions are shown as a function of temperature in Fig. 8.1. The measured temperature dependence of ΔH has been taken into account in calculating these pressures. First, we note that at Earthlike temperatures the equilibrium partial pressures for Mg and Ca minerals are very low. For the Mg case at $300K$ the equilibrium has $pCO_2 = 1.9 \cdot 10^{-5} bar$, or equivalently 19 $ppmv$ in a 1 bar background atmosphere. For the Ca case, it's even lower, amounting to a mere 0.1 $ppmv$. In equilibrium, weathering of Mg and Ca silicates would draw down atmospheric CO_2 to such low values that it would have little greenhouse effect. At present (and probably for most of Earth history), the pCO_2 is well in excess of these equilibrium values, so silicate weathering is always trying to reduce atmospheric CO_2 to nearly zero, though it never gets especially close to equilibrium because the system is kept out of equilibrium by outgassing of CO_2 from the interior. Weathering of iron silicates leads to a somewhat different story line. At $300K$ the equilibrium for the Fe case is 0.006 bar, or 6000 $ppmv$. This is well in excess of the present pCO_2, and probably in excess of any value attained in the past half billion years. The implication is that even if iron silicates were prevalent at the surface of the Earth, the weathering of Mg and Ca silicates keeps the atmospheric CO_2 too low for iron carbonates to form. On the other hand, we know that during the era of the Faint Young Sun the CO_2 must have been well in excess of 0.006 bar if the CO_2 greenhouse effect is to be strong enough to keep the Earth unfrozen. Under these circumstances, iron carbonates should have formed. Therefore, the presence of iron carbonates in ancient deposits serves as a proxy for high CO_2. The lack of iron carbonates in certain Archaean formations is often taken as evidence that CO_2 alone could not have been the answer to the Faint Young Sun paradox, but this interpretation should be treated with caution, since so little Archean surface rock is preserved and the temporal record is very sporadic.

The equilibrium pressures rise sharply with temperature At $700K$ the equilibria are 1200 $bars$, 45 $bars$ and 15000 $bars$ for the Mg, Ca and Fe cases, respectively. Because the Earth's interior temperature exceeds $700K$ not too far below the surface, this immediately implies that as carbonates and silica are engulfed by plate tectonics and subducted into the interior, CO_2

will be cooked out of the rocks and outgas through volcanoes and fissures, allowing the carbon to be returned to the atmosphere. Things are likely to work similarly for any planet with plate tectonics and a rocky crust, and there may be other (as-yet unknown) means of episodically engulfing crustal material and bringing it to a high enough temperature to release CO_2. The high-temperature equilibria also have implications for the state of the atmosphere and crust of Venus. The atmospheric surface pressure of Venus is about 90 $bars$ of nearly pure CO_2; this is well below the equilibrium pressures for Mg and Fe carbonates, so these minerals would be unstable at the surface of Venus, given sufficient SiO_2. In contrast, the equilibrium pressure for Ca carbonates reaches 90 $bars$ at $737K$, which is quite close to the actual surface temperature of Venus. It thus appears possible that the atmosphere of Venus is in equilibrium with a crustal reservoir of calcium carbonate.

There are no atmosphere-bearing planets in the Solar system at present with surface temperatures intermediate between those of Earth and Venus, Venus may have experienced such temperatures in the past in the course of a near-runaway greenhouse state, and extrasolar planets could well have an orbit and composition to put them in this range. At $400K$ the Ca equilibrium is still only 650 $ppmv$, though that for the Mg case amounts to 5% of a 1 bar atmosphere. By the time one gets to $500K$, however, large amounts of CO_2 are inevitably left in the atmosphere – 6 $bars$ for the Mg case and 0.12 bar for Ca case. Reasoning by analogy from the state of the present Earth atmosphere, when the system is out of equilibrium by virtue of outgassing from the interior, the actual atmospheric pCO_2 will be considerably in excess of the equilibrium value, though it is reasonable to conjecture that at high temperatures equilibrium might be approached more rapidly, which would allow the system to stay closer to equilibrium even in the presence of substantial outgassing.

The chemical equilibrium behavior of the silicate/carbonate/CO_2 system is straightforward, but it is probably the only thing that is straightforward about silicate weathering and its role in climate evolution. The picture of the CO_2 in an atmosphere as being in equilibrium with rocks near the surface may perhaps be valid for some planets with a high temperature surface, but in general it is a poor representation of what is going on. Certainly, this is the case for Earth, the atmosphere of which is far out of equilibrium with the surface. The inorganic carbon cycle on Earth and probably many other rocky planets with an Earthlike temperature is a dynamic equilibrium which involves both interior and crustal processes. The carbonates near the surface are engulfed in subduction zones and brought into the interior of the planet, where the temperature is high enough that equilibrium is reached quickly and CO_2 is driven almost entirely out of the carbonates. This CO_2 outgases through volcanoes and through submarine features like the mid-ocean ridge where new ocean crust is born. If no new carbonates were formed by reaction with silicates, atmospheric CO_2 on Earth would build up to very high levels. Current CO_2 outgassing rates have been estimated at $0.1 gigatonnes$ of carbon per year, or about $2 \cdot 10^{-4} kg/m^2$ of carbon. In 4 billion years this would pump 784,215 kg/m^2 of C into the atmosphere, which is equivalent to a CO_2 surface pressure of over 280 $bars$ – about a million times the amount observed in the atmosphere today. Since it is extremely unlikely that the outgassing rate was much lower in the past ages – indeed it was probably greater when the Earth was younger and had lost less interior heat – there clearly must be an effective removal mechanism which takes CO_2 out of the atmosphere. Formation of carbonates by reaction with silicates is by far the most likely candidate.

Rather than building up until the supply of interior carbon is exhausted, the atmospheric CO_2 only builds up to the point where the rate of formation of carbonates equals the rate of outgassing. For any given outgassing rate, then, determination of the amount of CO_2 in the atmosphere requires that we quantify the way the carbonate formation rate depends on CO_2 levels, temperature, and other aspects of the climate. This is unfortunately not a matter of simple

chemistry. For low temperature planets like Earth the carbon dioxide pressure at the surface is generally well in excess of the equilibrium value corresponding to surface temperature and composition, but the rate at which carbonates form and bring the system back toward equilibrium is not determined primarily by the kinetics of the chemical reaction. The quantity we need to understand is the weathering rate W which is the number of Moles of CO_2 per unit time which is converted to carbonate by reactions with silicates over the entire surface of the planet. We need to determine how W depends on temperature, precipitation, and other aspects of climate, as well as the nature of the surface of the planet we seek to understand.

As with so many facets of planetary climate, silicate weathering involves a conspiracy of CO_2 and water, and this is of central importance in the determination of weathering rates. Although the reactions as written in Eq 8.10 do not involve water, at Earthlike temperatures the reactions occur in solution when water comes into contact with rock, and should be thought of as a kind of dissolution of silicate minerals in the presence of the weak carbonic acid formed when CO_2 dissolves in liquid water. At low temperatures – certainly below $300K$ and perhaps below $400K$ – the dry reaction proceeds too slowly to be of significance over a time scale of a few billion years [1]. Given that the reaction is aqueous, it would be natural for the reader to conclude that silicate weathering on Earth would be dominated by undersea processes; after all, there is plenty of water there as well as plenty of silicate in the ocean floor. Contrary to expectations, however, the best estimates indicate that at present seafloor weathering amounts to only a tenth of the global total. The factors limiting seafloor weathering at present include the degree of acidity of the ocean, the low deep ocean temperatures, the composition of the ocean crust and the sluggish delivery of ocean water to new reactable surfaces (occurring primarily in hydrothermal systems today). One should not over-generalize from this state of affairs, since it is highly dependent on the current state of the climate, and in a radically different climate with much higher CO_2 and higher temperatures things could be quite different. Further, on Snowball Earth or on a waterworld with no crust exposed to the atmosphere, seafloor weathering would be dominant because it is the only form of weathering there is. Moreover, it is quite difficult to estimate the rate of seafloor weathering even in present conditions, and there are credible estimates suggesting that seafloor weathering accounts for a considerably higher proportion of the total than the standard picture would indicate. If seafloor weathering were to prove to be a considerable fraction of the total, then most of what we shall have to say subsequently about silicate weathering and climate regulation would be called into question. Climate evolution under the dominant control of seafloor weathering represents largely unexplored territory.

We shall adopt the conventional picture that silicate weathering for planets in a regime something like that of the Earth occurs primarily over land, as a result of rain washing over silicate bearing rocks. As a result, the rate of carbonate formation is expected to increase with the rate at which rain falls over weatherable silicate rocks; the rain both accelerates the reaction and carries away the soluble carbonates, exposing fresh silicate for further reactions. The functional form of this relation cannot be measured in the laboratory, and depends on the mineral, its physical structure and the presence of vegetation and other biological activity. There have been numerous attempts to estimate the precipitation dependence from various kinds of field measurements of weathering, but the process is still poorly constrained.

Laboratory measurements of aqueous phase silicate weathering show clearly that the rate of

[1] The kinetics of the dry phase reaction do not appear to have been quantified to any great degree. There is some laboratory evidence that CO_2 can be formed in dry reactions between carbonate and silica at temperatures above $500K$, and indeed the dry phase reaction must be taking place in Earth's interior to sustain outgassing. It is generally believed, though, that even at high temperatures carbonates cannot form at a significant rate without water.

carbonate formation increases with temperature according to the Arrhenius law, $\exp(-E/R^*T)$, where E is an empirically determined activation energy. When the range of temperatures about some base temperature T_o is not too large, the Arrhenius law can be simplified to the exponential form $W(T_o)\exp((T-T_o)/T_U)$, where $T_U = T_o^2 R^*/E$. The coefficient T_U varies amongst minerals and also depends on other conditions such as the acidity of the environment in which weathering occurs. Typical values in widespread use in weathering models lie in the vicinity of 10 K, for temperatures within a few tens of degrees of Earth's current surface temperature. It is generally assumed that the temperature-dependent reaction rates measured in the laboratory lead to the same temperature dependence of net weathering rate in the field, though it is far from clear that this should be the case. If rain remains in contact with weatherable rock for only a short time, then it would be expected that increasing the reaction rate would indeed increase the amount of carbonate formed; this is the prevailing view of what is going on in Nature. However, in circumstances where water remains in contact long enough for the reaction to come to equlibrium, the kinetics becomes irrelevant and the weathering rate should not be directly dependent on temperature, though it will still depend indirectly on temperature through the effect of temperature on the precipitation rate. Because of the steep exponential dependence of the reaction rate, the chemical kinetic effect is likely to become far less important as temperatures become much hotter than that of the present Earth. For example, with $T_U = 10K$, a reaction that takes one day to reach equilibrium at $300K$ would equilibrate in only 4 seconds at $400K$ and 178 microseconds at $500K$. Given the slow kinetics at Earthlike temperatures and below, it does seem a reasonable assumption that something like the Arrhenius law applies for such temperatures. For very hot conditions, as in a near runaway, the direct temperature dependence of W would need to be reconsidered, however.

A more problematic issue is whether W also depends directly on the partial pressure of CO_2 in the atmosphere. Note that this is a separate question from the *indirect* effect of CO_2 concentration on weathering, mediated by its effect on temperature and by the effect of temperature on precipitation. Normally, for gas-solid reactions, it would be expected that the reaction would proceed more rapidly if there were more molecules of the reactive gas around. This is precisely what happens in laboratory measurements of the weathering of specific silicate minerals (mostly feldspar) at temperatures of $400K$ or more. The directly measured high-temperature dependence is often extrapolated down to lower temperatures using the Arrhenius law. However, laboratory measurements at Earthlike temperatures, conducted in the presence of organic acids thought to be similar to those produced by land plants, very clearly show that weathering rate is essentially independent of the amount of CO2; such experiments have been conducted for CO_2 partial pressures ranging from about $0.3mb$ all the way up to 1 bar. The prevailing view in this subject seems to be that without land plants (or perhaps, without bacterial life modifying silicate surfaces) there is a power law dependence of weathering rate on CO_2 partial pressure, but that this dependence disappears once land plants have appeared on the scene. It is quite unclear whether lichens or bacterial life are sufficient to cause the transition in behavior, and it is even more unclear whether the supposed abiotic pCO_2 dependence really exists at low temperatures. One can also wonder whether abiotically produced acids could also eliminate the direct pCO_2 dependence

The weathering rate is affected by a number of other processes going on at the surface. Notably, once land plants are on the scene, changes in precipitation and temperature distributions will affect the distribution of land plant cover, and this will feed back on the weathering rate; this is a very difficult feedback to model. Besides, that, the weathering rate depends on the availability of weatherable surface area. This is affected by physical erosion rates, and can be greatly enhanced by mountain-building such as the rapid uplift of the Himalayas; erosion rates are also affected by glacier flow and by freeze-thaw cycles which can fracture rock. Volcanism is also important in providing fresh weatherable surfaces. On Venus today, weathering is likely to be slow even for reactions that can take place without liquid water, because most of the erosional processes that

produce fresh weatherable surface are absent or weak.

Note also that the conventional picture of silicate weathering on relatively cool planets like the present Earth presumes that the equilibrium atmospheric CO_2 is far below the prevailing atmospheric value, so that the weathering can be thought of essentially driving the atmospheric CO_2 towards zero. If, in contrast, the equilibrium has a significant amount of CO_2 left in the atmosphere, then one needs to take into account the actual equilibrium toward which the weathering reactions are driving the system. This becomes more and more of a significant factor as the temperature increases, and the importance of the effect also varies with the surface mineralogy, since the equilibria are dependant on what kinds of reactions are taking place. The following development adopts the conventional cool-planet formulation in which weathering reactions are slow enough relative to outgassing rates that the atmospheric CO_2 is always far above the equilibrium value.

Suppose that we have somehow managed to write the weathering rate as a function $W(P, T, pCO_2)$, where P is the rate at which precipitation falls over land. Then, if Γ is the outgassing rate, measured in the same units as W, equilibrium is determined by $W(P, T, pCO_2) = \Gamma$. Γ may change gradually over geological time, leading to long term changes in climate. The problem is then closed if CO_2 is the dominant greenhouse gas controlling climate, since then T and P can be determined as functions of CO_2,given other pertinent data such as the solar (or stellar) constant at the planet's orbit and the configuration of the continents. Continental configuration can strongly affect the weathering rate, because the size and placement of continents affects how much of the global precipitation falls on land and sustains continental weathering. We are led to a condition $W(P(pCO_2), T(pCO_2), pCO_2) = \Gamma$, which can be solved for pCO_2. Once that is known, everything else is known, and one can then explore the dependence of the resulting climate on slowly varying parameters such as the stellar luminosity, the continental configuration, and the outgassing rate. This is the basic idea behind all CO_2 weathering feedback models, and in such models we seek to determine the extent to which the weathering feedback can control temperature and keep it in a habitable range.

Without even writing down a specific form of W, we can obtain a very important general result in a special case. Namely, if the weathering rate is not *explicitly* dependent on pCO_2, and if the precipitation depends only on temperature and not directly on the supply of absorbed stellar energy, then the equilibrium condition is simply $W(P(T), T)) = \Gamma$. This means that for any given outgassing rate and stellar luminosity, *the temperature must adjust to a fixed value.* In particular, this temperature doesn't change as the star gets brighter over time. In this case the weathering feedback acts to control climate perfectly, and the planet's temperature changes only as a result of either the outgassing rate or the continental configuration changing. What happens in this regime is that the requirement of weathering balance fixes T, and then pCO_2 must take on whatever value is necessary to achieve this T. For example, when the Sun is dimmer, pCO_2 must take on a higher value so as to make up for the reduced solar energy. Conversely, when the Sun is brighter, pCO_2 must take on a lower value. The thermostat can break down on the cold end if the required amount of CO_2 exceeds the supply. The temperature regulation may also break down at the cold end if the planet falls into a Snowball state. Conventional wisdom has it that silicate weathering ceases in this regime, allowing CO_2 to build up to where the planet thaws. However, seafloor weathering will continue, and if it is effective enough it could prevent the planet from every warming up sufficiently to escape the Snowball state. On the hot end, the thermostat can break down if the amount of CO_2 required to achieve temperature T falls all the way to zero. At this point, further increases in luminosity will cause the temperature to increase. The planet does not immediately go into a runaway greenhouse at this point, but if the planet has an ocean (as it must, if we are to have silicate weathering at all) then the water vapor feedback will ultimately lead to a runaway once

the luminosity exceeds the runaway threshold.

Now let's return to the general case, including the direct effect of pCO_2 on weathering. Putting all the effects together, the weathering rate can be represented by the empirical expression

$$\frac{W}{W_o} = (\frac{P}{P_o})^a (\frac{p}{p_o})^b \exp \frac{T - T_o}{T_U} \tag{8.12}$$

where W is the weathering rate P is the rate at which rain falls over rocks (the runoff), p is the partial pressure of CO_2 in the atmosphere and T is the temperature of the surface at which weathering is taking place. W_o is the weathering rate for the reference state with runoff P_o, carbon dioxide partial pressure p_o and temperature T_o. α, β and T_U are empirically determined constants. The last of these represents the direct temperature sensitivity of the Urey-Ebelman reactions. Based on values that have appeared in the literature, we adopt $a = .65$, $b = .5$ and $T_U = 10K$.

To complete the determination of the weathering rate, we need to determine T as a function of the absorbed stellar radiation and pCO_2. In the calculations to follow, we do this using the polynomial OLR fits described in Chapter 4. The OLR curve chosen was based on Earth gravity, with an assumed relative humidity of 50%. In determining the absorbed stellar radiation, we'll assume an albedo of 0.2, which approximates Earth's, adjusted for the greenhouse effect of clouds. We also need to know how P depends on temperature. One reasonable choice would be to make it increase in proportion to Clausius-Clapeyron. However, calculations with comprehensive dynamic climate models tend to indicate that precipitation increases less rapidly than Clausius-Clapeyron, even before the limitation due to surface shortwave absorption is encountered. Thus, we'll take $P/P_0 = 1 + a_P \cdot (T - T_o)$, with $a_P = .03$. With these specifications we can solve $W/W_o = \Gamma/\Gamma_o$ and determine how the pCO_2 and temperature change as the luminosity is changed. We will see how effectively the weathering thermostat can offset the Faint Young Sun, now that we have included the explicit pCO_2 dependence. According to the conventional wisdom, we are now studying the case of a planet for which the continents are abiotic.

The temperature and CO_2 as a function of the stellar constant are shown in Fig. 8.2. T_o is taken to be near the Earth's present-day temperature, and the outgassing rate is held constant at the value that balances weathering at this temperature. These calculations do not include the effects of ice-albedo feedback, which properly should enter in the cold conditions encountered when the star is faint. For comparison, the figure also shows what the equilibrium temperature would be if there were no silicate weathering thermostat and the CO_2 remained fixed at its baseline value. The results show that the CO_2 rises to very high values at early times when the star is faint – nearly $100mb$, or about 10% of an Earthlike atmosphere. However, because the weathering increases so much at high pCO_2, the balance can be achieved with quite low temperatures. The temperature falls to $270K$, which is only slightly warmer than the $250K$ value it would have in the absence of a weathering thermostat. It thus appears that, without some modification of the picture by land plants or some other way to get rid of the explicit pCO_2 dependence of weathering, the weathering thermostat does not go very far to resolve the Faint Young Sun problem. *With the abiotic pCO_2 dependence, the silicate weathering thermostat leaves the Early Earth in a very cold state, unless the outgassing rate is considerably higher at that time.*

One can also see the moderating effect of the weathering thermostat on the warm side, as the star gets brighter. When the star is 30% brighter than its baseline value, the pCO_2 has fallen all the way to $.002mb$, or 2 *ppmv*, and the temperature has risen to $311K$. This compares to the temperature of $323K$ that would occur in the absence of the weathering thermostat. At this point, there is hardly any CO_2 left in the atmosphere, and there is limited scope for the weathering thermostat to defend the planet against further increases in the luminosity of its star.

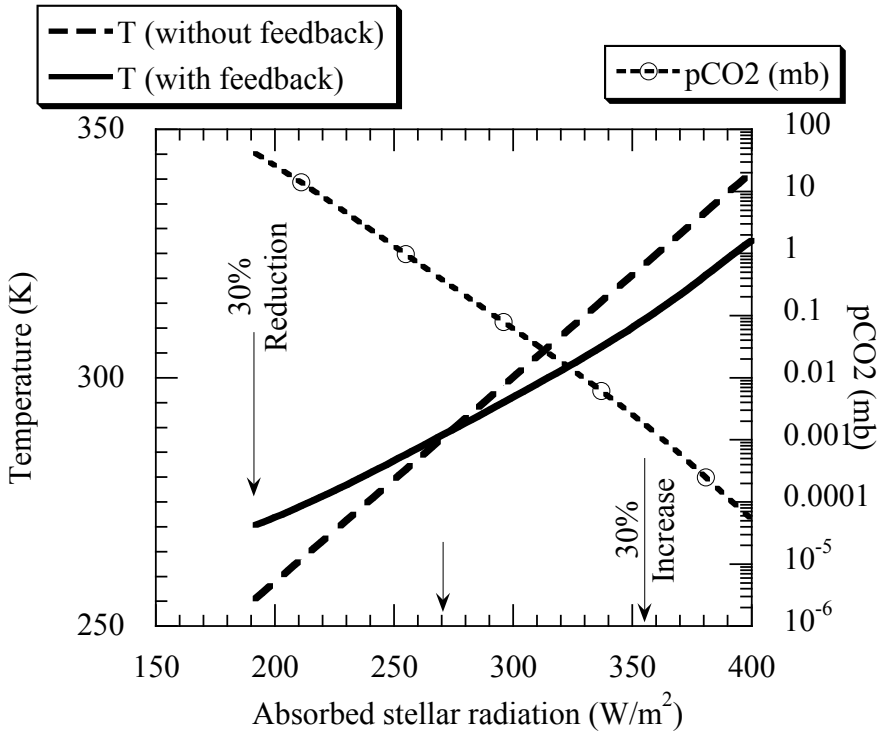

Figure 8.2: Effect of the abiotic form of the silicate weathering feedback on the variation of surface temperature with stellar luminosity. The luminosity in this graph is translated into absorbed stellar radiation per unit surface area of the planet, allowing for a 20% albedo. The pCO_2 curve gives the value of pCO_2 in equilibrium with a fixed outgassing rate for the case including the weathering thermostat.

If it really is true that land biota affect the explicit pCO_2 dependence as much as conventional wisdom suggests, then the implication is that land biota play a crucial role in planetary climate regulation. With land biota suppressing the explicit pCO_2 dependence of weathering, the climate regulation is nearly perfect. Without land biota, the silicate weathering thermostat moderates the influence of stellar luminosity on temperature change, but the regulation is not particularly tight, and on Earth one would have to invoke an increase in outgassing or some other effect in order to account for why the planet was not frozen during the Faint Young Sun – though, to be fair, the thermostat does bring the planet much closer to the temperature where a snowball could be avoided. For Earth, the question of when the thermostat became very efficient is tied in with the question of whether the suppression of pCO_2 dependence requires land plants (which only entered the picture about 300 million years ago), or whether bacterial colonization of land would be sufficient. The latter would extend the portion of Earth history for which the thermostat is efficient. The overall scenario including the evolution of land biota would go something like this: Early in the planet's history, there are no land biota and the thermostat is only moderately effective. The Faint Sun period would be very cold, though perhaps not completely frozen. As the luminosity increases, the planet would get considerably warmer, perhaps to conditions much warmer than the present. Then land biota evolve, greatly increasing the efficiency of weathering and eliminating the explicity pCO_2 dependence. At this point the planetary temperature drops sharply, but the reduction of weathering in cold, dry conditions prevents a Snowball state. Thereafter, the weathering thermostat becomes very efficient, and further changes in temperature are suppressed, until the required pCO_2 drops so low that temperature regulation ceases.

In Chapter 4 we discussed the classic dry-runaway greenhouse, in which the entire ocean evaporates into the atmosphere. However, if a planet fails to meet the criterion for a total dry-runaway (perhaps due to cloud or subsaturation effects), it may nevertheless get hot enough that it can lose a great deal of water vapor to space, despite retaining a liquid (but still hot) ocean. This is known as a *wet runaway* state, and it is the prevailing view of the path taken by Venus. Supposing this planet to have enough silicates to support silicate weathering, and supposing that there is enough reservoir of carbonate to support CO_2 outgassing from the interior, what do we expect the atmospheric CO_2 content to be? Do we have a steam-dominated atmosphere, or would the atmosphere also have significant amounts of CO_2 in it? Given the exponential dependence of weathering kinetics, it seems likely that the weathering reaction would be driven to equilibrium for a planet with surface temperatures much above $400K$, unless the outgassing rates are enormously greater than those prevailing on Earth today. Thus, one can estimate the CO_2 content during a wet runaway simply by looking at the equilibria for the Ebelmen-Urey reactions. The answer depends on temperature and surface mineralogy. At a temperature of $400K$ there would not be much CO_2 in the atmosphere unless the surface was dominated by iron carbonates, in which case one could have around $10bar$ of CO_2. By the time one reaches $600K$ there could be nearly a hundred *bar* of CO_2 even if the surface is dominated by magnesium carbonates. Calcium carbonates are very stable, however, so the temperature would have to go up to $700K$ before one had some tens of bars of CO_2 in the atmosphere. This is above the critical point for water, at which point the distinction between liquid water and water vapor disappears. This renders the distinction between a dry and wet runaway moot, and raises the interesting (and evidently unresolved) question of how the Ebelmen-Urey reactions behave in the presence of supercritical water. Is it like the dry reaction, like the aqueous reaction, or is it something completely different?

On a planet with oxygenic photosynthesis, the atmospheric CO_2 can also be affected by the burial of organic carbon and the release organic carbon by oxidation of the organic carbon pool – an example of oxidative weathering. Burial of organic carbon produced by oxygenic photosynthesis converts CO_2 (which is a greenhouse gas) into O_2 (which is not). This would cool the planet if it happens rapidly enough that the silicate weathering thermostat can't keep up. On a

well-oxygenated planet like the current Earth, organic carbon burial is relatively inefficient, since bacteria have had a few billion years to get very good at extracting energy by oxidizing any organic carbon that may be around. During the Great Oxygenation Event near the dawn of the Proterozoic, however, it is conceivable that oxygenic photosynthesis took over the planet so quickly that huge amounts of organic carbon were buried, drawing down atmospheric CO_2 and precipitating the Makganyene Snowball event. It is hard to imagine circumstances where something like this could happen under heavily oxygenated conditions. Massive release of CO_2 by oxidative weathering, on the other hand, can occur under modern oxygenated conditions. It is likely that the CO_2 release during the PETM event 55 million years ago came from oxidative weathering of the land carbon pool, and what is the present era of anthropogenic CO_2 release other than a form of oxidative weathering due to a particularly exotic form of biology? One ought to worry about whether the resulting warming could trigger an additional PETM-like land carbon release, which would add to the direct anthropogenic CO_2 release and compound our climate woes.

The CO_2 weathering feedback is but one of many possible climate feedbacks involving atmopheric reactions with crustal minerals, though to date it is the only one that has been worked out in detail. Other cycles that have been proposed include release of CH_4 from organic carbon by methanogenic bacteria on the young Earth, the regulation of SO_2 on Venus from dry reactions with surface carbonates, and the regulation of SO_2 on Early Mars and Early Earth by aqueous reactions producing sulfite minerals at the surface. On a planet without a substantial oceanic reservoir of water, the exchange of water with hydrated minerals could exert an important influence on atmospheric water vapor content; insofar as water affects the fluidity and melting-point of minerals, this can even feed back on the plate tectonics that affects the recycling of crustal material. N_2 doesn't easily form minerals on rocky planets, and so is unlikely to participate in major climate cycles, though it has been suggested that the biological formation of the ammonium ion NH_4^+ allows some drawdown of N into the Earth's mantle; this would somewhat affect the climate via Rayleigh scattering, pressure-broadening and lapse rate effects. On an icy body like Titan, however, N_2 readily forms "minerals" – they just happen to be called "ices" instead. Indeed, cryovolcanism based on various mixed NH_3-H_2O ices may well play a role in determining the amount of N_2 in Titan's atmosphere. The methane cycle on Titan is also likely to involve crustal and interior exchange processes. Photochemistry converts Titan's atmospheric methane to liquid ethane and various tarry sludges on the surface. At the time of writing the search is on for some way of closing the cycle by converting these substances back into methane.

8.4 Partitioning of constituents between atmosphere and ocean

How does the presence of a liquid ocean affect the composition of a planet's atmosphere? In this section, we'll consider the exchange of a substance such as the carbon in atmospheric CO_2 with the liquid ocean, neglecting any geochemical processes such as seafloor weathering which could remove dissolved components from the liquid and put them in long-term storage as solids in the oceanic crust. The primary example we have in mind is exchange of CO_2 between atmospheres and a water ocean, but some of the lessons apply more generally, especially to other reactive or nonreactive gases dissolving in water. The same general principles would apply to solvents other than water as well, though the general nature of ocean chemistry for oceans made of liquids other than water is at present essentially unexplored. The importance of oceans as a reservoir resides in the face that they have vastly more mass than the typical atmosphere (hence a lot of room to hold substances), but mix rapidly and can therefore exchange atmospheric constituents on relatively short time scales

of a millennium or so. This contrasts with the much more sluggish mixing associated with the solid crust and mantle, which prevents the still-greater mass of the rest of the planet from exchanging substances except on much longer time scales.

Within the general scenario outlined in the preceding section, the oceanic storage doesn't affect the atmosphere when the system is in equilibrium. In equilibrium, the amount of atmospheric gas entering the ocean equals the amount leaving, since –by definition– the oceanic reservoir is not changing when the system is in equilibrium. For atmospheric CO_2 concentration in equilibrium, for example, the ocean is just a pass-through reservoir, and in the traditional picture the atmospheric CO_2 concentration can be determined by silicate weathering on land without reference to how much of the planet's net carbon pool resides in the ocean water. When the system is out of equilibrium, however, uptake and release of $CO2$ by the ocean can have important consequences. For example, at the time of writing industrial humanity is releasing about 9 Gt of carbon per year into the atmosphere in the form of CO_2. How much of this stays in the atmosphere and adds to the greenhouse effect, and how much disappears into the ocean? To the extend that it disappears, how fast is the process and what is the rate-limiting step? Similarly, during the PETM event (see 1.9.1) an estimated 5000 Gt of carbon was released as CO_2 through oxidation of terrestrial organic carbon; if all of this CO_2 remained in the atmosphere, the resulting warming would be much greater than if a substantial portion disappeared into the ocean. The partitioning of carbon between atmosphere and ocean has similar consequences for deglaciation of Snowball Earth. In 10 million years, about 300,000 Gt of carbon in the form of CO_2 can be outgassed from the Earth's interior. As in the PETM case, the amount of warming resulting depends on the fraction which remains in the atmosphere. The Pleistocene glacial/interglacial CO_2 cycles present a more complex form of transient exchange of CO_2 with the ocean, involving both uptake and release, and probably assisted by the export of photosynthetically produced organic carbon from the upper ocean to the deep ocean.

As a warm-up to the full problem, let's first consider the simple case of a gas which does not undergo any chemical reactions after it dissolves in the ocean. The case of N_2 dissolving in a water ocean conforms well to this idealization. The solubility of a gas in a liquid is governed by *Henry's law*. For a gas A, with partial pressure p_A, Henry's law states that in equilbrium

$$p_A = K_H(T)c \tag{8.13}$$

where c is the concentration of substance A in the liquid and T is the temperature of the liquid. K_H is the Henry's Law constant, which has been measured and tabulated for a wide variety of substances. A small value of the Henry's law constant implies a higher solubility of the gas in liquid, since it takes less partial pressure to force a given concentration into solution. The units of K_H depend on the way the atmospheric content of A is represented and the way the concentration is represented; we will measure the gaseous partial pressure in Pa, and the concentration as mole fraction (number of molecules of A divided by number of molecules of solvent). At 298 K, for example, the Henry's law constant for N_2 dissolving in water is $9. \cdot 10^9 Pa$ using these units, since mole fraction is dimensionless. That means that our present atmospheric N_2 pressure of $8 \cdot 10^4 Pa$ would lead to a molar concentration of N_2 of $8.9 \cdot 10^{-6}$, or $8.9 ppmv$, within any body of $298K$ water that has been in contact with the atmosphere long enough to come into equilibrium.

The temperature dependence of the Henry's law constant follows an Arrhenius law, similarly to chemical equilibrium constants or Clausius-Clapeyron. Thus,

$$K_H(T) = K_H(T_o) \exp(-C_H \cdot (\frac{1}{T} - \frac{1}{T_o}) \tag{8.14}$$

The coefficient C_H is invariably positive. For N_2 in water, $C_H = 1300K$. Note that since K_H *increases* with temperature, the concentration in solution goes *down* as temperature *increases*, with

	N_2	O_2	CO_2	CH_4	H_2	Ar	NH_3	SO_2	H_2S
K_H (Pa)	$9.1 \cdot 10^9$	$4.3 \cdot 10^9$	$.16 \cdot 10^9$	$4.0 \cdot 10^9$	$7.1 \cdot 10^9$	$4.0 \cdot 10^9$	$9.1 \cdot 10^4$	$4.0 \cdot 10^6$	$6.39 \cdot 10^7$
C_H (K)	1300	1500	2400	1600	500	1500	4200	2900	2100

Table 8.1: Henry's law constants for selected gases dissolving in water. Values of K_H are given at $289K$.

partial pressure held fixed. In other words, the solubility of a gas goes down with temperature. This is somewhat non-intuitive, and is just the opposit behavior one has come to expect from the more familiar experiment of things like sugar dissolving in hot tea vs. cold lemonade. For a give atmospheric partial pressure, cold water holds gas than hot water. This is why trout like cold streams – they are more oxygen-rich than warm waters.

Henry's Law is another one of those magically universal thermodynamic relations, which applies across a vast range of circumstances. It applies as well for dilute solutions of N_2 in a liquid CH_4 ocean, or CO_2 dissolving in molten silicate, as it does for atmospheric gases dissolving in a conventional water ocean.

As an example, let's consider the partitioning of N_2 between Earth's present atmosphere and its ocean. The ocean contains $7.8 \cdot 10^{19} Mole$ of water, so at $298K$ the ocean would contain $6.9 \cdot 10^{14} Mole$ of N_2 using the equilibrium concentration calculated above. Now, using the surface pressure, the hydrostatic law, the surface area of the Earth and a mean molecular weight of 29 for air, we find that the atmosphere contains $1.8 \cdot 10^{17} Mole$, of which about 80%, or $1.4 \cdot 10^{17} Mole$ are N_2. Thus, the oceans at $298K$ would contain only 0.005 of the amount of N_2 in the atmosphere – a trivial proportion. This is an underestimate, since the oceans are on average closer to $275K$ than $298K$, but even allowing for the temperature effect, the presence of the oceans has little effect on the amount of N_2 in the atmosphere, as shown in the following exercise.

Exercise 8.4.1 Using the Henry's law temperature scaling coefficient for N_2, compute the proportion of N_2 in the ocean assuming a mean ocean temperature of $275K$

Henry's law data for a selection of gases is given in Table 8.1. The main distinction is between polar molecules like NH_3, SO_2 and H_2S, whose undisturbed state has a dipole moment, and non polar molecules like the rest in the table. Nonpolar gases are not very soluble, and have solubilities similar to N_2. CO_2 is signficantly more soluble than the rest, but not orders of magnitude so; we shall see shortly that there are other factors that have a far more profound impact on CO_2 uptake by oceans. The polar molecules are incredibly soluble, on the other hand, so on a planet with an ocean these substances would exist dominantly in aqueous solution, in raindrops, lakes and the ocean.

CO_2 behaves very differently from N_2, because once it dissolves in water it reacts to form carbonic acid (H_2CO_3) which dissociates to form bicarbonate ion (HCO_3^-) and a proton. The bicarbonate further dissociates into carbonate ion (CO_3^{--}) and another proton. The reactions in question are

$$\begin{aligned} CO_2(aq) + H_2O &\rightleftharpoons HCO_3^- + H^+ \\ HCO_3^- &\rightleftharpoons CO_3^{--} + H^+ \end{aligned} \tag{8.15}$$

	K_1	c_1	K_2	c_2	K_{sp}
35‰,2m	$1.83 \cdot 10^{-8}$	$508K$	$1.38 \cdot 10^{-11}$	$1910K$	$1.54 \cdot 10^{-10}$
35‰,3km	$2.28 \cdot 10^{-8}$	$1218K$	$1.61 \cdot 10^{-11}$	$2407K$	$2.56 \cdot 10^{-10}$

Table 8.2: The equilibrium constants are given at $298K$, assuming that concentrations are stated as mole fraction. The temperature dependence constants c_j are in K; the temperature dependence is of the form $\exp -c_j(1/T - 1/298)$. Temperature constants are not given for the solubility product constant K_{sp} because this quantity is not significantly dependent on temperature in the range 0-30C. Rows of the table give results at the stated depths in an ocean subject to Earth gravity.

whose equilibrium are described by the equations

$$\begin{aligned} \frac{[HCO_3^-][H^+]}{[CO_2(aq)]} &= K_1(T) \\ \frac{[CO_3^{--}][H^+]}{HCO_3^-} &= K_2(T) \end{aligned} \tag{8.16}$$

Water is essentially inexhaustible as a reactant, so the activity of water is absorbed into the definition of K_1. The activities indicated in square brackets differ somewhat from concentrations, but we shall ignore that refinement and take the activities to simply be the concentrations of the respective substances. Most conventionally, the concentrations are written in units of moles per liter, which for water is almost the same as moles per kilogram; the equilibrium constants would be stated in corresponding units. Here, we will instead measure concentrations as mole fraction, which is the ratio of the number of molecules of solute to the number of molecules of water in a given sample.

Like all equilibrium constants, K_1 and K_2 are temperature dependent, and follow an Arrhenius law; both constants increase with temperature. Values of the equilibrium constants plus their temperature dependence coefficients are given in Table 8.2. The constants also increase somewhat with increasing pressure. Sea water on Earth contains a great many ions from dissolved salts, measured in the aggregate as "salinity." These ions have a very strong effect on the activity of carbonate and bicarbonate ions (particularly the former). In the Table and in subsequent calculations, we avoid dealing explicitly by activity coefficients by absorbing them into effective equilibrium constants, which depend on salinity. With these effective equilibrium constants, calculations can be done as if the activities in the equilibrium relations were simply concentrations. It should be recognized, however, that the equilibrium constants are quite strongly affected by the salinity. For example, the effective value of K_2 for fresh water is an order of magnitude smaller than the value for typical sea water, which renders carbonate deposition in freshwater environments very different from what goes on in our ocean. The equilibrium constants might conceivably be affected even by the composition of seawater, which is something worth keeping in mind when thinking about oceans of the distant past, such as the very sulfate-rich anoxic oceans that could have prevailed at times during the Proterozoic and Archaean.

Henry's law gives the CO_2 concentration in terms of of the atmospheric partial pressure pCO_2, and pH gives the concentration $[H^+]$. Eq. 8.16 then states that for any given pH, the bicarbonate concentration is proportional to the dissolved CO_2 concentration, and in turn the carbonate concentration is proportional to the bicarbonate concentration; by Henry's law, the dissolved CO_2 concentration is itself proportional to pCO_2, so in the end the concentration of all carbon-bearing species are proportional to pCO_2, given a fixed pH. The partitioning amongst species is profoundly affected by the pH, though. K_1 is a very small number so the concentration of bicarbonate will be negligible compared to dissolved CO_2 unless the concentration of $[H^+]$ is

sufficiently small. For example, under fairly acidic conditions, with $pH = 5$, the proportionality constant $K_1/[H^+]$ is only .06 at $273K$. Since K_2 is even smaller than K_1, under these circumstances the concentration of carbonate is still smaller than that of bicarbonate. Thus, under sufficiently acidic conditions, essentially all of the dissolved inorganic carbon in the ocean is in the form of dissolved gas. As the pH rises, however the concentration of bicarbonate rises. At $pH = 6.2$ the bicarbonate concentration equals that of the dissolved gas, though the carbonate concentration is still a factor of 0.0006 smaller than that of bicarbonate. At $pH = 8.2$, which is the value prevailing in the present ocean, bicarbonate concentration is 100 times that of the dissolved gas, and carbonate concentration rises to 5.6% that of bicarbonate; in this case, most of the ocean inorganic carbon storage is in the form of bicarbonate. At still greater pH, carbonate comes to dominate the ocean carbon storage. A small change in the ocean pH makes a big change in the carbon storage in the ocean because the hydrogen ion concentration is exponential in pH. The situation is summarized in the left panel of Fig. 8.3.

Because CO_2 as a gas is not very soluble in water, one cannot store much carbon in the ocean when the pH is neutral or acidic. As pH increases, more of the dissolved CO_2 gets converted to bicarbonate and carbonate, and so the proportion of carbon in the ocean rather than the atmosphere rises dramatically. If there are N_o Moles of water in the ocean, then the number of Moles of carbon in the ocean is

$$pCO_2 \frac{N_o}{K_H}(1 + K_1/[H^+] + K_1K_2/[H^+]^2) \tag{8.17}$$

Now, if A is the surface area of the planet and g is its surface gravity, then the hydrostatic relation tells us that the number of Moles of carbon in the atmosphere is $pCO_2A/\overline{M}g$, where $\overline{M}$ is the mean molecular weight of the atmosphere. Hence the atmospheric fraction, or ratio of atmosphere to ocean carbon, is

$$f_{atm} = \frac{K_H A}{\overline{M} g N_o} \frac{1}{1 + K_1/[H^+] + K_1K_2/[H^+]^2} \tag{8.18}$$

The nondimensional coefficient at the front of this expression, which we shall call f_o, is a measure of the "size" of the atmosphere relative to the size of the ocean. It depends on pCO_2 only through the mean molecular weight of the atmosphere. When pCO_2 is small compared to the pressure of the rest of the atmosphere, $\overline{M}$ is fixed at the non-CO_2 mean value (29 in the case of present Earth air). As pCO_2 becomes large, $\overline{M}$ approaches 44, the value of a pure CO_2 atmosphere. f_o is the limiting atmospheric fraction in the case when bicarbonate and carbonate concentrations are negligible. For the Earth, $f_o = 2.03$ when CO_2 concentration is low, decreasing to 1.34 in the limit of a pure CO_2 atmosphere. As pH increases, the atmospheric fraction decreases sharply, as shown in the right panel of Fig. 8.3. In modern conditions on Earth, the atmosphere contains only about 2% of the total atmosphere/ocean carbon inventory. Adding carbon to the inventory generally changes the pH, so we are not yet fully equipped to say how much of the *added* carbon stays in the atmosphere.

To complete the solution to the problem we need to determine the pH of the ocean. This is the tricky part, since the answer ultimately depends on the supply of carbonate ion to the ocean from dissolution of carbonate on land and the sea floor. In order to determine the pH, we need to satisfy *charge balance* – the constraint that the net charge of positive and negative ions sums to zero. We will consider the simple case in which the only additional supply of carbonate comes from dissolution of $CaCO_3$ (limestone), which supplies both positive Ca^{++} ions and negative carbonate ions. In this case, the charge balance can be written

$$2[Ca^{++}] = [HCO_3^-] + 2[CO_3^{++}] + [OH^-] - [H^+] \tag{8.19}$$

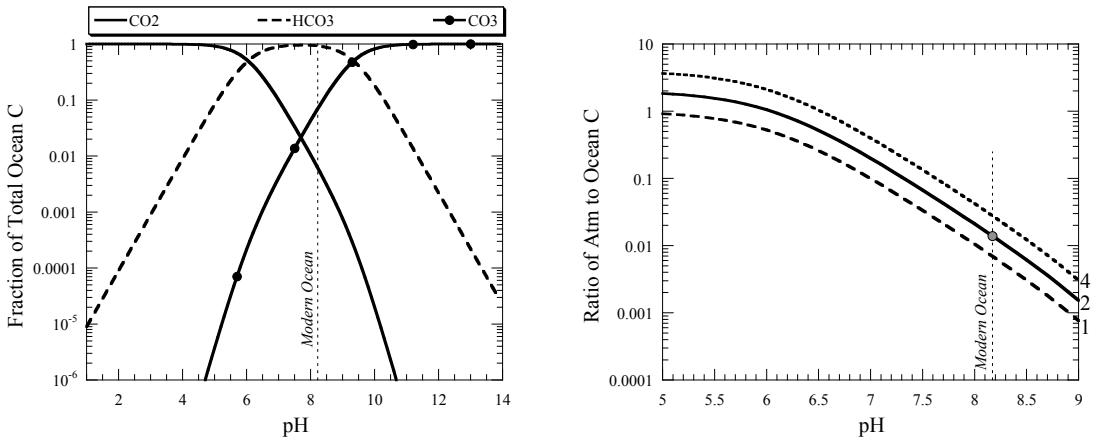

Figure 8.3: Left panel: The partitioning of total ocean dissolved inorganic carbon into dissolved CO_2, bicarbonate and carbonate, as a function of pH. Right panel: Ratio of atmospheric carbon to ocean inorganic carbon as a function of pH. The numbers to the right of the curves give the values of f_o, the atmosphere vs. ocean size parameter defined in the text. Smaller values correspond to a larger ocean, all other things being equal. For Earth's ocean $f_o \approx 1.87$. All calculations were done with an ocean temperature of $275K$, and for near-surface pressures.

assuming the activities written as square brackets are concentrations of the respective species. If pH is specified, then $[OH^-]$ can be computed from the water dissociation equilibrium relation in Eq. 8.8 (taking care to convert to the mole fraction units we shall use here). Further, if pCO_2 is given, the pH and carbonate/bicarbonate equilibrium relations determine $[HCO_3^-]$ and $[CO_3^{++}]$, whence the right hand side is known as a function of pH and pCO_2. This must balance the net charge of calcium ions. For any given $[Ca^{++}]$ and pCO_2, one can then iterate on pH using Newton's method until the charge balance is satisfied. Note that if there were no external source of carbonate from dissolution of limestone, then $[Ca^{++}] = 0$ and sea water in equilibrium with $300ppmv$ of atmospheric CO_2 would be mildly acidic, with a pH of 5.3 (See Problem **??**); inorganic carbon in the ocean would be overwhelmingly in the form of dissolved CO_2, and hence the ocean would store relatively little carbon. The actual pH of the present ocean is 8.17, and this alkaline state is maintained by the supply of dissolved carbonate and accompanying calcium ions, which wash in from the land and dissolve from the sea floor. This supply of carbonate is thus crucial to the ocean's ability to store carbon; without it, the ocean would be acidic and the bulk of dissolved inorganic carbon in the ocean would take the form of CO_2, whence a much greater fraction of the total atmosphere/ocean inventory would reside in the atmosphere.

Suppose next that we start with a situation in which balance is satisfied and then change pCO_2. This changes the value of the right hand side of Eq. 8.19, and disrupts the charge balance. In order to re-establish balance, the pH must change; how much it must change depends on what happens to $[Ca^{++}]$. Let's first consider the case in which the atmospheric CO_2 increases so rapidly that there is no time for new carbonate to be added to the ocean by dissolution on land or on the sea floor. This approximates the present situation, in which CO_2 has increased very rapidly owing to unrestrained burning of fossil fuels. In that case, $[Ca^+ +]$ remains fixed. To put some numbers to it, suppose that burning of fossil fuels were to add one trillion tonnes of C in the form of CO_2 to the atmosphere-ocean system. (This is a rather optimistic assessment of where we might hold the line with aggressive carbon emissions controls). That amounts to just under $2kg$, or .17 $Moles$, of

carbon per square meter of Earth's surface. It would increase the CO_2 concentration by $454 ppmv$ if it were to all stay in the atmosphere, increasing the added CO_2 re-equilibrates with the ocean, without benefit of input of additional limestone from dissolution. Carrying out the calculation, we find that the unperturbed preindustrial ocean/atmosphere system contains 46.64 trillion tonnes of carbon, almost entirely in the ocean. This increases by 1 trillion tonnes when the atmospheric pCO_2 is increased to $38.4 Pa$, or $384 ppmv$. Thus, when the equilibrium is achieved, only 104 $ppmv$ of the added carbon, or 23%, remains in the atmosphere. At this point, the pH of the ocean has decreased more than a full unit, to 7.08. The resulting acidification of the ocean can have a severe impact on the vast array of microscopic and macroscopic ocean life that uses carbonate to make shells. Details of the calculation, and further explorations, are carried out in Problem **??**.

Naively, one might have thought that since the ocean initially contains 98% of the carbon in the ocean-atmosphere system, the ocean would take up all but 2% of the carbon added to the atmosphere. This doesn't happen because the additional CO_2 depletes the supply of carbonate ion, allowing the pH of the ocean to go down. Carbonate ion is thus the bottleneck; it could almost be said to be "anti-CO_2." If one were to add an additional trillion tonnes of carbon to the atmosphere, the ocean would acidify further, and a greater proportion of the second trillion would stay in the atmosphere, as compared to the first trillion. (see Problem **??**).

Although nearly 80% of the first trillion tonnes of carbon we add to the atmosphere will eventually work its way into the ocean, transforming the environmental problem from one of global warming to one of ocean acidification, it will take approximately a thousand years for this equilibrium to be achieved. The time scale is set by the time required to mix dissolved carbon species into the deep ocean, since one needs the full mass of the ocean in order to deal with such a large amount of carbon. The deep ocean time scale becomes important because the exhaustion of carbonate ion keeps the upper ocean alone from being able to take up much additional carbon (see Problem **??**). The important role of the carbonate ion in limiting ocean carbon uptake, and its consequences for the magnitude of the global warming problem, was first recognized by Bolin and Ericsson. This breakthrough is commonly misattributed to Revelle and Suess, who in fact completely misinterpreted the effect and thought the ocean could handily take care of any likely amount of carbon humankind could throw at it. It will take about another 10,000 years for carbonate ion to be resupplied by dissolution, allowing the pH to rise gradually and the ocean to take up additional carbon over that span of time.

Finally, we'll consider the case in which the concentration of dissolved limestone is held in equilibrium with a reservoir of solid $CaCO_3$, rather like water vapor in equilibrium with ice at the saturated vapor pressure. This is approximately the state of affairs in the ocean on sufficiently long time scales. Dissolved $CaCO_3$ is held in a state of saturation by precipitation of solid or dissolution of solid, the latter of which might actually occur during rainfall over land, with the dissolved material washed into the oceans by rivers. Dissolution of calcium carbonate is described by the reaction and corresponding equilibrium constant

$$\begin{aligned} CaCO_3(s) &\rightleftharpoons Ca^{++} + CO_3^{--} \\ K_{sp} &= [Ca^{++}][CO_3^{--}] \end{aligned} \tag{8.20}$$

where K_{sp} is known as the *solubility produce constant*. The activity of $CaCO_3$ does not appear in the equilibrium expression since, by convention, the activity of a pure phase is set to unity. K_{sp} increases with pressure, but it is only very weakly dependent on temperature. Some representative values are given in Table 8.2. If there were no source of carbonate or calcium ion other than dissolution of limestone, then $[Ca^{++}] = [CO_3^{--}] \equiv x$, whence the equilibrium expression implies $x = \sqrt{K_{sp}}$. This is the concentration of dissolved limestone, once the water has come into equilibrium with the solid phase. Higher K_{sp} implies greater solubility.

In the real ocean, CO_2 also is a source of carbonate ion. As CO_2 goes up, the ocean acidifies and the concentration $[CO_3^{--}]$. However, this reduces the product $[Ca^{++}][CO_3^{--}]$, which means that if the unperturbed system was saturated, the new system is unsaturated, and so more $CaCO_3$ will dissolve to bring the system back into saturation. This will reduce the acidity of the ocean, *buffering* the pH. Limestone is nature's antacid, which is why it is the main component of antacid tablets used to combat overactive stomach acid. The buffering of pH allows the ocean to hold more carbon than it otherwise would, once sufficient time has passed for enough limestone to dissolve that a state of saturation is restored. Fig. 8.4 shows how the atmospheric fraction and ocean pH vary as a function of pCO_2, taking into account the buffering effect of limestone. We see that, despite the buffering effect, the atmospheric fraction rises with increasing pCO_2. For example, under preindustrial equilibrium conditions with $pCO_2 \approx 28Pa$, the atmosphere contains only on the order of a percent of the total carbon in the atmosphere-ocean system. However, if pCO_2 rises to $10^4 Pa$, which is on the order of the level needed to deglaciate a Snowball Earth, then the atmosphere contains about a third of the amount of carbon in the ocean, or a quarter of the total carbon. Were it not for this rise in atmospheric fraction, the amount of CO_2 that would need to be pumped into the system to attain an atmospheric pCO_2 of $10^4 Pa$ would be so huge as to definitively prevent deglaciation of the Snowball.

Other gases that form acids upon dissolving in water can have effects analogous to those we have been discussing in connection with CO_2, though none of these have been studied to nearly the same extent as the CO_2-bicarbonate-carbonate case. SO_2 is a case of notable interest, because it is a greenhouse gas and is also a significant component of volcanic outgassing. Upon dissolving in water, SO_2 establishes an equilibrium with HSO_3^- (bisulfite) and SO_3^{--} (sulfite). and forms a moderately strong acid. Compared to CO_2, SO_2 gas is extremely soluble in water, so oceans can take up a vast quantity of this substance even without conversion into bisulfite; the conversion has the potential to increase the uptake even further. Dissolution of sulfite minerals would play a role similar to that played by limestone in the carbonate system. If carbonate is also present in the system, it interacts in an interesting way with the sulfite system, since the dissolved SO_2 acidifies the ocean, prevents carbonate from precipitating and allows a greater proportion of the CO_2 to stay in the atmosphere. It has been suggested that the sulfite system, together with accompanying analogues of silicate weathering, could have played a role in the climate of Early Mars, and perhaps even of Early Earth. The study of climate-relevant geochemical cycles beyond the CO_2 system is in its infancy, and offers ample opportunities for the creative mind.

8.5 About Extreme Ultraviolet

Extreme ultraviolet (EUV) consists of electromagnetic radiation with wavelengths between 0.12 μm and 0.01 μm. EUV makes up only a tiny part of the spectrum of stars with photospheric temperatures under 10000 K – main sequence stars of spectral class B,A,F,G,K and M. (Recall that our Sun is a class G star). It nonetheless fuels the chemistry and physics of the outer atmosphere. EUV photons have sufficient energy to break up otherwise stable atmospheric compounds, allowing their components to combine into less stable forms that would not otherwise exist in appreciable quantities in the atmosphere. Further, because EUV photons are energetic enough to penetrate and interact with the electron clouds of atoms and molecules, the absorption cross section is so high that significant heating rates can be sustained despite the low EUV flux. This is not the case for the more abundant visible or near-ultraviolet photons, to which the tenuous outer atmosphere is largely transparent. For this reason, EUV absorption is an important source of energy available to sustain atmospheric escape. Because of the nearly ubiquitous role of EUV in what is to follow, it is useful to pause at this point and provide some general background on EUV radiation.

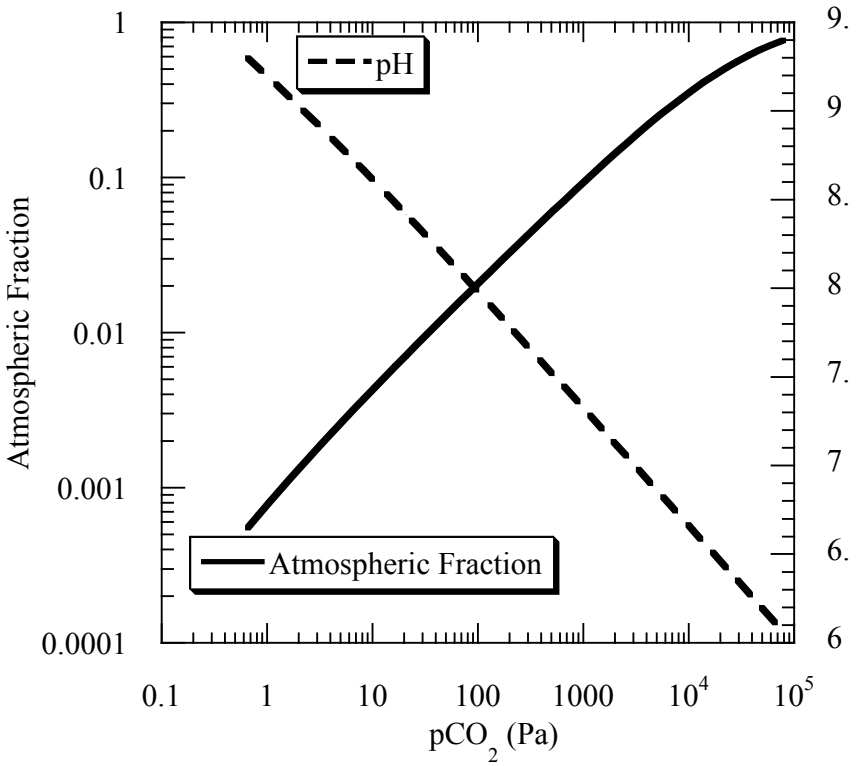

Figure 8.4: Atmospheric fraction and pH for a system buffered by dissolution of $CaCO_3$, which is presumed to be in a state of saturation. Calculations were carried out for Earthlike conditions, assuming an atmosphere whose non-CO_2 component has a molecular weight of 29. The number of Moles of non-CO_2 air in the system is held fixed as CO_2 is changed, at a value corresponding to 1 *bar* surface pressure in the absence of CO_2 (note that this pressure differes from the partial pressure of air in the mixture when pCO_2 is large.

EUV is not produced by blackbody radiation in a star's photosphere. Instead, it is produced high in the star's thin outer atmosphere – it's *corona* – where temperatures are brought to extremely high values by a variety of arcane and poorly understood heating mechanisms. The production of *EUV* in the hot corona has three important consequences. First, it allows the *EUV* energy flux to be far in excess of the blackbody value corresponding to the photospheric temperature. For example, at the Earth's orbit the photospheric blackbody flux from the Sun in the wavelength range from 0.055 μm to 0.060 μm would be a mere $7 \cdot 10^{-11} W/m^2$, whereas the observed solar *EUV* flux in this range is $9 \cdot 10^{-5} W/m^2$. The second important consequence is that the *EUV* fluctuates considerably in response to the solar activity cycle – loosely speaking, the sunspot cycle – because activity is intimately connected with the cycle of a star's magnetic field. The magnetic environment, in turn has a great effect on the corona. Thus, while the total solar luminosity varies very little over the course of the 11-year solar cycle, the net *EUV* flux varies by a factor of two or more – typically between $.003 W/m^2$ and $.007 W/m^2$ at Earth's orbit. In some sub-bands of *EUV* the variation is even more pronounced. The third consequence is that cool stars such as M-dwarfs can nonetheless have high sporadic *EUV* output, because the processes determining the coronal temperature are quite distinct from those controlling the photospheric temperature. M-dwarfs in particular have strong activity cycles, which give rise to the kind of events that effectively produce *EUV*. The strong and sporadic *EUV* output of such stars should give rise to novel aspects regarding the escape and chemistry of the atmospheres of planets orbiting these stars. One should also bear in mind that the Faint Young Sun could have had *EUV* output out of proportion to the dimness of the rest of the spectrum. In fact, while total stellar energy output increases with time, it is generally agreed that *EUV* output is higher for young Main Sequence stars. Solar *EUV* output is often assumed to be 3-6 times greater than at present in the first two billion years of the Sun's life as a star, though neither stellar models nor observations of other stars offer the possibility of precise estimates at present.

EUV is so strongly absorbed by the outer atmosphere that it can only be observed from outer space. The *EUV* output of stars other than our Sun is of extreme importance, but it was long thought that *EUV* astronomy was a dead-end, on account of absorption by interstellar hydrogen. It turns out, however, that there are sufficient fluctuations in the interstellar hydrogen density to permit a great deal of useful information to be obtained about stellar *EUV* output. Results from this emerging science will have a great bearing on the evolution of atmospheres of extrasolar planets.

8.6 A few words about atmospheric chemistry

Atmospheric chemistry influences climate evolution through determining the extent to which certain greenhouse gases can accumulate in the atmosphere. If there is a source of a gas such as CH_4 either from biological production or volcanic outgassing, the atmospheric concentration must be determined by a balance between source rate and chemical destruction rate. In addition, atmospheric chemistry plays a crucial role in the runaway greenhouse scenario, since permanent water loss requires water to be broken up into its constituent parts, and the hydrogen to escape before it can recombine into water or other substances too heavy to escape.

It would, of course, be preposterous to pretend that anything like an adequate treatment of atmospheric chemistry could be provided within the compass of a single section of a single chapter. This is a subject that, like atmosphere/ocean dynamics, requires a volume of its own. Indeed many tomes have been published on chemistry of planetary atmospheres, and many more pertinent specifically to present and past atmospheres of Earth, without coming close to exhausting

this rich and important subject. Still, as a problem in chemistry, atmospheric chemistry has certain peculiarities of its own, and the few words we provide here may help the orient the reader towards further study of the subject. This brief section is also intended to highlight where chemical reactions internal to the atmosphere fit into the bigger planetary climate picture portrayed in this book.

Section Under Construction

8.6.1 Photodissociation stirs the pot

Early attempts at resolving the Faint Young Sun problem for the early abiotic Earth relied on the greenhouse effect of NH_3 and CH_4, but these proposals quickly fell by the wayside once it was realized that the photochemical destruction rate of these compounds could not keep up with any likely sources. As an example, let's compute the photochemical lifetime of NH_3.

8.6.2 $OH-$ A radical's life

The OH radical is present in atmospheres in quantities so small as to be essentially unmeasurable, but it is so reactive it takes very little to have a profound impact on atmospheric chemistry, as the methane oxidation example illustrates. Because of its extreme reactivity, the OH radical has a short lifetime in the atmosphere. It is not made "to go" but is consumed on the spot without being significantly transported.

8.7 Escape of an atmosphere to space

Atmospheric escape calculations play a role in determining how a planet got to be the way we see it today, and what kind of atmosphere it may have had in its past; escape calculations can also indicate how long a planet can maintain a habitable climate, and can inform investigations into the mechanisms needed to maintain an atmosphere in a given state. For example, does there need to be an source of N_2 outgassing to maintain Titan's largely N_2 atmosphere? The answer to that hinges primarily on how rapidly a small, cold body like Titan can lose a relatively heavy gas like N_2. Similar questions apply to maintainance of CH_4 on Titan, since the processes which turn CH_4 into ethane and other compounds which accumulate on the surface liberate H_2, which must escape to space or accumulate in the atmosphere. Likewise, if Venus even had an ocean comparable to Earth's, and went into a runaway state, both the hydrogen and oxygen in the water must have been lost somehow, since there is very little water in Venus' atmosphere today. Could the hydrogen escape to space at a sufficient rate? The oxygen? The escape of water to space is what makes a runaway greenhouse essentially irreversible, so it enters into the habitability question for Venus and other planets subject to a runaway greenhouse. One picture of the climate of Early Mars posits a warm, wet climate supported in part by the greenhouse effect associated with a CO_2 atmosphere with a surface pressure of around 2 *bars*. How much of such an atmosphere could be lost to space in the available time? How much would have to be lost into chemical reaction with surface rocks instead? Is it possible that a body as small as Mars could remain habitable for billions of years, or does escape to space inevitably doom such a planet to the long chill? The contrast between Mars and Titan is stark. We expect atmospheres to be in some sense bound by gravity, but why is it then that Titan, with far weaker gravity than Mars, retains a 1.5 *bar* atmosphere while Mars retains practically none? One feels it must have something to do with the lower temperature of Titan, but that notion begs to be quantified. Even for Earth, escape

mechanisms are important: before the rise of atmospheric oxygen, there is the possibility that large amounts of hydrogen could accumulate in the atmosphere, if it does not escape rapidly to space. This is important because hydrogen and carbon dioxide can serve as feedstocks for synthesis of many prebiotic organic molecules.

8.7.1 Basic concepts

The problem of permanently removing a molecule from a planet's atmosphere is much the same as the problem of sending a rocket from Earth to Mars: one must impart enough velocity to the object, and in the right direction, to allow the object to overcome the potential energy at the bottom of the gravitational well, and still have enough kinetic energy left over to allow the object to continue moving away. This leads to the central concept of *escape velocity*, which is the minimum velocity an object needs in order to escape to infinity, provided no drag forces intervene. The escape velocity is obtained by equating initial kinetic energy to the gravitational potential energy. Let m be the mass of the object, v its speed and r its initial distance from the center of the planet. Let M_P be the mass of the planet and G be the universal gravitational constant. Then, equating kinetic to potential energy we find that $\frac{1}{2}mv^2 = GM_Pm/r$, so $v = \sqrt{2GM_P/r} = \sqrt{2gr}$ where g is the acceleration of gravity at distance r from the planet's center. In many situations of interest, the altitude from which the molecules escape is sufficiently close to the ground that g is only slightly less than the surface gravity. Important exceptions include small bodies with a massive atmosphere, such as Titan, or such as the Moon would be if it were given a massive atmosphere. In this section, we'll use g to represent the actual radially varying acceleration $g(r)$, and will use the symbol g_s when we need to refer specifically to the surface gravity. The acceleration at distance r is then $g(r) = g_s(r_s/r)^2$, where r_s is the radius of the planet's surface.

In order to allow a molecule to escape, enough energy must be delivered to the molecule to accelerate it to the escape velocity. The study of atmospheric escape amounts to the study of the various ways in which the necessary energy can be imparted. Since the kinetic energy of a molecule with mass m is $\frac{1}{2}mv^2$, light molecules like H_2 will escape more easily than heavier molecules like N_2, given an equal delivery of energy. Dissociation of a molecule like CO_2 or H_2 into lighter individual components also aids escape. Using the formula for escape velocity, we can define the *escape energy* of a molecule with mass m as mgr. For escape of N_2 from altitudes not too far from the Earth's surface, this energy is $2.9 \cdot 10^{-18} J$, or 2.9 *attoJ*[2]. For H_2 the escape energy is only 0.2 *attoJ*.

In the end, like so many things in planetary climate atmospheric escape is all about energy. There are four principle sources of energy that could potentially feed atmospheric escape:

- The general thermal energy of the atmospheric gas, which ultimately comes either from absorbed solar radiation or from heat leaking out of the interior of the planet.

- Direct absorption of solar energy in the outer portion of the planet's atmosphere, which may energize particles to escape velocity either directly or through indirect pathways,or which may manifest itself in a hydrodynamic escaping current of gas. The solar radiation responsibe for these mechanisms is typically in the extreme ultraviolet (EUV) portion of the solar spectrum because there is so little mass in the regions involved that absorptionis weak, whence there is a premium on absorbing individual photons which have a great deal of energy.

[2]One frequently sees the unit *electron volt* used for measuring small quantities of energy in a context like this. 1 $ev = 0.16$ *attoJ*

- Collisions with the energetic particles (usually protons) of the stellar wind which streams outward from the atmosphere of the planet's star.
- Kinetic energy imparted to the atmosphere by the impact of large objects.

For a portion of the atmosphere where collisions are frequent enought to maintain thermodynamic equilibrium, each degree of freedom gets an energy of $\frac{1}{2}kT$ on average. Some molecules will have more energy than this, and some will have less, but if the mean energy is coniderably less than the escape energy, only a very small proportion of molecules will have sufficient energy to escape. For $T = 300K$, the typical energy is only 0.002 *attoJ*. At this temperature only a small fraction of H_2 molecules would have sufficient energy to escape from Earth, and the escape rate becomes only moderately higher if the temperature of the escaping gas is raised to $1000K$. Heavier molecules like N_2 could hardly escape from a planet with Earth's or Venus' gravity at all, if the only source of energy were thermal motions. Even on a light body like Titan, the escape energy for N_2 is still 0.16 *attoJ* based on surface gravity, so escape will not be easy. Escape due to the energy associated with the thermal motions of particles in thermodynamic equilibrium is called *thermal escape*, or sometimes *Rayleigh-Jeans* escape. The ratio of escape energy to kT, namely $\lambda_c \equiv mgr/kT$, is an important parameter in the theory of thermal escape. In this formula, r is the radius of the atmospheric shell from which particles escape, g is the acceleration of gravity at that radius, and T is the typical temperature there.

For a particle to escape, it is not enought that it have reached escape velocity. It must also have a reasonable chance of escaping the gravitational well of the planet without suffering many collisions, since each collision will divert the particle from its outward path and rob it of some of its velocity. If the particle undergoes many collisions, it will undergo a random walk, leading to slow diffusion of the substance rather than rapid outward streaming. The portion of the atmosphere where particle collisions are so infrequent that a particle with sufficient energy has a good chance of escaping without collision is called the *exosphere*. In order to define the exosophere quantitatively, we first introduce the notion of *mean free path*, which is the mean distance traveled by a molecule between collisions. The mean free path depends on the total number density of particles, n, and the effective cross section area of the molecules, χ. For simplicity, we'll assume for the moment that all the molecules in the gas are identical. Two molecules are considered to collide if their centers approach within a distance of $2\sqrt{\chi/\pi}$. To estimate the mean free path, construct an imaginary cylinder with axis aligned with the direction of travel of the particle we are tracking, and having a radius $\sqrt{\chi/\pi}$. A collision is inevitable if this cylinder contains a particle from the rest of the gas, which becomes likely when $nV = 1$, where V is the volume of the cylinder. Writing $V = \ell\pi a^2 = \ell\chi$ we find that the mean free path is $\ell \approx 1/n\chi$. This estimate would be precise if the particles being collided with were stationary. In reality, the distance moved by the test particle before collision is affected by the fact that some of the other particles are moving toward the test particle, while others are moving away from it; a more precise calculation making use of the actual distribution of particle velocities in thermodynamic equilibrium yields the slightly modified result

$$\ell = \frac{1}{n\chi\sqrt{2}} \tag{8.21}$$

when all the particles have the same mass. When the background particles are very massive compared to the particle we are tracking, then they can be regarded as essentially stationary and the $\sqrt{2}$ factor in the denominator can be dropped; for most of the uses to which we will put the mean free path, this effect is of little importance.

The effective particle collision cross-section depends on the pair of molecules which are colliding, and is also a weak function of the energy of the collision, since molecules are not hard

	100K	300K	1000K
H-H_2	11.08	6.02	3.09
H-air	12.26	6.34	3.08
H-CO_2	13.63	7.46	3.85
H_2-air	144.50	125.96	108.36
H_2-CO_2	157.24	137.06	117.91
H_2-N_2	146.65	123.01	101.46
H_2O-air	194.71	142.21	100.78
CH_4-N_2	177.57	154.87	133.31
Ne-N_2	139.88	122.40	105.74
Ar-air	166.62	145.32	125.09

Table 8.3: Effective collision radius for various binary collisions, computed from diffusion data. The radius is given for three different temperatures, in units of *picometers*. ($1picom = 10^{-12}m$). The collision cross section for collision radius a is $\chi = \pi a^2$.

spheres but rather can be penetrated when the collision energy is large enough. Still, the effective collision radius does not differ greatly between one molecule and another. The collision radius can be inferred from measured diffusion rates, and a few typical values are given in Table 8.3. For the most part, the effective collision radius is between 100 and 200 *picometers*. An important exception is atomic hydrogen (H), which has an effective radius on the order of a mere 10 *picometers*. Curiously, the effective binary collision radius remains this small even when H is colliding with a much larger molecule. Evidently, the H atoms are like little bullets, which punch right through the outer electron clouds of the bigger molecules. This effect gives atomic hydrogen an anomalously large mean free path for a given density, which has a number of important consequences. Data for atomic oxygen (O) is hard to come by, but neon is believed to be an analog and should have a similar collision radius. Air at the Earth's surface with a temperature of $300K$ has a particle density of $2.4 \cdot 10^{25}m^{-3}$. Based on a collision radius of 125 *picom* the mean free path is about 0.6 μm. Adopting a scale height of $8km$, the particle density at an altitude of $100km$ falls to $9.0 \cdot 10^{19}m^{-3}$ and the mean free path increases to 16 cm – about the width of a hand.

The mean free path increases exponentially with altitude, because the particle density in a gravitationally bound atmosphere decreases exponentially with altitude, at a rate given by the density scale height (RT/g for the isothermal case). The exosphere is said to begin where the mean free path becomes sufficiently large that further collisions before escape are unlikely; this critical altitude is called the *exobase*. Commonly, it is defined as the altitude where the mean free path becomes equal to the scale height, since the exponential increase of mean free path with height means that if a particle doesn't collide with another within the first scale height, it is basically home free and unlikely to find another to collide with. When the exobase is not too far above the ground, in comparison to the planet's radius, g can be approximated by the surface gravity and one can immediately estimate the exobase altitude in terms of the temperature of the exosphere and the scale height for the dominant constituent of the exosphere. In applying this procedure, one must keep in mind that the temperature of the exosphere could be quite different from the temperature of the lower atmosphere, and the composition could differ greatly from the bulk composition of the atmosphere. We will have more to say about both these aspects of the exosphere a bit later. For the most part, one is interested in the exobase particle density, since that is what will determine the flux of particles to space. When one at least knows that the exobase is low enough that $g \approx g_s$, the exobase density is immediately given by the requirement that $\ell = H$, which implies that the

exobase particle density is

$$n_{ex} = \frac{g_s}{\chi R T_{ex}\sqrt{2}} = \frac{m g_s}{\chi k T_{ex}\sqrt{2}} \tag{8.22}$$

where T_{ex} is the exobase temperature, m is the actual mass (in kilograms) of the molecule which makes up most of the exosphere and k is the Boltzman thermodynamic constant.

For the more general case, one must extend the hydrostatic relation to account for the decay of gravity with altitude. This case is important for light constituents like H or H_2, which have a large scale height because low molecular weight implies large gas constant R. The exosphere can also be very extended even for heavier molecules, for small bodies with low surface gravity, such as Titan. Allowing for the inverse-square reduction of gravity with distance r from the center of the body, the equation of hydrostatic balance becomes

$$\partial_r p = -\rho g_s \frac{{r_s}^2}{r^2} \tag{8.23}$$

As in the treatment of the hydrostatic relation in Chapter 2, the equation is closed by using the ideal gas law, $p = \rho R T$, where R is the gas constant for the mixture making up the uppermost part of the atmosphere. In escape problems it is often more convenient to deal with particle number density rather than mass density. The particle density $n(r)$ is obtained by dividing ρ by the mass of a molecule, m. If the upper atmosphere is isothermal with temperature T_o, then the solution expressed as number density is:

$$n(r) = n(r_o)\exp(\frac{r_s}{H_s}(\frac{r_s}{r} - \frac{r_s}{r_o})), H_s \equiv \frac{R T_o}{g_s} \tag{8.24}$$

where r_o is some reference distance, generally presumed to be in the upper atmosphere. From this equation, it follows that the local scale height at radius r is $H(r) = (r/r_s)^2 H_s$.

Because of the attenuation of gravity, the density no longer asymptotes to zero at large distances from the planet. The limiting value is $n_\infty = n(r_o)\exp -(r_s/H_s)(r_s/r_o)$. Since this formula neglects all gravity but that of the planet, it is not surprising that one can fill infinite space with a finite density and have it stay put, since there is essentially zero gravity out there. To do so would require infinite mass, though, which means that if the planet is initially endowed with a finite-mass atmosphere, it will all leak away to space, given sufficient time. The estimate of that time scale is the objective of this section. In reality, the limiting density is not achieved, because the atmosphere is trucated by particle loss from the exobase. Still, it is important that there is a nonzero limiting density, since this implies that there is a nonzero limiting mean free path. Combined with the fact that the scale height increases with r, this can remove the exobase to infinity, meaning that there is no altitude at which the atmosphere can be considered collisionless. This situation won't arise for heavy constituents on reasonably massive bodies, since n_∞ is exceedingly small. Using a collision radius of 125 *picom*, the limiting mean free path is over $10^{167} m$ for O on Earth, based on a temperature of $300K$. This is well in excess of the size of the Universe. The limiting value for heavier constituents is even greater, and considerations for Venus turn out similarly. Even for Titan, the limiting mean free path for N_2 is about $10^{44} m$ based on a temperature of $100K$. The situation for the light species H and H_2 is more ambiguous. For example, if Titan had a pure H_2 atmosphere with a temperature of $100K$, the limiting mean free path would be only 850 m even if the surface pressure were a mere 0.1 Pa. This would certainly preclude a collisionless exosphere. Since H and H_2 typically appear in the atmospheres of terrestrial-type bodies as minor constituents mixed in with a heavier gas, further discussion is best deferred until after we have considered inhomogeneous atmospheres.

Exercise 8.7.1 Show that Eq. 8.24 reduces to the conventional hydrostatic relation derived in Chapter 2 when $r = r_s + z$ with $z/a \ll 1$, r_s being the radius of the planet.

To determine the exobase height, we need a model of the atmospheric structure which gives us the total number density $n(r)$ as a function of position. This can be challenging to do precisely, since $n(r)$ depends on the temperature and composition profile of the atmosphere; often, exobase heights for present-day Solar system planets are calculated from measured rather than theoretical density profiles. The other factor involved is the scale height at the exobase, which depends on the exosphere composition, position (through gravity) and temperature. Larger temperature increases the scale height and hence tends to move the exobase further out. The temperature of the exosphere is determined by a balance between heating by absorption of solar radiation (mainly ultraviolet), outgoing thermal infrared from deeper in the atmosphere, and infrared radiation. In some cases, there can be energy gain from collision with solar wind particles, and there can also be energy loss by outward streaming of mass in *hydrodynamic escape* (see Section 8.7.4). The radiative transfer in the exosphere is simplified by the fact that the atmosphere is optically thin, but is complicated by the fact that it is so tenuous that local thermodynamic equilibrium (and hence Kirchoff's laws) is not accurate. Still, when the exosphere is made of a good infrared emitter, the temperature tends to be on the order of a skin temperature, augmented a bit by solar absorption. Thus, the CO_2 dominated exobase of Venus has temperature between $200K$ and $300\ K$, and that of Mars is somewhat larger. The Earth's exosphere is unusually hot, since it is dominated by atomic oxygen which comes from photodissociation of the large O_2 concentration of the lower atmosphere. Atomic oxygen is a good ultraviolet absorber, but radiates infrared poorly, leading to high temperatures. Exosphere temperatures for N_2 or N dominated exospheres are a delicate matter, since N_2 neither absorbs nor emits well, and slight contamination by infrared emitters or good solar absorbers can make a big difference. Titan's N_2 dominated exosphere has an observed temperature of about $200K$.

Once the parameters of the outer atmosphere are settled, the exobase position is determined by solving $\ell(r)/H(r) = 1$ by iteration, where $H(r)$ is the scale height at position r and the mean free path $ell(r)$ is inversely proportional to $n(r)$. To get some numbers on the table for discussion, let's adopt a simple model atmosphere in which $n(r)$ is computed based on Eq. 8.24 with a uniform equivalent lower atmosphere temperature T_o all the way down to the planet's surface, where the surface pressure p_s is specified. The exobase temperature is specified separately. Exobase altitudes for some hypothetical single-component atmospheres are given in Table 8.4. The atomic oxygen case for Earth is meant to serve as a crude representation of the oxygen-dominated exosphere of Earth. In this approximation, the atmospheric structure is computed as if all the Earth's oxygen were in the form of O, and ignores the fact that the O_2 is only converted to O at altitudes above $100\ m$ as well as ignoring the effect of other gases on the vertical structure. We'll be able to do better later when we take up mixed atmospheres, but it is interesting to note that the exobase height of $400\ km$ in this approximation is not too far from the true exobase height ($500\ km$) computed on the basis of the observed O density in Earth's upper atmosphere. The N_2 case may be thought of as approximating an Early Earth situation in which there is little O_2 available to feed an oxygen-dominated exosphere. The two Venus cases represent approximations to the present state of Venus, and a hypothetical past near-runaway state with a pure steam atmosphere. The first Mars case assumes a dense CO_2 atmosphere such as might have prevailed on Early Mars, while the second is an approximate to the situation where the exosphere is dominated by atomic oxygen arising from photodissociation of CO_2. The first Titan case approximatess the present, while the second gives some indication of what would happen if the N_2 were to dissociate into molecular nitrogen (a somewhat implausible situation, but one which is included to allow us later to put a generous bound on nitrogen loss from Titan). Finally, the N_2 lunar atmosphere gives

Planet	p_s (bar)	T_o	T_{ex}	z_{ex} (km)	t_{loss} (GYr)	$w_{J,H}$, m/s
Earth,N_2	1.0	300	300	221	$7 \cdot 10^{286}$	$4.97 \cdot 10^{-7}$
Earth, O	0.2	300	1000	401	$4 \cdot 10^{42}$	7.94
Venus, CO_2	90.0	500	300	304	$> 10^{300}$	$1.7 \cdot 10^{-5}$
Venus, H_2O	1.0	400	300	542	$7 \cdot 10^{147}$	$3.458 \cdot 10^{-5}$
Mars, CO_2	2.0	250	300	352	$3 \cdot 10^{81}$	36.41
Mars, O	1.0	250	1000	1312	4000	808.13
Titan,N_2	1.5	100	200	774	$3 \cdot 10^{15}$	268.83
Titan,N	1.5	100	200	2454	5000	365.18
Moon, N_2	1.0	260	300	6145	4	613.54

Table 8.4: Characteristics of the exosphere and loss rates for various hypothetical single-component atmospheres. The planet and the atmospheric composition are given in the leftmost column. p_s is the surface pressure in $bars$, T_o is the effective mean temperature of the atmosphere below the exobase, T_{ex} is the exobase temperature, zex is the altitude of the exobase above the surface (in kilometers), t_{loss} is the time needed to lose the atmosphere by thermal escape (in billions of years), and $w_{J,H}$ is the Jeans escape coefficient for atomic hydrogen (in m/s). The hydrogen escape coefficient assumes that hydrogen is a minor constituent at the exobase. The mean free path was computed using a fixed molecular collision radius of 125 *picom* in all cases.

us an indication of what an atmosphere on Earth's Moon might have looked like if it retained or gained an atmosphere after formation. The lower atmosphere temperature approximates the temperature the Moon would have with little or no greenhouse gases in its atmosphere.

Relative to the planetary radius, the estimated exobases are all fairly close to the ground with the exception of the atomic oxygen case on Mars, the N_2 case on Titan, the atomic nitrogen case on Titan, and the N_2 case on the Moon. In the first two cases, the altitude of the exobase is on the order of a third of the planetary radius, but in the latter two cases the exobgse extends far out into space. The effect is particularly pronounced in the Lunar N_2 case. Note that the exobase extends much farther out than in the Titan N_2 case, even though the Moon has somewhat higher surface gravity than Titan. This happens because we have assumed a greater atmospheric temperature for the Lunar case, consistently with its closer proximity to the Sun. This remark underscores the importance of lower atmospheric temperature in determining the characteristics of atmospheric escape: perfectly apart from the exospheric temperature, a hotter lower atmosphere has a larger scale height, and therefore can extend further out to where the gravity is lower and the atmosphere can escape more easily. This is not much of an issue for bodies as massive as Earth or Venus but for smaller bodies it can be quite a significant effect.

Let's take stock of what we know so far. To determine the rate of escape of a consitutent, we need to know the height of the exobase, the number density of that constituent at the exobase, and the proportion of particles whose energy exceeds the escape energy computed at the exobase. The definition of the exobase involves the temperature at the exobase, through the definition of scale height, so we must know a temperature for the exobase as well. This temperature might or might not also serve to characterize the distribution of particle velocity, depending on circumstances. Note that the concept of "exobase" is itself a severe idealization. The picture this calls to mind is of a distinct surface separating lower altitudes where collisions are frequent enough to maintain thermal equilibrium and higher altitudes where particles undergo ballistic trajectories without collision. This would be nearly the case for evaporation of a liquid into a vacuum, since there is a near-discontinuity in density in that situation. For gases, the transition is gradual, and it would be better to talk in terms of an "exobase region" involving a continuous profile of collision frequency

and some escape to space from each layer – more toward the top, less toward the bottom. Modern calculations of atmospheric escape do indeed employ this level of sophistication, but the refinement alters estimates based on the ideal picture only by a factor of two or so. We'll see soon that this is not a serious threat to our main conclusions.

To proceed further, we need a probability distribution for molecular energy. The simplest case is one in which the molecules near the exobase can be regarded as being in thermodynamic equilibrium. This leads to what is called *Rayleigh-Jeans* or *thermal* escape. It is by far the simplest theory of atmospheric escape, but it is also the most useless; its main utility is to show that thermal escape is not a significant means of removing atmospheric constituents with the possible exception of light species such as He or molecular hydrogen (and even those only to a limited extent and in limited circumstances). The calculation proceeds as follows. For a gas in thermodynamic equilibrium at temperature T, the probability of a molecule with mass m having speed v is proportional to $\exp(-\frac{1}{2}mv^2/kT)$. This is the *Maxwell-Boltzmann distribution.* Note that if the gas is a mixture of molecules with various m, this formula still applies for the velocity distribution of each species separately, with the corresponding m used in the formula. The Maxwell-Boltzmann distribution has the important property that the proportion of molecules with energy much greater than kT becomes exponentially small. To determine the escape flux, one must integrate the Maxwell-Boltzmann distribution to determine the proportion of particles that have enough energy to escape, taking into account also the fact that particles are moving isotropically in all directions and that it is only the part of energy associated with radially outward motion that contributes to escape. If $n_{ex,m}$ is the number density at the exobase of a species whose molecules have mass m, then the flux of particles to space is $w_{J,m}n_{ex,m}$, where

$$w_{J,m} = \frac{1}{2\sqrt{\pi}}(1 + \lambda_c(m))e^{-\lambda_c(m)}\sqrt{(\frac{2kT_{ex}}{m})} \tag{8.25}$$

where $\lambda_c(m) = mg(r_{ex})r_{ex}/kT$ is the escape parameter defined previously. The Jeans flux coefficient $w_{J,m}$ has the dimensions of a velocity, and consists of the typical thermal velocity at the exobase reduced by an exponential factor that accounts for the proportion of molecules whose energy exceeds the escape energy. The total escape flux from the planet, expressed as molecules per second is $4\pi r_{ex}^2 w_{J,m}n_{ex,m}$. For calculations of the lifetime of an atmospheric constituent, it is convenient to introduce the escape flux per unit surface area of the planet, for which we will use the notation Φ. Thus, $\Phi = w_{J,m}n_{ex,m}(r_{ex}/r_s)^2$.

The penultimate column of Table 8.4 gives the characteristic loss time of the dominant constituent of the hypothetical atmosphere by Jeans escape. This loss time is obtained by dividing the Jeans loss rate for the dominant constituent into the total number of particles in the atmosphere. With the exception of the Lunar N_2 case, all the loss times are far too long to allow significant loss over the lifetime of the Solar system. For that matter, most of the loss times are well in excess of the lifetime of the Universe, and not even the extreme assumption of total decomposition of the atmosphere into lighter atomic constituents changes this conclusion. For the most part, the main constituents of terrestrial-type atmospheres cannot escape to any significant degree by thermal means. In particular, it is impossible to lose a primordial Venusian ocean by Jeans escape of H_2O. Even if we split off the O, the Earth atomic oxygen case says that it would be impossible to lose any significant quantity of the oxygen in the water by Jeans escape. The one case in which thermal escape is of interest for a heavy constituent is the warm Lunar case. This case may seem somewhat fanciful but it is of considerable relevance ot the question of habitable moons, such as might belong extrasolar gas giants which orbit their primaries at Earthlike distances. Jeans escape is a real threat to the atmospheres of small moons if they are warm enough to support liquid water.

To get a feeling for the magnitude of the thermal escape of atomic hydrogen in the regime

where H is a minor constituent of the exosphere, let's suppose that the molar concentration of H is 10% at the exobase. Later we'll learn how to relate the exobase concentration to the composition of the lower atmosphere. Let's take first the Earth case with an O dominated exosphere. Using the assumptions of Table 8.4 the total number density at the exobase, obtained by plugging the exobase position and gas constants into Eq. 8.24, is $2.4 \cdot 10^{14}/m^3$, whence the H number density is $2.4 \cdot 10^{13}/m^3$. Multiplying this by the Jeans escape coefficient from the table and normalizing to surface area, we find an H escape flux $\Phi = 2.15 \cdot 10^{14}/m^2 s$. If the H ultimately came from decomposition of sea water, then each two atoms of H that escape account for the loss of one water molecule and the generation of one atom of oxygen. Converting this to mass, we find that in four billion years you could lose about $365,000kg$ of water from each square meter of the Earth's surface, equal to a depth of about $365m$. You could not come close to losing an ocean on Earth or Venus this way, even with a hot exosphere; with a colder exosphere such as on Venus or the Early Earth, the loss rate would dwindle to practically nothing. If we want to get rid of a primordial Venusian ocean, we'll have to look at hydrogen-dominated exospheres, and find means other than thermal escape to pump the hydrogen into space.

Hydrogen, in the form of H_2 is one of the substances commonly outgassed from volcanoes on Earth and probably on other geologically active planets. In Earth's present highly oxygenated atmosphere, this hydrogen rapidly oxidizes into water, so there is little opportunity for free hydrogen to accumulate. On the anoxic early Earth, however, the accumulation of hydrogen would be limited by the rate of escape to space. For a cold N_2 dominated exosphere, the exobase density of atomic hydrogen is $6.7 \cdot 10^{15}/m^3$ if the molar concentration is 10%. Using the Jeans escape coefficient from Table 8.4, the hydrogen escape flux would be a mere $\Phi = 3.32 \cdot 10^9/m^2 s$. It has been estimated that the volcanic outgassing rate of H_2 on the Early Earth could have been on the order of 10^{15} molecules per second per square meter of Earth's surface, which is many orders of magnitude in excess of the Jeans escape flux. Thus, if Jeans escape were the only escape mechanism for hydrogen, hydrogen would accumulate to very high concentrations on the Early Earth. In reality, it would only accumulate to the point where the exosphere became hydrogen-dominated, whereupon other, more efficient escape mechanisms would take over. Still, we have learned from this exercise that hydrogen has the potential to build up to high values on an anoxic planet, that a cold exobase plays an important role in allowing this to happen, and that the exosphere is likely to have been hydrogen-dominated.

Another case of interest is hydrogen loss from Titan. In this case, the hydrogen is supplied by decomposition of CH_4 in the atmosphere, and the loss is important because the atmospheric chemistry would change quite a bit if hydrogen stuck around in the atmosphere. Let's suppose a 10% atomic hydrogen concentration at an N_2 dominated exobase on Titan. From Table 8.4 , we see that the Jeans escape coefficient for atomic hydrogen is very large in the Titan case, owing to the low gravity. The exobase particle density is about $1.3 \cdot 10^{14}/m^3$ based on the table, whence the assumed hydrogen density is $1.3 \cdot 10^{13}/m^3$ and the escape flux is $\Phi = 6 \cdot 10^{16}/m^2 s$. Assuming a CH_4 molar concentration of 30% over a layer $15km$ deep in Titan's lower atmosphere, there are $1.3 \cdot 10^{30}$ hydrogen atoms per square meter of Titan's surface, stored in the form of CH_4. The calculated Jeans escape rate would be sufficient to remove this entire inventory in under a million years. The precise rate of hydrogen loss depends on the rate of decomposition of methane and the rate of delivery of hydrogen to the exobase, but it seems quite likely that Jeans escape can get rid of the hydrogen resulting from methane decomposition.

Before taking on more complicated and effective means of escape, there is one more basic concept we need to take on: diffusion and gravitational segregation of atmospheric species. Let's track the position of a molecule of species A moving with typical speed v, and colliding with background molecules from time to time. The typical distance the molecule moves between collisions

is the mean free path ℓ computed earlier. If we idealize the collisions as causing a randomization of the particle's direction, then the particle will undergo a random walk. For particles undergoing random motions of this sort, the flux is proportional do the gradient of particle concentration; the process is called *diffusion*, and the proportionality constant is the *diffusion coefficient*, which we shall call D. It is closely related to the heat diffusion coefficient we have introduced in previous chapters. In addition, molecules or atoms in a gravitational field will accelerate downward under the action of gravity until the drag force due to collisions with the rest of the gas equals the gravitational force. This equilibration happens quickly, so that the particles attain a *terminal fall velocity* w_f. The terminal velocity is proportional to the local acceleration of gravity, and leads to a downward particle flux which is the product of the fall speed with the particle density.

The diffusion coefficient has units of length squared over time, and by dimensional analysis must be proportional to the product of mean free path ℓ with the typical thermal velocity $\sqrt{kT/m}$ where m is the particle mass of the species we ar tracking. Since ℓ is inversely proportional to the *total* particle density n, the diffusion coefficient increases in inverse proportion to n. For this reason, it is often expressed in terms of a *binary diffusion parameter* b, via the expression $b \equiv Dn$. For any given pair of species, b is a function of temperature alone. For ideal hard-sphere collisions between a particles with masses m_1 and m_2 and radii r_1 and r_2, the binary collision parameter is given by the expression

$$b = \frac{3\sqrt{2\pi}}{64}\frac{v}{\chi} \tag{8.26}$$

where χ is the collision cross section area based on radius $(r_1 + r_2)/2$ and $v \equiv \sqrt{kT/\bar{m}}$ is the thermal velocity based on the harmonic mean of the masses, $\bar{m} = m_1 m_2/(m_1 + m_2)$. For an ideal hard-sphere gas the binary parameter, and hence the diffusion coefficient, increases with the square root of temperature. For actual gases, however the effective collision diameter goes down somewhat with temperature, leading to other empirical temperature scaling laws, generally with temperature exponents in the range of .7 to 1. As with collision radius, atomic hydrogen is an exception, having a temperature scaling exponents between 1.6 and 1.7 for collisions with most species. The magnitude of the binary parameter for atomic hydrogen is also greater than one would expect from hard-sphere theory, since the effective collision radius is that of atomic hydrogen even when it is colliding with a substantially larger particle. Thus, atomic hydrogen has anomalously large diffusion, which increases anomalously strongly with temperature. The diffusion coefficient of atomic hydrogen is even larger than one would expect solely on the basis of it's low mass and hence large thermal velocity.

Next, we remark that the fall speed scales with the velocity acquired by the particle through gravitational acceleration in the time between collisions. Thus w_f scales with $g\ell/\sqrt{kT/m}$. The ratio D/w_f thus is proportional to $(kT/m)/g = R_A T/g$, where R_A is the gas constant for the species A. This is just the isothermal scale height H_A that the species would have in isolation. In fact a more detailed calculation shows that for an isothermal ideal gas, the ratio D/w_f is not just proportional to H_A, but is actually exactly equal to it. (As a short cut to this result, we may argue that it is implied by the requirement that the scale height be equal to the usual hydrostatic result when the species A dominates the atmospheric composition). In the following, we'll keep things simple by restricting attention to the isothermal case, which will be sufficient for our purposes.

Let n_A be the particle density of species A, which may be one of many constituents of the gas. Putting together the flux due to diffusion and the gravitational settling, the net flux of the species (upward positive) is

$$F_A = -w_f n_A - D\frac{dn_A}{dr} \tag{8.27}$$

Equilibrium is defined by a state of zero flux. In that case, the particle density is governed by the

equation

$$\frac{dn_A}{dr} = -\frac{1}{H_A} n_A, H_A \equiv \frac{w_f}{D} = \frac{R_A T}{g(r)} = \frac{kT}{m_A g(r)} \tag{8.28}$$

where m_A is the mass of a molecule of species A. Thus, the particle density decays exponentially with scale height H_A. Note that this is identical in form to the particle density given by hydrostatic balance, except that the equation we have just derived applies to the particle density of each species separately, and not just to the particle density of all species together. In fact, in equilibrium each species acts as if it were in hydrostatic equilibrium separately, and has the same hydrostatic scale height as if the other gases weren't there at all. Since we are assuming thermodynamic equilibrium, all species are characterized by the same temperature, and of course all species are subject to the same gravitational acceleration. Thus, the scale height varies inversely with the molecular weight of the species. Since the density of light species decays less sharply with altitude than the density of heavier species, the atmosphere will tend to sort itself out in the vertical according to molecular weight. Light constituents will congregate near the top of the atmosphere like escaped helium balloons at the top of a circus tent.

This can only happen, however, if the mixing is dominated by molecular diffusion. In the lower atmosphere, mixing is overwhelmingly due to turbulent fluid motions, which treat all species equally and keep the mixing ratios uniform, in the absence of strong sinks or sources by chemistry or phase change. Since the molecular diffusivity is inversely proportional to total particle density, it will increase roughly exponentially with altitude, and will therefore come to dominate turbulent mixing at sufficiently high altitudes. The altitude where diffusive segregation begins to set in is called the *homopause* (sometimes *turbopause*) and the lower part of the atmosphere where mixing ratios of nonreactive substances are uniform is called the *homosphere*. It is very difficult to get an *a priori* estimate of turbulent mixing rates – indeed at one time it was expected that the Earth's stratosphere would be diffusively segregated. More often than not, observations of atmospheric composition provide the most reliable estimate of the degree of turbulent mixing. For the present Earth, the homopause is near $100km$, at which point the observed particle density is $n_h(Earth) = 1.2 \cdot 10^{19}/m^3$.

Above the homopause, then, atmospheric constituents segregate in the vertical according to molecular weight. The region above the homopause is also typically (though not necessarily) where atmospheric molecules begin to be exposed to ultraviolet photons sufficiently energetic to break up even the more stable components into lighter constituents, which also will stratify according to molecular weight. For Earth the scale height for N_2 is 9.1 km, for CO_2 is $5.8km$, for O_2 is $8.0km$, for atomic oxygen is $15.9km$, for H_2 is $127.3km$ and for atomic hydrogen is $254.5km$, all based on a temperature of $300K$. Numbers for Venus are similar. The first implication of these numbers is that the scale height for hydrogen is so large that even a small concentration of hydrogen at the homopause can cause the atmosphere to become hydrogen dominated a small distance above the homopause. For example, if an N_2/H_2 atmosphere contains 1% molar concentration of H_2 at the homopause, then $50km$ up the concentration has risen to 63% and $70km$ up it is 93%. If the H_2 is converted to atomic hydrogen above the homopause, the segregation is even more effective. Similarly, if we take an Earthlike atmosphere that is 80% N_2 and 20% O_2 at the homopause, and then convert the O_2 into atomic oxygen, we wind up with 33% atomic oxygen near the homopause. By $20km$ up, the concentration reaches 56%, and at $40km$ it is 76% and thoroughly dominates the atmospheric composition. Finally, if we take an Early Earth CO_2/N_2 atmosphere consisting of 10% CO_2 at the homopause, then the CO_2 concentration falls to 0.5% $50km$ up, whereby we expect the outer atmosphere to be N_2 dominated. It is possible that the dissociation of CO_2 into CO and O could lead to an atomic oxygen dominated exobase, but the fact that this does not happen on Venus today suggests strongly that the dissociation is too weak for this to happen, or

the recombination of the two species is too efficient [3].

Given an estimate of the turbulent diffusivity D_{turb}, the homopause density can be estimated directly from the scaling of the diffusion coefficient. Specifically, since $D \approx \ell(kT/m)^{\frac{1}{2}}$ then setting $D = D_{turb}$ and using the expression for the mean free path ℓ implies

$$n_h \approx \frac{1}{\chi D_{turb}}(\frac{kT}{2m})^{\frac{1}{2}} \tag{8.29}$$

In cases where no observations bearing on D_{turb} are available, assuming D_{turb} to be the same as for Earth is probably as good an assumption as any. In that case, the homopause density is related to Earth's value by the formula

$$n_h \approx \frac{\chi(Earth)}{\chi}(\frac{T}{T(Earth)}\frac{m(Earth)}{m})^{\frac{1}{2}} n_h(Earth) \tag{8.30}$$

Given the weak variation of the factors multiplying $n_h(Earth)$ in typical cases, a good rule of thumb for use in making crude estimates of escape fluxes is to simply assume the homopause density to be the same as Earth's, though of course where observations are available it is better to use the observed value. For most escape calculations, it is not necessary to know the homopause altitude, though it can be estimated from the lower atmosphere scale height if it is desired. The homopause altitude only becomes important when it is high enough that gravity is significantly attenuated relative to surface gravity.

The homopause density provides the essential point of reference for most atmospheric escape problems involving multicomponent atmospheres. Since most of the atmospheric mass is in the troposphere, in order to understand how long it takes to lose some part of an atmosphere, we need to relate the escape flux to the tropospheric composition, which in turn requires us to relate the composition of the exobase to the tropospheric composition. This proceeds via the intermediary of determining the homopause composition. To take the simplest case first consider gases that do not undergo condensation or significant chemical sinks or sources within the lower atmosphere – for example O_2 and N_2 on Earth. These will be well-mixed throughout the homosphere, so if there is about 20% molar O_2 in the troposphere, there will be about 20% molar O_2 at the homopause. To get the O_2 particle density, we multiply this ratio by the homopause density, which we know how to determine. This then gives the supply of O_2 molecules, which dissociate above the homopause into atomic oxygen, allowing us to determine the atomic oxygen concentration at the exobase using the scale height for atomic oxygen alone to extrapolate from the homopause to the exobase. (A more precise calculation would require modelling of dissociation and reaction rates above the homopause). Things would work similarly for other gases that are nonreactive/noncondensing in the lower atmosphere.

For condensing gases, such as water vapor mixed with air on Earth, CH_4 mixed with N_2 on Titan, or water vapor mixed with CO_2 on Early Venus, there is an additional step on the way to determining the composition of the homopause. Condensible gases do not have uniform concentrations in the homosphere, because of the limitations imposed by Clausius-Clapeyron. Let's take water vapor on Earth as an example. Water vapor makes up a few percent of the lower

[3] Venus at present does have a region above the exobase which is dominated by atomic oxygen, but this layer is too tenuous to affect the exobase height. The question of the circumstances in which oxygen can build up to a hot Earth-type exobase is a delicate and difficult one, which hinges on details of atmospheric chemistry and atmospheric composition. An early water-rich Venus atmosphere would have another source of oxygen through dissociation of water vapor, which could conceivably lead to an oxygen-dominated exobase. Calculations performed to date do not seem to bear out this possibility, but the situation has not been thoroughly explored and there is plenty of room for surprises

atmosphere at present, but most water vapor entering the stratosphere must make it through the cold tropical tropopause. The temperature there is around $200K$, and the corresponding water vapor mixing ratio, by Clausius Clapeyron, is $1.6 \cdot 10^{-5}$, given a tropopause pressure of $100mb$. Though there are slight additional water vapor sources in the stratosphere from oxidation of methane, the tropopause concentration is a good estimate of the water vapor concentration that will be found at the homopause. The tropopause acts as a *cold trap*, dehumidifying the upper atmosphere and strongly limiting the opportunity for water vapor to escape or for hydrogen to build up in Earth's upper atmosphere through decomposition of water vapor.

The cold trap temperature is defined as the lowest temperature encountered below the homopause, and to determine it precisely, one must carry out a full radiative-convective calculation of the atmospheric structure. On an adiabat, temperature would go down indefinitely with height until absolute zero were reached; it is the interruption of temperature decay by the takeover of radiative equilibrium in the stratosphere that usually determines the cold trap temperature. In the absence of a full radiative-convective equilibrium calculation, the skin temperature of the planet often provides an adequate crude estimate of the cold trap temperature. Once the cold trap temperature is known, the maximum possible partial pressure of the condensible at the cold trap is given by Clausius-Clapeyron. However, it is the molar concentration of the condensible we need, since this is the quantity that is preserved as air is mixed up toward the homopause without further condensation. To determine the molar concentration we need the partial pressure of the non-condensible gas at the cold trap. This is obtained using the tools provided in Chapter 2. One computes the adiabat starting from a specified surface pressure, surface temperature, and surface condensible concentration – following the dry adiabat with constant condensible condensation until the atmosphere becomes saturated, and following the moist adiabat thereafter until the cold trap temperature is reached. The usual procedure for computing the moist adiabat then yields the necessary molar concentration. All other things being equal, as more non-condensible is added to the atmosphere, the cold trap concentration goes down owing to greater dilution of the condensible substance. The precise functional form of the dilution depends on the thermodynamic constants of the condensible and noncondensible substances under consideration.

As an example, let's look at the cold trap water vapor concentration that would be encountered during a dry runaway greenhouse in a CO_2-H_2O system. Recall that in a dry runaway the surface gets so hot that the entire ocean is evaporated into the atmosphere, and there is no liquid water at the surface; in this case, the shutoff of silicate weathering should allow any outgassed CO_2 to accumulate in the atmosphere, resulting in atmospheres consisting of CO_2 and water vapor in proportions determined by the abundance of these substances in the planetary composition (less whatever water may have already escaped). Results for various sizes of oceans and various CO_2 abundances are given in Table 8.5, based on a cold trap temperature of $200K$. The saturation vapor pressure and the adiabat were computed using the ideal gas equation of state and the idealized exponential form of Clausius-Clapeyron; these are not quantitatively accurate for the pressures and temperature under consideration, but they suffice to delineate the general behavior of the cold trap concentration. We see that, for any given inventory of water, the cold trap concentration approaches unity (pure steam) when there is little CO_2 present, but that the cold trap concentration falls to very small values as the CO_2 inventory approaches values similar to that of Venus. or the CO_2 equivalent of Earth's crustal carbonates. Also, for any fixed partial pressre of CO_2 at the surface, the cold trap concentration increases as the water inventory increases. Still, for a $90bar$ ocean (about half the mass of Earth's), and with a $90bar$ inventory of CO_2, the cold trap concentration is only 2.4%. Thus, unless the CO_2 inventory on a planet is very low or the water inventory is very high, the cold trap is likely to impose a significant barrier to water loss during a dry runaway scenario. Even if the water inventory is initally high, as water is lost the cold trap becomes a progressively more severe impediment, making it hard to lose the last $90bars$ worth of

	$0bar$	$1bar$	$10bar$	$30bar$	60 bar	90 bar
25 bar	1	.83	.23	.016	$3.3{\cdot}10^{-4}$	$5.7{\cdot}10^{-5}$
50 bar	1	.90	.42	.11	.0090	$9.3{\cdot}10^{-4}$
100 bar	1	.95	.61	.28	.092	.024

Table 8.5: Table of water vapor molar concentrations at a $200K$ cold trap, for a CO_2-water atmosphere. The column headers give the partial pressure of CO_2 at the surface. For each row, the mass of the ocean is held fixed at the indicated amount. The mass of the ocean is expressed as the pressure that would be exerted by the ocean if the water were condensed out into a liquid layer. For a planet with $g = 10m/s^2$, a 100 bar ocean corresponds to a mass of $10^6 kg/m^2$, or a depth of about 1 km. The 25 bar and 50 bar cases were computed with a surface temperature of $540K$, while the 100 bar case was computed at $570K$ so as to allow for a more massive water content without bringing the surface too close to saturation. Note that, as discussed in Chapter 2, the equivalent pressure of ocean differs somewhat from the partial pressure of water at the surface, since the mixing ratio of water is not uniform above the altitude where condensation first occurs.

ocean, and even harder to lose the last $50bars$.

For a single-component condensing atmosphere such as a water-vapor dominated runaway atmosphere on Venus or a condensing CO_2 atmosphere on Mars, one no longer has to consider the cold-trap issue, however. If there is only a single atmospheric component, then perforce knowing the total homopause density tells us the particle density of the atmospheric substance, regardless of how much condensation it has undergone in the troposphere.

Besides condensation traps, there can be chemical reactions which affect the homopause concentration. Notably, H_2 has little chance to escape in the modern oxygenated Earth, because it oxidizes to the heavier, condensible H_2O before it has a chance to reach the homopause.

Now let's revisit the problem of hydrogen loss from Early Earth, a runaway-state Venus, and Titan. We'll assume a mixture of hydrogen with some other gase in a known proportion at the homopause, and then use the scale heights of the two gases to compute the changing composition as the exobase is approached. This allows us to say when hydrogen dominates the exobase, and what the resulting exobase height is. An important complication is the anomalously small collision cross-section of atomic hydrogen, and we must remember to take this into account when computing the mean free path for hydrogen-dominated exospheres.

For the anoxic Earth, we wish to determine how high the homopause concentration has to be in order for the escape flux to equal the volcanic outgassing. We simplify the problem by assuming H_2 to be well mixed below the homopause, but to dissociate into atomic hydrogen just above the homopause. Thus, if we know the homopause concentration of atomic hydrogen, the well-mixed tropospheric H_2 density is half this value. Start by assuming the atomic hydrogen density at the homopause to be 20%, and that the balance of the atmosphere is N_2. Using the scale heights for the two gases, when we compute the exobase position taking into account the varying composition with height, we find that the exobase is completely hydrogen dominated, and that the exobase has moved out to an altitude of $1853km$ (based on an exobase temperature of $300K$). The escape flux from this extended pure hydrogen exobase is $\Phi = 10^{12}/m^2 s$, which is still three orders of magnitude below the estimated volcanic outgassing rate of H_2. Unless some more effective escape mechanism intervenes, hydrogen should build up to extremely high concentrations in the lower atmosphere.

For Venus, we assume an all water-vapor lower atmosphere. We take the homopause density to be $1.2 \cdot 10^{19}$ and assume that one half of the water vapor there breaks up into atomic hydrogen

and oxygen. To avoid dealing with a three-component atmosphere,we'll somewhat arbitrarily ignore the resulting oxygen (perhaps it recombines into O_2 which has such a small scale height that not much of it reaches the exobase) and compute the exobase from a homopause composition consisting of one third water vapor and two thirds atomic hydrogen; in addition, we'll assume a $300K$ exobase temperature. The exobase is again found to be hydrogen dominated, and at the relatively high altitude of 3050 km above the surface. The escape flux is $\Phi = 1.1 \cdot 10^{14}/m^2 s$, which would remove the hydrogen in one bar of water vapor in 200 million years. This is significant, but in 2 billion years one could only remove ten bars of ocean. By this means one could get rid of an ocean only about a tenth the mass of Earth's, though one could get rid of more if one could justify using a higher exobase temperature. Assuming that the water vapor at the honopause dissociates completely into atomic oxygen and atomic hydrogen changes these numbers very little, since the exobase is still hydrogen dominated.

It should be remarked that it is hard enough to get rid of an ocean's worth of hydrogen on a runaway Venus, but getting rid of an ocean's worth of oxygen by escape to space is completely out of bounds and none of the other escape mechanisms we will consider come close to closing the gap. The only hope of getting rid of the oxygen resulting from runaway followed by hydrogen escape is to react the oxygen with crustal rocks. Even this is problematic, since a great volume of crustal rock must be made available in order to take up the oxygen from an appreciable ocean. Whether this is indeed possible is one of the outstanding Big Questions. There is no data that absolutely forces us to assume that Venus indeed started with an ocean, so it remains possible that Venus was quite dry from the very beginning.

In our earlier calculation of hydrogen loss from Titan we found that a 10% hydrogen concentration at the exobase was sufficient to sustain a large thermal escape rate. How low does the homopause concentration have to be in order to keep the exobase N_2-dominated? To answer this, we again make use of the scale heights of the two gases to compute the exobase composition simultaneously with the exobase height. In this case, we find that with a $300K$ exobase, the homopause mixing ratio of hydrogen must be 10^{-6} or less in order to keep Titan's exobase N_2 dominated. With that homopause concentration, the escape flux is $\Phi = 8.8 \cdot 10^{15}/m^2 s$, which is somewhat less that our previous estimate (mainly because of the different means of estimating the exobase density). The main conclusion to be drawn from this exercise is that it only takes a tiny hydrogen concentration at the homopause to sustain the large escape rates we computed earlier. If the hydrogen concentration is increased to the point that the exosphere begins to become hydrogen dominated, then the exobase in fact moves out to infinity, because of the large scale height and low gravity. In that regime, hydrogen is likely to escape hydrodynamically (Section 8.7.4) rather than thermally.

8.7.2 Diffusion limited escape

The efficiency of escape of material that reaches the exobase is not necessarily the controlling factor determining atmospheric mass loss. For mass to escape from the exobase, it must first be delivered to the exobase, and in many circumstances the rate of transport of mass to the exobase is the limiting factor. When a minor consituent of an atmosphere is escaping, it must first diffuse through the dominant component on its way to the exobase, and even if the escape from the exobase is very effective, mass cannot escape faster than the rate it which it can diffuse up to the exobase. In such cases we can put an upper bound on the rate of escape without knowing much about the precise means of escape from the exobase. This upper bound is the rate of *diffusion-limited escape*. It has the virtue that it can be computed in a very simple and straightforward fashion.

We consider the diffusion in a gravitational field of a substance A with number density $n_A(r)$ through a background gas with density $n(z)$ satisfying $dn/dz = -h/H$. The equilibrium distribution of A was determined earlier by setting the flux to zero, but now we will determine its distribution assuming a constant nonzero flux. If we let b be the binary diffusion parameter for substance A in the background gas, then Eq. 8.27 for the flux can be re-written

$$F_A = -\frac{1}{H_A}\frac{b}{n}\cdot n_A - \frac{b}{n}\frac{dn_A}{dr} = -b\frac{n_{A,e}}{n}\frac{d}{dr}\frac{n_A}{n_{A,e}} \tag{8.31}$$

where $n_{A,e}$ is the equilibrium distribution of substance A, which satisfies $dn_{A,e}/dr = -n_{A,e}/H_A$. As expected, the flux vanishes when $n_A = n_{A,e}$.

For the density distribution to be time-independent, the net flux through a spherical shell, $4\pi r^2 F_A(r)$ must be independent of r. We'll normalize this constant flux to the surface area, writing $\Phi = (r/r_s)^2 F_A$ as we did for the Jeans flux. For a given constant Φ, Eq. 8.31 defines a first order differential equation for n_A. The upper boundary condition for this equation is applied at the exobase, and states that the flux delivered to the exobase must equal the escape flux from the exobase. The escape flux can be written $w_* n_A(r_{ex})$, where w_* is the escape flux coefficient associated with Jeans escape or some other mechanism. Thus, the upper boundary condition can be written $(r_{ex}/r_s)^2 w_* n_A(r_{ex}) = \Phi$. This determines $n_A(r_{ex}$ in terms of Φ, and we must then solve the equation to get $n_A(r)$ and adjust Φ so that the lower boundary condition on n_A at the homopause is satisfied. Now, when w_* becomes very large, molecules are removed essentially instantaneously when they reach the exobase. In this case $n_A(r_{ex}) \to 0$ and we can take a shortcut to determine the limiting flux.

For simplicity we'll assume that the layer is isothermal, so that b is constant. Multiply Eq. 8.31 by $n/n_{A,e}$ and integrate from the homopause to the exobase to yield

$$\Phi \cdot \int_{r_h}^{r_{ex}} \frac{r_s^2}{r^2}\frac{n}{n_{A,e}}dr = b \cdot \frac{n_A(r_h)}{n_{A,e}(r_h}, \tag{8.32}$$

which makes use of the assumption $n_A(r_{ex}) \approx 0$. The integral involves only known quantities, so this equation defines the limiting flux in terms of the homopause density of the escaping substance. Let's suppose that the exobase has low altitude in comparison with the radius of the planet. In this case, gravity is nearly constant and $n/n_{A,e}$ varies like $\exp(-(1/H - 1/H_A)z)$ where z is the altitude. The ratio decays exponentially if the scale height of the background gas is less than the scale height of minor constituent, i.e. if the minor constituent is lighter than the background gas. If moreover the layer between homopause and exobase is thick enough that the ratio decays to zero at the exobase, then the integral is simply $n/n_{A,e}$ at the homopause divided by $(1/H - 1/H_A)$. Therefore, under these assumptions, which are quite widely applicable, the diffusion-limited escape flux takes on the simple form

$$\Phi = b\frac{n_A(r_h)}{n(r_h}\cdot(\frac{1}{H} - \frac{1}{H_A}) = Dn_A(r_h)\cdot(\frac{1}{H} - \frac{1}{H_A}) \tag{8.33}$$

where D is the diffusivity at the homopause. The escape of the minor constituent cannot exceed this flux no matter how effective the escape mechanism may be.

Exercise 8.7.2 Derive an expression for the diffusion limited flux in the case when the diffusing constituent has greater molecular weight than the backgound gas. How does the limiting flux depend on the layer depth in this case?

As a first simple example of diffusion limited escape, let's take a look at the escape of hydrogen from an anoxic early Earth. Suppose that H_2 diffuses through pure N_2 above the homopause. At $300K$, the binary parameter for this pair of species is $2 \cdot 10^{21}/ms$. Then, noting that n_{H_2}/n is the molar concentration η_{H_2} of hydrogen at the homopause and that the scale height for H_2 is much greater than the scale height for N_2, Eq. 8.33 implies that $\Phi \approx (2 \cdot 10^{21}/H_{N_2})\eta_{H_2} = 2.2 \cdot 10^{17}\eta_{H_2}/m^2 s$. We'll assume that H_2 has nearly uniform concentration below the homopause,as is reasonable in the absence of oxygen. With this assumption, we can directly determine how high the concentration has to go in order to lose the volcanic hydrogen source, assuming the loss to be diffusion limited. Thus, $\eta_{H_2} = 10^{15}/2.2 \cdot 10^{17} = .0045$. Thus, while Jeans escape would allow H_2 to build up to very high concentrations in the tropospahere, diffusion-limited escape of H_2 could hold the concentration to well under a percent. Of course, for the diffusion limit to be reached, we would need a much more efficient means than Jeans escape of removing H_2 once it reaches the outer atmosphere. If the H_2 were to dissociate into atomic hydrogen near the homopause,the diffusion limited escape flux would be much greater and the equilibrium diffusion limited concentration would be much lower, since the binary coefficient for atomic hydrogen diffusing through N_2 is 10^{24} at $300K$.

Exercise 8.7.3 Suppose that an N_2/H_2 atmosphere has some initial hydrogen concentration which is escaping at the diffusion limited rate, but which is not being replenished by volcanic outgassing or any other source. Compute the exponential decay time of the hydrogen in the atmosphere assuming (a) H_2 diffuses through N_2 above the homopause, or (b) H_2 dissociates and diffuses as H through N_2 above the homopause.

As a second example, let's consider diffusion limited water escape for a "dry runaway" state on a planet like Venus. For the dry runaway, we assume that the entire ocean is evaporated into the atmosphere as water vapor, and that in the absence of liquid water a dense CO_2 atmosphere accumulates because of weak or absent silicate weathering. In this case, water may escape in the form of H_2O diffusing through CO_2. The binary parameter for this pair is $8.4 \cdot 10^{20}/ms$ at $300K$ based on the hard sphere approximation with 140 *picom* radii. Plugging in the appropriate scale heights for Venus Eq. 8.33 implies that $\Phi \approx 7.8 \cdot 10^{16}\eta_{H_2O}/m^2 s$, where η_{H_2O} is the water vapor concentration at the homopause. For the present application, we are interested in how long it takes to lose the hydrogen in an ocean, assuming it diffuses upwards as water vapor which then dissociates at high altitudes. We have seen earlier that η_{H_2O} is determined by the cold trap concentration in a dry runaway, which can range from miniscule values of a few parts per million to values approaching unity as wet-runaway conditions are approached. As a point of reference, let's assume that the cold trap concentration is 10%, which is typical of hot, moist conditions. Then, in one billion years, $2.4 \cdot 10^{32}$ molecules of water could be lost per square meter of the planet's surface. This amounts to $7.2 \cdot 10^6 kg/m^2$, or just over a $7km$ depth of ocean. It would thus appear that diffusion limitation is not a major impediment to loss of an ocean if the cold trap concentration is 10% or more. When the cold trap concentration falls below 1%, though, the diffusion limitation becomes a serious bottleneck. On the other hand, if water dissociates near the homopause and hydrogen escapes by diffusing as H or H_2, the diffusion limited escape rate becomes much greater, and lower cold trap concentrations can be tolerated without making it difficult to lose an ocean, as is illustrated in the following exercise

Exercise 8.7.4 For the conditions given in the preceding paragraph, compute the diffusion limited escape flux assuming water dissociates into H_2 at the homopause, so that $\eta_{H_2} = \eta_{H_2O}$, the latter being the cold trap concentration. Compute the lowest value of η_{H_2O} that permits the hydrogen in a 7 km deep ocean to escape. Do the same for the case in which water dissociates into H (in which case $\eta_H = 2\eta_{H_2O}$). In both cases you may use the hard-sphere formula in computing the

binary parameter b, assuming a collision radius of 140 *picom* for the H_2 case and a collision radius of 7 *picom* for *both* colliding species for the atomic hydrogen case.

Further examples of diffusion limited escape are developed in the Workbook for this chapter.

The concept of diffusion-limited escape applies in a straightforward fashion only to the escape of a minor-constituent, which makes up a small proportion of the diffusing layer. The removal of substantial quantities of a major constituent from the top of the diffusing layer has the potential to create large, unbalanced pressure gradients, which would drive an upward mass flux far in excess of the diffusive flux. There is no magic concentration threshold defining a "major" constituent, but certainly one should begin to worry about the induced flow when the concentration exceeds 10% or so. This regime is largely unexplored territory, but is somewhat related to the hydrodynamic escape mechanism which we shall discuss a bit later in this section.

8.7.3 Non-thermal escape

Using Planck's constant, the energy of an EUV photon with wavelength 0.05 μm is $4 \cdot 10^{-18} J$. This is sufficient to dissociate the components of many molecules, and to knock off electrons from just about anything – a process called *ionization* which produces charged particles in the outer atmosphere. In fact, ionization is the principle means by which absorbed EUV heats the outer atmosphere, since ejection of an electron increases the kinetic energy of the ion left behind, as well as imparting energy to particles the electron subsquently collides with. When easily ionized species are used up, the heating is correspondingly reduced.

The energy of such a photon is somewhat in excess of the escape energy for atomic oxygen on Earth, and so if this energy can be converted into kinetic energy of an atom, it can lead to escape at rates far higher than one would get from Jeans escape. This idea underlies the subject of nonthermal escape.

The term *nonthermal escape* represents a whole zoo of mechanisms united only by the common theme that they are not thermal – that they rely on the strong deviations from the Maxwell-Boltzmann distribution that are possible when collisions are infrequent. The study on nonthermal escape is rather like, to paraphrase a remark of Stanislaw Ulam, the study of the physiology of non-elephants. We will discuss a few examples to give the general flavor of the issues involved, but the reader should be aware that we are barely scratching the surface of this difficult and interesting subject.

The Maxwell-Boltzmann distribution corresponding to a given temperature is maintained through frequent collisions, which redistribute energy amongst the molecules making up the gas. If an individual particle in the gas acquires some new kinetic energy as a result of absorption of an EUV photon or a chemical reaction which releases energy, the energy may initially be much greater than the typical energy kT characteristic of the temperature of the gas. It is only after several collisions that the extra energy is equitably redistributed amongst the other particles – a process known as *thermalization*. If collisions are infrequent, the thermalization time can be very long, leading to a large population of particles that have anomalously large energy. Indeed in such a case, the energy distribution deviates greatly from the Maxwell-Boltzmann distribution, and the gas is no longer characterized by a temperature in the usual sense of equilibrium thermodynamics. The anomalously energetic atoms or molecules are often referred to as the "hot" population, as in "hot oxygen" or "hot hydrogen". The hot atoms may have enough energy to escape before they thermalize, or they may remain gravitationally bound but later impart their energy to lighter species which can escape. This is the general idea of nonthermal escape, and the reader can

probably understand already why there are many ways in which it can happen.

When the nonthermal escape is an indirect result of the accumulation of gravitationally bound hot atoms, the calculation of escape rate is very complicated, since one needs first to model the number of hot atoms, and the distribution of their energy. Ultimately, the energy is supplied by EUV absorption (or perhaps solar wind interactions, to be discussed separately), so that the flux of EUV provides an upper limit to the rate at which any constituent can escape. However, the actual rate depends on how the delivered energy is spread amongst the escaping constituents. If the energy is concentrated in relatively few particles, than an escape flux can be sustained, whereas if the energy is spread too thinly or is wasted on heavy particles, there may be little escape. Another important consideration is that only EUV delivered above the exobase can lead to nonthermal escape; energy deposited at lower altitudes will instead thermalize through collisions. Determining the proportion of EUV which is deposited above the exobase requires consideration of the number of particles above the exobase and the EUV absorption cross section of the species making up the exosphere.

Photons have little momentum, so the absorption of an EUV photon cannot directly increase the kinetic energy of the molecule or atom absorbing it to any great degree. Rather, the increase of kinetic energy takes place through *ionization* or *dissociation*. It the first case, a fast electron is ejected in one direction while the heavier positive ion moves more slowly in the opposite direction, with such a speed as to satisfy the momentum balance. It takes energy (the *ionization energy*, which differs according to species) to pry loose an electron, so the total kinetic energy supplied to the electron and ion is the energy of the photon minus the ionization energy. Dissociation is similar, except that the molecule breaks apart into neutral or charged heavy constituents instead. Dissociation also requires energy, so that the energy available to increase the kinetic energy of the fragments is the energy of the photon diminished by the dissociation energy.

As a concrete and relatively simple example, let's consider photodissociation of N_2 on Mars, which has been suggested as a possible means of losing the nitrogen in a hypothetical nitrogen-rich primordial Martian atmosphere. This problem is of interest because Mars at present has little N_2 in its atmosphere, and we'd like to know whether this means that Mars must have formed with very little nitrogen (unlike Earth or Venus), or whether the smaller size of the planet allowed its initial nitrogen endowment to escape.

Assuming the Martian exobase to be reasonably close to the ground, an N atom requires $2.92 \cdot 10^{-19} J$ of energy to escape. This is much greater than the typical thermal energy $kT = 4. \cdot 10^{-21} J$. Now, the dissociation energy of N_2 is about $1.94 \cdot 10^{-18} J$, so assuming the excess absorbed energy to be equally distributed between the two dissociated atoms, a photon with energy in excess of $1.94 \cdot 10^{-18} J + 2 \cdot 2.92 \cdot 10^{-19} J$, i.e. $2.52 \cdot 10^{-18} J$. Using Planck's constant, this corresponds to EUV photons with wavelengths shorter than .078 μm. Next, we must determine the rate at which such photons collide with N_2 molecules and cause dissociation. For simplicity, we'll assume a pure N_2 atmosphere, so we need not worry about collisions with other molecules. If the exospehere is optically thin in the EUV, then the proportion of incident photons which are absorbed is the product of the absorption cross section with the number of particles per square meter in the exobase. The latter is approximately the product of the exobase density with the scale height, i.e. $1/\chi\sqrt{2}$, where χ is the molecular collisional cross section area. We are left with the rather tidy result that *the proportion of photons that are absorbed in the exosphere is the ratio of the EUV cross section to the molecular collision cross-section.* In the relevant part of the EUV the absorption cross section for N_2 is about $3 \cdot 10^{-21} m^2$, so the proportion of incident photons which are absorbed in the N_2 exosphere is 4.6%. Note that this fraction is independent of the total mass of the atmosphere, so that we cannot increase the rate of escape by increasing the total amount of N_2 in the atmosphere. In general, this is one of the chief factors limiting the effectiveness of

nonthermal escape due to photodissociation.

Based on satellite observations the flux of sufficiently energetic EUV photons at Earth's orbit is currently about $8 \cdot 10^{14}/m^2 s$, which scales to about $4 \cdot 10^{14}/m^2 s$ at the orbit of Mars. Allowing for the proportion which are absorbed in the exosphere, this leads to an escape of $4.6 \cdot 10^{12}$ N atoms per square meter of the planet's surface per second or $1.45 \cdot 10^{29}$ atoms per square meter of surface in a billion years. One bar of N_2 on Mars contains $1.16 \cdot 10^{30}$ atoms per square meter, so we conclude that EUV induced nonthermal escape has the potential to remove about a half a bar of N_2 from Mars over the lifetime of the Solar system, or more if the EUV flux were higher in the early Solar system.

The above calculation is sufficient to show that escape from EUV-induced dissociation is a potentially important factor worthy of more sophisticated study, but it is highly oversimplified and leaves out many considerations that could substantially reduce the escape. First, the dissociated N atoms are not always in the lowest energy electron configuration (the ground state). The energy that goes into excited electron states is not available to feed kinetic energy sustaining escape. To crudely take this into account, in the above calculation we used the energy for dissociation into the most favored excited electron configuration; the dissociation energy into ground-state N atoms would be only $1.57 \cdot 10^{-18} J$, though such dissociations are believed to be rare. Second, photons with energy higher than $2.5 \cdot 10^{-18} J$ can cause ionization of the N_2 molecule rather than dissociation. Only a small proportion of the molecules will directly dissociate. Some of the N_2^+ ions will later dissociate and lead to escape when they recombine with the free electrons released by ionization – a process called *dissociative recombination* – but calculation of the proportion that do so is quite involved. Third, in an atmosphere that is not pure N_2 there is a stew of other dissociation products, both neutral and ionized, that need to be considered, as well as the reactions between them.

On the other hand, dissociation due to EUV photons with energy less than $2.52 \cdot 10^{-18} J$ leads to a population of gravitationally bound N atoms which have typical energies much larger than the typical thermal energy. Such populations of energetic particles are referred to as "hot", as in "hot nitrogen" or "hot oxygen", even though they are not in thermodynamic equilibrium and their temperature is not strictly defined. Sometimes a "temperature" is defined for such populations just as a means of characterizing the energy. For example, if a photon with energy $2.3 \cdot 10^{-18} J$ causes dissociation, it leaves behind two N atoms with an energy of $0.17 \cdot 10^{-18} J$ each. Setting kT equal to this energy yields a "temperature" of over $12000K$. This is merely a way of summarizing how the energy of the hot population compares the the much lower background thermal energy (characterized by a temperature of around $300K$ for the Martian exosphere). The hot population does not escape itself, but it represents a reservoir of energy that can be transferred to other species (especially lighter ones), and which can lead to the escape of those. In this mode, the outer atmosphere acts like a storage-beam particle accelerator, accumulating energetic particles for later use.

Other dissociations can also lead to escape or accumulation of hot atoms. N_2 has an unusually large dissociation energy. The reactions $O_2 + h\nu \rightarrow O + O$, $CO_2 + h\nu \rightarrow CO + O$, $H_2O + h\nu \rightarrow H + OH$ and $OH + h\nu \rightarrow O + H$ all have dissociation energies of around $0.8 \cdot 10^{-18} J$. Dissociation of these by EUV leaves more energy left over to feed escape.

Dissociative recombination, particularly involving oxygen, is an important indirect means by which EUV absorption can lead to escape or accumulation of hot atoms. For example, when the recombination of the O_2^+ ion with an electron leads to the dissociation of the result into two O atoms in their ground states, then $1.12 \cdot 10^{-18} J$ are released, or $0.56 \cdot 10^{-18}$ per atom. At present, such reactions take place on Venus and Mars as a consequence of the photodissociation of CO_2, but in earlier epochs on Venus water vapor could also have been involved. The escape energy for

O on Venus is $1.43 \cdot 10^{-18} J$, so dissociative recombination does not directly lead to escape of O on Venus, but rather to the accumulation of hot oxygen that can later allow a smaller part of the oxygen (or a greater part of a lighter species) to escape. Significant amounts of oxygen are indeed observed escaping at present from Venus, but it is not thought that the nonthermal mechanisms could have gotten rid of any significant part of the oxygen in a primordial Venusian ocean. On Mars, however, dissociative recombination directly leads to the escape of an oxygen atom, since the escape energy is only $0.33 \cdot 10^{-18} J$.

A further set of complications arises from the fact that charged particles interact with the planet's magnetic field, if it has one. Charged particles tend to spiral tightly along magnetic field lines, and so they can escape easily only when they are on the relatively limited proportion of field lines that are *open* – which have one end leading out into outer space. When the magnetic field is important, there is thus a great premium in generation not just of energetic particles in general, but energetic *neutral* particles. For this reason, a great deal of attention in work on nonthermal escape has been lavished on processes that allow energetic ions to deliver their energy to neutral particles.

8.7.4 Hydrodynamic escape

Hydrodynamic escape is basically a more efficient means of deploying the energy available to the atmosphere in order to assist escape. The energy involved still comes from EUV absorption or the general thermal energy of the atmosphere, but instead of this accumulating in a more or less random set of motions, in some circumstances the energy can sustain a mean outward escaping flow which carries fluid to space without wasting energy on motions directed toward the planet or on a population of molecules with velocities too small to escape. In these circumstances, there is no longer an exobase from which particles escape directly to space. Instead, there is an outflow that acts like a collisional fluid out to distances so great that the atmosphere is no longer gravitationally bound. Hydrodynamic escape plays a very central role in hydrogen-rich outer atmospheres (including those that arise from dissociation of water vapor), and so we will accord it a great deal of attention. The phenomenon can also play a role in escape of heavier species in the case of small bodies or very hot planets. Hydrodynamic escape is a fascinating and important subject, and one which is very ripe for further research. It is also a subject where the student can attain a nearly complete level of understanding on the basis of some simple principles of thermodynamics and mechanics. Therefore, it is a subject which we will delve into at some considerable length.

Equations for 1D compressible transonic flow

The starting point for our discussion is Newton's Law – force equals mass time acceleration – written for the radial direction measured outward from the planet's center. Let r be the radial position, and suppose that the only nonvanishing velocity is the radial velocity $w(r)$. We suppose further that the system is in a steady state, so that winds, temperature and so forth are time-independent when measured at any fixed position. This does not mean that acceleration vanishes, however, since the acceleration must be measured following the path of an outward-moving fluid particle, whose position can be written $r(t)$. Since $w \equiv dr/dt$, the acceleration following a fluid parcel is $dw/dt = (dw/dr)(dr/dt) = w(dw/dr)$. Let r_s be the radius of the planet's surface [4] , and

[4] Any other convenient reference radius can be used in place of r_s.

$g_s = g(r_s)$ be the surface gravity. Then Newton's law (expressed per unit volume) becomes

$$\rho w \frac{dw}{dr} = -\frac{dp}{dr} - \rho g_s \frac{r_s^2}{r^2} \tag{8.34}$$

When $w = 0$ this reduces to the hydrostatic balance given in Eq. 8.23. Atmospheres are never exactly at rest, and so the hydrostatic *approximation* we have been using throughout this book amounts to an assumption that the accelerations on the left hand side of the equation are negligible compared to the individual terms on the right hand side. What we are up to now is the business of figuring out what happens when the radial acceleration becomes large enough to disrupt the hydrostatic balance. Basically, we only need to solve the radial momentum equation subject to suitable boundary conditions; the resulting solution determines the outward mass flux. However, there are a number of subtleties concerning the circumstances in which a steady solution can exist, and the nature of the boundary conditions that can be applied. Therefore, we proceed to the solution through a number of intermediate steps.

To obtain a solution, the momentum equation must be supplemented by mass conservation and thermodynamic relations. For steady flow, conservation of mass requires that the mass flux must be independent of r. Defining the area of the shell at radius r as $A(r) = 4\pi r^2$, the mass flux $\rho(r)w(r)A(r)$ is constant. It is convenient to define the mass flux per unit surface are of the planet,

$$\Phi \equiv (\rho w A(r))/A(r_s) = \rho w (r/r_s)^2, \tag{8.35}$$

which is of course also constant. The thermodynamic relations needed consist of the equation of state ($p = \rho RT$ in the present discussion) and the corresponding equation for potential temperature (θ) or entropy ($c_p \ln\theta$). In the adiabatic case, the entropy is independent of r and is fixed by the boundary conditions. In the general case including heating, we need an equation for the radial variation of entropy, which we will bring in later.

The transition from flow speeds slower than that of sound (subsonic flow) to flow speeds greater than that of sound (supersonic flow) plays an important role throughout the following, so we will need to know the speed of sound. The sound speed will be denoted by the symbol c in this section, since there is little risk of confusion with the speed of light here. For an ideal gas with gas constant R and temperature T, the speed of sound is given by $c^2 = \gamma RT$, where $\gamma = c_p/c_v$ (see Problem **??**). With a little manipulation, Eq. 8.34 can then be re-written to yield a powerful constraint on the circumstances in which a one-dimensional flow can smoothly make the transition from subsonic to supersonic. We start by dividing the equation by ρ and working with the resulting pressure gradient term $\rho^{-1}dp/dr$. Using the definition of potential temperature, this term can be re-written

$$\frac{1}{\rho}\frac{dp}{dr} = \frac{p}{\rho}\frac{d\ln(p/p_o)}{dr} = c^2\frac{d\ln(\theta)}{dr} - c^2\frac{d\ln(\rho)}{dr} \tag{8.36}$$

Next, we need to use mass conservation to eliminate ρ. Specifically, we take the log of $Phi = \rho w A(r)/A(r_s)$, then take the derivative with respect to r, and solve for $d\ln(\rho)/dr$. Upon substitution of the result into the momentum equation and rearranging terms we find

$$(1 - \frac{c^2}{w^2})w\frac{dw}{dr} = c^2\frac{d\ln(A/\theta)}{dr} - g_s\frac{r_s^2}{r^2} \tag{8.37}$$

The ratio w/c is the *Mach number*, for which we will use the symbol M. Eq. 8.37, called the *transonic rule*, implies that the right hand side must vanish at the point where $M = 1$, if we require that dw/dr be finite there. This relation is valid even in the presence of diabatic heating due to radiation, thermal diffusion or any other means; diabatic heating causes θ to vary with r,

and the entropy equation needs to be used to obtain this gradient in terms of the heating rate. The point where $M = 1$ is called the *sonic point*, or sometimes the *critical point*.

If gravity is set to zero, Eq. 8.37 also constrains the flow of fluid in a tube with cross section area $A(r)$, with r being the distance along the axis of the tube (see Problem **??**). In that context, it implies that if one wants to create an adiabatic supersonic jet by feeding subsonic flow into one end of a tube, then things must be arranged so that the sonic point occurs at a constriction of the tube where the area has a local minimum. In particular, one cannot make a supersonic nozzle in the intuitive shape of a cone with the point snipped off. This realization was the basis of the design of the de Laval nozzle [5].

If the heating vanishes in the vicinity of the transonic point, θ is constant there and the transonic condition becomes

$$c^2 = \frac{1}{2} g_s \frac{r_s^2}{r} = \frac{1}{4} w_{esc}^2(r) \tag{8.38}$$

where w_{esc} is the escape velocity at radius r. Thus, the transonic rule states that at the point of transition between subsonic and supersonic flow, the speed of sound must be half the escape velocity. Using the expression for c^2, this condition determines the temperature at the sonic point once the position is given. Except for small bodies with low surface gravity, the sonic point must be very far out from the planet if one is to avoid temperatures temperatures far higher than are likely to be sustainable given the supply of energy. For example, with H_2 on Earth the sonic point temperature would be over $4800K$ if it were placed at 1.1 Earth radii from the planet's center. At 30 radii out, where the gravity is weaker, the sonic point temperature falls to $177K$. For heavier molecules, the temperatures are much higher. For N_2 the sonic point temperature would be $2500K$ even at 30 Earth radii. Numbers for Venus would be similar. Because it is difficult to sustain such high temperatures, hydrodynamic escape of gases much heavier than H_2 from Earth- or Venus-sized bodies is not likely, except perhaps for planets in orbits where they receive much more radiation from the primary star than does Earth or Venus. For smaller bodies, escape of heavier gases starts to come more within the realm of possibility; for N_2 on Mars, the sonic point temperature is $503K$ at 30 Mars radii, and on Titan it is $139K$ at 30 Titan radii. Before long, we will learn how to compute how much absorbed solar radiation is necessary to sustain such temperatures.

The next step is to derive an energy equation, which we do by rewriting the pressure gradient term in the momentum equation in a different form from that used in the preceding discussion. The first law of thermodynamics states that $c_p dT - \rho^{-1} dp = \delta Q$, where δQ is the heat added per unit mass, as defined in Chapter 2. If we divide by dr then the first law can be used to rewrite the pressure gradient term $\rho^{-1} dp/dr$ in the momentum equation. Assuming c_p to be constant, the result can be put into the form

$$\frac{d}{dr}\left[\frac{1}{2} w^2 + c_p T - \frac{1}{2} 2 g_s r_s \frac{r_s}{r}\right] = \frac{\delta Q}{dr} \tag{8.39}$$

Note that $2 g_s r_s$ is the square of the escape velocity from the surface, and when multiplied by r_s/r it becomes the square of the escape velocity from radius r. Let us defer for the moment the business of making sense of the term $\delta Q/dr$ – which is a bit of a mathematical monstrosity – and

[5] Gustaf Patrick de Laval (1845-1913) was a Swedish inventor who developed the de Laval nozzle as a way of making a more powerful steam turbine. Subsequent developments led to rotary separation of oil and water, which found even greater commercial applications in the dairy industry, to the problem of cream separation. Centrifugal cream separators were the mainstay of his company, Alfa Laval, which exists to this day. de Laval also invented the first commercially viable milking machine. The de Laval nozzle was first used for rocketry by Robert Goddard, and makes a cameo appearance in Homer Hickam's book, *Rocket Boys* (which became *October Sky* when it reached the silver screen).

explore the adiabatic case, for which $\delta Q = 0$. In this case, the expression

$$E = \frac{1}{2}w^2 + c_pT - \frac{1}{2}2g_sr_s\frac{r_s}{r} \tag{8.40}$$

must be independent of r. E is the energy per unit mass of the fluid, the three terms representing kinetic energy, internal thermal energy, and gravitational potential energy. Note that when the kinetic energy is negligible and the gravitational term is expanded about r_s by writing $r = r_s + Z$ with $Z/r_s \ll 1$, E reduces to the dry static energy $c_pT + gZ$ derived in Chapter 2. The neglect of w is in fact why this quantity is called *static* energy.

Eq. 8.40 is much like the energy balance we used to determine the escape velocity, except this time the thermal energy c_pT is also in play. The energy must be positive at infinity for escape to be possible, so the threshold for escape is defined by $E = 0$; evaluating this at $r \approx r_s$ and assuming w to be small there, we find that escape is possible when $c_pT > g_sr_s$ near r_s. For atmospheres having temperature low enough that Jeans escape is small, the thermal term is negligible compared to the gravitational term, though, since it can be easily shown that c_pT/g_sr_s is the same order of magnitude as the escape parameter λ_c defined earlier.

Exercise 8.7.5 Write down the relation between c_pT/g_sr_s and λ_c.

The large r behavior is important, since it provides our boundary condition where the atmosphere meets the near-vacuum of outer space. The combination of mass conservation and energy conservation yields two possible kinds of large-r behavior in the adiabatic case. Since $\rho w r^2$ is constant, then if w remains finite at large r, $\rho \to 0$ there. If the far-field flow is adiabatic, then ρ is proportional to $(p/p_o)^{1/gamma}$, and so vanishing density implies vanishing pressure, which turn (by the formula for potential temperature) implies that $T \to 0$ at large r. Since w is finite but T gets small, the speed of sound approaches zero and the flow is supersonic at infinity. Thus, the branch that is supersonic at infinity has vanishing pressure, density and temperature at infinity, and patches smoothly to outer space. Moreover, since the flow is supersonic, information cannot travel upstream back toward the surface of the planet, so conditions at infinity do not affect conditions lower down in the escaping atmosphere. So far we haven't used the energy conservation equation directly, but only the adiabatic assumption; the supersonic conditions and vanishing of temperature and pressure at infinity can survive the addition of a moderate amount of heating in the exterior region. If there is no heating there, then using Eq. 8.40 energy conservation tells us that if w is nonzero at infinity, it in fact becomes *constant* there. At large r, both the temperature and gravitational potential energy vanish, meaning that all the energy of the lower atmosphere is converted to kinetic energy of the escaping flow.

On the other hand, w may vanish at large r. In this case, Eq. 8.40 immediately implies that T asymptotes to a constant at large r, whence density and pressure also asymptote to constants. Finite temperature with vanishing velocity implies that this solution is subsonic at infinity. This is not a viable steady state, because it requires that the interplanetary medium exert a back-pressure on the atmosphere to hold it in. The solution also has the unphysical property of filling interplanetary space with gas having a finite density. What happens if we try to set up a subsonically escaping atmosphere? Note that for subsonic flow, information can propagate upstream towards the planet's surface. Therefore, if there is no back-pressure to hold back the flow, it is likely that the escaping outer region of the atmosphere will accelerate to supersonic conditions, while sending a signal upstream that modifies the upstream boundary condition in such a fashion that the transonic rule is satisfied.

Adiabatic Escape

For an escaping atmosphere with nonzero outward mass flux, it is therefore generally believed that the solution must be supersonic at large r in order to be physically realizable. If the atmosphere starts with small w at the base of the escaping flow (as is generally required by the large density there), then the relatively high temperature at the base yields a large sound speed, implying that the lower atmosphere is subsonic. *Therefore, an escaping atmosphere will have a sonic point, where the transonic rule will need to apply.* This allows us to compute the temperature at the base of an adiabatic escaping atmosphere. There is no exobase for atmospheres undergoing hydrodynamic escape, but we will take the base to be at a position where a single gas, light enough to escape hydrodynamically, dominates the atmospheric composition. This position, which we'll denote by r_b, could be at the planet's surface in some cases, but more typically would be somewhat above the homopause. In any event, it is generally not terribly far above the planet's surface, as compared to the planet's radius, so r_b/r_s is typically of order unity. We equate the energy at r_b (assuming kinetic energy $\frac{1}{2}w^2$ to be negligible there) to the energy at the critical point r_c as follows:

$$\begin{aligned} c_p T_b - \frac{1}{2} 2 g_s r_s \frac{r_s}{r_b} &= \frac{1}{2} c(T_c)^2 + c_p T_c - \frac{1}{2} 2 g_s r_s \frac{r_s}{r_c} \\ &= \frac{5-3\gamma}{4(\gamma-1)} \frac{1}{2} 2 g_s r_s \frac{r_s}{r_c} \end{aligned} \tag{8.41}$$

The second equality makes use of the transonic rule in order to eliminate the temperature and sound speed at the critical (i.e. sonic) point. This determines T_b in terms of the critical point position, the size and gravity of the planet, and the characteristics of the gas. It is a remarkable fact that for a gas made of spherical atoms with no internal degrees of freedom, the base temperature T_b becomes independent of the critical point position, since $\gamma = \frac{5}{3}$ for such gases. This is particularly important given that hydrodynamic escape of atomic hydrogen is of primary interest.

For more complex gases, T_b decreases gently as the sonic point is moved outward, approaching a limiting value at large distances. Assuming $r_s/r_c \ll 1$, and $rs/r_b \approx 1$, then $T_b \approx g_s r_s/c_p$. Note that this limiting basal temperature is precisely the same threshold temperature for escape which we computed earlier on the basis of simple energetic grounds. Eq. 8.41 becomes invalid if we move the critical point all the way to r_b, since we assumed the kinetic energy to be negligible at low levels when deriving the equation. In the limit where the critical point approaches the base, we get the base temperature directly from the transonic rule, which tells us $c^2 = \gamma R T_b = \frac{1}{2} g_s r_s (r_s/r_b)$, or $T_b = \frac{1}{2}(\gamma-1)^{-1} g_s r_s/c_p$ if $r_s/r_b \approx 1$. Hence, the maximum base temperature exceeds the minimum by a factor of $\frac{1}{2}(\gamma-1)^{-1}$, which is 1.3 for H_2, 1.25 for N_2 and 1.7 for CO_2. However, the high end of the temperature range where the critical point approaches the base is of little physical relevance, since it corresponds to the case where the lower atmosphere has somehow directly been given enough kinetic energy to escape.

Except for small bodies, the threshold temperature for escape is very high: over $3000K$ for atomic hydrogen on Earth, or $4386K$ for H_2 – rising to an outrageous $60,000K$ for N_2. The required temperatures are high simply because the potential temperature is constant, so that the atmosphere is expanding and cooling along the dry adiabat; in order to still be hot enough to match the temperature at the sonic point, it must thus start with an exceedingly high temperature. On Titan,where gravity is weaker adiabatic escape of atomic hydrogen requires a basal temperature of only $167K$, while H_2 requires $244K$ and N_2 requires $3352K$. For the Moon, escape of H_2 would require a minimum temperature of $198K$; this is somewhat cooler than the Titan case because the greater surface gravity of the Moon is more than made up for by its smaller radius. Adiabatic hydrodynamic escape of hydrogen from Titan or the Moon seems within the realm of possibility,

but for larger,denser bodies it would be possible only in exotic high energy conditions. In fact, high-temperature adiabatic escape is much more relevant as a theory for the solar wind than it is for escape of most planetary atmospheres. Indeed, the fluid dynamics we have outlined above was first developed to explain the solar wind, and only later adapted for planetary purposes. For Earth or Mars sized bodies, it is conceivable that the required temperatures could be attained for H or maybe H_2 in the aftermath of a giant impact, or perhaps while a magma ocean is still extant. In general, though, hydrodynamic escape of planetary atmospheres must be sustained by absorption of solar (or stellar) radiation at lower altitudes, which offsets the adiabatic cooling of the expanding atmosphere and allows the sonic condition to be met without unreasonably high temperatures at the base. The absorption acts like the boost phase of a rocket, launching the atmosphere toward escape velocity.

At this point we encounter a troubling difficulty regarding the lower boundary condition: for adiabatic hydrodynamic escape, we have little or no ability to control the temperature at the base of the escaping atmosphere, since the basal temperature is nearly (or completely) insensitive to the position of the sonic point. The density (or equivalently, the pressure) at the sonic point is a free parameter, but its value affects only the density at the base, not the temperature. We have already shown that when the atmosphere is too cold, then it runs out of energy if it tries to flow outward, so that the system must instead settle into the hydrostatically balanced solution with no outflow. But what happens if the atmosphere is too *hot*? What goes wrong when the lower atmosphere is too hot is that the entire atmosphere is too hot to have a sonic point, so it is either supersonic everywhere or subsonic everywhere. The first solution violates the condition that the velocity in the lower atmosphere start off small, while the second solution fails to meet the customary boundary condition at infinity. Yet, it is implausible that an atmosphere with a somewhat lower temperature that is happily escaping will suddenly lose its ability to escape if more energy is added to the system by increasing the temperature. Recognizing that in subsonic flow, the conditions at infinity can send a signal upstream modifying conditions near the planet's surface, it seems most likely that in the too-hot case, a supersonic flow will develop at infinity, which will reduce the air temperature near the planet's surface in such a way as to allow the transonic rule to be satisfied. This situation does not seem to have been explored by numerical simulation at the time of writing, however.

To illustrate how the calculation of escape flux works, let's consider adiabatic blowoff of a pure H_2 atmosphere from the Moon, assuming the surface pressure to be 1 bar. In this case, we assume the atmosphere is heated from below by solar absorption at the lunar surface, idealizing the atmosphere as being transparent to solar radiation. The surface temperature is just the equilibrium no-atmosphere blackbody temperature, since H_2 nearly transparent to infrared at these pressures. Under these conditions, we can take the base of the escaping atmosphere to be right at the surface. Now, Eq. 8.41 tightly constrains the allowable range of surface temperatures for an escaping atmosphere. Since w was assumed small at the base in the derivation of Eq. 8.41, the formula becomes invalid if we move the sonic point all the way to the surface. However, if we put it fairly close, at $r/r_s = 1.5$, then $T_b = 270.5K$, while the transonic rule tells us that ths sonic point temperature is $T_c = 163K$. As T_b approaches the minimum temperature of $197.7K$, the sonic point moves to infinity and T_c falls to zero. We'll assume the Moon to be in an orbit for which the surface temperature lies in this range; interestingly, the position of the actual Moon would satisfy this condition if the Moon had enough atmosphere to even out temperature fluctuations. Now, to determine the escape rate, we need the density at the sonic point. Since potential temperature is constant, the temperature ratio T_c/T_b determines the pressure at the sonic point in terms of the surface pressure, and from this and the temperature we get the required density, which is $.026kg/m^3$ when $T_b = 270.5K$ and the sonic point is at $r_c/r_s = 1.5$, or $.00064kg/m^3$ when $T_b = 208.7K$ and the sonic point is at $r_c/r_s = 10$. The velocity at the sonic point is just the sound speed there

(known because we know the temperature), and the mass flux per unit planetary surface area is $\Phi = \rho_c(\gamma R T_c)^{\frac{1}{2}}(r_c/r_s)^2$. This is $56.8 kg/m^2 s$ for $T_b = 270.5K$ and $24.0 kg/m^2 s$ for $T_b = 208.6K$. Note that for any given r_c, the critical point density ρ_c is proportional to the surface pressure, so changing the surface pressure just changes the escape flux proportionately. These are very large escape fluxes. The mass of the hypothetical lunar atmosphere is about $62,000 kg/m^2$, and would be lost in just over a half hour in the hotter case and just over an hour in the colder case. We have thus learned that for a Moon-sized body in an orbit where the solar radiation is similar to that received at Earth's orbit today, H_2 would be lost by adiabatic blowoff almost at once. As the temperature decreases toward the threshold temperature, then the sonic point moves out to infinity, and the escape flux very gradually reduces; in fact, it can be easily shown that the escape flux approaches zero as the surface temperature approaches the threshold temperature. Hydrodynamic escape is a process with dramatic thresholds: for a surface temperature of $208K$ the atmosphere is lost in a matter of hours, whereas the escape flow shuts off completely when the temperature drops below $198K$ (though other escape mechanisms may still cause loss of the atmosphere).

Exercise 8.7.6 Following the reasoning in the preceding paragraph, for an arbitrary body derive a formula for the escape flux Φ in terms of the surface temperature and surface density. Show that the escape flux is proportional to surface density, and that the escape flux approaches zero as the surface temperature approaches the minimum surface temperature for adiabatic hydrodynamic escape.

Re-interpretation of the transonic rule in terms of energetics

We can gain a better appreciation of the meaning of the transonic rule, and of what happens when it is violated, by expressing the energy conservation relation in terms of the Mach number. This requires expressing c^2 in terms of the Mach number, or equivalently writing T in terms of the Mach number, which can be done by making use of mass conservation, the ideal gas equation of state, and the expression for potential temperature. In particular, if we let (p_o, T_o, ρ_o, M_o) be the state of the atmosphere at some point r_o, then the four relations

$$T = \theta \cdot (\frac{p}{p_o})^{R/c_p}, \rho = \rho_o \frac{T_o}{\theta}(\frac{p}{p_o})^{1/\gamma}, \Phi = \rho c M \frac{r^2}{r_s^2} = \rho_o c_o M_o \frac{r_o^2}{r_s^2}, \frac{c}{c_o} = (\frac{T}{T_o})^{\frac{1}{2}} \tag{8.42}$$

can be solved for temperature, yielding

$$T = \theta \cdot \left[(\frac{\theta}{T_o})^{\frac{1}{2}} \frac{r_o^2}{r^2} \frac{M_o}{M}\right]^{\beta} \tag{8.43}$$

where $\beta \equiv 2(\gamma - 1)/(\gamma + 1)$. This result is valid whether or not the flow is adiabatic. In the adiabatic case, θ is independent of r and $\theta = T_o$ if we define the potential temperature with regard to the pressure p_o at the reference point. With the relation 8.43 in hand, Eq. 8.40 can be written

$$\begin{aligned} E &= c^2 \cdot (\frac{1}{2}M^2 + \frac{1}{\gamma - 1}) - \frac{1}{2} 2 g_s r_s \frac{r_s}{r} \\ &= \gamma R \theta \cdot [(\frac{\theta}{T_o})^{\frac{1}{2}} \frac{r_o^2}{r^2} \frac{M_o}{M}]^{\beta} \cdot (\frac{1}{2}M^2 + \frac{1}{\gamma - 1}) - \frac{1}{2} 2 g_s r_s \frac{r_s}{r} \end{aligned} \tag{8.44}$$

The right hand side can be considered to be a function $E(M, r/r_s)$ with $2 g_s r_s$ and conditions at r_o as parameters. This expression is valid even in the presence of heating, in which case θ is a function of r. For the sake of generality, we have written the expression in terms of an arbitrary reference

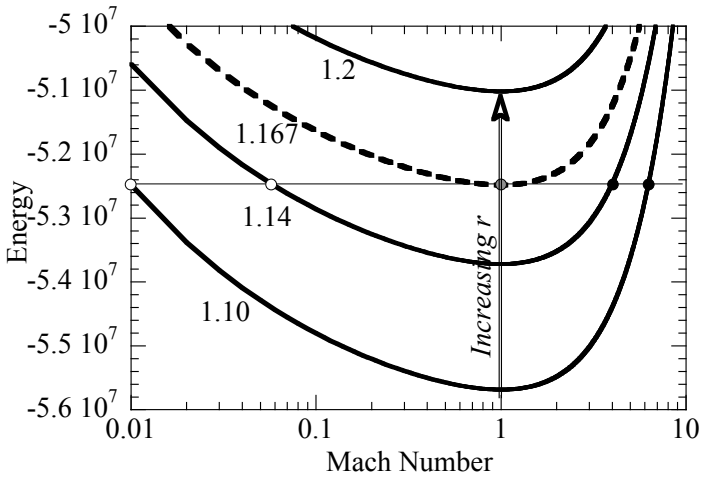

Figure 8.5: **CAPTION

point r_o, but most commonly we will take r_o to be a sonic point, at which $M_o = 1$ by definition and at which moreover the transonic rule is satisfied. The transonic rule gives T_o in terms of r_o and the gravitational acceleration, so in this case the only free parameters governing the shape of the curve are r_o (which we'll call r_c in this case) and θ. When the flow is moreover adiabatic throughout the escaping atmosphere, $\theta = T_o$ and the curve is governed by a single parameter, for any given planet.

Recall from Chapter 2 that γ is determined by the number of excited degrees of freedom of the atoms or molecules making up the gas. For spherical particles with no internal degrees of freedom, $\gamma = 5/3$, while $\gamma \to 1$ for complex molecules with very many excited degrees of freedom, though for most atmospheric gases $\gamma > 1.29$. Hence, $0 < \beta \leq \frac{1}{2}$, with the lower limit not being very closely approached in practice. Because $\beta < 2$, it follows that $E(M, r/r_s) \to \infty$ as $M \to \infty$. In addition, since β is positive, it follows that $E(M, r/r_s) \to \infty$ as $M \to 0$. For fixed r/r_s, the function has a single minimum at $M = 1$, as can be verified by differentiation of the expression with respect to M. It is the shape of the energy curve, and the occurrence of the minimum at $M = 1$, that underlies the transonic rule.

Exercise 8.7.7 Carry out the suggested differentiation, verify that the minimum of E occurs at $M = 1$, and write down the expression for $E_{min}(r/r_s)$.

Some representative energy curves are sketched in Fig. 8.5. For any given energy E_o, there are three possible situations regarding the flows that satisfy $E(M, r/r_s) = E_o$:

- There can be a pair of solutions, one of which is subsonic and the other supersonic.
- There can be a single solution, with $M = 1$
- There can be no solutions at all.

Now let's consider what happens if we start with a subsonic solution and track it's evolution as we increase r, which changes the energy curve. The gravitational potential term shifts the curve

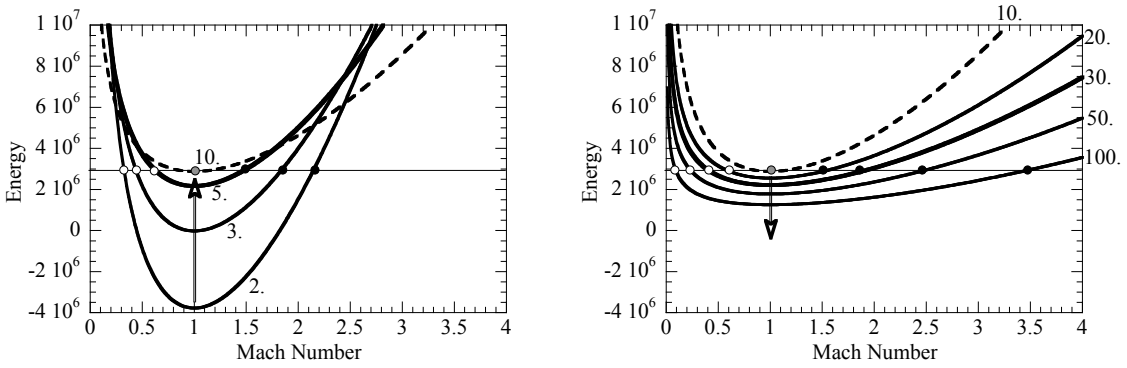

Figure 8.6: Sequence of energy curves for the situation in which the transonic rule is satisfied for $r/r_s = 10$. Left panel: The energy curves move upward as r/r_s is increased from 2 to 10. Right panel: The energy curves move downward as r/r_s is increased further from 10 to 100.

upward without changing its shape, while the r-dependent factor in the first term of Eq. 8.44 flattens the curve and moves it downward as r is increased. The situation is depicted in Figure 8.5, where we start with $T = 300K$ and $M = .01$ at $r/r_s = 1.1$. We already know from our previous results that this temperature is too low to allow the transonic rule to be satisfied for Earthlike gravity, so what happens as r is increased? Examining the intersection points between the initial energy and the energy curve, we see that the Mach number increases until it reaches the sonic point at $r/r_s = 1.167$. When r is further increased, the energy curve continues to move upward, however, so there are no solutions compatible with the initial energy. This is the generic behavior when we start from too low a temperature and the transonic rule is not satisfied. The atmosphere can still approach $M = 1$ from either the subsonic or supersonic side, but the upward movement of the energy curve through the sonic point means that the gradient of $M(r)$ and $w(r)$ have square-root singularities there – a consequence of the violation of the transonic rule. From a physical standpoint, what is important is not so much the singularity as the fact that solutions cease to exist altogether once r is moved past the sonic point.

An examination of Eq. 8.43 shows that the temperature decreases as the sonic point is approached; this increases the Mach number by decreasing the speed of sound. In fact, where the Mach number is small enough that kinetic energy is negligible, the energy equation reduces to conservation of dry static energy $c_pT + gz$, and tells us that the atmosphere ascends along the dry adiabat. This means it can't ascend far, because the temperature falls to zero at a finite height for the dry adiabat. Incorporation of the kinetic energy term causes the atmosphere to run out of energy before the temperature falls to zero. This is the situation we are fighting if we try to make an atmosphere hydrodynamically escape while starting from a realistic base temperature: without some additional supply of energy, the atmosphere runs out of energy before it gets very far. In this case, energetics requires the atmosphere to settle into a state of rest without a mean outflow. When starting from a low temperature on a planet with Earthlike gravity, energy must be deposited in the upper atmosphere in order to allow it to escape. Generally, this energy is supplied in the form of extreme ultraviolet.

Next we'll examine the geometry of the energy curves in a case where the transonic rule is satisfied and adiabatic hydrodynamic escape is possible. The situation is depicted in Fig. 8.6. To pass from a subsonic to a supersonic state, the energy curve first comes up to where the subsonic and supersonic solutions coalesce at the sonic point, but then moves down again as r is

further increased, allowing the solution to continue on the supersonic branch. The transonic rule is equivalent to the requirement that when expressed as a function of r, the minimum of the energy curve $E(M = 1, r/r_o)$ has a maximum at $r = r_c$, so that the curve first goes up to the sonic point then turns around and heads back downwards again.

Exercise 8.7.8 Verify this property by using the expression for $E(M = 1, r/r_o)$ and locating its minimum by taking the derivative with respect to r. To keep things simple, you can restrict attention to the adiabatic case $\theta = const.$.

Escape driven by *EUV* heating

It is now time to bring heating into the picture. Heating alters the preceding construction in two ways: Firstly, instead of the expression E being independent of r, it will vary on account of the deposition of energy from solar radiation or other sources. This moves the energy curve up or down with r, without changing its shape. Secondly, heating will cause θ to vary with r, which changes the shape of the energy curve. It remains true, however, that the energy curve for any r has a unique minimum where the Mach number is unity, and that in order to effect the transition from subsonic to supersonic flow, the line defining the available amount of energy at any given r must start above the minimum, be brought to tangency at the minimum (either by moving the curve upward or moving the line downward or some combination of the two), and then moving the line defining available energy to some distance above the minimum as r is increased further.

In order to incorporate heating into the energy equation, we rewrite the diabatic term on the right hand side of Eq. 8.39 as follows:

$$\frac{\delta Q}{dr} = \frac{\rho \delta Q/dt}{\rho dr/dt} = \frac{\dot{q}}{\rho w} = \frac{(r/r_s)^2 \dot{q}}{\Phi} \tag{8.45}$$

where $\dot{q}$ is the heating rate *per unit volume*. Now let $F(r)$ be the flux of energy due to means other than the bulk fluid flow, with the convention that inward fluxes are taken as positive. We will mostly deal with radiative flux, but F could equally well represent flux due to molecular diffusion of heat. In terms of the flux, the heating rate is given by $\dot{q} = r^{-2} d(r^2 F)/dr$, so given that Φ is constant the heating term in the energy equation becomes

$$\frac{\delta Q}{dr} = \frac{d}{dr}[\frac{r^2}{r_s^2}\frac{F}{\Phi}] \tag{8.46}$$

Since this is the gradient of a flux, Eq. 8.39 tells us that it can be combined with the previous expression for E to yield the revised conservation law

$$\frac{1}{2}w^2 + c_p T - \frac{1}{2} 2 g_s r_s \frac{r_s}{r} - \frac{r^2}{r_s^2}\frac{F}{\Phi} = E_o \tag{8.47}$$

where E_o is a constant. Note that F/Φ is an energy flux divided by a mass flux, and therefore has dimensions of velocity squared. For a given variation in energy flux, reducing the mass flux Φ leads to a greater radial variation in the energy density E (the first pair of terms in the equation), because fluid has more time to accumulate energy when it is moving slowly.

The treatment of the radiative part of the heating is a bit tricky. As in our earlier calculations of planetary temperature, we are doing a globally averaged energy budget assuming uniform atmospheric conditions over the globe. That requires figuring the amount of *EUV* flux intercepted by the *planet's atmosphere*, and distributing it uniformly over the sphere. The assumption of

effective redistribution of solar heating over the outer atmosphere is highly questionable, but it is quite customary in hydrodynamic escape calculations and is in any event the only approximation which allows us to make headway without very extensive and complex numerical simulations. That notwithstanding, there is an additional issue that did not arise in our earlier treatments of globally averaged energy budgets. As usual, the planet intercepts a disk of light of some radius, and the intercepted flux is spread uniformly over the surface of sphere of the corresponding radius. The difference in the present case is that the portion of the atmosphere that is dense enough to absorb significant amounts of EUV radiation can extend many planetary radii out from the surface. Therefore, the radius of the disk of intercepted radiation depends on the optical thickness (hence density) of the atmosphere. Let's suppose that the atmosphere is transparent to EUV for radii farther out than some radius r_{abs}, but that the closer-in atmosphere absorbs strongly. Then, if $F_\circledast$ is the incoming flux of EUV from the sun (analogous to the solar constant), the intercepted power is $\pi r_{abs}^2 F_\circledast$. If we want this to be the total power entering the system, then the corresponding uniform *radial* flux $4\pi r^2 F(r)$ through a shell of radius r must be held constant at the value $\pi r_{abs}^2 F_\circledast$ for $r > r_{abs}$. Therefore, we model radiative absorption by stipulating that $r^2 F(r) = \frac{1}{4} r_{abs}^2 F_\circledast$ for $r > r_{abs}$ while allowing the flux to decay to zero as radiation is absorbed deeper in the atmosphere.

Eq. 8.47 imposes a powerful constraint on the mass flux that can be sustained by a given level of heating. We'll suppose as usual that w is small at the base of the escaping atmosphere, but this time we'll also assume that the base is cool, so that $c_p T$ is negligible compared to the gravitational term at the base. This is the typical situation, in which the atmosphere is strongly gravitationally bound and the escape parameter λ_c is small. Unlike the adiabatic escape case, we do not endow the base of the atmosphere with enough thermal energy to sustain the outflow, but rather deposit it gradually through absorption of stellar flux, generally in the extreme ultraviolet spectrum. Equating energy at the base (where radiative flux falls to zero) and at infinity yields the relation

$$\frac{1}{\Phi}\left(\frac{1}{4}\frac{r_{abs}^2}{r_s^2} F_\circledast\right) = \frac{1}{2} w_\infty^2 + \frac{1}{2} 2 g_s r_s \frac{r_s}{r_b} \tag{8.48}$$

Since $w_\infty^2 > 0$, this imposes an upper bound on the mass flux Φ for any given amount of radiative absorption. Note that this energetic constraint survives the addition of heat diffusion to the flux term, since diffusion only redistributes energy in the vertical and does not add new energy to the system. If $(r_{abs}/r_s)^2 \approx 1$ and $r_s/r_b \approx 1$ as is typically the case, than the constraint is simply $\Phi \leq \frac{1}{4} F_\circledast / g_s r_s$. It is important to recognize that this bound on the escape flux applies only in the low temperature limit, in which $c_p T$ at the base is negligible compared to the gravitational potential. For any finite temperature, the escape flux can exceed the limiting flux, by an amount that increases with temperature. As the base temperature approaches the temperature at which adiabatic escape becomes possible, the escape flux can become arbitrarily large, limited only by the density at the base.

The physical content of the constraint is simple: The escaping atmosphere carries kinetic and potential energy with it, and this outward energy flux must be matched by the supply of radiant energy absorbed within the escaping atmosphere. The amount of energy flux escaping due to mass flow is negligible from a standpoint of planetary energy balance; that energy loss is still by far dominated by infrared emission. However, from the standpoint of the energy budget of the outer atmosphere *alone*, the energy loss due to mass outflow can be the dominant term – at least for gases like H_2 which are poor infrared emitters. Infrared emission, to the extent that it occurs at all, can be thought of as stealing energy from the supply of EUV heating available to sustain escape.

In order to complete the solution, we need to know how the potential temperature varies with radius. Once the potential temperature is known, we have enough thermodynamic information

to compute the profiles of pressure and density as well. The radial variation of the potential temperature is obtained from the entropy equation:

$$\frac{d}{dr}c_p \ln\theta = \frac{1}{T}\frac{\delta q}{dr} = \frac{1}{T}\frac{\rho\delta q/dt}{\rho w} = \frac{1}{T}\frac{(r/r_s)^2\dot{Q}}{\Phi} \tag{8.49}$$

The heating term $(r/r_s)^2\dot{Q}$ can be written as the radial gradient of a flux as before, but because of the factor $1/T$ appearing in the entropy equation, this equation cannot be integrated to yield a pointwise relation between entropy and flux the way we did for energy. The entropy change between a point r_A and r_B depends on the shape of the heating curve between those points, and not just the amount of heat added; heat added at low temperature has a greater effect on entropy than heat added at high temperature.

We'll now exhibit some numerical solutions for the case in which the heating $\dot{Q}$ is a known function of r. Radiative heating depends on the density and temperature so strictly speaking it must be solved for together with the atmospheric structure; the extension to this case is straightforward once one understands how to solve the simpler problem. Given the heating, the numerical solution proceeds as follows. One starts by choosing the critical point position r_c/r_s, from which one can compute the critical point sound speed and temperature. Then, one integrates the differential equation 8.49 in a direction toward the planet. Since the escape flux Φ appears as a parameter in this equation, one must guess a value of Φ to carry out the integration. The chosen value of Φ also fixes the density at the critical point since the velocity at the critical point is the local speed of sound. At each step of the integration of Eq. 8.49, one obtains value of $\ln\theta$, but to proceed one also needs to update the value of T. This is done by solving the energy equation, Eq. 8.47 (rewritten in terms of Mach number using Eq. 8.44), for the Mach number, following the subsonic branch. The Mach number, in turn, determines the new temperature and allows the integration to proceed further. As the integration proceeds, a point may be encountered in which solutions to the energy equation no longer exist, in which case the chosen value of Φ is not realizable. If this situation doesnt arise, the integration is continued until the base r_b is reached. One now knows the value of θ there, which for positive heating will be less (usually *much* less) than the value at the critical point. The integration has already completely determined the temperature structure of the atmosphere, and he ratio of potential temperature at r_b to the value at r_c determines the proportionality constant between the density at r_b and the (known) density at r_c. This procedure yields a family of solutions, with r_c and Φ as parameters. When Φ is large, Eq. 8.49 says that the potential tempreature becomes constant, in which case we recover the adiabatic escape solutions which typically have very high temperatures at the base. As Φ is made smaller, heating causes the potential temperature at the base to be much smaller than the potential temperature at the critical point, which results in cooler temperatures at the base. When Φ is made too small, however, the temperatures are driven to excessively low values and one encounters at some point a supersonic transition at a radius where the transonic rule is not satisfied, whereupon the solution ceases to exist. Carrying out this procedure requires only the integration of a first order differential equation and the solution of the energy equation at each step using Newton's method. It is quite straightforward to implement.

Before showing a specific family of solutions obtained using the above procedure, it is useful to identify a few relevant nondimensional parameters. The energy-limited escape flux identified earlier provides a convenient scale for nondimensionalization of Φ. Let's call the limiting flux Φ_*. It can be written $\frac{1}{2}F_\circledast/w_{esc}^2$, where w_{esc} is the escape velocity from the surface, namely $\sqrt{2g_s r_s}$. We can define a characteristic temperature T_* such that $c_p T_* = w_{esc}^2$. The heating can be written $\dot{Q} = (d/dr)(r^2/r_s^2 F)$, so it can be nondimensionalized by multiplying both sides of the entropy equation by r_s (amounting to taking the planetary radius as the unit of length) and then writing

$F = F_{\circledast} \cdot (F/F_{\circledast})$ If the temperature is nondimensionalized against T_*, then the entropy equation depends on the *EUV* flux, the escape velocity and the escape flux only through the dimensionless combination Φ/Φ_*. Specifically, the nondimensional entropy equation becomes

$$\frac{d}{d\tilde{r}} \ln \theta = \frac{1}{\tilde{T}} \frac{\tilde{r}^2 \tilde{\dot{Q}}}{\tilde{\Phi}} \tag{8.50}$$

where $\tilde{r} \equiv r/r_s$, $\tilde{\Phi} \equiv \Phi/\Phi_*$, $\tilde{T} \equiv T/T_*$ and $\tilde{\dot{Q}}$ is the nondimensional heating profile. Similarly, if one nondimensionalizes the energy equation by dividing by w_{esc}^2, one finds that the same three quantitites appear in the energy equation only in the form Φ/Φ_*. Thus, for any given *shape* of the *EUV* heating profile, Φ/Φ_* determines the basic behavior of the atmospheric structure; changing the values of any of the terms in this nondimensional parameter just uniformly rescales the temperature and density profiles.

In Figure 8.7 we show a family of solutions for a hydrodynamically escaping H_2 atmosphere on Earth. In this calculation, r_c/r_s is held fixed at 30 while the nondimensional escape flux is varied. The calculations were carried out with $r_b/r_s = 1.1$. The *EUV* heating profile was chosen such that

$$\frac{r^2}{r_s^2}\dot{Q} = \frac{d}{dr}\frac{r^2}{r_s^2}F = Q_o \exp -(\frac{r - r_b}{r_{EUV}}) \tag{8.51}$$

The heating profile determines the flux $(r/r_s)^2 F(r)$, which is needed for use in the energy equation. The flux integration shows that when the heating is shallow, $Q_o \approx \frac{1}{4} F_{\circledast}/r_{EUV}$. The calculations shown in the figure were done with $r_{EUV}/r_s = 1$, yielding a shallow, low-level heating.

Looking at the results, we see that when Φ/Φ_* is made large, the temperature is monotonically decreasing, and at very large escape fluxes the curve looks like the adiabatic escape solution as expected. As $\Phi_*/\Phi \to 1$, however, the temperature at the base falls toward zero, and the temperature rises from cool values to a maximum before decaying approximately adiabatically toward the critical point. At lower values of Φ/Φ_*, solutions fail to exist in the purely radiatively heated case. Although the energy constraint expressed in Eq. 8.48 permits lower escape fluxes when the velocity at infinity is nonzero, the resulting flows cannot be made to satisfy the transonic rule in the configuration under investigation.

The cool-base solutions constitute the desired regime for hydrodynamic escape of hydrogen from Earth or Venus. In these solutions, *EUV* heating takes the hydrogen entering at the cool temperatures of the lower atmosphere, and heats it as it flows outward, to the point where it is hot enough and far enough out to escape. The *EUV* heating region is like the boost phase of a rocket, where burning of fuel launches the rocket and accelerates it to escape velocity.

What we have achieved by introducing radiative heating is the ability to sustain hydrodynamic escape with realistic low level temperatures. For the cool-base solutions, the escape flux is on the order of Φ_*, so we can use this quantity to estimate the significance of the escape flux. For Earth conditions, where $F_{\circledast} \approx .004 W/m^2$, $\Phi_* = 1.6 \cdot 10^{-11} kg/m^2 s$, corresponding to a flux of hydrogen atoms of $10^{16}/m^2 s$. This greatly exceeds the Jeans escape rate we computed earlier, and is more than sufficient to keep up with the volcanic outgassing. For the Venus case, where $F_{\circledast} \approx .008$ owing to the closer orbit, we find a mass flux of $3.72 \cdot 10^{-11} kg/m^2 s$, or an atomic hydrogen escape flux of $2.24 \cdot 10^{16}/m^2 s$. In a billion years, this could get rid of the hydrogen from a $10 km$ deep ocean (of course, one still needs to find something to do with the resulting oxygen). The reader should be cautioned that these numbers should not be taken as definitive; a more sophisticated calculation taking into account heat diffusion and a more realistic model of deposition of radiation could reduce the estimated fluxes. Still we have shown that hydrodynamic escape easily has the potential to get rid of the required amount of hydrogen on Earth or Venus.

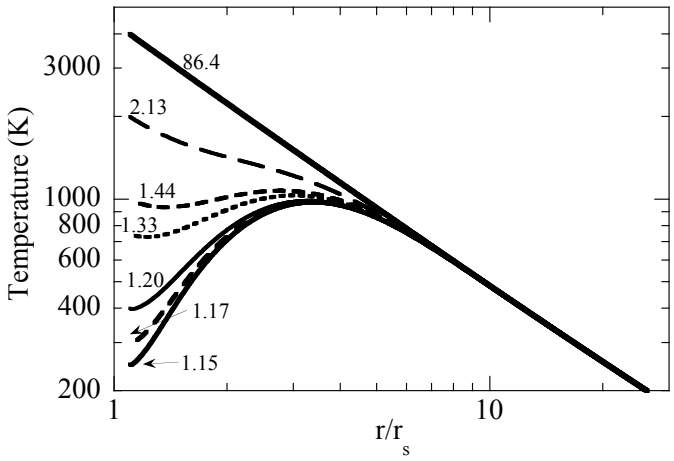

Figure 8.7: Temperature structure of an excaping H_2 Earth atmosphere subject to radiative heating at low levels. The critical point is held fixed at $r/r_s = 30$. The numbers on curves give the value of the nondimensional escape flux, Φ/Φ_*.

In the presence of radiative heating, we can match the desired temperature boundary condition at the base of the escaping atmosphere by varying the escape flux with fixed r_c, but the problem now is that we cannot independently specify the density there. For any given base temperature,this is fixed by the other parameters in the problem. Examination of the density solutions (not shown) reveals that in the family of curves given in Figure 8.7, the density decreases rapidly as the base temperature increases. In dimensional terms, at $250K$ the density is $2.5 \cdot 10^{21}/m^3$, falling to $3.6 \cdot 10^{20}/m^3$ at $300K$, and to $2.3 \cdot 10^{19}/m^3$ at $400K$. The strong temperature dependence arises mainly because the density scale height is smaller for low temperatures, so that the atmosphere has to start with higher density in order to match with the density needed at the critical point. The additional degree of freedom needed to match the density condition at the base while keeping the temperature fixed at some desired value is provided by varying the critical point position r_c. Some numerical experimentation, backed up by a bit of tedious algebra applied to the energy and entropy equations, reveals the following behavior for the density at the base:

- Except for the case of an ideal monatomic gas with $\gamma = \frac{5}{3}$, the density at the base can be made arbitrarily large for fixed base temperature by moving r_c outward; in fact, for large r_c/r_s the base density increases linearly with r_c, with a slope that increases as the base temperature increases.

- For any gas, the base density can be made small by moving r_c toward the surface, though for ideal monatomic gases the decrease in density only becomes effective as r_c moves into the heating region. However, if the base density one is trying to match gets too small, then the required r_c approaches the surface, meaning that the atmosphere escapes only because it is supersonic already at low levels – i.e. that it escapes by virtue of having been given high kinetic energy at low levels. These solutions are of little physical interest, as they do not patch to an atmosphere in hydrostatic balance at low altitudes.

- For ideal monatomic gases, solutions satisfying the transonic rule cease to exist once r_c is moved moderately outside the heating region. This feature is connected with the insensitivity

of T_b to the critical point position, expressed in Eq. 8.41 (applied to the base of the *adiabatic* region rather than all the way to the base of the heating region). As a result, monatomic gases appear to be subject to a *maximum* allowable base density. The implications of this peculiarity for the important case of monatomic hydrogen escape seem not to have been remarked on or explored in the literature. We speculate that when the base density is "too large," the flow tries to become subsonic out to infinity, whereafter upstream influence in some way acts to alter conditions at the base. Because of the way the EUV heating affects the transonic condition, the behavior of density in the monatomic case is likely to be highly sensitive to the radial distribution of the heating, and in particular to the extent to which some heating persists out to great distances from the planet's surface.

- For low base temperatures, the escape flux remains near the limiting flux Φ_*, and does not increase significantly as the base density increases. Increasing the base density in this regime moves the critical point outward, but does not increase the flux. This behavior contrasts with both the high-temperature regime and with Jeans escape, and comes about because the escape flux is constrained by the energy supplied by EUV heating.

Our understanding of hydrodynamic escape so far can be summarized in the following recipe for estimating the escape flux due to this mechanism: First compute the characteristic temperature T_*, defined by $c_p T_* = 2 g_s r_s$. Is the temperature at the base comparable to or greater than this value? If so, you can use the adiabatic escape formulation as an estimate, which yields the proportionality constant between the escape flux and the atmospheric density at the base, which can take on whatever value is dictated by the lower boundary condition. The proportionality constant was computed in Exercise 8.7.6. On the other hand, if the temperature at the base is much less than T_* your atmosphere is in the low temperature limit, and EUV absorption plays a critical role in sustaining escape. In this case there is a threshold base density which must be exceeded in order to permit hydrodynamic escape from a level in the atmosphere which is nearly in hydrostatic balance. The threshold density must be determined by integrating the entropy equation inward from some chosen critical point, and moving the critical point inward until the velocity at the base becomes unacceptably large. This is the only hard part of the calculation. However, once you know that the base density exceeds the threshold, the escape flux in the low temperature regime can be estimated by simply evaluating the limiting flux, $\Phi_* \approx \frac{1}{4} F_\circledast / g_s r_s$, where $F_\circledast$ is the EUV flux impinging on the planet.

A rather startling aspect of the radiatively heated escape calculation is that the limiting escape flux Φ_* is independent of the molecular weight of the substance making up the escaping atmosphere. From this, it would appear that a given amount of absorbed EUV could make a kilogram of oxygen escape just as easily as a kilogram of hydrogen. Continuum hydrodynamics appears to know nothing about the fact that the fluid is made up of discrete particles, so it is not immediately obvious how the molecular weight should enter the problem. In fact, the microscopic nature of the fluid enters through c_p, which, through equipartion of energy, *does* know about the existence of particles. A kilogram of H_2 has more degrees of freedom than a kilogram of O_2, and this is reflectied in a higher value of c_p for H_2. The specific heat enters into the determination of the density at the base, and it is this density boundary condition which tells us that it is essentially impossible for heavy species to escape hydrodynamically, unless the temperature becomes very large. Using dimensional analysis, we can use our H_2 escape calculations to make some inferences about what happens if we try to make a heavier species escape. The value of c_p for atomic oxygen is nearly ten times smaller than the value for H_2, so that the characteristic temperature T_* for oxygen is more than ten times that of H_2. That means that to achieve the same particle density for oxygen as was achieved at $250K$ in the H_2 case, the base temperature has to be raised to over $2500K$. An escape calculation with atomic oxygen shows that one can indeed achieve a temperature of $250K$

at the base by making Φ/Φ_* sufficiently close to the limiting value, but when one does so the base particle density is $10^32/m^3$ and the pressure there is over *three million bars*!. There is not likely to be any planet that can meet such a high density threshold.

Effects of heat diffusion

EUV heating is not the only diabatic effect of importance in the outer atmosphere. The other main heating and cooling effect that needs to be taken into account is the redistribution of heat by diffusion. Heat diffusion is much like diffusion of mass, and becomes strong when the mean free path is large. When the temperature gradient is dT/dr, the flux of heat is $\kappa dT/dr$, where κ is the same kind of thermal conductivity we encountered in earlier chapters, only this time applied to a gas rather than a solid. One of the many important effects of heat diffusion is that it allows heat from deposition of *EUV* at high levels to be diffused down to the lower atmosphere, which is often necessary to sustain hydrodynamic escape. Indeed, the radiative escape calculations we presented above in some sense implicitly assume some diffusion, since it is unlikely that enough *EUV* would penetrate to r_b to allow the fairly strong heating we assumed there in our calculation.

In contrast with radiative heating, heat diffusion fundamentally changes the mathematical character of the problem. With radiative heating only, the energy equation 8.47 is an algebraic equation that is solved for the Mach number. It doesn't involve any derivatives. However, with heat diffusion, the heat flux F appearing in the equation is no longer just a function of r, but instead involves a term $\kappa dT/dr$. Thus, diffusion *changes the energy equation from an algebraic equation for the Mach number to a differential equation for temperature*. This is true no matter how small the thermal conductivity, though the equation will become very ill posed when κ is small. One can also see the singular effect of diffusion by examining the entropy equation. With only radiative heating, the highest order derivative is on the left hand side, and constitutes a first order differential equation for θ. With diffusion, the heat heating term on the right hand side has a contribution from diffusion, which introduces the gradient of $\kappa dT/dr$. The entropy equation thus becomes instead a second order differential equation for T. The extra derivative can provide some additional scope for meeting both a temperature and density boundary condition at the base of the atmosphere, though this advantage is somewhat constrained by the fact that the heat diffusion is weak near the base, where the density is relatively large.

The second order diffusion equation based on the entropy equation can in principle be used as the basis of a solution method in the diffusive case. A more common approach to incorporating heat diffusion proceeds from a variation on the reasoning we used in deriving the transonic rule. Instead of relating the pressure gradient to the potential temperature, as we did in our derivation of 8.37, we can instead write $p = \rho RT$ and use the same manipulations as previously to eliminate $d\rho/dr$ using conservation of mass. In that case, we obtain the following alternate form of Eq. 8.37

$$(1 - \frac{c_T^2}{w^2})w\frac{dw}{dr} = -R\frac{dT}{dr} + c_T^2\frac{d\ln(A)}{dr} - g_s\frac{r_s^2}{r^2} \tag{8.52}$$

where $c_T^2 \equiv RT$. The left hand side looks similar to the previous form, except that the *isothermal* sound speed c_T (Problem **??**) appears in place of the adiabatic sound speed. From this equation one would be tempted to draw the paradoxical conclusion that the critical point should be defined with regard to the isothermal sound speed rather than the adiabatic sound speed. This would be a fallacy. It's true that the right hand side must vanish where w equals the isothermal sound speed, but this does not give us a useful transonic rule since we do not *a priori* know the value of dT/dr. This contrasts with the original form, Eq. 8.37, where we know $d\theta/dr$ from the entropy equation, and in particular that it vanishes if there is no heating near the critical point. But that's

not the end of the story. If the diffusion is nonzero everywhere, then the energy equation, Eq. 8.47 can be solved for dT/dr in terms of w, T, r and the radiative flux; the resulting expression has no derivatives, and so when substituted into Eq. 8.52 it does not change the location of the singular point of the equation where the coefficient of the derivative vanishes. It would appear that the introduction of even infinitesimal heat diffusion discontinuously changes the mathematical character of the problem, so that we are left with a transonic rule that is applied at the critical point defined by isothermal rather than adiabatic sound speed. In the diffusive case, the most common approach to obtaining a steady solution is to simultaneously integrate Eq. 8.52 and Eq. 8.47, which jointly define a coupled set of differential equations for the pair (w, T). The integration is supplemented by a transonic condition applied at the point defined by the isothermal sound speed. This is a technically correct approach to the problem, but as a numerical scheme it is sure to become badly behaved in one way or another when the heat diffusivity becomes small. This may account for the difficulty some researchers have reported in finding consistent solutions to the escape equations in the presence of both heating and diffusion.

There is one special case, however, where one can take a short-cut to avoid the complexities and the problematic features of the integration sketched out above. When the heat diffusion is so strong that it keeps the outer atmosphere isothermal, then dT/dr vanishes, and Eq. 8.52 provides a usable constraint on the conditions at the critical point defined with reference to the isothermal sound speed. The isothermal case is worth pursuing, as it provides the opposite extreme to the no-diffusion case we considered previously, and thus serves to highlight the effects of diffusion. Numerical solutions to the full problem for Earth and Venus have temperature variations that somewhat resemble the cool-base nondiffusive cases shown in Fig. 8.7, so the isothermal limit cannot be relied on for an accurate estimate of the actual escape flux, however.

As for the adiabatic case, the isothermal case admits a very simple explicit solution for the escape flux. When the atmosphere is isothermal, we don't need to invoke the First Law of Thermodynamics to derive a conservation law from the momentum equation. Since T is constant, substituting $p = \rho RT$ immediately yields the result

$$\frac{1}{2}w^2 + c_T^2 \ln \rho - \frac{1}{2} 2 g_s r_s \frac{r_s}{r} = const. \tag{8.53}$$

Note that this energy expression is a constant even though no explicit heating term appears in the equation. Moreover, c_T is a constant as well, because the atmosphere is isothermal. By equating the values of this expression at the base and at the critical point, we find

$$\ln \frac{\rho(r_c)}{\rho(r_b)} = -\frac{1}{2}\frac{2 g_s r_s}{c_T^2}\left(\frac{r_s}{r_b} - \frac{r_s}{r_c}\right) - \frac{1}{2} = -2\left(\frac{r_c}{r_b} - 1\right) - \frac{1}{2} \tag{8.54}$$

As usual, we have assumed that the kinetic energy is small at the base. The second equality arises from application of the isothermal form of the transonic rule. Eq. 8.54 links the density at the base to the density at the critical point once the critical point position is specified. The result implies that $\rho(r_c)/\rho(r_b)$ gets exponentially small as the critical point is removed to infinity. Further, the escape flux is $\Phi = \rho(r_c) c_T (r_c)(r_c/r_s)^2$, so Φ is known once the position of the critical point and the density there are known. So far, the nature of the isothermal solution is rather similar to the adiabatic case, in the sense that the base temperature and base density can be fixed by adjusting r_c and $\rho(r_c)$, whereafter the escape flux is also determined. An important difference with the adiabatic case is that the base temperature can be made quite cool, given that the temperature is uniform and its value is set by the (low) speed of sound at the distant critical point. The quantitative behavior of the escape flux in a few illustrative cases is explored in Problem **??**.

Because diffusion only redistributes heat and is not in itself a source of energy for escape, it would be bizarre if introduction of strong diffusion were to allow an atmosphere to escape at low

temperature without any further conditions being met. Indeed, we are not quite done with this problem. The solution must be completed by making use of the energy constraint linking escape flux to *EUV* heating. For the isothermal case, this comes in via the entropy equation. For an isothermal medium $\ln\theta = -(R/c_p)\ln\rho + const.$ Further, since the temperature is constant, when Eq. 8.49 is integrated over an interval of r, the heating term integrates out to a difference in the fluxes at the endpoints. Putting the two results together, integrating the entropy equation from the base to the critical point, and dividing through by R yields the expression

$$-\Phi \cdot \ln\frac{\rho(r_c)}{\rho(r_b)} = \frac{1}{4}\frac{r_{abs}^2}{r_s^2}\frac{F_\circledast}{c_T^2} - \kappa\frac{r_c^2}{r_s^2}\frac{dT}{dr}|_{r_c} + \kappa\frac{r_b^2}{r_s^2}\frac{dT}{dr}|_{r_b} \tag{8.55}$$

where we have assumed that the net radiative flux vanishes at the base and becomes constant above r_{abs}. r_{abs} is also assumed to be below r_c. Although the system is nearly isothermal, the diffusive heat fluxes out of the boundaries are not necessarily small, since the small (but nonzero) gradients are multiplied by a large diffusivity. If, however, we assume that the diffusion only redistributes heat provided by radiation and doesn't import heat from the lower atmosphere or export it through the critical point, then the last two terms on the right hand side of Eq. 8.55 can be dropped. Further, the isothermal transonic rule says that $2c_T^2 = g_s r_s^2/r_c$, while Eq. 8.54 allows us to rewrite the log of the density ratio in terms of r_c as well. With these substitutions, the flux becomes simply $\Phi = \frac{1}{4}F_\circledast/g_s r_s$, assuming $r_b/r_s \approx r_{abs}/r_s \approx 1$ and $r_c/r_b \gg 1$. The latter assumption is in fact necessary to assure that the kinetic energy at the base is indeed negligible, as was assumed in the derivation of Eq. 8.54. This flux is precisely the same as the low-temperature limiting flux derived earlier for the nondiffusive case. This satisfactory result provides a consistency check on our reasoning; that the two results should be the same was a foregone conclusion, given the requirements of energy conservation and the assumptions made regarding the energy available to drive the escape.

As in the low-temperature nondiffusive case, then, Φ is fixed by the gravity and *EUV* heating, whence evaluating Φ at the critical point implies that $\rho(r_c)$ decreases algebraically (specifically,like $r_c^{3/2}$) as the critical point is moved outwards. Eq. 8.54 then implies that the density at the base can be made arbitrarily large by moving the critical point outward, since the exponential growth of $\rho(r_b)$ trumps the algebraic decay of $\rho(r_c)$. Conversely,there are limits to how small the base density can be made by moving the critical point inward, since one ultimately violates the condition that the flow be nearly at rest near the base. The qualitative behavior is the same as for the nondiffusive case, in that the *EUV* flux fixes the escape rate, but that one must have enough density at the base if the escaping solution is to exist at all. Unlike the nondiffusive case, though, nothing special happens for monatomic gases.

Wrap-up and summary of hydrodynamic escape

An important feature of the blowoff state is that the escaping flow of a light gas like hydrogen can carry heavier minor constituents along with it, if the outward velocity of the hydrogen is greater than the characteristic fall speed of the heavy constituent. This works only if the concentration of heavy constituents is small enough that it doesn't increase the mean molecular weight of the gas sufficiently to choke off escape. To determine which species can escape, we note that the fall speed (relative to the background current) of a species with molecular weight M is $w_f = (R^*T/Mg)(b/n)$, where b is the binary diffusion parameter for the heavy species diffusing through hydrogen and n is the number density of the hydrogen. For a heavy gas diffusing through a much lighter gas, the binary parameter is nearly independent of M, so that the fall speed is inversely proportional to M. Species heavier than a threshold value do not escape at all, whereas lighter ones escape at a

rates which depend linearly on $1/M$ – in contrast to Jeans escape, which depends exponentially on molecular weight. This differential escape implies a characteristic pattern of enrichment, which is most readily detected in noble gases like Xenon, which are not complicated by chemical reactions. The effect may not be very important for climate evolution, but it provides the main means of determining whether an atmosphere ever experienced a blowoff state in its past.

The one-dimensional treatment of hydrodynamic escape we have given may be elegant, but it suffers from a glaring physical inadequacy: The escape is energized by EUV absorption from the planet's star, which illuminates only the dayside. Yet, despite the fact that the thin outer atmosphere has little thermal inertia to even out the dayside/nightside constrast, the escaping flow has been modelled as spherically symmetric. In reality, the escape is likely to take the form of a complex three dimensional flow, with an outward directed jet centered on the subsolar point while some of the outward directed circulation on the dayside will close in the cold nightside exosphere instead of escaping to space. While more comprehensive treatments of hydrodynamic escape have added much sophistication in terms of atmospheric chemistry and radiative transfer, none at the time of writing have taken on the grand challenge of modelling the three-dimensional structure of an escaping atmosphere. Another challenge is that the traditional hydrodynamic escape formulation assumes local thermodynamic equilibrium and treats the fluid as a continuum, whereas the actual dynamics become nearly collisionless when one goes sufficiently far out in the atmosphere. If the transition to nearly collisionless dynamics occurs past the transonic point, this may matter little, given that information cannot propagate upstream in supersonic flow. However, if the transition occurs below the transonic point, the very notion of "speed of sound" breaks down and the controlling role of the transonic rule is likely to change substantially. Dealing with the transition from continuum to collisionless flow in this case is a considerable challenge. Given the nonlinearities and thresholds in the escape problem, it is likely that major revisions in our conception of escape rate are in store once somebody rises to these challenges.

8.7.5 Erosion by solar wind

Solar wind erosion is a form of nonthermal escape energized by solar wind particles instead of EUV photons. The corona is basically the exosphere of the Sun, and the solar wind is nothing more nor less than hydrodynamic escape of the Solar atmosphere, which is primarily hydrogen ionized to protons. The mechanism is general and applies to virtually all stars, though we will not attempt to discuss here how the stellar wind characteristics vary from star to star, nor the way the stellar wind changes as a star proceeds trhough its lifecycle.

The solar wind consists almost entirely of protons (95this is likely to be the case for all main sequence stars. Solar wind particles are extremely energetic; they fly out with speeds of 300 to 600 km/s or even more, having energy between $8 \cdot 10^{-17} J$ and $3.2 \cdot 10^{-16} J$. The solar wind is very tenuous; at the Earth's orbit, it has a density ranging up to 10^7 particles per cubic meter, leading to a particle flux of $4 \cdot 10^{12}/m^2 s$, though it also fluctuates down to values as low as a tenth as much. The flux at other orbits can be estimated using the inverse-square law. The energy flux in the Solar wind is on the order of $10^{-3} W/m^2$ at Earth's orbit, which is comparable to the total EUV energy flux. Stellar winds are stronger when a star is young, and decrease over time, with the most pronounced decay occurring in the first billion years of the star's like for G class stars like the Sun. The precise nature of the time evolution is not definitively settled, and ties in with many of the same issues that determine the long term evolution of EUV output.

Based strictly on energetic considerations, it would appear that the solar wind has the potential to cause a great deal of atmospheric erosion even for planets as massive as Earth or

Venus, and still more so for Mars. The escape energy for an oxygen atom from Earth or Venus is about $1.7 \cdot 10^{-18} J$, so a single solar wind proton has enough energy to knock loose about 120 oxygen atoms if it could be optimally deployed. Taking into account the flux of solar wind protons (reduced by a factor of 4 to allow for averaging over the surface of the planet), this would lead to the loss of 20 bars of oxygen on Venus in the course of a billion years, or about half that much from Earth where the solar wind flux is weaker. For Venus, even this upper bound is only sufficient to remove the oxygen in a 200 meter deep ocean. The oxygen loss estimate for Earth is only relevant to the time after the atmosphere was appreciably oxygenated, but it does indicate that solar wind erosion cannot be *a priori* ruled out as a factor in evolution of the atmosphere. On Mars, we are principally interested in the process that could lead to the loss of a dense primordial CO_2 atmosphere; the solar wind flux is lower at the orbit of Mars, but the excape energy is less. A purely energetic estimate suggests that solar wind erosion could potentially lead to a loss of 8 $bars$ of CO_2 in the course of a billion years, which would be more than sufficient to leave Mars in its present state.

However, there are a number of factors that greatly limit the ability of collision with solar wind protons to directly erode the heavier species of an atmosphere. The first is that, in a collision between a proton and a heavier particle, only a small part of the proton's energy is transferred to the heavy particle. For in-line collisions, conservation of energy and momentum during a collision between a light particle of mass m_1 and a stationary heavy particle of mass m_2 implies that the fraction of incident energy transferred to the heavy species is only $4(m_1/m_2)$. For protons colliding with CO_2, this amounts to about 10%. Moreover, the light proton reverses in direction in a collision, and therefore tends to escape the planet's gravity well before it can lead to additional escape. Since the typical solar wind proton has enough energy to cause 217 CO_2 molecules to escape Mars, even the limited energy depositied could cause nearly 22 molecules to escape, but only if the subsequent collisions of CO_2 or its components with the rest of the atmosphere caused 100% efficient escape. This is an unrealistic upper bound, since much of the energy of subsequent collisions will be lost to multiple collisions with simply heat the atmosphere a bit. Thus, if solar wind protons could collide with the Martian atmosphere, the actual loss of CO_2 over a billion years would be somewhere between 0.8 $bars$ and 0.08 $bars$. The former, based on the upper bound, would still be significant, but the latter would relegate solar wind erosion to the role of a minor player. Similar considerations reduce the estimates of oxygen loss from Earth of Venus.

In reality, the effect of planetary and solar electromagnetic fields render the problem far more complicated. For planets with a strong planetary magnetic field, such as Earth, the planetary field very effectively shields the atmosphere from impacts with solar wind protons, and limits erosion to small values. Titan has no magnetic field of its own, but it benefits from shielding by Saturn's magnetic field during the part of its orbit when it is effectively in the solar wind shadow caused by Saturn's field. The same phenomenon is likely to apply generically to moons orbiting gas giants. Venus has essentially no planetary magnetic field, and Mars has only a very weak one arising from remnant magnetizationof the crust. However, buildup of ionized species in the outer atmosphere still leads to electromagnetic fields that are sufficient to deflect most incident protons and keep them from interacting significantly with the atmosphere. For this reason, direct collisions with protons is not generally considered to be a significant source of erosion of heavy species even for nonmagnetized planets like Mars or Venus.

Heavier ions are less deflected by magnetic fields, and so can penetrate more deeply into the atmosphere. Collisions with heavy ions are also more effective at transferring energy to heavy species. For this reason, it is not the solar wind itself, but secondary acceleration of heavy ions by the solar wind that play the greatest role in solar wind erosion. This is where things get complicated. First of all, you need a supply of heavy ionized species; these are created by ionization

due to the *EUV* flux. On Mars and Early Venus, the heavy ions of principle interest are ionized oxygen atoms, in the Martian case arising from decomposition of CO_2 and in the Early Venus case arising from decomposition of H_2O if the planet is in a runaway state. The second step is accelerating the oxygen ions to an energy where they can cause escape. This step is not done by collisions with solar wind protons, but rather by the forces exerted by the electromagnetic field carried by the solar wind. The energy still ultimately comes from the energy of the solar wind, but it is transferred by the intermediary of large scale electromagnetic interactions. The calculations necessary to determine the flux and energy of heavy ions are very complex, since they require a model of the electromagnetic field of the solar wind as well as the tracking of a large shower of ions injected into the field. The final stage of the problem is figuring out what happens when an accelerated oxygen ion collides with the atmosphere. When the shower of ions hits the atmosphere, a certain fraction of the target will splashed out backward with sufficient energy to escape the gravitational well. This process is known as *sputtering*. Sputtering is not nearly 100% efficient at converting incident energy to escaping particles, and the computation of sputtering efficiency is difficult, depending, among other things, on the degree to which the collisions cause the target molecules to dissociate.

Because of the complexities and considerable uncertainties of such calculations, we are in the regrettable position of not being able to guide the reader through any simple robust estimates of the actual likely mass loss due to solar wind erosion; it is a subject for experts and cognoscenti, but a very important one. The articles cited in the Further Readings section for this chapter will provide some introduction to the subject, and the range of numbers that have emerged to date. Most estimates of loss of CO_2 from Mars due to solar wind sputtering suggest that 0.1 to 0.2 *bars* could be lost in this way, though the somewhat controversial calculations of Kass and Yung put the number as high as 1 *bar*. It thus appears that solar wind erosion is a potentially significant factor for loss of a primordial dense Martian atmosphere, but the the general sentiment is that most of the mass loss would have to occur by other means (probably impact erosion, to be discussed shortly). Solar wind erosion does not seem to provide a viable mass loss mechanism for either the hydrogen or oxygen in a runaway Venus state, despite the planet's proximity to the Sun and even assuming that the Early Venus, like that of today, had no planetary magnetic field. The only situation in which solar wind erosion could have contributed to loss of a Venus ocean is if the loss occurred in the first hundred million years of the life of the Solar system, during which time it has been speculated that the solar wind may have been over a thousand times as intense as it is today. Even then, the loss occurs not by direct erosion or sputtering, but rather by providing enough additional heating to cause hydrodynamic escape of hydrogen with sufficient velocity to drag oxygen along with it. This mechanism bears more resemblance to hydrodynamic escape, with solar wind substituted for *EUV* as an energy source, then it does to the sputtering mechanisms that have been discussed in connection with Mars. The shielding effect of the Earth's magnetic field limits solar wind erosion to very small values, so that it is not a major factor in atmospheric evolution for Earth.

As in the case of escape driven by other processes, the trickle of escape of heavy species driven by solar wind interactions is pertinent to a rich variety of interesting questions bearing on the observed structures of outer planetary atmospheres today. Many of these questions are important in interpreting the isotopic composition of atmospheres in terms of past atmospheric evolution. However, for the purposes of this book we are principally interested in those mechanisms that cause enough escape to substantially effect the evolution of a planet's climate or habitability. The best guideline we can provide at this point is that, for bodies orbiting G-class stars like the Sun, solar wind erosion should be kept in mind as a sigificant factor in climate evolution for bodies the size of Mars or smaller, which are unshielded by a planetary magnetic field. M-dwarfs, despite being cool, have much higher stellar wind fluxes than the Sun, and could therefore sustain

significant atmospheric escape from somewhat larger bodies, especially if they are in close orbits.

8.7.6 Impact erosion

The two main cases in the solar system where theories of atmospheric evolution call for massive atmospheric loss are the problem of water loss on Venus and the loss of a hypothetical dense CO_2 atmosphere on Mars. For Titan the question is the converse – accounting for the lack of atmospheric N_2 loss, and that is plausibly accounted for by the solar wind shielding provided by Saturn's magnetic field and the low *EUV* flux at such large distances from the Sun. There may have been blowoff early in Titan's history, but on an icy body there are plenty of volatile reservoirs available to restock an atmosphere. Hydrodynamic escape of hydrogen from photodissociated water provides a plausible mechanism for the Venus case, but we still don't have a good way to get rid of a 2 *bar* Early Mars atmosphere, except for the remote possibility that solar wind erosion could do the trick. If there is no way to lose such a dense greenhouse atmosphere from Early Mars, then explanations for the apparently warm and wet early climate of that planet must be sought elsewhere. It is likely, however, that impact erosion could provide much of the needed loss mechanism. Impact erosion could be similarly important for Mars-sized bodies elsewhere in the Universe. One should be cautious about concluding that the way things happened in the solar system are the way things must happen elsewhere, but still the fact that Earth and Venus retain thick atmospheres while Mars does not suggests that Mars-sized bodies may be susceptible to loss of atmosphere. One would like to know what the possible mechanisms are, and whether there are circumstances in which a Mars-sized body could retain an atmosphere at temperatures warm enough to be habitable and without the shielding effect of a nearby giant planet.

The impact history in the inner portion of a planetary system, where rocky planets form, can be divided up into five broad stages:

- The *Early Accretion* stage, in which small planetesimals are colliding to form larger objects, which in turn aggregate into a broad spectrum of still larger objects.
- The *Late Accretion* stage, in which most of the aggregation is complete, but there are still a number of planet-sized bodies in nearby orbits. According to simulations, Lunar to Mars sized bodies are common at this stage, and one or more giant impacts are likely. In our own Solar System, there is evidence that Earth, Venus and Mars all experienced a giant impact at some point
- The *Sweep* stage, in which each planet has attained nearly its ultimate size, and is sweeping up much smaller debris in the vicinity of its orbit. The impacts in this stage are still frequent, but involve collisions with bodies much smaller than the planet. There are no giant impactors left.
- The *Late Heavy Bombardment*, in which a second swarm of impactors originating from perturbations of mass stored in more distant orbits encounters the inner system. In our Solar System the Late Heavy Bombardment occurred around 3.8 billion years ago. It is not known what caused this bombardment, where the mass came from, or whether such bombardments are a generic feature of the late stages of planetary system formation. Indeed, it is not even completely clear whether the Late Heavy Bombardment is really distinct from the sweep stage. In any event, the period of impacts that could substantially erode atmospheres came to a close with the Late Heavy Bombardment, about one billion years after the beginning of the formation of our Solar System. This number can be taken as a rough guide to the time

scale of impact erosion in other planetary systems, though in systems without a Late Heavy Bombardment the period when inner planets are subject to impact erosion probably would end up to 300 million years earlier.

- The *Steady State Bombardment* stage, in which nearby sources of impactor mass have been used up. During this stage there are infrequent collisions by objects drawn from the pool of cometary and asteroidal material. Large objects of this class are only occasionally flung into planet-crossing orbits by long term chaotic gravitational interactions, though inconsequential encounters with very small objects (as in meteor showers) occur on a nearly continuous basis.

Impact erosion is a crucial factor in the process by which a planet forms and retains an atmosphere as it accretes by collision of planetesimals, but in the following we will be principally concerned with impacts that occur in the later stages of the process –the sweep phase and the Late Heavy Bombardment (if it is present). In this stage the planet has attained close to its ultimate mass and most of the impactors are much less massive than the target planet but there is still enough impactor mass available to cause significant atmospheric erosion. We will also offer a few remarks on the consequences of giant impacts which preceed the sweep stage. The long period of steady state impacts which follow the sweep stage and continue to this day can have cataclysmic consequences such as the mass extinction triggered by the Cretaceous-Tertiary impact, but these impacts are too small and too infrequent to cause much atmospheric erosion.

In impact erosion, the source of energy available to accelerate parts of the the atmosphere to escape velocity is the kinetic energy of the impactor. This energy depends on the mass of the impactor and the velocity with which it encounters the planet. When an impactor is dropped onto a much more massive planet from a great distance, it will strike the planet with a velocity equal to the planet's escape velocity, provided the impactor is initially at rest relative to the planet. This is a simple consequence of conservation of kinetic plus potential energy. It can be shown that because objects in neighboring orbits travel at different speeds, gravitational interactions cause the typical impact speed to be somewhat greater than the escape velocity. However, the enhancement is seldom as much as 20%, so throughout the following we will simply take the escape velocity as the characteristic impact speed. The scaling of impact velocity with escape velocity eliminates much of the intuitive effect of planetary size on impact erosion: though the atmosphere of a small planet is less gravitationally bound than that of a large planet, the typical impactor energy is also less for a small planet.

The case of satellites is quite different, since there we need to take into account the gravity of the parent body as well as that of the satellite. Consider the case of Titan, which orbits Saturn at a distance of $1.22 \cdot 10^9 m$. At this distance, the gravitational acceleration of Saturn is .025 m/s^2. The corresponding escape velocity from Saturn for an object starting at this orbit is 7.9 km/s, which is also the speed an object dropped from infinity reaches under the gravitational influence of Saturn, upon reaching Titan's orbit. In comparison, the escape velocity from Titan is only 2.6 km/s. Under the joint influence of both gravitational fields, the impactor speed would be approximately $\sqrt{2.6^2 + 7.9^2}$, or 8.3 km/s. In reality, the encounter speed could be somewhat greater, depending on the geometry of the impact, since one should also take into account the orbital speed of Titan, which is 5.5 km/s. The enhancement of impactor velocity for satellites allows the atmosphere to be eroded by a much smaller total mass of impactors than would be required if the body were a planet in an orbit of its own. This effect is overwhelmed, however, by the fact that a satellite must compete for impacts with the much larger planet it is orbiting. Impacts are received in proportion to the cross section areas of the bodies, so the satellite receives many fewer impacts than it would if it were a planet in its own orbit, all other things being equal. For Titan, the ratio of impacts is 0.002, so this effect ups the required mass of available impactors in the orbit by a factor of 500. In

the estimates to follow, we'll use the term "Titan-P" to refer to a Titan-like body in a planetary orbit of its own, reserving the name "Titan" for the actual satellite subject to the effects of Saturn.

In a similar vein, the classic theory of impact erosion from planets assumes impactors in fairly circular orbits originating not far from the planet being, impacted, with the planet itself assumed to be in a fairly circular orbit. Recent simulations of planet formation indicate that there can be stages in which planetesimals have highly eccentric orbits. Indeed the catalog of extrasolar planets is rife with systems having very eccentric orbits. When the impactor or planet has an eccentric orbit, the impact velocity can considerably exceed the planet's escape velocity, much as was the case for impacts on satellites. In this case, the planet's parent star plays much the same role as a satellite's parent planet. If eccentric impacts are common, they tend to tilt the balance in favor of greater atmospheric erosion from small planets, since the eccentric impacts decouple impact energy from the strength of a planet's gravity. The understanding of this effect is rapidly evolving, and so we will merely flag it here as a subject worthy of the readers' attention. Our discussion in the following will be mostly based on the classic scaling of impact energy.

The typical impact speed is much greater than the typical speed of sound. For example, an Earthlike body would have an impact speed of over 11,000 m/s, whereas the speed of sound in air at $300K$ is only 347 m/s. Since pressure and density modifications can travel no faster than the speed of sound, the impactor carves a cylinder into the atmosphere leaving a near-vacuum in its wake. The walls of the cylinder have little time to close in, and the flow just ahead of the impactor cannot know about the presence of the planet's surface far below. The situation is not like a piston slowly compressing a column of air. Instead, the drag force on the impactor is solely dependent on the instantaneous density and temperature just ahead of its position at any time, and on the velocity and cross-section area of the impactor.

Any amount of energy delivered to the planet could, in principle, cause some amount of atmosphere to be accelerated to escape velocity. The key question in determining how much atmosphere *actually* escapes is the volume of atmosphere over which the energy is diluted. Striking a match releases about 1000 J of energy, which is enough to cause 16 milligrams of air to escape from Earth. It doesn't happen, though, because the energy released is too diluted to cause any gas to escape. The fluid dynamics governing partioning of energy upon impact is complicated and difficult to simulate accurately. Nonetheless, simulations and theory point to a few simple principles upon which estimates of impact erosion can be based.

We can distinguish three classes of impacts:

- Small impacts, which dissipate their energy in the atmosphere. Simulations indicate that these do not cause much atmospheric escape.
- Intermediate sized impacts which dissipate most of their energy when striking the ground or ocean, but which are still small compared to the target planet and therefore do not significantly accelerate the planet as a whole. In this case, the impactor still has a great deal of energy left upon striking the solid or liquid surface of a planet. The the energy is released in a concentrated burst which can lead to considerable erosion of the portion of the atmosphere in the vicinity of the impact.
- Giant impacts, in which the impactor has mass comparable to the target planet. These can cause total blowoff of an atmosphere since the shocked planet acts like a piston which can accelerate nearly the entire mass of the atmosphere to escape velocity.

The energy required for an impactor to reach the surface with most of its energy left depends on how massive the atmosphere is. Any impactor at all will reach the surface of the Moon, but

hardly anything gets through to the surface of Venus – which is why the surface of Venus is characterized by just a few but very large impact craters. The drag force exerted on a high speed sphere of radius r is $C_D\rho_a\pi r^2U^2$, where U is the speed of the body, rho_a is the atmospheric density and C_D is a dimensionless number which asymptotes to values near unity for very supersonic flow. When passing through an atmosphere of thickness H (estimated as the density scale height), the work done against the object is $C_D\rho_a\pi r^2U^2H$. We'll assume that the initial velocity is high enough that the additional work done by the force of gravity after the projectile encounters the atmosphere can be neglected. Then, equating the work done by drag to the kinetic energy $\frac{1}{2}MU^2$ for a body of mass M, we find that the condition for the body to reach the surface with some of its initial energy left is $M > 2C_D\rho_a\pi r^2H$. Since $C_D \approx 1$, this condition says that the mass of the body must exceed twice the mass of the cylinder of atmosphere having the same cross section area as the body. Writing $M = \frac{4}{3}\pi r^3\rho_i$, where ρ_i is the impactor density, we find $r > \frac{3}{2}(\rho_a/\rho_i)H$ for $C_D = 1$. For Earth's present atmosphere, this yields a critical radius of a mere $3.5m$, assuming a silicate impactor having a density of $3000kg/m^3$. For the atmosphere of Venus, the critical radius is a more impressive 340 m.

Exercise 8.7.9 Estimate the critical radius for a silicate impactor to penetrate the $7mb$ atmosphere of present Mars.

Let's focus next on the effect of intermediate impacts. When the impactor reaches the surface, most of its energy is turned into heat, which creates a shock wave which travels outwards from the point of impact, accelerating the atmosphere. Some of the energy also goes into vaporizing the solid surface, which adds mass to the gas that must be ejected, and steals energy that could otherwise be used to feed escape; this is a refinement we shall not pursue quantitatively. Fluid mechanical simulations and analytic shock wave solutions indicate that once a critical energy is reached, essentially all the atmosphere in a narrow cone above the impact is ejected, as shown in the left panel of Fig. 8.8. As the mass of the impactor is increased, the angle of the cone widens, until it reaches the point where all the atmosphere above a plane tangent to the sphere at the point of impact is blown off. Further increases in the mass of the impactor cause little or no additional escape, until the size class of giant impacts is approached. There is a narrow range of impactor masses between the mass where a narrow cone is first blown off and the mass at which the whole tangent slice is blown off. Therefore, one can get a reasonable impact of impact erosion by non-giant impacts by simply assuming that all impactors with a mass below that required to blow off a tangent slice cause no loss, whereas all impactors with mass above this critical mass cause loss of one tangent slice of the atmosphere.

Now we are prepared to answer the key question of how the susceptibility to impact erosion scales with the size of the planet and the thickness of its atmosphere. There are two ingredients to this estimate: the critical impactor mass needed to blow off a tangent slice of the atmosphere, and the fraction of atmospheric mass eroded by each such impact. The critical mass is important because it says how big an impactor has to be to cause significant erosion; this affects the erosion rate because there are many more small impactors than there are big impactors. To estimate the mass of atmosphere above a tangent plane, we represent the atmosphere as a uniform density layer with density ρ_a and depth H equal to the scale height at a mean atmospheric temperature T. With this approximation, the mass above a tangent plane is simply ρ_aH^2a, where a is the radius of the planet. The energy needed to cause this amount of mass to escape to space is $\frac{1}{2}v_e^2\rho_aH^2a$, while the energy of an impactor of mass m is $\frac{1}{2}mv_i^2$ where v_i is the speed of the impactor. However, since $v_i \approx v_e$, the velocity terms cancel and we find that the critical impactor mass is $m_c \approx \rho_aH^2a$, i.e. the mass of atmosphere above the tangent plane. Note that ρ_aH is the mass of atmosphere per unit area of the planet's surface. We will find it convenient to use this quantity as our basic

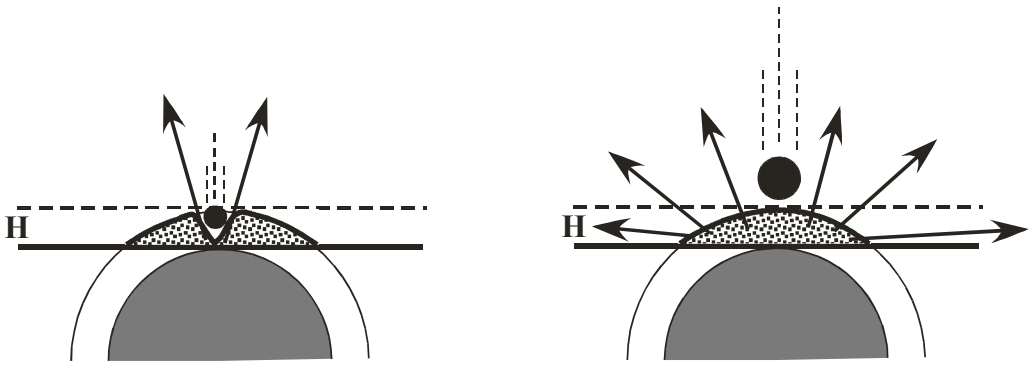

Figure 8.8: Portion of atmosphere subject to impact erosion by non-giant impacts. The left panel shows the portion potentiallly eroded by small impacts; very small impactors dissipate their energy before reaching the ground and cause little or no erosion. The right panel shows the limiting erosion by a larger impactor, which can erode all the atmosphere above the tangent plane to the planet's surface at the point of impact. Increasing the mass of the impactor does not yield much further erosion, until the giant-impact class, able to significantly accelerate the entire target planet, is reached.

measure of the amount of atmosphere remaining on the planet. In terms of the surface pressure $\rho_a H \approx p_s/g$, according to the hydrostatic relation.

For fixed $\rho_a H$,the critical mass scales with $Ha = RTa/g$, but since $g = \frac{4}{3}G\rho_p a$ where ρ_p is the mean density of the planet, we find that the critical mass is

$$m_c \approx \frac{3}{4}\frac{RT}{G\rho_p}(\rho_a H) \tag{8.56}$$

Thus, *for fixed atmospheric mass per unit area, the critical impactor mass is independent of the size of the planet.* While the critical mass formula does not discriminate by size of planet, it does say that massive atmospheres (in the sense of large $\rho_a H$) are more difficult to erode than tenuous atmospheres, because larger impactors are needed to trigger erosion in the more massive case, but there are fewer large impactors than small impactors. For a given mass spectrum of impactors, erosion of a massive atmosphere is initially slow and intermittent, accelerating and becoming more steady as erosion proceeds and the atmosphere becomes less massive. For a satellite, m_c must be reduced by a factor of $(v_e/v_i)^2$, which is about 0.1 for Titan.

The quantity $\frac{3}{4}RT/G\rho_p$ has the dimensions of a length squared; we'll use the symbol ℓ^2 to refer to it. The length scale ℓ depends on the density of the planet and the composition and temperature of the atmosphere, but not on the mass of either the planet or the atmosphere. It varies little over a wide range of planetary situations. For a $1bar$ N_2 atmosphere at $280K$ on Earth, $\ell = 233km$. For a $2bar$ Early Mars CO_2 atmosphere at $280K$, $\ell = 220km$. For a 1.5 *bar* N_2 atmosphere on Titan at $80K$, $\ell = 213km$. Some typical values of critical impactor mass and size are given in Table 8.6. Note that the critical mass is mostly controlled by the mass path $\rho_a H$, rather than the size of the planet. The Early Mars case with a 2 *bar* atmosphere and the

	m_c (kg)	r_c,silicate (km)	r_c,ice (km)	N_e	m_{tot}/m_{Earth}
Earth, $1bar$ N_2,$280K$	$5.5{\cdot}10^{14}$	3.5	5.2	3003	$1.0\cdot10^{-3}$
Mars, $2bar$ CO_2,$280K$	$2.6{\cdot}10^{15}$	5.9	8.7	951	$0.7\cdot10^{-3}$
Mars, 100 $mbar$ CO_2,$220K$	$1.0{\cdot}10^{14}$	2.0	2.9	1210	$0.18\cdot10^{-3}$
Venus, $90bar$ CO_2, $700K$	$9.2{\cdot}10^{16}$	19.4	28.3	1623	$7.1\cdot10^{-3}$
Venus, $1bar$ N_2, $280K$	$6.4{\cdot}10^{14}$	3.7	5.4	2582	$0.94\cdot10^{-3}$
Titan-P,$1.5bar$ N_2,$80K$	$5.0{\cdot}10^{15}$	7.4	10.8	585	$0.6\cdot10^{-3}$
Titan ,$1.5bar$ N_2,$80K$	$5.0{\cdot}10^{14}$	3.4	5.0	585	$94.\cdot10^{-3}$
Super-Earth, $1bar$ N_2,$280K$	$3.2{\cdot}10^{14}$	3.0	4.3	8778	$2.3\cdot10^{-3}$

Table 8.6: Table of impact erosion parameters for various bodies. m_c is the critical mass required to blow off a tangent slice, and the r_c columns give the corresponding impactor radii (in km) for silicate or icy impactors. N_e is the number of impacts with $m > m_c$ which are required to deplete most of the atmosphere, and m_{tot} is the total available impactor mass in orbit required to yield this number of impacts. The calculations of m_{tot} depend on the mass distribution; results in this table assume a power law distribution with $q = 1.5$ and the maximum impactor mass m_+ equal to one tenth the mass of Earth's Moon. The "Titan-P" case gives results for a hypothetical Titan-like body in an orbit of its own, while the "Titan" case includes the effects of Saturn, including the competition with Saturn for available inpactor mass. The "Super-Earth" case is for a planet with the same density as Earth and with a mass of 5 Earth masses.

Titan-P case actually require larger impacts to erode than the present Earth case. The massive atmosphere of Venus is hard to erode, but if Venus ever went through a period when it had a 1 *bar* atmophere, it would be essentially as easy to erode as Earth. Generally speaking, erosion of Earthlike atmospheres is sustained by impactors with radii of a few kilometers or more, and somewhat larger impactors are required for the Early Mars case. As the Early Mars atmosphere erodes down to surface pressures of $100mb$, one can make do with impactors of somewhat over a third the size, or one twentieth of the mass. The table also includes a Super-Earth case based on the extrasolar planet Gliese 581c, which has a mass about five times that of Earth. The critical impactor size is slightly lower than for Earth, because the surface gravity is higher and hence a 1 *bar* atmosphere on the Super-Earth has less mass per unit area than a 1 *bar* atmosphere on Earth.

For each impactor exceeding the critical mass, the fraction of atmosphere eroded is $\pi a\rho_a H^2/(4\pi\rho_a H a^2)$, which is $\frac{1}{4}H/a$ or equivalently $\frac{1}{4}\ell^2/a^2$. Thus,*the fraction of atmosphere eroded per supercritical impact decreases quadratically with the radius of the planet, all other things being equal.* This is the main reason that small planets are more susceptible to impact erosion than large planets. The characteristic number, N_e, of supercritical impacts needed to substantially erode the atmosphere is $4a^2/\ell^2$. If the impactors arrive in sequence and the atmosphere has a chance to adjust back to uniform coverage of the planet between impacts, then this number of impactors would erode the atmosphere down to a mass of $1/e$ of its initial mass, given that after each impact there is less mass left to erode and each supercritical impact just takes away a fixed fraction of what is there. The values of N_e for some planets of interest are given in Table 8.6. This number is primarily controlled by the size of the planet, ranging from about 600 for Titan to about 3000 for Earth and about 9000 for a Super-Earth. The colder Mars case with a thin atmosphere requires more impactors than the hot Mars case with a $2bar$ atmosphere because the scale height is smaller in the former case.

To complete the story, we must estimate the total mass of impactors that must hit the planet in order to get the number of supercritical impacts (N_e) required for subtantial erosion. We will carry out this estimate for a late stage in planetary formation, when the planet in question is by

far the largest thing near its orbit and the remaining debris near the orbit is all small compared to the planet; our planet is at this stage the big kid on the block, subject to small to intermediate impacts as it sweeps up a late veneer of the remaining debris. We can define a catchment basin for the planet, consisting of the range of orbits for which the debris is more likely to impact the planet under investigation rather than some other planet. As time goes on, essentially all of the mass in this catchment basin will eventually impact the planet. (This part of the story will be slightly modified for satellites). It is this total mass we shall estimate. For given total mass in the catchment basin, a small planet like Mars will take longer to sweep up the debris, than a larger planet such as Earth, in proportion to the relative cross section areas. Thus, for a small planet, erosion by intermediate impacts will carry on for a longer time than for a large planet. For Mars, the late stage of the erosion process will last 3.5 longer than it would for Earth, all other things being equal. This time scale comparison may be a significant factor in accounting for the present tenuous atmosphere of Mars, since Mars can regenerate an atmosphere if it loses it early on when it is still tectonically active, but not if the loss occurs later, when the planet's interior has frozen out and has ceased outgassing volatiles.

To proceed further we must make some assumption about the mass distribution of impactors, because of the role of the critical mass m_c in determining how much atmosphere gets blown off by an individual impact. The optimal distribution for erosion would have all the impactors of equal size m_c. Making impactors smaller reduces the erosion because the impactor is unable to blow off a tangent slice, and making the impactors larger wastes impactor mass because a large (but not giant) impact can't blow off more than a tangent mass. Information about the mass distribution and total available mass of impactors comes to us mainly from the cratering record of rocky planets, and among those primarily from the Moon and Mars (which have well-preserved surfaces not much subject to erosion). The estimates are highly uncertain, and uncertainties in the impactor distribution almost certainly overwhelm uncertainties in the detailed fluid dynamics governing how much mass is blown off by an individual impact. The mass spectrum of impactors can also be estimated from the mass distribution in today's asteroid belt. These estimates are generally compatible with estimates derived from the cratering record. Information about the *timing* if the impacts, and the rate of decay in the inner Solar system, comes from looking at the cratering record in younger resurfaced terrain, and in the Lunar case, from direct radiometric dating of crater crater samples returned to Earth for analysis.

Let $N(m)$ be the number of impactors with mass greater than m. This is the function we need to know. The mass spectrum $n(m)$ is given by $dN/dm = -n(m)$. Equivalently, $n(m)dm$ is the number of impactors in the mass range between $m - dm/2$ and $m + dm/2$. The distribution of crater radii on any individual body has been found to approximately obey an r^{-3} power law, where r is the crater radius. This power law captures the basic crater distribution for moons of Jupiter and Saturn, as well as for the inner planets, though the details of the deviations from the ideal power law are different between the outer solar system and the inner solar system. The corrections to the r^{-3} law are strikingly similar between Mercury, the Moon and Mars. This provides strong evidence that the entire inner Solar System was subject to the same population of impactors. Because crater radius scales with a power of impact energy, the crater distribution implies a power law distribution of impactor energy. Since impact velocity is approximately constant for any given body, the impact energy is proportional to the impact mass for a given body, implying a power law distribution for impactor mass. Specifically, numerical simulations and study of thermonuclear bomb craters imply that crater radius scales approximately with $E^{1/3}$ (i.e. $m^{1/3} for impactors$. If the crater diameter distribution is $n(r)$, we get the corresponding mass distribution by writing

$$n(r)dr = n(m^{1/3})d(m^{1/3}) = frac13n(m^{1/3})m^{-2/3}dm \tag{8.57}$$

from which we identify $frac13n(m^{1/3})m^{-2/3}$ as the mass spectrum. The r^{-3} crater radius power

law thus implies an $m^{-5/3}$ power law for impactor mass. Use of more detailed fits to the crater data along with alternate crater-size models, as well as direct fits to asteroidal mass distribution, yield exponents between 1.5 and 1.8.

Suppose now that $n(m) \propto m^{-q}$ for some exponent q. The total mass of impactors is $\int m \cdot n(m)dm$, and the blowup of $n(m)$ at small m does not cause the total mass to diverge as long as $q < 2$. On the other hand, the *number* of small impactors is infinite for $q > 1$, as is the case for the observed distribution. However, for $q > 1$ the total mass of *large* impactors diverges if the power law continues out to infinite mass. Thus, to make physical sense, the power law must be truncated at some mass m_+, which represents the largest mass impactor in the population. With this assumption, the total mass in the distribution is finite, and we can write a normalized distribution as

$$n(m) = \frac{2-q}{m_+}\frac{m_{tot}}{m_+}(\frac{m}{m_+})^{-q} \tag{8.58}$$

where, m_{tot} is the total mass of impactors. It is presumed that $n = 0$ for $m > m_+$.

Exercise 8.7.10 Verify that m_+ is the total mass of impactors implied by the distribution in Eq. 8.58. Find the cumulative distribution $N(m)$ and discuss how this behaves for $m_+ \to \infty$ with m_{tot} fixed.

The total number of impactors with mass greater than m_c is

$$N(m_c) = \int_{m_c}^{m_+} n(m)dm = \frac{m_{tot}}{m_+}\frac{2-q}{q-1}((\frac{m_c}{m_+})^(1-q) - 1) \tag{8.59}$$

From this, we set $N(m_c)$ equal to N_e, which tells us the required total mass m_{tot}, given q and m_+. The special case of satellites is treated by reducing m_c according to the estimate of impactor velocity enhancement, and multiplying the value of m_{tot} by the ratio of area of the primary to area of the satellite, so as to take into account the proportion of total available impactors that hit the satellite rather than the primary.

The results of this calculation of m_{tot} are given in the final column of Table 8.6. These calculations were carried out with $q = 1.5$ and $m_+ = 7.35 \cdot 10^{21} kg$, which is one tenth the mass of the Earth's moon. It takes rather little impactor mass in the late stage veneer to deplete the atmosphere of an Earthlike planet – only a tenth of a percent of Earth's mass, which is not an unreasonable amount to be left over after the assembly of an Earth-sized planet. An important result is that if Mars were to start out with a $2bar\ CO_2$ atmosphere (as suggested by some climate calculations based on evidence for warm,wet early conditions), its atmosphere would not be much more subject to erosion than Earth's. The mass of available impactors required to erode such a Martian atmosphere would be fully 70% of the corresponding mass for Earth. The main reason the estimates are so similar is that a $2bar$ atmosphere on Mars has much more mass per unit area than Earth's atmosphere, requiring a higher critical mass of impactor as compared to Earth. A more tenuous Martian atmosphere is much more erodable than Earth's, as illustrated by the 100 mb Mars case in the table. Similarly, if Venus had an Earthlike atmosphere, its atmosphere would be essentially as erodable as Earth's, whereas the actual dense Venus atmosphere requires about seven times as much available impactor mass to erode. The hypothetical Super-Earth case is only a bit less subject to erosion than Earth, in this case because a 1 bar atmosphere on a large planet has less mass per unit area than Earth's atmosphere. The importance of the atmospheric mass effect shows also in the hypothetical planetary Titan case, which, owing to its very massive atmospherre, requires nearly as much available impactor mass to erode as does the 2 bar Early Mars case. The real Titan, in contrast, is very difficult to erode, requiring an available impactor mass of nearly a tenth of Earth's mass, owing to the competition with Saturn for impacts.

The essential puzzle posed by the results of Table 8.6 is that it looks quite plausible that Earth's atmosphere would be subject to loss by impact erosion in the sweep stage, and that a dense Early Mars atmosphere would not be appreciably less erodable than Earth. How, then, to account for the present tenuous Martian atmosphere, while Earth has a substantial atmosphere remaining? One potential factor is that Earth's atmosphere *was indeed* lost by impact erosion, but was regenerated by outgassing from the interior. Consistent with this picture, we note that while Mars requires nearly as much available impactor mass a Earth, this impactor mass is delivered over a much longer time, owing to the smaller cross section of Mars. Combined with the relatively early shutdown of tectonic activity and hence outgassing on Mars (owing to its small size) it could be that the essential difference between the planets resides not so much in ability to *hold* an atmosphere as in ability to *regenerate* an atmosphere. A severe difficulty with this picture, however, is the abundance of N_2 in Earth's atmosphere. A CO_2 or water vapor atmosphere could be easily regenerated, but it is not easy to hide enough N_2 in the mantle to allow this component to be regenerated. And recall that Venus has even more N_2 in its atmospherere than Earth, suggesting that even if Venus went through an early stage with far less CO_2 in its atmosphere, it did not suffer total atmosphere loss by impacts during that stage. Could it be that there is an ability to sequester a bar or two of N_2 in a planet's mantle? Could it be that Earth started out with much more N_2 in its atmosphere and that what we have today i the small bit left over after substantial impact erosion? Or could it be that the mass of impactors was not in fact sufficient to deplete Earth's atmosphere and that the tenuous Martian atmosphere has some other explanation? Perhaps it never generated a dense atmosphere, because it never received enough oxygen-bearing material to turn carbon into carbonate and CO_2. Perhaps Mars lost its atmosphere in a chance giant impact which got rid of Martian N_2, whereas Earth's Moon-forming impact was not big enough to get rid of all the N_2. If a giant impact got rid of most of the primordial N_2 on Mars, then perhaps the rest could have been gotten rid of by nonthermal escape and solar wind erosion. But if Mars lost its atmosphere too early (especially its ability to generate a dense CO_2 atmosphere) then it becomes hard to account for the large, extensive water-carved channels on Mars, some of which suggest persistence of active surface hydrology up to 3.5 billion years ago, with episodic recurrence of less extensive river networks extending billions of years later. More precise dating of these hydrological features, which will come ultimately with sample return missions from Mars, will go far to help resolve these puzzles. Still, the Mystery of the Missing Martian Atmosphere is likely to remain one of the Big Questions for a long while to come.

How do giant impacts fit into the picture? Giant impacts do not come in a continuous stream, but Lunar to Mars-sized bodies are common enough in the late stages of planetary formation that it is likely that one or more giant impact occurs before the planet attains its final size. The very existence of the Moon provides evidence that Earth experienced a giant impact, while the anomalous retrograde rotation of Venus has been taken as evidence that a giant impact occurred there as well. The Martian crust exhibits a striking dichotomy between rugged thick-crusted and heavily cratered Southern Hemisphere highlands and smoother, thinner Northern Hemisphere lowlands; this has sometimes been taken as having resulted from a giant impact, though one smaller in relative scale than Earth's Moon-forming impact. A single giant impact can blow off an entire atmosphere, but this is not inevitable; depending on the energy of the impactor, there can be a substantial proportion of the original atmosphere left. The issues in reconciling the histories of Earth and Mars are essentially the same as for the sweep stage impact erosion: how to account for the story of N_2 on Earth (or Venus, for that matter)? And how to account for the hydrology of Early Mars if a giant impact blew off the primordial Martian atmosphere but the planet was unable to regenerate a new CO_2 atmosphere by outgassing?

It should be kept in mind that impacts can also be an important *source* of volatiles. Comets can directly bring in volatiles such at CO_2, water and methane. Moreover, the high pressure

and temperature during the impact shock can cook water vapor out of hydrated minerals such as serpentine. Similarly, impacts can release CO_2 from carbonates in the crust. Carbonates are not primary planet-forming substances, but could be formed in the process of accretion through reaction between primary forms of carbon, oxygen, water and silicates. If this happens, CO_2 can be retained within carbonate even when some atmospheric loss event has blown away an earlier atmosphere, and this CO_2 can be released as a result of later impacts at a rate that can be much faster than the release associated with volcanic activity.

8.8 For Further Reading

Chapter 9

A peek at dynamics

Chapter Under Construction. There are some unmarked missing pieces.

9.1 Overview

So far, we have studiously avoided discussing the circulations of atmospheres or oceans, or indeed fluid mechanics of any type, with the exception of a brief foray into compressible one-dimensional hydrodynamics in Section 8.7.4. This is not because the subject is unimportant, but rather because the subject is *too important* to be relegated to the kind of superficial discussion we could accord it while doing justice to the rest of the physics governing the fluid envelopes of planets. This chapter provides a glimpse at what the reader has been missing. It highlights what the reader needs to keep in mind when learning atmosphere/ocean fluid dynamics, and for the student who has already acquired some familiarity with that subject connects fluid mechanical effects with the key planetary climate phenomena that have been the subject of this book. It is, in essence, a sampler of some of the many ways that large scale fluid dynamics affects planetary climate.

This being the final chapter (for now) of a long journey, we will also take stock of how well we have done at coming to an understanding of the Big Questions introduced in Chapter 1. We wrap up with a reminder of the great breadth of largely unexplored problems the reader is already equipped to take on. The universe of problems becomes all the richer once planetary fluid dynamics is brought into the picture.

9.2 Horizontal heat transport

From the standpoint of planetary climate, the most important effect of large scale fluid flow in a planet's fluid envelope is that it carries heat from one place to another, helping to even out the large geographical temperature variations that would otherwise be caused by variations in insolation between poles and tropics or dayside and nightside. There are two limits in which one does not need a detailed model of the horizontal heat transport. If the heat transport is very effective, then the temperature can be considered independent of geographical position and so the temperature profile can be determined using a single-column radiative-convective model driven by global mean top-of-atmosphere insolation. Venus is in this category. At the opposite extreme, the atmosphere

may be so ineffective at transporting heat that each geographical position can be considered to be in local radiative-convective equilibrium. Again, the climate can be determined using one-dimensional radiative-convective calculations, but this time one must run one such calculation for each geographical position and time period for which the temperature is desired. The weak-transport limit most commonly occurs for thin atmospheres, which simply haven't enough mass to transport heat effectively. Mars is an example of such a planet. Many planets of interest lie between the extremes, however. Earth is one such. For the intermediate cases, it is necessary to model the atmospheric and oceanic heat transport.

The discussion in this secton will emphasize atmospheric transports. Oceans also transport energy, and oceanic transports are often represented in simplified models in ways quite similar to the treatment of atmospheric transports to be given below. However, the nature of energy-transporting circulations in Earthlike oceans is different in key respects from the atmospheric case. The features that give oceanic circulations a distinct character include the near-incompressibility, the contribution of both salinity and temperature to buoyancy, primary driving by wind-stress and thermal exchange at the top boundary, and the presence of continental barriers to flow. One shouldn't be too hasty to generalize from the nature of Earth's oceans, though. On a waterworld, there would be no continental barriers, and it is not inevitable that salinity should have an important effect on density. Moreover, deep atmospheres which are optically thick both in the infrared and solar spectrum may share some features of the top-driven circulation of the ocean, at least so far as thermal exchanges go. To top it off, the very distinction between liquid ocean and gaseous atmosphere disappears when the pressure exceeds the critical point discussed in Chapter 2. The special features peculiar to oceanic circulations will be left for the reader to pursue elsewhere. The heat transporting circulations to be discussed below are abstracted from those occurring in the Earth's atmosphere. They typify the major heat-transporting mechanisms found in the atmospheres of all the Solar system bodies with substantial atmospheres, including Jupiter,Saturn, Venus and Titan; even Mars, with its thin atmosphere, exhibits basically the same features in one form or another. Similar features would apply to a great variety of hypothetical exoplanet atmospheres.

For atmospheres containing condensible substances, as is the case for water vapor on Earth, precipitation is an inevitable byproduct of horizontal heat transport, which occurs in part through tranport of latent heat between the portion of the planet where the substance is picked up by evaporation and the portion where it is deposited through condensation. This behavior is a horizontal version of the precipitation that occurs in connection with vertical transport of latent heat in a convecting column. Thus, the problem of determining precipitation distribution is intimately connected with – though not identical to – the problem o horizontal energy transport. The challenge there is to figure out what portion of the heat transport is due to sensible heat transport, and what part is due to latent heat.

The first thing to note about atmospheric energy transport is that the energy content of most known or contemplated atmospheres is dominated by the heat content, with transports of energy stored in the form of kinetic energy playing at best a minor role. For example, the energy per unit mass involved in changing the temperature of a dry Earth air parcel by ΔT is $c_p \Delta T$, while the entire kinetic energy content is $\frac{1}{2}U^2$ where U is the root-mean-square wind speed in the air parcel. The ratio of kinetic to heat energy is $\frac{1}{2}U^2/c_p\Delta T$, which works out to 0.006 if $U = 20m/s$ and $\Delta T = 30K$. Another way of putting it is that, if all the energy in a 20 m/s jet were dissipated and converted to heat, it would only raise the temperature of the air parcel by $0.2K$. Even this is an overestimate of the significance of heating by kinetic energy dissipation, since friction is weak throughout most of the atmosphere, so that it would take quite a long time to frictionally dissipate the kinetic energy. It is possible to envision circumstances where substantial parts of the

energy transport were in the form of kinetic energy, but the following discussion will adopt the conventional line of thinking, which neglects kinetic energy terms in the energy budget. *This does not mean that atmospheric and oceanic currents are themselves unimportant.* The kinetic energy stored in the form of organized winds may be negligible, but it is important that some of the planet's energy is in this form, since it is the winds that carry heat from one place to another.

9.2.1 A little fluid mechanics

In order to discuss the horizontal redistribution of heat by fluid motions, it is necessary to introduce a few basic fluid mechanical concepts. The development of this material requires more extensive use of partial differential equations and multivariable vector calculus than we have employed previously. It is useful background, but the reader who feels unequipped for this foray into the world of fluid mechanics can skip ahead to Eq. 9.11 and pick up from there.

We will adopt longitude λ and latitude ϕ as horizontal coordinates, and pressure p as the vertical coordinate. The latitudes and longitudes appearing in the fluid equations below are measured in radians. SInce we have adopted pressure as our vertical coordinate, all partial derivatives with regard to latitude, longitude and time are to be understood as occurring at constant pressure. At each point on the sphere, we decompose the horizontal velocity into a component u which points along the local latitude circle (eastward positive), and v which points along the local longitude circle (northward positive). u is called the *zonal wind*, and v the *meridional wind*. Note that u and v are conventional wind speeds measured in distance moved in the respective directions per unit time; they are *not* angular velocities such as latitude or longitude change per unit time, though they can easily be related to angular velocities. Since pressure is the vertical coordinate, the vertical motion is characterized by the *pressure velocity* ω, which gives the rate of change of pressure with regard to time, as seen by an observer moving along with a fluid parcel. Negative values of ω denote upward motion. Note that even for a flat surface, ω is not generally zero at the ground, since surface pressure generally is time-dependent.

The first expression we need is for the *material derivative*. Let $B(\lambda, \phi, p, t)$ be any quantity that depends on space and time; it could, for example, be an energy density, or the concentration of an atmospheric constituent. Then, its material derivative dB/dt is defined as the rate of change of B as seen by an observer who is blown along by the three-dimensional time-dependent wind field, meaning that the observer's velocity at any given point in space and time is equal to the wind velocity at the observer's current location. A straightforward application of the chain rule for differention shows that

$$\frac{d}{dt} \equiv \partial_t + \frac{1}{a\cos\phi} u \partial_\lambda + \frac{1}{a} v \partial_\phi + \omega \partial_p \tag{9.1}$$

We will also need an equation which represents the conservation of mass – the *mass continuity equation*. For hydrostatic flow, the time derivative drops out of this equation when it is written in pressure coordinates, leading to a simple and convenient form. When there is no mass loss due to condensation, the mass continuity relation reads

$$\frac{1}{a\cos\phi} \partial_\lambda u + \frac{1}{a\cos\phi} \partial_\phi v \cos\phi + \partial_p \omega = 0 \tag{9.2}$$

This expression says that the flow is nondivergent when written in pressure coordinates. When the relative mass loss due to condensation is small, as is the case for the present Earth, this formula remains approximately valid. When there is more substantial condensation, as for CO_2 on Early

Mars and CH_4 on Titan today, then a source and sink term due to formation of precipitation and evaporation/sublimation of falling rain or snow needs to be put on the right-hand side, making the flow divergent. The divergence introduced by strong condensation can introduce novel fluid mechanical effect which are essentially unexplored, but in the following we will stick to the simpler case in which condensation takes away minimal atmospheric mass.

Using Eq. 9.2, the expression for dB/dt can be recast in a flux form as follows:

$$\begin{aligned}\frac{dB}{dt} &= \partial_t B + \frac{1}{a\cos\phi}u\partial_\lambda B + \frac{1}{a\cos\phi}v\cos\phi\partial_\phi B + \omega\partial_p B \\ &= \partial_t B + \frac{1}{a\cos\phi}\partial_\lambda uB + \frac{1}{a\cos\phi}\partial_\phi vB\cos\phi + \partial_p\omega B\end{aligned} \tag{9.3}$$

Suppose now that B has a source $\mathfrak{S}_B$, so that $dB/dt = \mathfrak{S}_B$. Then, the vertically integrated conservation equation reads

$$\begin{aligned}\int_0^{p_s}\partial_t B\frac{dp}{g} + \frac{1}{g}[\omega(p_s)B - (\frac{1}{a\cos\phi}u\partial_\lambda p_s + \frac{1}{a\cos\phi}v\cos\phi\partial_\phi p_s)B] \\ + \frac{1}{a\cos\phi}\partial_\lambda\int_0^{p_s}uB\frac{dp}{g} + \frac{1}{a\cos\phi}\partial_\phi\int_0^{p_s}vB\frac{dp}{g}\cos\phi \\ = \int_0^{p_s}\mathfrak{S}_B\frac{dp}{g}\end{aligned} \tag{9.4}$$

in which we have divided both sides by g to emphasize that the vertical integral is in fact a mass-weighted integral, given the hydrostatic assumption. Now we make use of the fact that, by definition, $\omega(p_s) = dp_s/dt$, which makes the term in brackets reduce to $B\partial_t p_s$. This term is just what is needed to cancel the the term from the time dependence of the limit of integration p_s, which appears when one brings ∂_t outside the integral. The conservation equation can then be rewritten as

$$\begin{aligned}\partial_t\int_0^{p_s}B\frac{dp}{g} + \frac{1}{a\cos\phi}\partial_\lambda\int_0^{p_s}uB\frac{dp}{g} + \frac{1}{a\cos\phi}\partial_\phi\int_0^{p_s}vB\frac{dp}{g}\cos\phi \\ = \int_0^{p_s}\mathfrak{S}_B\frac{dp}{g}\end{aligned} \tag{9.5}$$

Finally, we take the zonal mean of the equation – i.e. the average along latitude circles – denoting zonal mean quantities with angle brackets. The resulting zonal mean conservation equation is

$$\partial_t\langle\int_0^{p_s}B\frac{dp}{g}\rangle + \frac{1}{a\cos\phi}\partial_\phi\langle\int_0^{p_s}vB\frac{dp}{g}\rangle\cos\phi = \langle\int_0^{p_s}\mathfrak{S}_B\frac{dp}{g}\rangle \tag{9.6}$$

To obtain the zonal-mean energy equation, we simply apply Eq. 9.6 to the moist static energy per unit mass, $\mathfrak{M}$, defined in Section 2.7.3. We will restrict attention to the case in which the concentration of the condensible component is small, so that the simplified form of the moist static energy applies and we need not worry about the loss of moist static energy through precipitation. The source term of the column integral of $\mathfrak{M}$ takes on a simple form, because $\mathfrak{M}$ is an exact differential. Specifically, suppose that the net vertical flux of energy by all means is $F(p)$ at pressure level p. Then, the energy source per unit mass is $S_\mathfrak{M} \equiv -gdF/dp$, with the convention that downward fluxes are positive. The vertically integrated source appearing in the transport equation then simply the difference between F at the bottom and top of the atmosphere. Another way of putting it is that, as discussed in Section 2.7.3, the change in the moist static

energy content of a column is given by the net energy put into the column through the upper and lower boundaries. Recall, however, that if we weren't working in the dilute limit, it would also be necessary to allow for the loss of energy through precipitation.

At the top of the atmosphere, let $F_{\circledast,top}(\phi,t)$ be the net zonal mean solar flux into the atmosphere (downward positive), and $OLR(\phi,t)$ be the zonal mean outgoing thermal infrared (upward positive, as usual). At the surface, let $F_{\circledast,s}(\phi,t)$ be the zonal mean solar flux exiting the bottom of the atmosphere; the difference with $F_{\circledast,top}(\phi,t)$ gives the atmospheric solar absorption. The heat flux out of the bottom of the atmosphere consists of turbulent as well as infrared radiative terms; call this net flux simply F_s, with the convention that a positive value indicates a transfer of heat from the atmosphere to the underlying surface. With these definitions the vertically integrated source term, which gives the rate of change of moist static energy in a column, becomes

$$\langle \int_0^{p_s} \mathfrak{S}_B \frac{dp}{g} \rangle = F_{\circledast,top} - OLR - F_{\circledast,s} - F_s \tag{9.7}$$

Substituting this into Eq. 9.6, and using the moist static energy in place of B, we obtain the energy balance equation

$$\partial_t E_{atm} + \frac{1}{\cos\phi} \partial_\phi \Phi_{atm} \cos\phi = \langle F_{\circledast,top} \rangle - \langle OLR \rangle - \langle F_{\circledast,s} \rangle - \langle F_s \rangle \tag{9.8}$$

where we have defined the atmospheric energy flux as

$$\Phi_{atm} \equiv \frac{1}{a} \langle \int_0^{p_s} v \mathfrak{M} \frac{dp}{g} \rangle \tag{9.9}$$

and the mean heat storage in the atmospheric column as

$$E_{atm} \equiv \langle \int_0^{p_s} \mathfrak{M} \frac{dp}{g} \rangle \tag{9.10}$$

It is convenient to introduce the coordinate $y = \sin\phi$, since then the element of area is simply $2\pi a^2 dy$. The atmospheric energy budget equation then becomes

$$\partial_t E_{atm} + \partial_y \Phi_{atm} \cos\phi = \langle F_{\circledast,top} \rangle - \langle OLR \rangle - \langle F_{\circledast,s} \rangle - \langle F_s \rangle \tag{9.11}$$

Note that $\cos\phi$ can be expressed as $\sqrt{1-y^2}$ if desired.

To complete the analysis, the atmospheric budget must be coupled into the surface budget. The simplest case is the one in which the surface has negligible thermal inertia, as in the case of a solid. Typically, in this case the surface budget can be considered to be in equilibrium, in which case $\langle F_{\circledast,s} \rangle + \langle F_s \rangle = 0$. When this assumption is valid, Eq. 9.11 involves only the top-of-atmosphere radiation budget and the atmospheric heat storage.

For a planet with an ocean, we need to allow for the thermal inertia of the surface and oceanic heat transports, both of which can through the surface budget out of local equilibrium. As a simple example of this situation, let's suppose that the oceanic thermal inertia can be represented by a mixed layer of uniform depth H. Suppose further that ocean currents transport heat meridionally at a rate Φ_{ocean}, defined analogously to Φ_{atm}. Most of this heat transport actually occurs below the mixed layer, and is communicated from the deep ocean to the mixed layer by vertical exchange of water. We will not go into the oceanic heat transport processes in any detail, and will be content with merely summarizing the effects on the heat transport, which will be sufficient for

the present purposes. Given that the ocean is heated and cooled only at the surface, the energy balance equation for the ocean becomes

$$\partial_t c_{p,w} H T_{mix} + \partial_y \Phi_{ocean} \cos\phi = \langle F_{\circledast,s} \rangle + \langle F_s \rangle \tag{9.12}$$

where $c_{p,w}$ is the specific heat of the liquid making up the ocean and T_{mix} is the mixed layer temperature.

Adding the ocean and atmosphere budgets, the surface exchange terms drop out, and we obtain the net column energy budget

$$\partial_t (E_{atm} + c_{p,w} H T_{mix}) + \partial_y \Phi \cos\phi = \langle F_{\circledast,top} \rangle - \langle OLR \rangle \tag{9.13}$$

where the net horizontal heat flux is $\Phi \equiv \Phi_{atm} + \Phi_{ocean}$. Note that the energy input and loss is now written entirely in terms of the top-of-atmosphere radiative exchange. This equation says that any top of atmosphere imbalance must be compensated by some combination of horizontal heat flux and change in the atmosphere/ocean energy storage. When the ocean thermal inertia dominates over the atmosphere, then any imbalance that is not accomodated by horizontal heat transport goes into changing the mixed layer temperature. When the ocean thermal inertia is small compared to that of the atmosphere, we are back in the same regime as the solid-surface case, in which the residual imbalance instead goes into changing the atmospheric heat storage.

9.2.2 Some observations

Since the top-of-atmosphere energy budget is purely radiative, the net top-of-atmosphere imbalance in the vicinity of each latitude and longitude point of a planet can be observed by satellite-borne instruments. If such data are averaged over a sufficiently long time period that the energy storage in each ocean-atmosphere column can be considered stationary, then the imbalance provides a direct measure of how much heat must be imported into the column or exported from the column in order to balance the budget. While it gives the net ocean/atmosphere dynamic heat flux, it does not say how the required fluxes are partitioned between the two fluids. Since the seasonal cycle repeats almost exactly from one year to the next, an annual average usually suffices to determine the annual mean dynamic heat flux, even on a planet with substantial thermal inertia in the atmosphere or ocean. The rapid increase in atmospheric CO_2 on Earth due to industrial CO_2 emissions has thrown off the Earth's radiation budget by a few W/m^2, but this is not a large enough effect to seriously compromise estimates of dynamical heat fluxes based on annual mean observations. Top-of-atmosphere data is not in itself sufficient to determine the seasonal cycle of heat transport, however. For example, the January mean energy imbalance for Earth would show a large net input of energy into the Southern Hemisphere, and a large net deficit in the Northern Hemisphere. This imbalance is not made up by a massive energy transport between the hemispheres; instead, it mostly reflects transient uptake of heat by the ocean in the summer hemisphere, and transient heat release from the ocean in the winter hemisphere.

So far, geographically resolved top-of-atmosphere radiation budget measurements are only available for the Earth. Zonal mean data from the ERBE satellite observations are shown in Fig. 9.1. The data are plotted as a function of the coordinate $y = \sin\phi$ because $dy = \cos\phi d\phi$ is proportional to the element of area on the sphere, so that the area under the energy imbalance curve between two values of y is the net energy imbalance integrated over that region. The net imbalance is very nearly symmetric about the equator. At the equator, the atmosphere receives about 60 W/m^2 more solar energy than it emits locally as infrared. At each pole, the situation is reversed, with the polar regions emitting about $100 W/m^2$ more than they receive as sunlight.

There is a net atmosphere ocean transport from the regions with $|y| < .6$ to the higher latitude regions. In terms of latitude, the division between the two regions occurs at $\pm 37^o$. The net top-of atmosphere imbalance, F_{toa}, can be quite well fit by the parabola $F_{toa} = 62.216 - 7.112y - 169.96y^2$. The global mean of this quantity is about $5W/m^2$ out of balance, which is comparable to the measurement error of the ERBE instruments. Note that the interhemispheric asymmetry is very weak. This is a quite mysterious result, and surprising in view of the considerable asymmetry in the continental distribution between the hemispheres. This behavior is as-yet unexplained, and the range of circumstances under which one should expect symmetric heat transport are not known.

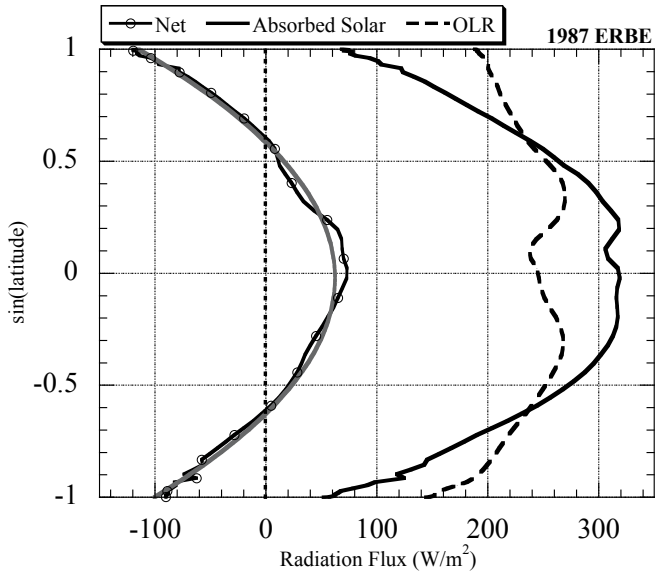

Figure 9.1: Top of atmosphere energy balance from the ERBE satellite dataset, annually averaged for 1987. The grey solid line is a quadratic fit to the net top of atmosphere imbalance.

Detailed analysis of surface fluxes indicate that over most of the planet, the heat transport implied by Fig. 9.1 is dominated by the atmosphere. Moreover, modelling studies indicate that if the oceanic heat transport is suppressed, the atmosphere takes up much of the slack. Hence, we are somewhat justified in focusing attention on the atmospheric transport mechanisms, though the reader should be aware that the oceanic heat transport can also have a significant effect on the climate, though we shall not discuss it here.

There are two main styles of atmospheric heat-transporting circulations, both of which occur on Earth, and both of which to varying degrees account for meridional heat transport on other planets as well. In the low latitudes, the heat transport is dominated by a zonally symmetric overturning *Hadley circulation*, which varies gradually over the course of the seasonal cycle. At higher latitudes, the heat transport is dominated by wavelike mobile *synoptic eddies* which have spatial scales on the order of a few thousand kilometers and propagate from west to east at a rate of 10-20 m/s. The synoptic eddies are evident on conventional weather maps as mobile high/low pressure systems, with their familiar systems of warm and cold fronts and associated weather.

Fig. 9.2 shows the climatological mean January and July Hadley circulation for the years 1960 to 1970, diagnosed from zonally averaged horizontal wind observations. The circulation is characterized by a mass streamfunction in the latitude-height plane. The winds blow along the

streamlines, and the mass flux between any two contours (in units of mass per unit time) is given by the difference between the streamfunction values on the two contours. Thus, the atmospheric currents transport mass more rapidly where the contours are more closely packed.

On the Earth, the Hadley cell is confined to the latitude range from $30N$ to $30S$, and effectively defines the tropics. The circulation reverses between the solstices. In each solstice, the summer hemisphere portion of the tropics receives more solar radiation than it radiates away locally, while the winter hemisphere portion of the tropics suffers a net loss of energy, both through an excess of infrared emission relative to local solar absorption, and because synoptic eddies transport heat out of the winter tropical regions. Heat transports by the Hadley cell make up the difference, carrying energy from the summer to the winter hemisphere across the equator. For climates like that of the present Earth, moisture is crucial to the way the Hadley cell transports heat. Observations show that above the boundary layer the temperature profile is very nearly identical between the ascending and descending branch of the Hadley circulation; the profile in both regions remains very close to the moist adiabat.

In the Earthlike regime, the transport of moist static energy comes from the contrast in humidity between the ascending and the descending branch. The ascending branch is near saturation, while there is little or no deep convection in the descending branch, whence there is little supply of moisture to the descending air. In consequence, the descending branch approximately conserves the moisture mixing ratio of the upper troposphere, and is very dry when it nears the ground. Ordinarily, dry descent would follow the dry adiabat, and hence by warmer than the ascending branch upon reaching the ground. Instead, the dynamical heat transport in the Hadley cell is efficient enough to keep the descending branch on the same temperature profile as the moist adiabat. This is possible because the subsiding air, which warms as it compresses, loses heat by infrared radiative cooling at the rate required to balance the budget. Indeed, the radiative cooling on the dry branch provided an important constraint on the strength of the circulation, since the radiative cooling must be balanced by compressional heating. The net result of this process is that the low level air in the subsiding branch has the same temperature as the low level air in the ascending branch, but is much drier; hence it has much lower moist static energy. By this process, the Hadley cell exports air with high moist static energy in the upper-level outflow, while it imports low level air with lower moist static energy. The Hadley cell would behave similarly for any planet that had a significant condensible component in its atmosphere, notably for the case of methane on Titan.

On a dry planet – for example Snowball Earth, which is dry because it is cold – the energy transport in the Hadley circulation must work differently. It must work by transporting dry static energy, and that requires that there be a difference in dry static energy between the upper level outflow and the lower level inflow. This requires a horizontal temperature gradient, at least at the outflow and inflow levels. On a dry planet, the circulation itself looks much the same as that shown in Fig. 9.2, but the thermal structure of the tropics is rather different than Earth's present tropics.

For a watery planet like Earth, the thermal inertia of the ocean also affects the amount of heat that must be transported by the Hadley circulation, and hence its strength. That is because the ocean is out of equilibrium in the course of the seasonal cycle. In the winter hemisphere, it is able to release stored energy to help make up the energy deficit of the atmospheric column, and this reduces the amount of dynamical heat transport required to balance the budget.

Synoptic eddy heat transport to be discussed here

An interesting feature of the heat imbalance data in Fig. 9.1 is that the energy transport is seamless at the edge of the Hadley circulation. There is no evidence of a break in the transport

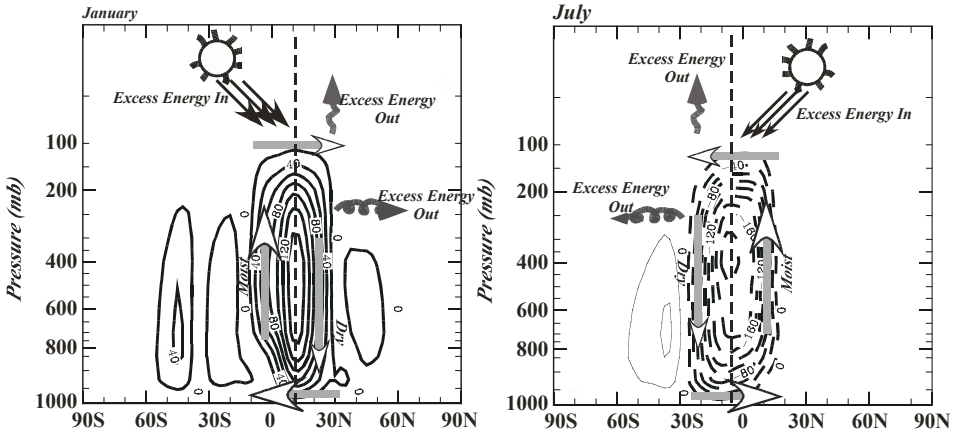

Figure 9.2: The January and July Hadley cell structure, based on 1960 to 1970 data for the Earth. The contours give the mass streamfunction in units of $10^9 kg/s$. Dashed contours are negative, and the arrows indicate the sense of the circulation.

mechanism, and the synoptic eddy transport picks up smoothly at the edge of the tropics as the Hadley transport dies out. This is probably because synoptic eddies penetrate significantly into the subsiding branch of the Hadley circulation, and indeed are an integral part of the heat and moisture budget there.

9.2.3 Scale analysis of heat transport

9.2.4 Formulation of energy balance models

In some sense, every climate model is an energy balance model. Full general circulation models could be described as energy balance models in which the heat transports are computed by solving the three-dimensional equations governing fluid motion. What is generally meant by the term "energy balance model," however, is a model in which some assumption is invoked to allow the heat transport to be computed rapidly without actually solving for the underlying flow. As part of this simplification, it is usually assumed that the thermal state of the atmosphere and ocean at each latitude can be represented by a single temperature T, most commonly taken to be the surface temperature.

The most common form of energy balance model represents the atmosphere-ocean heat transport as a thermal diffusion, namely

$$\Phi \cos\phi = D \cos\phi \partial_\phi T = (1 - y^2) D \partial_y T \tag{9.14}$$

The diffusivity D may depend on y, either directly or through the intermediary of T and its gradient.

To close the problem, it is necessary to write the OLR as a function of T, which requires making an assumption about the vertical distribution of temperature and humidity. For the atmosphere, this is typically done by assuming that the atmosphere is on the moist adiabat and that the relative humidity has some fixed value, as we did in computing the $OLR(T)$ fits in Table **??**. This also permits the atmospheric part of the vertical heat storage to be written as a function of T; the mixed layer part of the oceanic heat storage is by its nature proportional to T, if the mixed layer temperature doesn't deviate too much from the overlying air temperature. More problematically, one needs to make some assumption about the effects of clouds on albedo and OLR. There is really no satisfactory way to do this within an energy balance model, and so we will not discuss cloud effects in any great detail. Clouds are often taken into account through an adjustment of the albedo, but this can lead to erroneous conclusions if one tunes the adjustment to one climate and then applies it to an altered climate.

With the above simplifications, the energy balance model now reads

$$\partial_t E(T) = \partial_y[(1-y^2)D\partial_y T] + (1-\alpha)L_\circledast f(y,t) - OLR(T) \tag{9.15}$$

where α is the planetary albedo, and f is the flux distribution factor discussed in Chapter 7. For a rapidly rotating planet, the diurnally averaged form of f would typically be used, but for a planet whose day is comparable to its year the instantaneous value would generally be more appropriate. The albedo in general will depend on y, either directly, or indirectly through the effect of temperature on ice cover. If we write $\mu(T) \equiv dE/dT$, then the left hand side of Eq. 9.15 becomes $\mu\partial_t E$. The equation thus takes on the form of a one-dimensional time-dependent diffusion equation with a source term, and can be solved numerically using the same methods discussed in Chapter 7. In the following, we will confine attention to steady solutions.

9.2.5 Equilibrium solution of diffusive energy balance models

In a steady state, the zonal-mean diffusive energy balance model becomes

$$\frac{d}{dy}(1-y^2)D\frac{dT}{dy} = OLR(T) - (1-\alpha)L_\circledast f(y) \tag{9.16}$$

We have replaced the partial derivatives with ordinary derivatives, since this is now an ordinary differential equation. There are two distinct cases in which the steady state form of the equation is appropriate. First, steady state conditions apply approximately if the seasonal temperature fluctuations are weak, as would be the case for a planet with a deep mixed layer ocean or one with low obliquity in a nearly circular orbit. In that case, the flux factor $f(y)$ would be taken to be the annual mean flux factor such as displayed in Fig. 7.6. Even if the seasonal fluctuation are not small, the steady state equation will still govern the annual mean temperature if $OLR(T)$ and α are approximately linear in T. Alternately, steady state conditions apply instantaneously at each point of the seasonal cycle if the response time of the system is short compared to the planet's year, as would typically be the case for a planet with an all-solid surface having an atmosphere which is not too thick. In that case, $f(y)$ would be the diurnally averaged flux factor at any given point in the seasonal cycle, if the planet is rapidly rotating. If the response time is very short, or the planet's day is very long, then one might instead use the instantaneous flux rather than the diurnal average. The flux factor varies with time in the course of the seasonal cycle, but because the response time of the system is short the temperature has time to come into equilibrium with the solar flux received at each individual point of the seasonal cycle.

Eq. 9.16 is a second order equation and therefore requires two boundary conditions. Given that the physical problem is on a sphere, which has no physical boundaries, the notion of boundary

condition may seem somewhat peculiar. However, although the sphere itself has no boundaries, our *coordinate system* for the sphere has boundaries at the poles, $y = \pm 1$. As boundary conditions, we require that no energy leak out of the system through the poles, namely that $(1 - y^2)DdT/dy$ vanish at $y = \pm 1$ It might seem that this condition is tautologically satisfied in view of the factor $(1 - y^2)$, but the condition can be violated if dT/dy blows up at one or more pole. The no-flux condition is thus equivalent to the requirement that dT/dy be well-behaved at the poles. In the special case where the forcing is symmetric about the equator, one of the polar conditions can be replaced by the requirement that $dT/dy = 0$ at $y = 0$.

With the specification of the boundary conditions, Eq. 9.16 becomes a nonlinear two-point ODE boundary value problem. The nonlinearity creeps in through $OLR(T)$, and perhaps also through the temperature dependence of the albedo. One needs to find a solution that simultaneously satisfies both boundary conditions, but numerical integration starting at one of the boundaries requires the specification of *two* initial conditions. In this situation, iterative methods can be used to find a solution. One first defines the auxiliary variable $\Phi \equiv (1 - y^2)D\frac{dT}{dy}$, and writes the problem as a pair of first-order equations in the two variables Phi and T. The equation for T is $dT/dy = D^{-1}\Phi/(1 - y^2)$, which becomes singular at $y = \pm 1$. The simplest way to deal with this problem is to start the numerical integration at a value y_o which is very close to, but not at, the pole; one must use a corrspondingly small numerical integration step in order to resolve the near-singularity. With that trick in mind, one carries out the integration starting at y_o, using $\Phi = 0$ and a *guess* T_p of the value of temperature at the pole. Upon integrating to the other pole, it will generally be found that Φ is not zero there. One completes the solution by iterating on T_p using Newton's method until the value of Φ at the opposite pole is acceptably small. In the case of a symmetric climate, the procedure is the same except that one integrates only to the equator and imposes the condition $\Phi = 0$ there instead.

If the OLR is written in the linearized form $OLR(T) = a_o + a_1 \cdot (T - T_o)$ and the albedo is independent of T, then Eq. 9.16 becomes a linear system. Note that it remains linear even if the albedo depends on y. In the linear case, the problem can be solved by superposing two suitably chosen solutions, without the need for an iteration. Ice-albedo feedback would make the problem nonlinear, but we'll see shortly how one can deal with that through iterating on the ice margin.

Some numerical solutions of the linearized energy balance model are shown in Fig. 9.3. The OLR coefficients are chosen to correspond to a moist Earthlike atmosphere with 50% relative humidity and a CO_2 concentration of 300 $ppmv$, and a diurnally/annually averaged form of f corresponding to the Earth's present obliquity was assumed. The left panel shows results for a waterworld with uniform albedo. With $D = 0.25W/m^2$, the tropical temperature is much warmer than observed on Earth, and the gradient in temperature between pole and equator is large. As D is increased, the temperature becomes more uniform. For $D = 0.5W/m^2$, the temperature gradient is fairly realistic, but but the tropics is still too warm, and the pole somewhat less so. For $D = 1W/m^2$, the tropics is somewhat cooler, but the poles are now too warm and the temperature gradient too weak, as compared to observations. When polar ice is introduced in the right panel of Fig. 9.3, the polar temperature drops, but because of heat diffusion the tropical temperature is dragged down as well. With polar ice, the case $D = 1W/m^2$ now looks fairly realistic. Note that the heat diffusion smooths out the discontinuity of temperature one would otherwise get at the ice margin. When diffusivity is weak, one sees a stronger temperature contrast across the ice margin. For weak diffusivity, the polar ice also has less effect on tropical temperatures, because the two regions are more thermally decoupled.

The results for the polar ice case shown in Fig. 9.3 are inconsistent, because the ice-margin temperature is not at the freezing point. For $D = 0.25W/m^2$ the ice margin temperature is below freezing, and the ice margin will tend to advance and cover more of the globe. For $D = 1W/m^2$

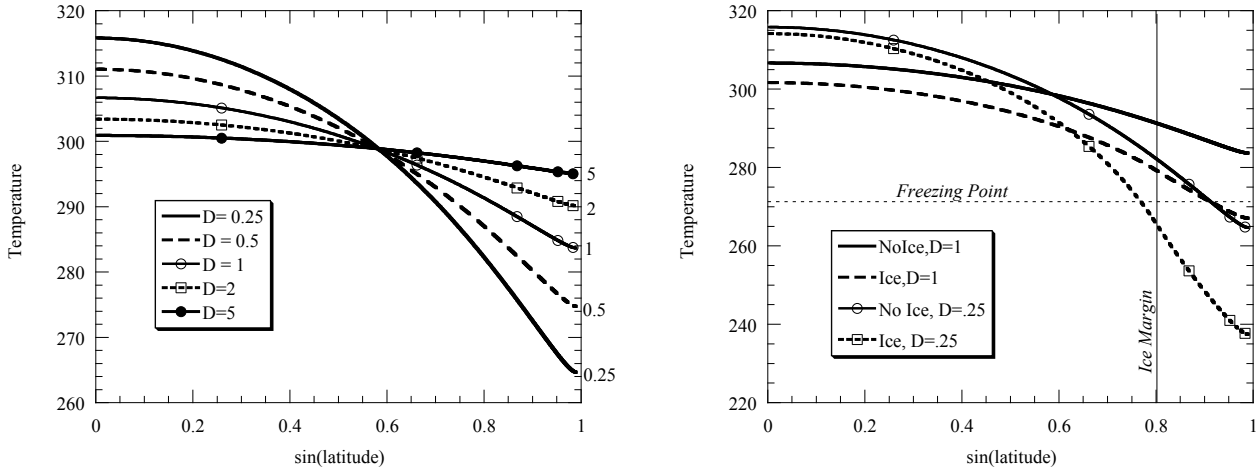

Figure 9.3: Solution to the equilibrium energy balance model for a fixed albedo of 0.15 (left panel) and for a case with ice where $|y| > .8$ (right panel). The ice albedo is 0.6. Results are shown for several different values of the diffusivity coefficient D (in W/m^2), indicated on the curves.

the ice margin temperature is above freezing, so the ice margin will tend to retreat toward the poles. In order to find the equilibrium ice margin, one must perform a series of calculations, and determine the ice margin temperature as a function of the ice margin position; the equilibria are determined by the intersection of the curve with the freezing point temperature. There are also two limiting cases where the ice margin temperature can consistently be different from the freezing point. If the ice "margin" is at $y = 0$, then the planet is in a Snowball state, and a state with the temperature at $y = 0$ below freezing is consistent since the ice has no more room to advance. At the opposite extreme, if the ice "margin"is at $y = 1$ the temperature is free to be above freezing, since that corresponds to an ice-free world with no more room for the ice to retreat.

A calculation of the ice margin temperature curve is shown in Fig. 9.4. For the highest diffusivity case, $D = 2W/m^2$, the curve is monotonically increasing in y_{ice}. There is a stable snowball state for $y_{ice} = 0$ and a stable ice-free state for $y_{ice} = 1$. There is also an additional partially ice-covered equilibrium state, but it can readily be seen that this is an unstable equilibrium. A displacement of the ice margin poleward causes the ice margin temperature to be above freezing, leading to further retreat until the planet falls into the ice-free state. Similarly, a displacement toward the equator causes the planet to fall into the Snowball state. As the diffusivity is reduced toward $0.5W/m^2$, the curve develops a local maximum, but the dip does not cause further intersections with the freezing line until D is increased further. For the low diffusivity case $D = 0.25W/m^2$, the ice-free state no longer exists, but there is now a stable state with a small polar ice cap.

The general behavior is qualitatively like that we saw in the ice-albedo bifurcation diagram based on the zero-dimensional energy balance model in Chapter 3. Now, however, we have the advantage that the behavior is more closely tied to the actual distribution of solar radiation over the planet's surface, and the magnitude of the heat transport. Note that the case $D = 0.5W/m^2$, which is in some sense the most like the real Earth in terms of its temperature structure, just misses having a stable small polar ice cap. This suggests that our present climate state may be rather fragile, with modest effects from clouds or ocean heat fluxes pushing the bifurcation behavior one

way or another. Seasonal effects, which we have also left out of the picture, may also play a decisive role.

Based on general energy balance reasoning, we can anticipate that high obliquity states will be less favorable to polar ice, because the poles receive more solar energy in those cases; conversely, low obliquity states will be more favorable to polar ice. For any given obliquity, low diffusivity will favor polar ice if the planet is fairly warm, since less tropical heat invades the polar region. On the other hand, if the planet is quite cool, the situation is reversed, and high diffusivity can cause the planet to fall into a Snowball state, since some of the heat from the tropics can be diffusively bled off into the bitterly cold poler regions, causing the tropics to freeze over. A more complete exploration of the ice-albedo bifurcation diagram associated with the diffusive energy balance model is carried out in Problem **??**

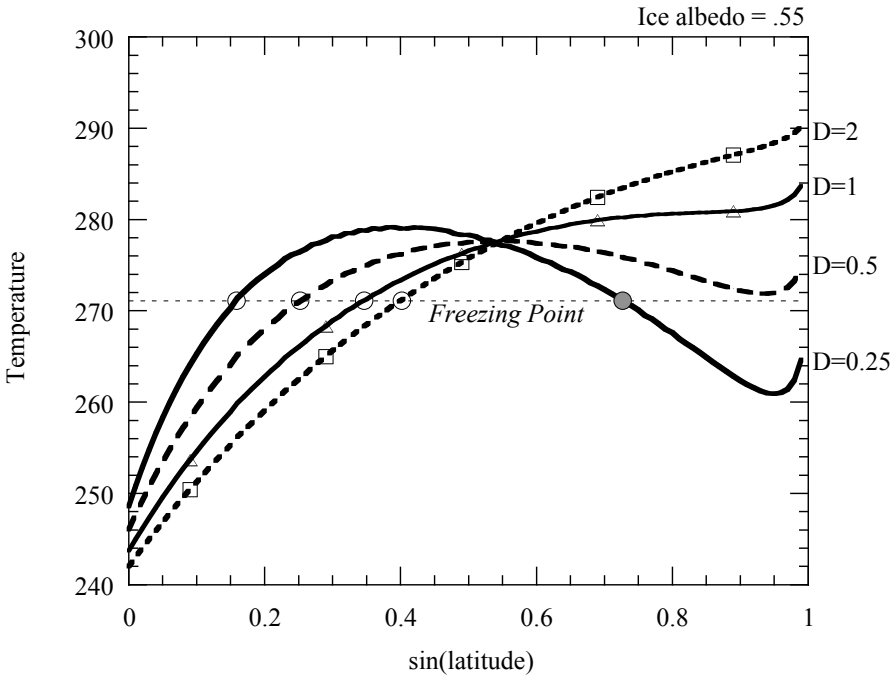

Figure 9.4: The temperature at the ice margin as a function of the ice margin position, computed for four different values of the diffusivity (values indicated on curves, in W/m^2K.). The ice albedo was taken to be 0.55 and the oceanic albedo was fixed at 0.15. Unstable equilibrium points are indicated with open circles, while the stable equilibrium point is indicated with a shaded circle. In all cases the Snowball state ($y_{ice} = 0$) is also a stable equilibrium. Except for the lowest diffusivity case, the ice-free state is likewise a stable equilibrium.

9.2.6 Limitations of diffusive energy balance models

9.2.7 Alternative approaches to modeling heat flux

9.3 Dynamics of relative humidity

9.4 Dynamics of static stability

9.5 Big questions:How are we doing?

Chapter 10

Appendix: Notation