ResearchOnline@JCU

This file is part of the following reference:

McLean, John D. (2017) *An audit of uncertainties in the HadCRUT4 temperature anomaly dataset plus the investigation of three other contemporary climate issues.* PhD thesis, James Cook University.

Access to this file is available from:

https://researchonline.jcu.edu.au/52041/

The author has certified to JCU that they have made a reasonable effort to gain permission and acknowledge the owner of any third party copyright material included in this document. If you believe that this is not the case, please contact ResearchOnline@jcu.edu.au and quote **https://researchonline.jcu.edu.au/52041/**

JAMES COOK UNIVERSITY AUSTRALIA

An audit of uncertainties in the HadCRUT4 temperature anomaly dataset plus the investigation of three other contemporary climate issues

John D. McLean
B. Arch (Univ. Melbourne, 1979)

A thesis submitted for the degree of Doctor of Philosophy

College of Physics, James Cook University, Townsville

August 2017

Acknowledgements

This thesis is dedicated to the memory of the late Robert ("Bob") Carter, former professor of geology at James Cook University, who first suggested that I should attempt a Doctorate of Philosophy and was my initial supervisor.

I thank my supervisor, Professor Peter Ridd, for his encouragement, general guidance and patience.

I also thank Dr Phil Clarke of Warrandyte, Victoria, for his proofreading of an early draft and indicating where in his opinion the text could be improved.

Statement of the Contribution of Others

No other person made any contribution to the work contained in this thesis.

Copyright Statement

Every reasonable effort has been made to gain permission and acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.

As far as I am aware, all owners of copyright material are duly acknowledged in the list of references (cited papers) or footnotes (less stringent material). Chapter 9 of this thesis is reformatted copy of a published paper of which I was the sole author, the copyright of which is held jointly with the online journal in which it appears.

Abstract

This thesis is in two parts. The first discusses many uncertainties associated with the widely used HadCRUT4 global temperature dataset. The second part deals with three other issues in climate science, viz (i) the possible relationship between cloud cover and global average temperature; (ii) a better indicator of the El Nino-Southern Oscillation; and (iii) the probability of severe coral bleaching along the Great Barrier Reef prior to 1998.

Part 1

The HadCRUT4 dataset provides information about temperatures over as much of the Earth's surface for which historical temperature recordings are available. It is regularly cited by government authorities and by organizations such as the Intergovernmental Panel on Climate Change (IPCC) and the United Nations Framework Convention on Climate Change (UNFCCC) and yet very little appears in the scientific literature about its errors and uncertainties other than from within the organizations responsible for its collation. Authors of the IPCC's Fifth Climate Assessment Report (2013) admitted during the review process for that report that no audit of the HadCRUT4 dataset or any associated dataset had been undertaken.

Given that governments are deciding energy and climate policies on claims based on the HadCRUT4 dataset, an independent audit of its accuracy and uncertainties was undertaken using data from 1850 to 2015. The audit covers a broad range of issues but leaves the quantifying of the impact of such errors to others, save for some general comments about the direction of changes in error margins.

Bespoke software revealed many areas of concern. Data coverage was found to vary between 12% and 91% of the Earth's surface, which means that any declared "global average" assumes that temperatures over the part of the world for which no data was available were identical to that average. It also assumes that meaningful trends can be calculated even though the coverage varies. Also revealed was that when coverage

was low it was found to be concentrated on particular regions, meaning for example that less than 13% of a hemisphere might account for more than 60% of the data coverage for that hemisphere in a given month.

Sample size was another problem. A single observation station in the entire Southern hemisphere reported data in the first three years of the HadCRUT4 record and only nine were reporting by 1859. The data is processed and presented on a monthly basis as values for each grid cell, each of which covers 5° latitude x 5° longitude. More than 30% of the grid cells derived from sea surface temperature do so from 1850 to about 1950 on the basis of from one to five measurements in the entire month.

Outliers were also discovered in the data, even in the 30-year period from 1961 to 1990 over which long-term average temperatures are calculated and sometimes in the longer period over which standard deviations were calculated. Their presence in these periods firstly widens the range of acceptable values (i.e. includes the inclusion of other outliers) and secondly distorts the crucial long-term average temperatures.

Numerous other inconsistencies and uncertainties were identified, including differences between dataset values that should be identical, unexplained differences between land and sea temperatures, inconsistent sources of coastal temperature data, questionable "bulk adjustment" to sea surface temperature data, poor data quality control and possible errors in the processing of land and sea temperature data prior to their submission for inclusion in the dataset.

Over 25 findings are presented in the summary to part 1of the thesis, the most serious being that HadCRUT4 global averages prior to 1950 are of limited value because for almost all of the period from 1850 to 1950 the coverage of the Earth's surface was less than 50%. The summary also proposes how a new historical dataset with fewer uncertainties and inconsistencies might be constructed from some but not all of the existing data.

Part 2

Three topics are covered in this section. The first is a published paper that attributes the pattern in HadCRUT4 global average temperature anomalies since 1950 firstly to a change in the El Nino-Southern Oscillation after 1977 from having few El Nino events to having many, then from 1987 to 1997 to a general reduction in cloud cover. Since 1997 the trend in global average temperature is minimal but low-level cloud cover has decreased while mid to upper level cloud cover have correspondingly increased.

The second topic focuses on an alternative to the commonly used Troup Southern Oscillation Index (SOI), which is used to indicate the state of the El Nino-Southern Oscillation (ENSO) and is based on sea-level pressure (SLP) at just two locations - Darwin (Australia) and Tahiti (French Polynesia). It proposes that an ensemble index be created from data at six other locations, three in the eastern Pacific and three in the western Pacific or further west. The ensemble index is shown to have three significant advantages. Firstly, it provides better coverage of the entire region where mean sea level pressure is directly related to ENSO conditions. Secondly, and somewhat related, is that data noise such as that caused by small-scale weather conditions at one or two locations is less likely to have a major impact on the ensemble SOI than in the Troup-SOI. Thirdly, and less significant, it is refining an existing index rather than introducing a new one.

The third topic addresses the often-repeated claim that severe and extensive coral bleaching on Australia's Great Barrier Reef (GBR) driven by warm water first occurred in 1998. Regular aerial surveys of the GBR did not commence until the late 1990s so to address the question of earlier severe bleaching four different techniques are used.

The first approach uses statistical and probability argument based on known recent bleaching events and Willis Island temperature data to conclude that the probability of bleaching in one or more years of the period from 1922 to 1997 is at least 0.6.

The second approach considers previous instances of El Nino events during summer months, such events being related to bleaching overseas and along the GBR in recent years. El Nino events during eighteen summer months over the period 1939 to 1997 were identified, the strongest occurring in 1983 when extensive coral bleaching was reported in the eastern Pacific near Panama and in waters off Indonesia.

The third considers the number of warm summer days and the maximum length of sequences of warm days according to temperature records from the observation station on Willis Island, about 350 km east of the GBR. Warm summers prior to 1998 include those which ended in 1944 and 1964, the latter coming immediately after a short El Nino event that weakened in January of that year.

The final approach was to analyse sea temperature records in the ICOADS database that applied to summer and the GBR, with the data divided into three bands, each spanning 5° of latitude. An examination of the proportion of warm days revealed that of the 37 observations made in the northern third of the reef during March 1970 33 include reports of temperatures ≥29°C and 25 of temperatures ≥30°C. Sequences of warm days were also examined, this revealing sea temperatures ≥30°C on nine successive days in the middle third of the reef commencing 25 December 1923 and seven successive days in the northern third of the reef in late March 1983.

Contrary to various claims, severe coral bleaching on the GBR prior to 1998 seems likely with a probability of ~0.6 that bleaching occurred in one or more summers from 1922 to 1997, with the other methods suggesting it was most likely in the summers ending in 1983 and 1963.

***

Contents

List of Tables

List of Figures

Part 1

An audit of the Issues of the HadCRUT4 temperature dataset and associated data files.

Chapter 1: Introduction

1.1 Introduction

The money spent addressing climate change is sometimes truly staggering. In 2009 it was estimated that the US had spent $79 billion since 1989 on policies related to climate change[1]. In 2015 climate change was said to be a $1.5 trillion industry[2]. In March 2016 it was announced that the Australian government had committed AUD$2.55 billion for carbon abatement within the country[3] and a further $1 billion over five years to support developing countries build resilience to climate change and reduce their carbon dioxide emissions. A meeting of the IMF in October 2016 talked of mobilizing USD$100 billion in annual financing flows from rich countries to developing economies by 2020 and mentioned one estimate suggesting that around USD$90 trillion will need to be invested by 2030 in infrastructure, agriculture and energy systems, to meet the requirements of the Paris Climate Treaty of 2015[4].

Underpinning all this spending is the proposition that temperatures have risen to a dangerous level and that mankind is to blame. The source of the temperature data in question is often the HadCRUT4 dataset, which is cited by the Intergovernmental Panel on Climate Change (e.g. IPCC, 2013) and by other government and non-government organizations. Despite the great reliance on this data it appears that no independent in-depth audit or analysis of its accuracy has ever been undertaken, save for a paper that addresses some, but not all, issues.

The author became aware of the situation when he was Expert Reviewer of IPCC 5AR, the IPCC's fifth climate assessment report, published in 2013. In response to

[1] http://scienceandpublicpolicy.org/images/stories/papers/originals/climate_money.pdf
[2] http://joannenova.com.au/2015/07/spot-the-vested-interest-the-1-5-trillion-climate-change-industry/
[3] http://www.abc.net.au/news/2016-05-05/government-spends-$500m-reducing-carbon-emissions/7388310
[4] http://newsroom.unfccc.int/unfccc-newsroom/the-paris-agreement-will-soon-enter-into-force-now-we-need-to-move-the-money/

two comments about an independent audit of key data the IPCC authors responded as follows[5]:

> **Comment 2-1106** ... Did the IPCC seek an independent audit of HadCRUT3 (of [sic] HadCRUT4) data prior to citing it? If not, why not? ... [John McLean, Australia]
>
> **IPCC Author:** Rejected. The assessment is charged with undertaking a holistic literature review. Not undertaking audits of each and every dataset which would be an impossible task with solely voluntary effort on the part of the (C)Las [i.e. the report authors].

And ...

> **Comment 2-1256** Did you audit HadSST3 before using it? ... [John McLean, Australia]
>
> **IPCC Author:** Rejected. Details of data quality control for HadSST3 is presented in Kennedy et al. (2011c) which is referenced in the previous paragraph. It is not the job of IPCC to audit each and every dataset in substantive detail.

Regarding the response to the second comment it should be noted that "Kennedy at al (2011c)", is a paper with five authors, four of whom at the time of writing the paper were employees of the Hadley Centre, the creators of the HadSST3 dataset, which means that the evaluation was not independent. Further, the issue is not about "data quality control" *per se* but accuracy and reliability.

In 2010 the UK parliament undertook an inquiry into the so-called "Climategate" emails, the unauthorised disclosure of many emails from people at the UK's Climate Research Unit (CRU). Conclusion three of the inquiry's report[6] said:

---

[5] Review comments and responses at http://www.ipcc.ch/pdf/assessment-report/ar5/wg1/drafts/Ch02_WG1AR5SOD_RevCommResponses_Final.pdf

[6] Available via http://www.publications.parliament.uk/pa/cm/cmsctech.htm

> A great responsibility rests on the shoulders of climate science: to provide the planet's decision makers with the knowledge they need to secure our future. The challenge that this poses is extensive and some of these decisions risk our standard of living. When the prices to pay are so large, the knowledge on which these kinds of decisions are taken had better be right. The science must be irreproachable.

I argue that it's not simply the scientific knowledge that needs to be irreproachable but also the data that underpins the science and justifies actions to address the matter. In the context of issues of climate change, if the HadCRUT4 is shown to be inaccurate or uncertain it might change the manner in which climate issues are addressed, either to increase efforts, decrease them, or redirect them into other areas.

1.2 The structure, tools and limits of this thesis

The aim of Part 1 of this thesis is to try to establish the general credibility of the HadCRUT4 temperature dataset. It is an initial analysis of areas of concern ranging from macro-scale items, such as data coverage and the number of samples (e.g. number of reporting observation stations or sea surface observations), down to the micro-scale issues of data outliers and temperature adjustments at individual stations. Some issues will impact "big picture" issues like hemispheric or global averages. Others will impact "small picture" issues such as temperature patterns for very small regions or even individual observation stations. While the small issues might individually be minor, in sufficient number and of sufficient magnitude even they can impact the "big picture issues".

This thesis makes little attempt to quantify the uncertainties exposed by this investigation, save for some brief mention of the impact certain issues might have on error margins, because numerous issues are discussed and it would be an enormous the task to quantify the uncertainties associated with the many instances of each. It has been left to others to quantify the impact of incomplete data, inconsistencies,

questionable assumptions, very likely data errors and questionable adjustments of the recorded data.

The primary data sources for this thesis are the HadCRUT4 dataset and associated HadSST3 and CRUTEM4 datasets. All three datasets are updated each month, typically to add very recent data but sometimes to revise historical data, but for consistency the end-of-year datasets for 2015, which is to say the state they were after the update in January 2016, are used throughout.

The chapters of Part 1 are organised in a sequence that first addresses the large scale issues with the HadCRUT4 dataset - coverage, data quantity, the calculation of long-term average temperatures - followed by chapters on lesser but still important issues, including various inconsistencies and issues specific to each of land-based temperature data from observation stations or sea surface temperature data. Two subsequent chapters prior to the summary chapter look in detail at the source data from which the various datasets are constructed because flaws in the source data could easily be carried through to the completed dataset.

1.3 Basic concepts

1.3.1 Introduction

The understanding of certain basic concepts is necessary in order to fully appreciate several matters raised in this thesis so this section is an introduction to those issues.

1.3.2 Data and information sources

The HadCRUT4 temperature anomaly dataset is derived from the data used in two other datasets, the HadSST3 dataset and the CRUTEM4 dataset. The former is of

temperature anomalies for sea surface temperatures and is created and maintained by the UK's Met Office Hadley Centre for Climate Science and Services (more commonly just "Hadley Centre"). The latter is derived from data obtained by land-based observation stations and processed by the Climatic Research Unit of the University of East Anglia (and commonly known as "the CRU").

The HadSST3 web page directs users wanting details to Kennedy et al (2011a) and Kennedy et al (2011a) but a more detailed reference for information about the processing is Rayner et al (2006). The primary sources for detailed information about the CRUTEM4 dataset are Jones et al (2012) and Osborn & Jones (2014). Information about the composite dataset, HadCRUT4, which is constructed from the data for each of these two datasets can be found primarily in Morice et al (2012). Detailed discussion of these references follows later in this chapter.

Only McKitrick (2010) seems to offer much in the way of independent examination of the temperature anomaly dataset and its construct, albeit based on HadCRUT3, the version prior to HadCRUT4.

1.3.3 Grid-based system

Like the CRUTEM4 and HadSST3 datasets, the HadCRUT4 dataset is grid-based with data values for each grid cell (or flagged as missing). The three datasets all use grid cells of 5° latitude x 5° longitude, with full coverage of the Earth's surface requiring 36 grid cells from north to south and 72 from west to east, a total of 2592 grid cells.

The advantage of using a grid-based methodology is that there is no bias towards grid cells with a greater amount of raw data over those with less raw data. On the other hand one disadvantage is that the cell boundaries are fixed and each cell is processed in isolation from its neighbours despite the possibility that an observation station in a neighbouring grid cell might be physically closer to more of the target grid cell than the stations within that cell.

1.3.4 Grid cell types

The HadCRUT4 data are from two basic sources, fixed observation stations (almost exclusively on land) or measurements of sea surface temperature, but the grid cells are of three types: (1) "Land" grid cells, which exclusively cover regions of land, for which the data comes from the CRUTEM4 dataset, (2) "sea" grid cells, which exclusively cover regions of sea grid and for which the data is from the HadSST3 dataset, and (3) "coastal" grid cells for which the data will be a merging of HadSST3 and CRUTEM4 data if both have data for that cell and otherwise whichever is available.

When both sources are used for coastal grid cells the approach in HadCRUT4 is to weight the two values according to the fractional area of each, with a minimum land data weighting (i.e. weighting for CRUTEM4 data) of 25%. This contrasts with the approach from the HadCRUT3, the previous form of HadCRUT data, in which the data from the two sources was weighted in inverse proportion to error variance. The change goes some way towards recognising that some islands are only a few percent of the total area covered by the grid cell.

The number of grid cells and the percentage of the Earth's surface that each cell type covers are shown in Table 1-1 and a map to indicate the grid cells is shown in Figure 1.1. HadSST3 data can apply to both "sea" and "coastal" and therefore at a maximum can cover 81.9% of the Earth's surface, while the maximum cover of CRUTEM4 data will include both "land" and "coastal" grid cell types and is therefore 46.2%

	Cell Type	Cell Count	Percent of cells	Percent of Earth's Surface
Northern Hemisphere	Land	293	22.61%	24.32%
	Sea	560	43.21%	41.79%
	Coastal	443	34.18%	33.89%
Southern Hemisphere	Land	270	20.83%	11.94%
	Sea	752	58.02%	65.91%
	Coastal	274	21.14%	22.15%
Global	Land	563	21.72%	18.13%
	Sea	1312	50.62%	53.85%
	Coastal	717	27.66%	28.02%

Table 1-1 Grid cell types and the number grid cells, percentage of grid cells and the percentage of the Earth's surface they cover.

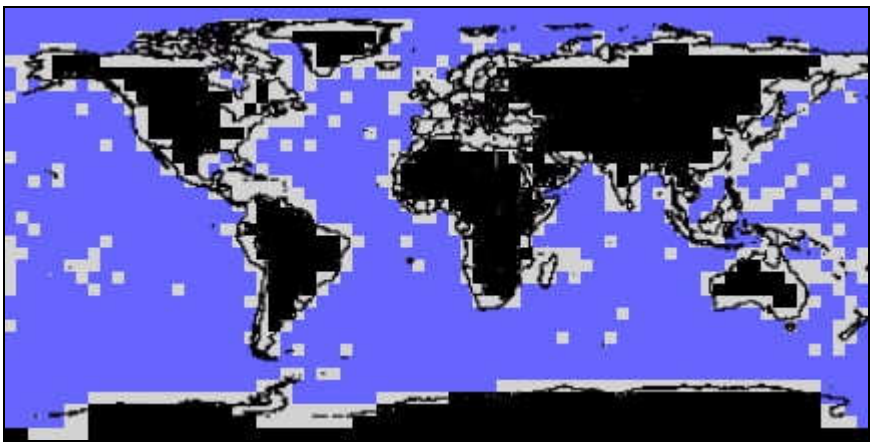


Figure 1.1 Map of the three types of grid cells, with 'land' grid cells in black, 'sea' grid cells in blue and 'coastal' grid cells in grey.

1.3.5 Concept of temperature anomalies

The HadCRUT4 dataset and its two source datasets (CRUTEM4 and HadSST3) are based on temperature anomalies because this method takes into account variations in

temperature caused particularly by latitude and altitude but also by exposure (e.g. to winds).

In general terms, all temperature anomalies are calculated according to

$$T_{anom} = T_s - T_{base}$$

where T_s is a specific temperature and T_{base} is the base or reference temperature.

For HadCRUT4, CRUTEM4 and HadSST3 datasets the data is in reference to specific months with T_s being the mean temperature in a given month and T_{base} the long-term average for the same calendar month, producing the temperature anomaly for the given month.

For CRUTEM4 data, i.e. data obtained from observation stations, both the base temperature and the anomaly are calculated from the mean monthly temperatures, which are the average of the mean daily minimum temperature and the mean daily maximum temperature across the month. The base temperature, T_{base} when calculating the anomaly for a given month is the average of the monthly mean temperatures for the same calendar month across the period from 1961 to 1990 inclusive (e.g. when calculating a temperature anomaly for January the base temperature is the average in each January from 1961 to 1990 inclusive). More than one station might be located in any given CRUTEM4 grid cell and so the cell's value in a given month is the average of the temperature anomalies for all reporting observation stations within that grid cell in that month.

A different approach is used to calculate sea surface temperature anomalies for the HadSST3 dataset. While the dataset is expressed on a grid cell size of 5° x 5° and by month, the values are derived from temperature anomalies calculated using 1° x 1° grid cells (here called sub-cells for convenience) and 5-day intervals (also known as pentads). Each calendar month has a notional six pentads except for August, which is assigned seven, making 73 pentads per 365-day year. Anomalies are separately calculated for each sub-cell and pentad, with the base temperature, T_{base}, estimated

mathematically and subsequently modified by actual SST measurements made during the period from 1961 to 1990 within the sub-cell and pentad.

The sea surface temperature anomaly expressed in the HadSST3 dataset at 5° x 5° grid cell resolution and in whole months is determined via a two-step process. The location and day of the measurement are first converted to a sub-cell and pentad, then the long-term average for the sub-cell and pentad used to calculate the sub-cell anomaly. The second step is to average and interpolate the anomalies in order to convert it to HadSST3 5° x 5° grid cells at monthly intervals, the interpolation being necessary at the start and end of the given month if it fails to align with pentad boundaries.

1.3.6 Ensemble Approach

In previous versions of the HadCRUT dataset the published dataset was calculated directly from the relevant HadSST and CRUTEM datasets, each of which were directly calculated from the input data.

In recognition that some uncertainties are unknown or even unknowable, the latest version of the HadSST dataset, i.e. HadSST3, was created as an "ensemble" of 100 different datasets, each of which was constructed according to different assumptions, particularly about the proportion of the use of two alternative methods for measuring sea surface temperature that produce different results and necessitate adjustment to data obtained by one method to make it compatible with the other method. The CRUTEM4 raw data required no such modification but mathematical simulating of uncertainties and potential biases was used to produced multiple datasets.

HadCRUT4 dataset uses the same technique of 100 different datasets, each constructed according to varying assumptions used for the HadSST3 and CRUTEM4 data, with the mean value of the values in each dataset published as an 'ensemble mean'. While the use of multiple datasets might be useful for establishing the sensitivity to certain assumptions ultimately the "ensemble" datasets are the averaging

of data derived from (presumably) one correct and ninety-nine incorrect datasets, although it is impossible to determine which grouping a given dataset belongs in, and whose assumptions often conflicted with each other.

The use of "ensemble mean" datasets makes difficult any comparisons between datasets because of the averaging of values. When comparisons are used in this thesis it is recognised that the averaging might cause small differences between the datasets and therefore certain tolerance thresholds (e.g. 0.05°C) are set and only differences of greater magnitude are considered to be genuine differences.

1.3.7 Calculations for a (near) spherical Earth

The overlaying of a two-dimensional grid onto the Earth's shape of an oblate spheroid of the Earth has implications for the calculation of both the data coverage of the Earth's surface and the average temperature in each hemisphere. Each grid cell covers 5° latitude x 5° longitude but lines of longitude converge as the latitude moves from the equator to the North or South Pole. Weighting the data by the cosine of the latitude of the centre of the grid cell takes the area covered by each grid cell into account, the weighting itself being used to determine relative coverage and in combination with the temperature anomaly data to determine the average anomaly over a number of grid cells.

Coverage in a given month is therefore determined by summing the weightings (cosines of the latitude of the centre) of all reporting grid cells in that month and expressing it as a percentage of the sum of the weightings of all grid cells across the Earth's surface. The percentage coverage for a given area in a given month, Tcov, is therefore given by

$$Tcov = 100 * \frac{\sum (\text{cosine } (x)) \text{ for all reporting grid cells}}{\sum (\text{cosine } (x)) \text{ for all grid cells}}$$

where x is the degrees of latitude of the centre of the grid cell and both the numerator and denominator refer to grid cells over the same area (e.g. hemisphere or region)

The calculation of the average temperature anomaly from grid cell values is similar to the above but now the temperature data is incorporated to produce a weighted average temperature anomaly. The average temperature anomaly for a hemisphere, for example, is given by:

$$T_{hemi} = \frac{\sum (\text{cosine}(x) * T_{(x,y)})}{\sum \text{cosine}(x)}$$

where $T_{(x,y)}$ is the grid cell value for the cell centred at latitude and longitude (x,y), cosine(x) is the cosine of the angle of latitude of the centre of the grid cell and the summing is across all reporting grid cells in the hemisphere.

The approach used for calculating T_{hemi} can be generalised to determine the average temperature anomaly for any selected region by simply applying it across only the selected grid cells.

The calculation of the global average temperature anomaly[7] is based on the hemispheric average temperature anomalies. In the case of HadSST3 and HadCRUT4 datasets the global average is simply the mean of the two hemispheric values, calculated this way to accept that the data coverage for the two hemispheres is usually different. In contrast the average global temperature anomaly for the CRUTEM4 dataset is calculated as

$$T_{global} = ((2 * T_{NHav}) + T_{SHav}) / 3$$

where T_{global} is the global average temperature anomaly and T_{NHav} and T_{SHav} are the average anomalies for the Northern and Southern hemispheres respectively, the

[7] In the author's experience the term "global average temperature anomaly" is ambiguous. It has often been wrongly understood to be an anomaly of average global temperature. The correct meaning of the term is the global average of the temperature anomalies.

different weightings for each hemisphere being an attempt to compensate for the very different amount of land (v. sea) in each.

1.3.8 Historical annual average temperature anomalies

For general reference in relation to what follows Figure 1.2 shows the annual average HadCRUT4 global average temperature anomaly, along with the corresponding data for its two sources, HadSST3 and CRUTEM4. Because of the number of reporting HadSST3 grid cells and their location on the Earth's surface, the HadCRUT4 global averages are slightly biased towards the value of the HadSST3 global averages, hence the closer proximity to that graph line than to the CRUTEM4 line.

When examined in detail Figure 1.2 shows a changing difference between the CRUTEM4 and HadSST3 global averages, with CRUTEM4 generally lower than HadSST3 during 1865-1895 but greater from about 1990 onwards. Chapter 5 will discuss these differences in more detail.

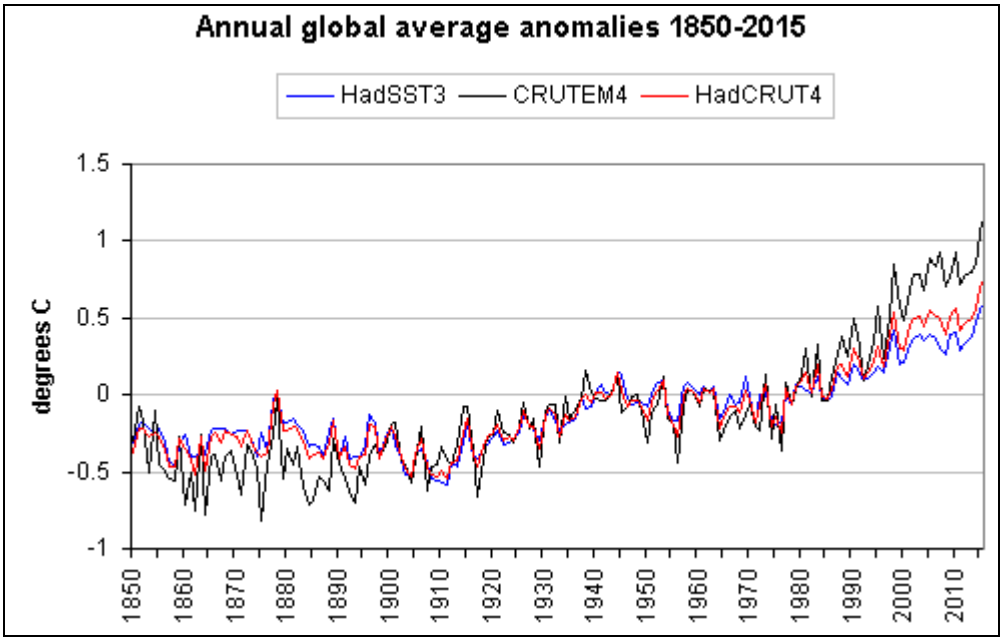


Figure 1.2 HadCRUT4 annual global averages plotted with the annual global averages of CRUTEM4 and HadSST3 data on which HadCRUT4 is based.

1.4 Comments on key literature

As mentioned earlier, no published papers seem to address all areas of concern about the HadCRUT4 dataset. Certain issues raised in subsequent chapters of this thesis are not even mentioned in papers written by people directly involved with creation and updating of it or of the two associated datasets. McKitrick (2010) provides independent discussion of some but not all of the issues but we'll look first at the papers from employees of the CRU and the Hadley Centre before returning to McKitrick (2010).

The definitive papers for the creation of the CRUTEM4 dataset are Jones et al (2012) and Osborn and Jones (2014), with Jones (2016) discussing the reliability of the data. The last-mentioned will be discussed in the final chapter so that comments can refer to material within this thesis, so focus for the moment is on the two other papers.

Jones et al (2012) talks at length about data supplied by national meteorological services (NMSs). It mentions only briefly that this data was considered to be better quality than the data used in CRUTEM3 and HadCRUT3 datasets. The improvement in quality can be attributed largely to national meteorological services (e.g. Australia's Bureau of Meteorology) now adjusting the data themselves rather than having the CRU or the Global Historical Climatology Network (GHCN) undertake this work.

Jones et al (2012) also fails to adequately explain some changes in regard to important issues related to the CRUTEM4 datasets. The minimum number of years of data over which the long-term average temperatures are calculated is 14 from the period from 1961 to1990 in CRUTEM4 but was 15 in CRUTEM3 and no reason for the reduction is given. The calculation of the global average temperature anomaly for CRUTEM3 was a simple mean value of the two hemisphere averages but for CRUTEM4 it is calculated as (2/3 * NH + 1/3 * SH) and again no reason is given for the change.

The paper also contains many inconsistent, ambiguous and even incorrect terminologies. The term "monthly averages [of temperatures]" is used when the "calendar month averages" are meant; and temperature anomalies are not "from a

common period" but determined using the average temperature calculated over that common period. The terms "temperature" and "temperature anomaly" are used as if they were synonymous, as are "anomaly" and "average anomaly". Also confusing are "temperature averages" when referring to average temperature anomalies or even weighted average temperature anomalies.

Osborn and Jones (2014) is generally better at describing processing changes from previous versions of the CRUTEM4 dataset but like Jones et al (2012) has numerous errors with terminology, most typically using the word "temperature" when "temperature anomaly" would be correct.

The web page for the HadSST3 data[8] lists Kennedy et al (2011a) and Kennedy et al (20011c) as the primary references for the dataset. These papers in fact discuss the measuring and sampling uncertainties (Kennedy et al, 2011a) and the biases and homogenisation (Kennedy et al, 2011b); they do not detail the derivation of the long-term averages or the processing of the observations from which the dataset is created. The processing is described on a web page[9] where no references are given for either the description of quality control or the gridding process (i.e. deriving an anomaly from one or more observations). The only direction as to where further information can be found is the comment "The gridding process proceeded as in HadSST2" but the absence of links or references to specific papers is unhelpful.

The web page for HadSST2 data, the version prior to HadSST3, gives Rayner et al (2006) as the principal reference. This paper provides a detailed discussion of several issues including the use of 1° x 1° grid cells and pentads, and the long-term averages that it refers to as the "monthly climatology". It also discusses various changes that have occurred over time in regards to sea surface temperature recording, including changes in ships' routes and changes in the methodology of measurement, and mentions the uncertainties associated with the "inadequate sampling in grid cells", which is a point that this thesis will address in chapter 3.

[8] http://www.metoffice.goc.uk/hadobs/hadsst3/
[9] http://www.metoffice.ogov.uk/hadobs/hadsst3/description.html

One criticism of Rayner et al (2006) also applies to Kennedy et al (2011a and 2011b), namely that too much information is given scant discussion, with readers referred to other papers. Information given this way is unfortunately very fragmented and discussed in different styles and contexts, leaving the reader uncertain of whether every topic of this complex issue has been properly addressed.

McKitrick (2010) is probably the most extensive independent audit to date, albeit not of the HadCRUT4 dataset but the previous version, HadCRUT3. Some comments in the paper have been made redundant by changes of basic procedures for HadCRUT4 but others are still pertinent and will be mentioned in the chapters about the number of observation stations and the measurement and subsequent adjustment of sea surface temperature.

1.5 Summary

As general background and context to subsequent chapters, this chapter has discussed the data sources for the HadCRUT4 dataset, the grid-based system that is used, the concept and use of temperature anomalies, the different cell types, the general concept of coverage and the method used to calculate the hemispheric and global temperature anomalies. It went on to illustrate the historical annual average temperature anomalies according to HadCRUT4, CRUTEM4 and HadSST3 datasets before expressing some concerns about the quality of the primary documentation for each of these.

Chapter 2: Coverage

2.1 Introduction

The geographic coverage of HadCRUT4 data has varied greatly over the period since 1850. This is rarely mentioned when discussing HadCRUT4 temperature data but is important because hemisphere and global averages are calculated without complete data. In these circumstances the implicit assumption is that the average temperature anomaly of all regions for which data is unavailable would be the same as the average for the data that is available.

The concept of coverage as used by the HadCRUT4 dataset, and for that matter CRUTEM4 and HadSST3 datasets, is nothing more than notional because the temperature at every point on the Earth's surface is not recorded and observation stations might be located hundreds of kilometres from their nearest neighbours.

This chapter will discuss what "coverage" means in the context of HadCRUT4, the variation in HadCRUT4 coverage over time, the links between the magnitude of coverage and the change in average temperature anomalies, and how the uneven coverage at various times might have distorted the average temperature anomalies.

2.2 The calculation of coverage

As discussed in the previous chapter, the data coverage is calculated according to the sum of the weightings of reporting grid cells as a fraction (or percent) of the sum of the weightings of all grid cells, the weightings being the cosine of the latitude of the centre of each grid cell. In essence, HadCRUT4 coverage is in terms of entire grid cells regardless of the distribution of observations, either from land or sea, made within that grid cell.

The impact on coverage of decreasing or increasing grid cell size was investigated by using CRUTEM4 station metadata with its precise location of each observation station by latitude and longitude to one decimal place. Figure 2-1 shows the global coverage of CRUTEM4 data when grid cells 2.5° x 2.5° (latitude x longitude), 5° x 5° and 10° x 10° are used.

The difference in coverage with these different cell sizes is not only related to the grid cell size but depends also on the distribution of stations within the grid cells. Three observation stations in a single 5° x 5° grid cell might occupy one, two or three grid cells of 2.5° x 2.5° if this smaller size was used, in which circumstances the coverage will be one two or three of the smaller grid cells. Rather than decreasing the grid cell size it might be increased to 10° x 10°, which is equivalent to four 5° x 5° grid cells. It would only require one of the 5° x 5° grid cells to contain one or more observation stations in order that the 10° x 10° grid cell had at least one observation station and the entire cell being regarded as contributing to the dataset coverage.

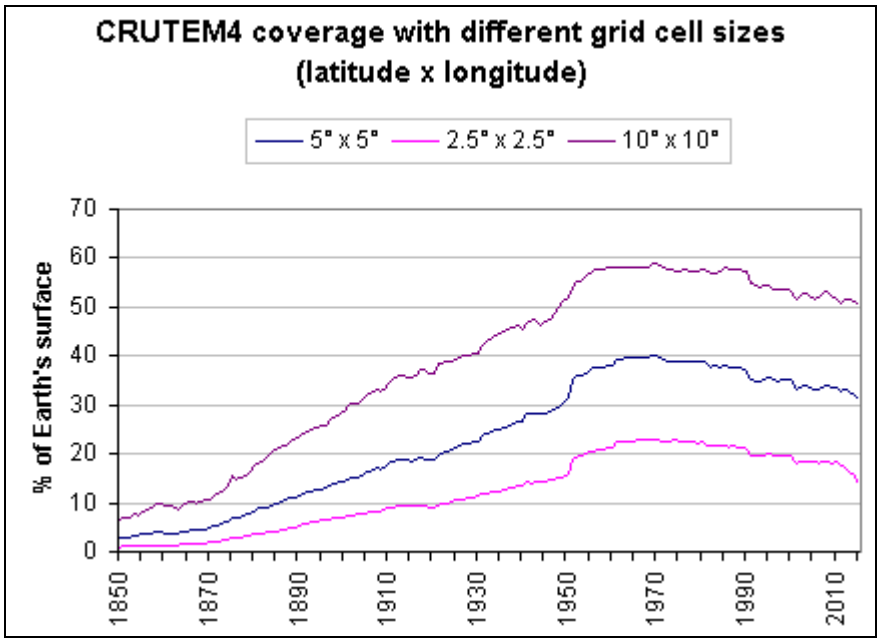


Figure 2.1 Variation in coverage of CRUTEM4 data when grid cells of 2.5° x 2.5° (latitude x longitude), 5° x 5° and 10° x 10° are used.

None of the grid cell sizes illustrated in Figure 2.1 is any more correct than the others, so the adoption of 5° x 5° grid cells for HadCRUT4 is somewhat arbitrary and a

compromise between cells that span thousands of kilometres from north to south and east to west, grid cells that are so small that coverage will be poor.

The HadCRUT4 data coverage is therefore notional and has no meaning outside of that datasets and the CRUTEM4 and HadSST3 datasets. A claim of 75% coverage for example actually means data is available from grid cells that together amount to 75% of the Earth's surface.

2.3 Inconsistency in HadCRUT4 data coverage

The previous section described how coverage is related to grid cell size and illustrated this using CRUTEM4 data. This section focuses on the variable notional coverage of the HadCRUT4 dataset over time according to its 5° latitude x 5° longitude grid cell size.

Figure 2.2 shows the annual average global and hemispheric coverage from 1850 to 2015. The coverage of the Northern Hemisphere during the period 1850-2015 was always greater than or equal to the coverage of the Southern Hemisphere save for a period from about 1860 to 1880. Coverage of the Southern Hemisphere only very briefly exceeded 50% prior to 1925, dropped again during World War II, recovered to above 50% in 1950 and only consistently exceeded two thirds of the hemisphere after 1961.

Global coverage did not exceed 75% until about 1960 and in 2015 was 83.4%, having peaked at 88.1% in 1979, the same year that Southern Hemisphere peaked at 82.7% and nine years after Northern Hemisphere coverage peaked at 92.3%.

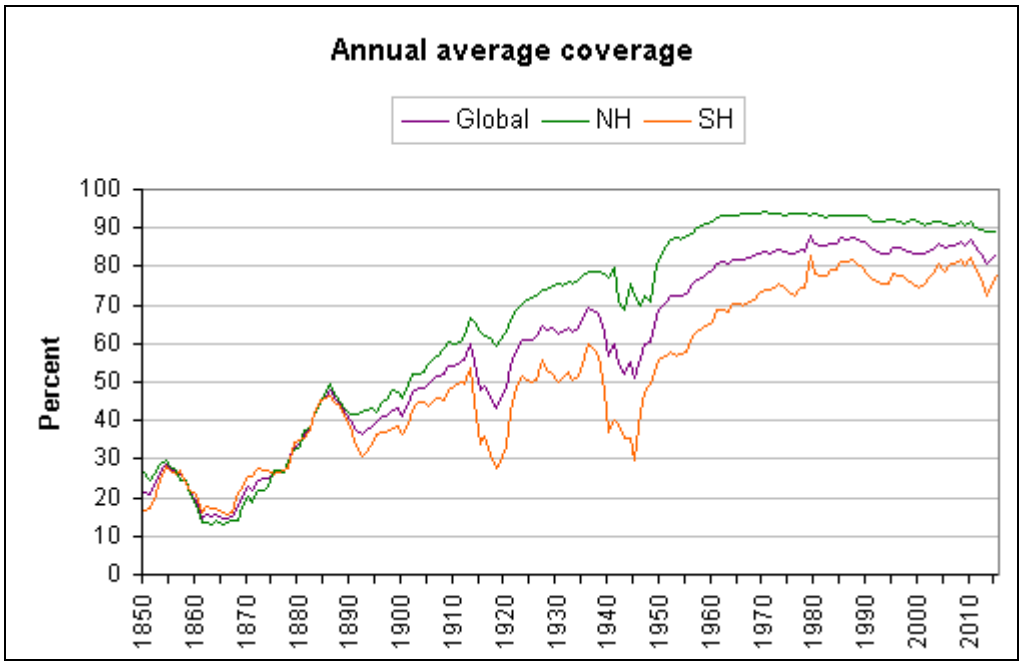


Figure 2.2 Global and hemispheric HadCRUT4 data coverage over time, the global data being as percentage of the Earth's total surface and the hemispheric data as a percentage of the total surface of the hemisphere.

When coverage is less than 100% there is an implicit assumption that had the missing data been present the global and hemispheric average temperature anomalies would be identical to those derived from the available data. This assumption is unsustainable because, as we will see, at various times more data was available from certain regions of the Earth's surface data than from other regions, skewing the anomalies according to those regional weather conditions. Temperatures vary more in the mid and high latitudes than they do in the tropics and therefore even the location from which data is missing has implications for how the large-scale averages might differ if the coverage was 100%.

The global coverage of sea surface temperature data and observation station data was also examined. The Earth's surface comprises 29% land and 71% water but on a grid cell basis ~28% of grid cells are "coastal" and their data could be sourced from temperature measurements over land and/or sea (see Chapter 1). Figure 2.3 shows the annual average coverage for the CRUTEM4 and HadSST3 datasets. HadSST3 coverage in December 2015 is 69% whereas CRUTEM4 coverage is 31%, meaning that HadSST3 accounts for a much greater amount of the coverage in that month, but

magnitude of the difference in coverage has been far from constant. Oceans have a buffering effect on temperature change, shown clearly by sea surface temperatures that vary less from month-to-month than do temperatures over land, so different ratios of coverage from the two sources will impact hemispheric and global HadCRUT4 average temperature anomalies.

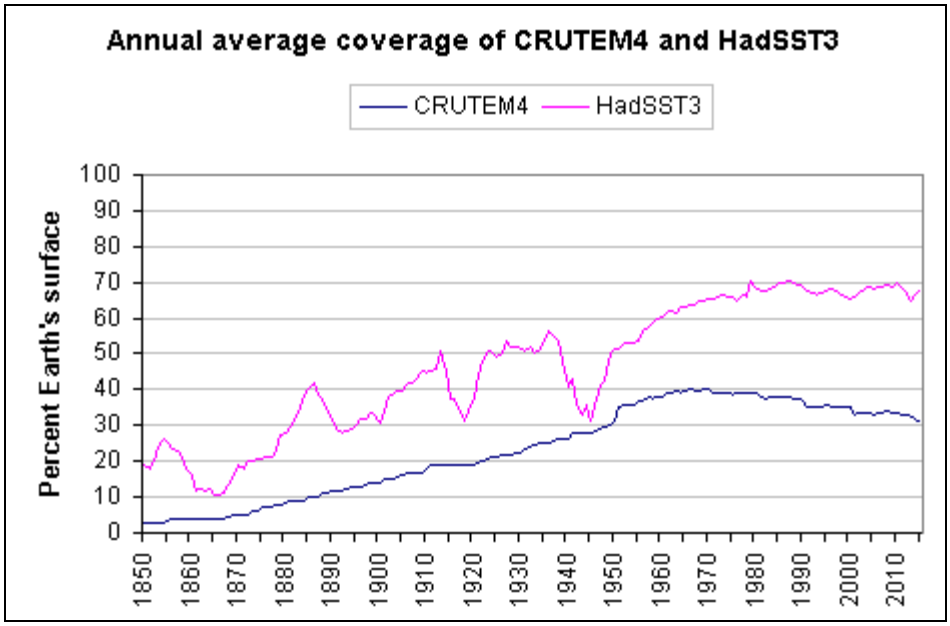


Figure 2.3 Annual average coverage of data of the two datasets associated with HadCRUT4, namely CRUTEM4 and HadSST3

The CRUTEM4 and HadSST3 coverage of the Northern Hemisphere is shown in Figure 2.4, the maximum possible coverage (based on grid cell types) being ~75% for CRUTEM4 and ~58% for HadSST3. The corresponding graphs for the Southern Hemisphere are shown in Figure 2.5, with maximum coverage of ~34% and 88% for lCRUTEM4 and HadSST3 respectively. As noted in chapter 1, the greater global coverage of HadSST3 means that HadCRUT4 global averages are closer to HadSST3 than CRUTEM4 averages.

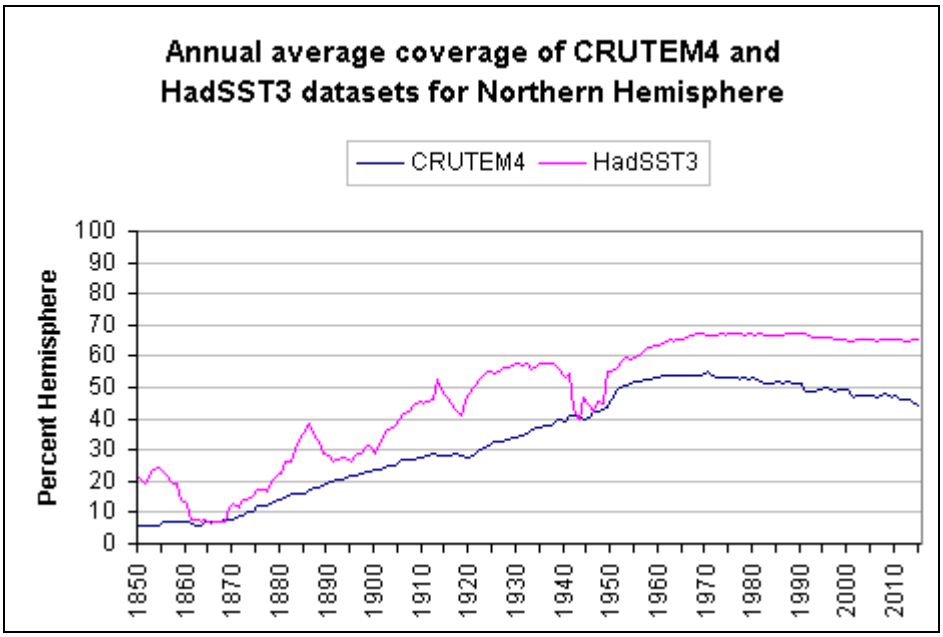


Figure 2.4 Annual average coverages of CRUTEM4 and HadSST3 datasets in the Northern Hemisphere

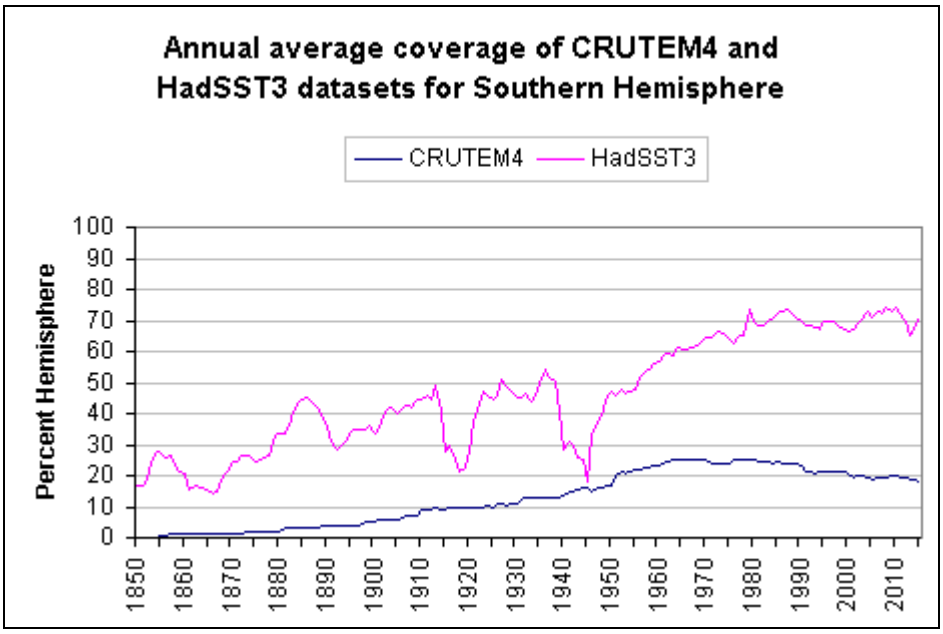


Figure 2.5 As for Figure 2.4 but for the Southern Hemisphere.

2.4 Monthly variation in HadCRUT4 global coverage

This chapter as so far discussed and illustrated annual average coverage but this masks the variation in coverage in each month.

Figure 2.6 shows the minimum and maximum coverage in each year of the record. The annual range in coverage is greatest in the Southern Hemisphere where it usually falls between 10% and 20% of the hemisphere for the 49 years from 1945 to 1993. Prior to 1945 the annual variation was generally below 10% of the hemisphere. Coverage decreased sharply with the outbreak of war reducing shipping movements and then increased quickly when wars ended.

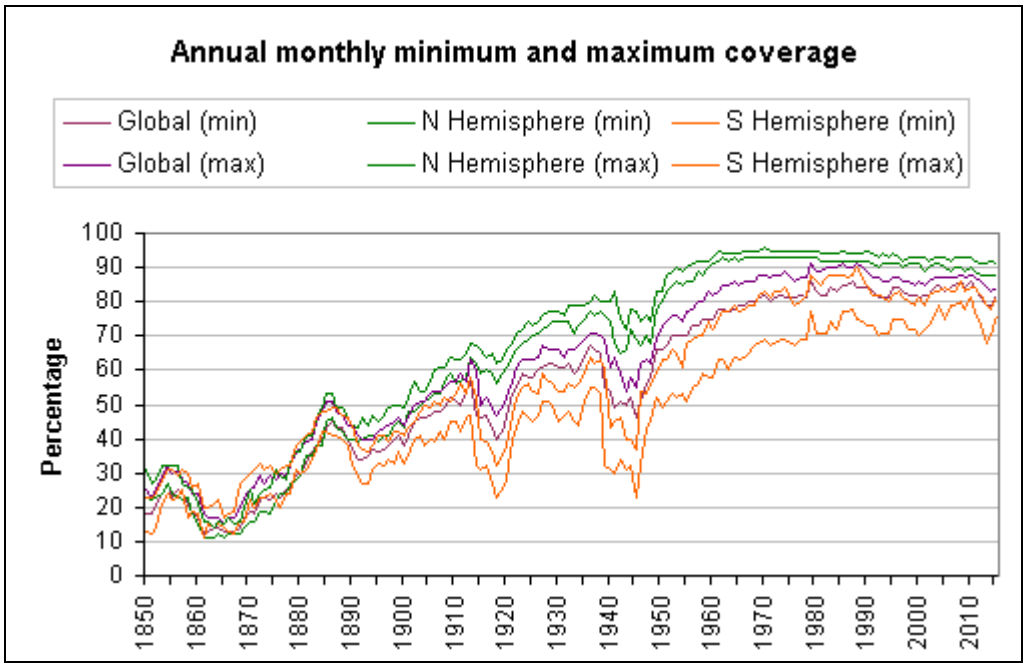


Figure 2.6 Maximum and minimum monthly coverage in each year on a global and hemispheric basis

In general the annual range of coverage is a consequence of variations in shipping traffic at certain times of the year or, in some cases over land, the temporary absence of staff to read observation station thermometers, which has tended to happen more in

Polar regions. Attribution of the cause of the variation in coverage is made more complex by the weighting based on the cosine of the latitude of the centre of the grid cell, meaning that a failure to report data in just a few grid cells in the tropics might mean the same shortfall in coverage as a failure to report from eight or even ten times that number in Polar regions.

2.5 Coverage and variation in month-to-month average temperature anomaly

The absolute month-to-month variations (i.e. change in value from one month to the next) in HadCRUT4 average global temperature anomalies is somewhat related to coverage, especially in the early years when coverage is low. Month-to-month variation over land is greater than at sea where the warming and cooling of the mass of the ocean takes longer. Figure 2.7 (a) shows the relationship between CRUTEM4 global averages and the month-to-month variation and Figure 2.7 (b) does likewise for HadSST3 data. Both figures show instances of the inverse relationship between coverage and temperature variability, particularly with low coverage generally meaning greater month-to-month variations.

The situation with the month-to-month variation in HadCRUT4 average temperature anomalies is more complicated because while the general relationship still broadly holds, it is influenced by the number and locations of grid cells using CRUTEM4 or HadSST3 data or both in each month.

In January 1863, when the HadCRUT4 global average temperature anomaly rose by 1.02°C over the previous month, global coverage was 15% and in March 1869, when global coverage was 20%, the HadCRUT4 global average fell by 0.859°C from the previous month. These extreme month-to-month changes in average temperature do not occur later in the record when coverage was more complete. From January 2000 to December 2015 global coverage has varied between 79% and 88% and the month-to-month variation has ranged from -0.279°C to 0.305°C, with 141 of the 192 months

(i.e. 73.4% of months) showing variations in global average temperature anomaly of within ±0.1°C.

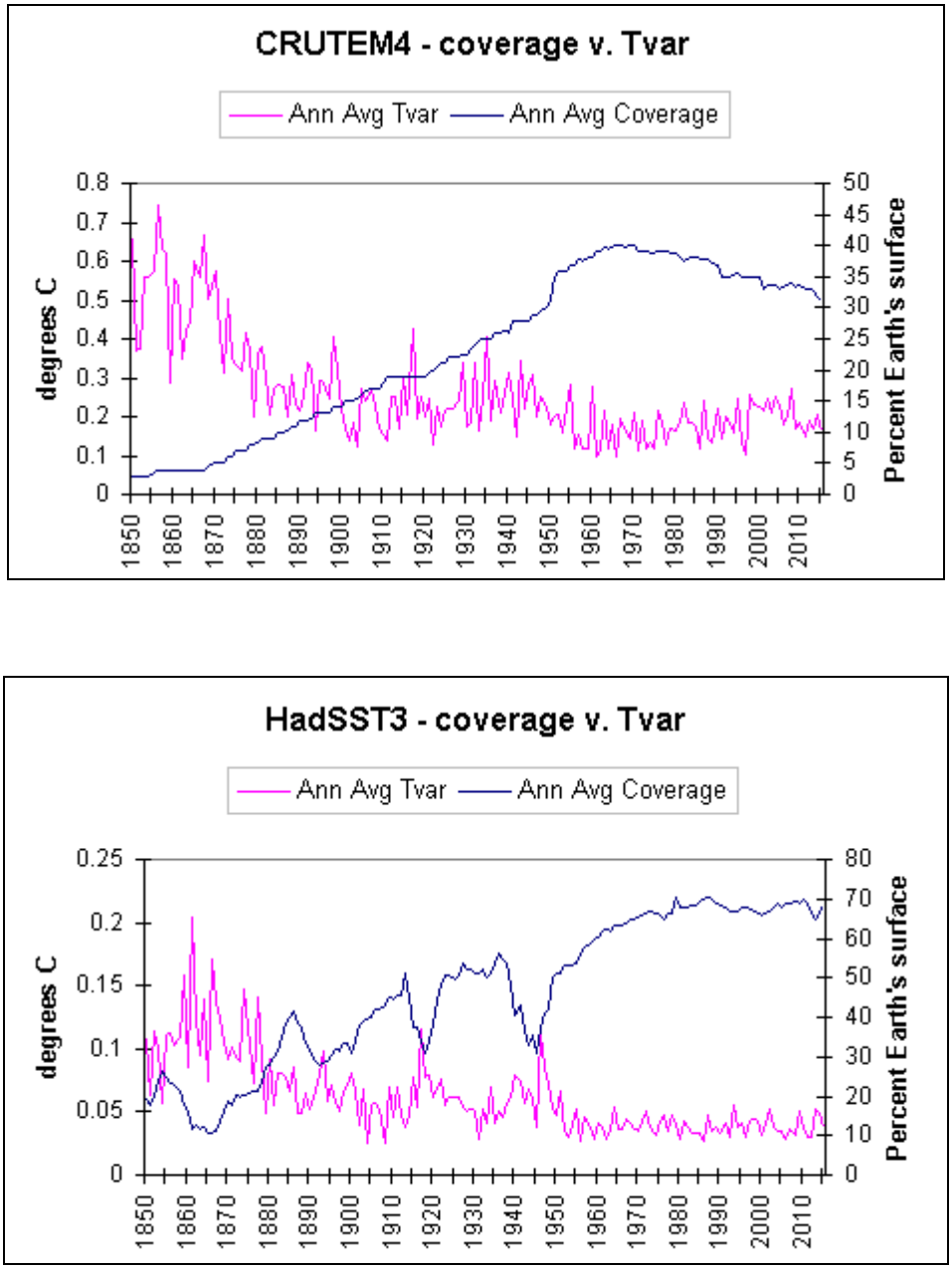


Figure 2.7 Annual average coverage and annual average Tvar (i.e. absolute month-to-month variation in average global temperature anomaly) for CRUTEM4 (top) and, at a different Y-axis scale, HadSST3 (bottom)

2.6 Inhomogeneities in coverage - (a) Latitude bands

Figure 2.2 showed how HadCRUT4 coverage has varied over time. An important question is whether the regions from where data was obtained were sufficiently dispersed as to provide a reasonable estimate of global conditions. Figure 2.8 shows that this was not the case and in fact the contribution of all reporting grid cells in each latitude band to the total coverage in each hemisphere has varied greatly over time. (As expected, the weighting according to the cosine of the grid cell centre means that coverage will differ even when data is reported from the same number of grid cells in each band.)

In the Northern Hemisphere (Figure 2.8 top), during the period from 1860 to 1890 latitude bands 30°-40°N and 40°-50°N made disproportionately high contributions to the hemisphere's coverage and hence to average hemispheric temperature anomaly, while the bands 10°-20°N and 20°-30°N made disproportionately low contributions to that coverage. In this hemisphere the coverage stabilised in about 1950 and the contributions are commensurate with the weighting factor.

In the Southern Hemisphere (Figure 2.8 bottom), during the period from 1855 to 1900, the contributions from bands 30°-40°S and 40°-50°S were disproportionately high, but the contribution of the latter band fell sharply from about 17.0% to 3.7% during 1913 to 1918, which is below the contribution it would make if an equal number of grid cells had data in each latitude band. Unlike the Northern Hemisphere, the contributions from Southern Hemisphere latitude bands do not stabilise by 1950, in fact they do not stabilised by 2015.

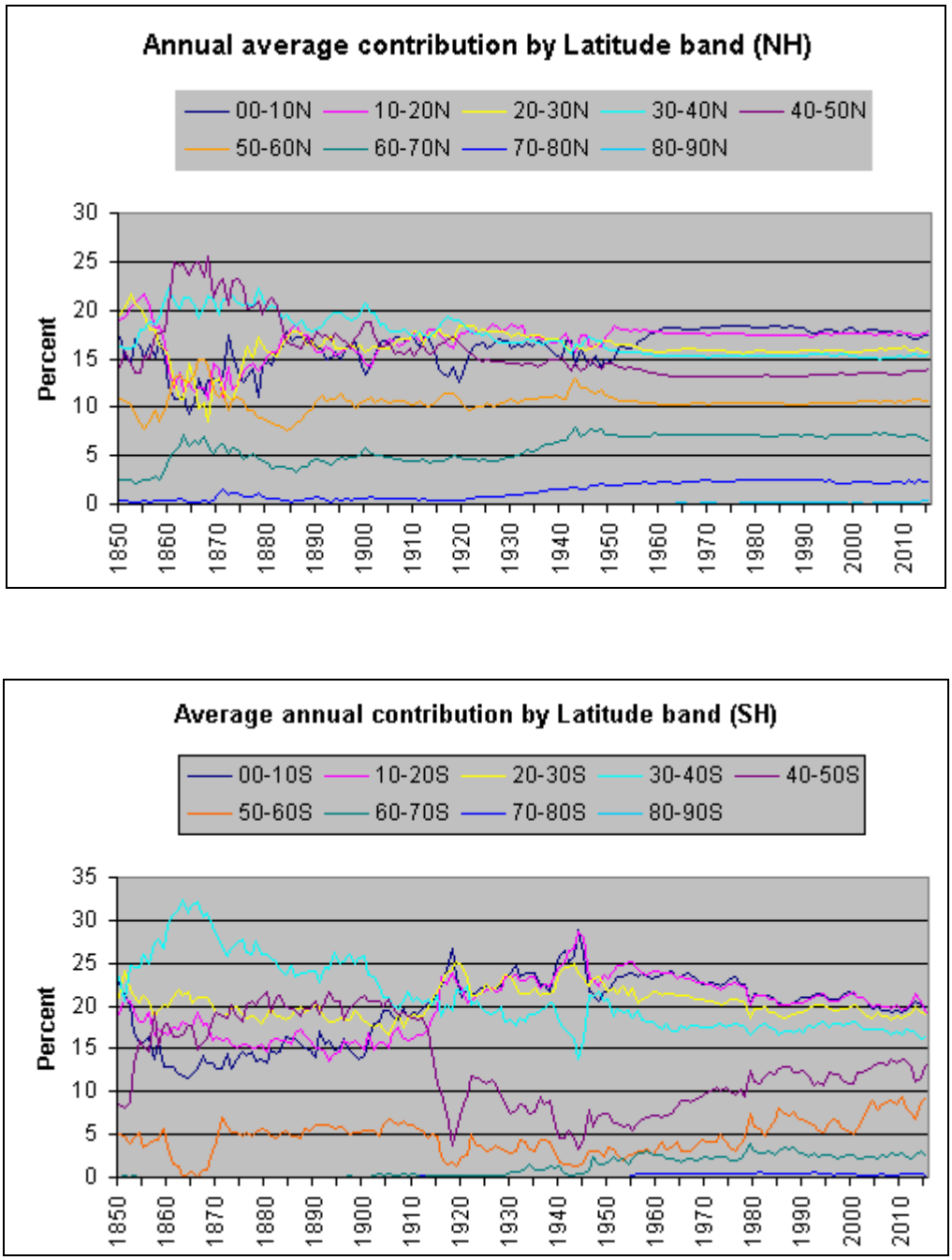


Figure 2.8 Percentage contributions of each latitude band to the annual average total coverage of each Hemisphere (Top: Northern Hemisphere, bottom: Southern Hemisphere).

Some of the changes in the contributions to Southern Hemisphere coverage can easily be accounted for. The contribution of latitude bands 30°-40°S and 40°-50°S from 1850 to 1910 is largely attributable to ships following trade routes from Europe to Indonesia and the "Far East" and the transport of immigrants to Australia and New Zealand. By the mid-1850s steam-driven shipping was common in the Atlantic but

wind-powered shipping continued to be used on the Europe-Australia route because the supply of adequate coal was uncertain, the downside being that the ships continued to travel south of Africa because the Suez Canal, which opened in November 1869, was unsuitable for wind-powered vessels.

The bands 30°-40°S and 40°-50°S cover a region where the cold waters of the Southern Ocean meet warmer subtropical water (Toole & Warren, 1993, Belkin & Gordon, 1996), meaning that average Southern Hemisphere temperature anomalies from 1850 to 1900 disproportionately include the anomalies from this region of mixing waters. The number of reporting grid cells in the region within latitudes 40°-50°S and longitudes 50°W-150°E (i.e. South American coast to Australia's east coast) averaged 56.2 of a possible 80 across each month of 1913, when the average temperature anomaly for the region was -0.78°C, but averaged 1.2 per month of the possible 80 in 1919, when the region's average temperature anomaly was +0.01°C.

2.7 Inhomogeneities in coverage - (b) Longitude bands

The same kind of disproportionate input for latitude bands can be seen in the annual average contribution of 20-degree longitude bands to each hemisphere's total coverage. Figure 2.9 shows the percentage contribution of each 20°-longitude band to Northern Hemisphere SST coverage, the upper figure dealing with longitudes west of Greenwich (longitude = 0.0°) and the lower with longitudes east.

The top portion of Figure 2.9, showing West longitude bands in the Northern Hemisphere, indicates a strong bias towards the 20°-40°W band, which covers much of the mid-Atlantic, and a less disproportionately high contribution from the 0°-20°W band that covers much of the United Kingdom, western France, Spain and the eastern portion of the Atlantic Ocean. The contributions to total coverage are very similar around 1920 and almost equal (all at 5.55% of hemisphere) from 1960 onwards.

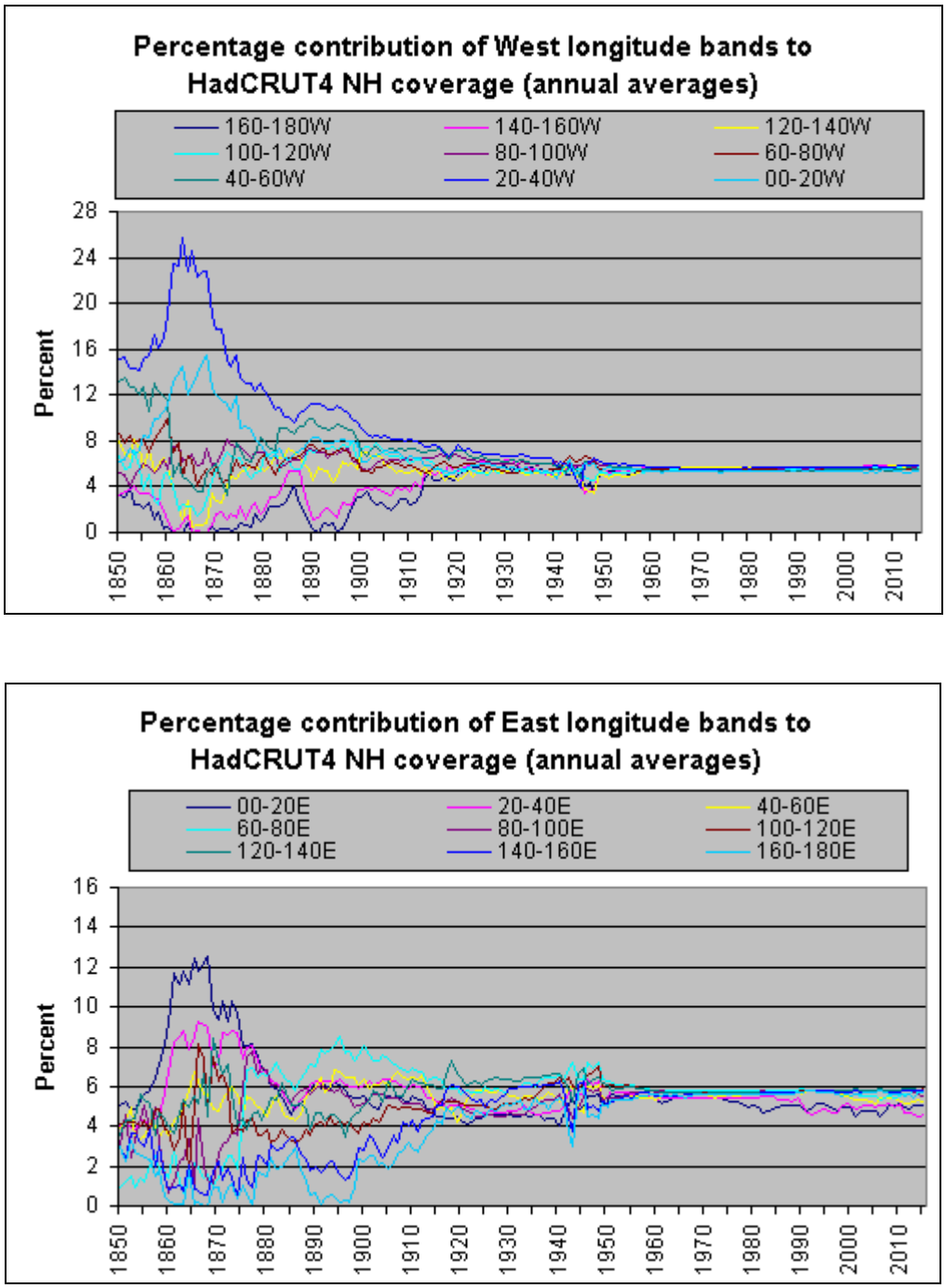


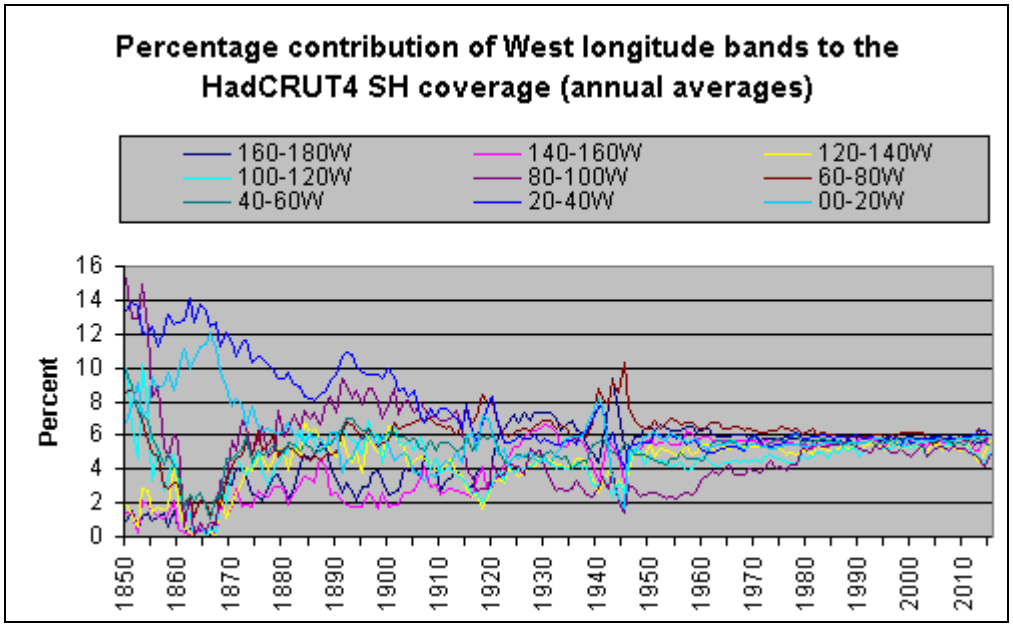
Figure 2.9 Percentage contributions of east (bottom) and west (top) longitude bands to HadCRUT4 coverage of the Northern Hemisphere.

The bottom portion of Figure 2.9 is of East longitude bands in the Northern Hemisphere and shows a very variable situation until about 1960. The maximum contribution by any band is about 50% of the maximum contribution from any West longitude band, which means a more even distribution of reporting grid cells across the bands. Of particular note is that the northern Pacific Ocean, covered by bands

140°-160°E and 160°-180°E made very little contribution to total NH coverage prior to 1900 and only slowly increased that contribution over the next 60 years.

The percentage contributions of longitude bands to the Southern Hemisphere coverage are more irregular (Figure 2.10) than those for the Northern Hemisphere. The dominant West bands in the 1860s are 20°-40°W and 0°-20°W, which cover part of the Atlantic Ocean and are probably a reflection of the shipping routes at the time. The sharp decrease in the contribution to coverage of the 80°-100°W band in the late 1855's is likely due to both changes in shipping routes and of a decrease in SH coverage at the time (see Figure 2.1). The contribution of the three East longitude bands between 120E and 180E increased from 1850 to about 1900 when data from Australian observation stations is included in the HadCRUT4 dataset for the first time. Unlike with the Northern Hemisphere, abrupt shifts in contributions to Southern Hemisphere coverage are evident during the years of the two world wars, this being a consequence of the greater area of ocean in the hemisphere and the changes to shipping traffic caused by the starting and ending of the wars.

While Figure 2.9, for the Northern Hemisphere, shows near-stable and equal contributions from about 1960 onwards, Figure 2.10, for the Southern Hemisphere, shows uneven contributions continuing to the end of the period shown (i.e. 2015).



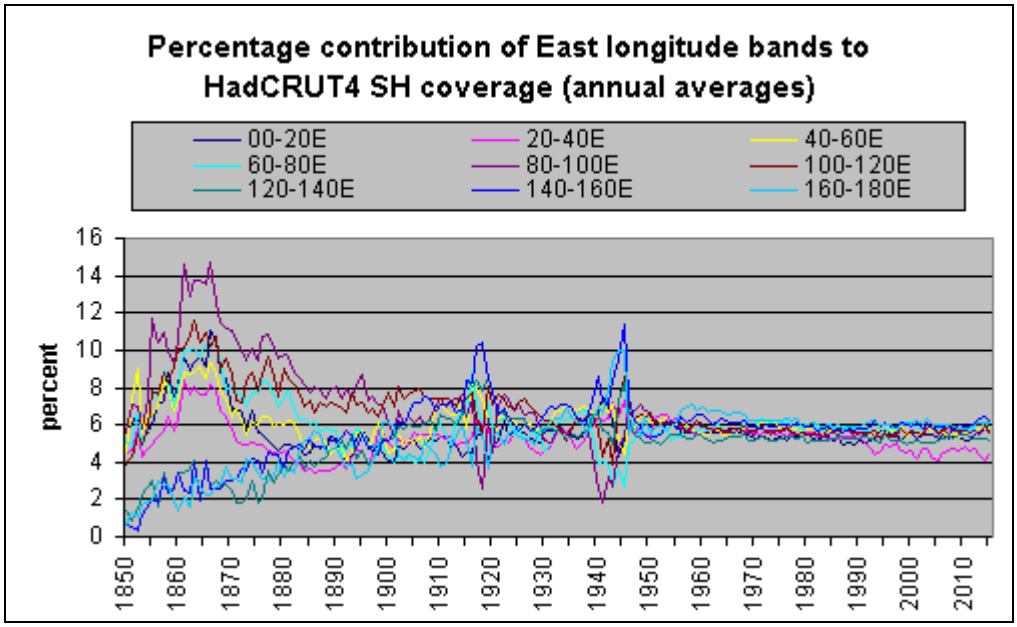


Figure 2.10 Percentage contributions of east (bottom) and west (top) longitude bands to HadCRUT4 coverage of the Southern Hemisphere.

2.8 Inhomogeneities in coverage - (c) regional bias NH

The studies of latitude and longitude bands described above suggest that during the period 1850-1870 grid cells covering Europe and the North Atlantic ocean contributed to a greater proportion of northern hemisphere coverage, and therefore the average

temperature anomaly, than the proportion of the area of the regions to the surface area of the hemisphere. To put it another way this region of the Northern Hemisphere provided more temperature data than did other parts of the hemisphere for the calculation of the average anomaly, meaning that the average is biased towards this area. Given that the global average is calculated from the average of the two hemispheres then the global average is also skewed by this biased coverage.

For the HadCRUT4 dataset this is confirmed by an analysis of coverage of Western Europe and a portion of the Atlantic just north of the equator (regions A and B in Figure 2.11). According to the HadCRUT4 method of determining coverage, the region in question covers just 12.5% of the Northern Hemisphere (not very obvious from the Figure which uses Mercator's projection and excludes the Antarctic) yet on an annual average basis the grid cells in the region accounted for more than 55% of the coverage of the hemisphere during eight years of the 1860s, in four of those years accounting for more than 60% (Figure 2.12). An analysis of monthly data shows that the regions accounted for more than 50% of the hemisphere's coverage in all 96 months from January 1861 to December 1868. In 56 of those months (58.3% of 96) it exceeded 60% and in 13 of the months (13.5%) it exceeded 66.6%.

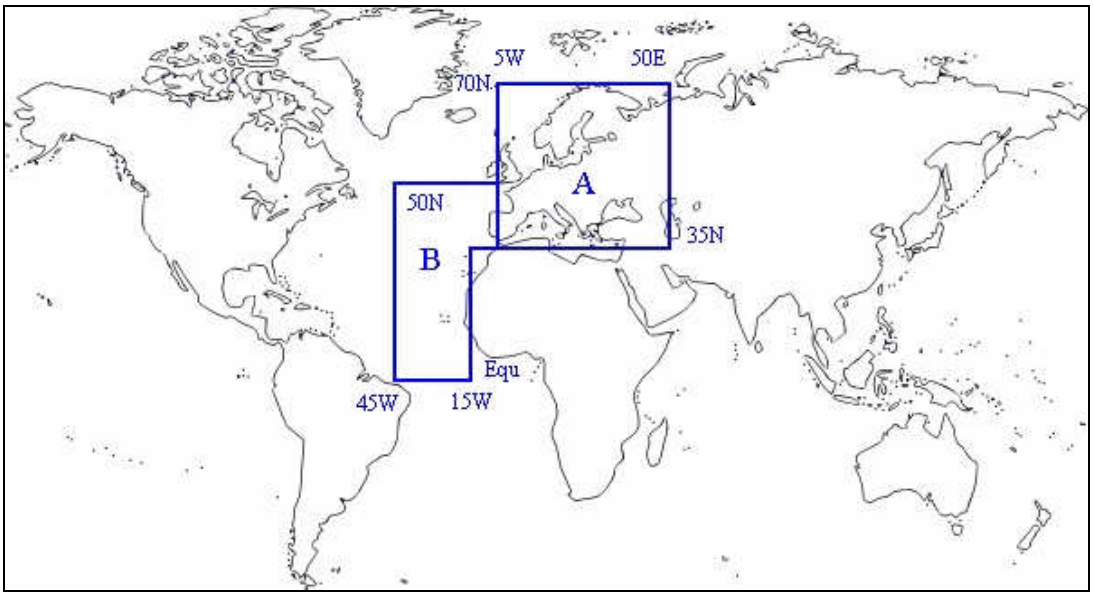


Figure 2.11 Two regions of the Northern Hemisphere whose coverage is discussed in this section

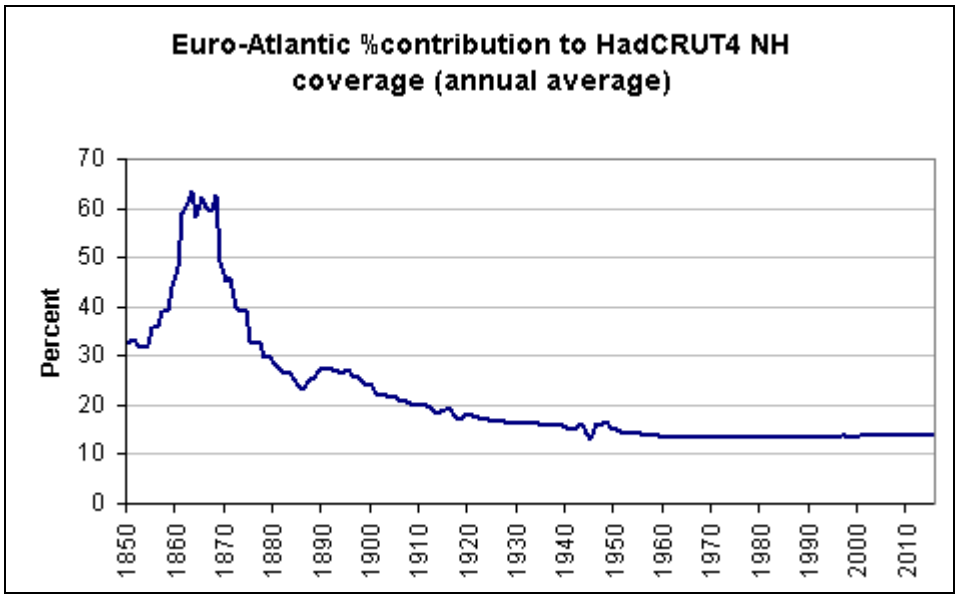


Figure 2.12 Annual average percentage contribution of HadCRUT4 northern hemisphere coverage made by the composite of the A and B regions shown in Figure 2.11

The CRUTEM4 data from observation stations for most of the continent of Europe (region A in Figure 2.11) was also examined, this region, according to the CRUTEM4 method of determining coverage, being ~9.3% of the land area of the northern hemisphere. Figure 2.13 shows the annual average percentage contribution of the region to the CRUTEM4 northern hemisphere coverage.

In the early years of the CRUTEM4 record the contribution of Europe to the total coverage of the hemisphere is much greater than the region's proportion of the hemisphere's total land area. From 1858 to 1868 the annual average contribution from this region always exceeds 50% of the coverage reported for the hemisphere. The analysis of monthly data showed that from January 1860 to December 1867, after rounding to a single decimal digit, this region always accounted for 50% or more of the Northern hemisphere coverage.

Earlier in this chapter it was shown that global and hemisphere coverage was quite low in the early parts of the HadCRUT4 data record (i.e. 1850 to at least 1900) and now it has been shown that average temperature anomalies for the northern hemisphere were disproportionately weighted towards these areas that had good

coverage (i.e. most grid cells contain data) when other coverage in other areas was poor (i.e. far fewer grid cells with data).

The bias towards Europe could be particularly distorting because Europe was emerging from the Little Ice Age at the time and temperature anomalies were likely to be abnormally low. In the case of CRUTEM4 there is the further problem that Europe is densely populated and data from its observation stations are likely to have suffered from urbanisation over time. In some cases the urbanisation might still be present in the station data but in other cases the data might have been adjusted to remove these non-meteorological contributions to temperature. Chapter 6 will discuss how distortion, causing excessively low temperature anomalies in the early part of the record and excessively high temperatures in the more recent date, or both, is quite possible.

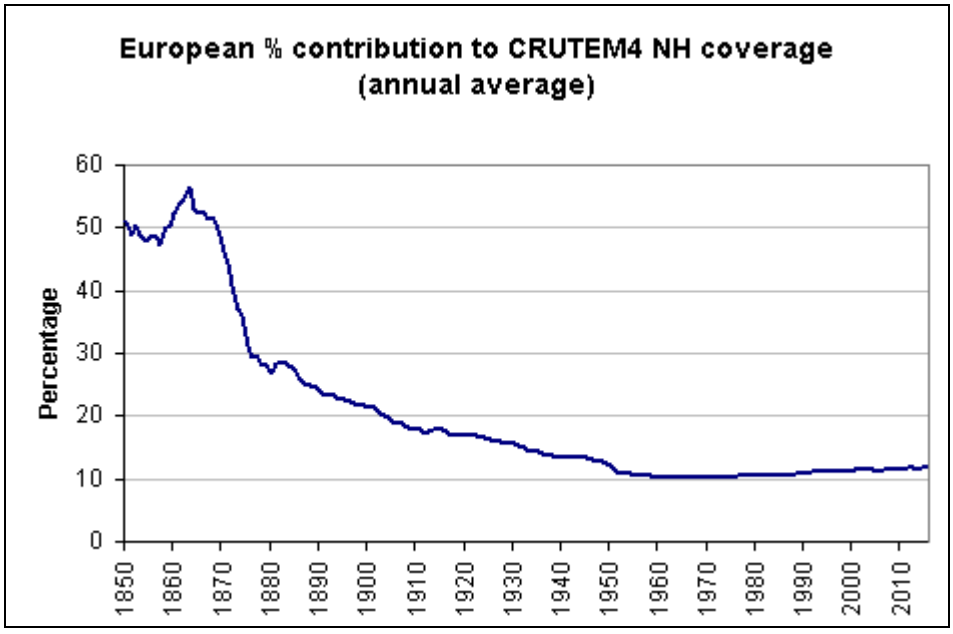


Figure 2.13 Annual average percentage contribution of CRUTEM4 data for Europe (region A in Fig. 2.11) to the total CRUTEM4 coverage of the northern hemisphere

2.9 Inhomogeneities in coverage - (d) regional bias SH

Coverage of the Southern Hemisphere was likewise biased, particularly during the late 1800's, and hemispheric average temperatures were skewed in favour of the regions of greatest coverage. (The region responsible for the bias was identified by determining the grid cells that reported data in more than 200 of the 360 months from January 1850 to December 1879 and focussing on the dominant distribution pattern that this revealed.) Unlike the Northern Hemisphere, with distinct large rectangular "blocks" of coverage, one encompassing Europe and one for a large portion the North Atlantic Ocean that touched on the equator (Figure 2.11), Southern Hemisphere coverage focussed heavily on shipping trade routes from Europe, around the Cape of Good Hope to Indonesia (Dutch interests) and to the Far East (British interests) (Figure 2.14).

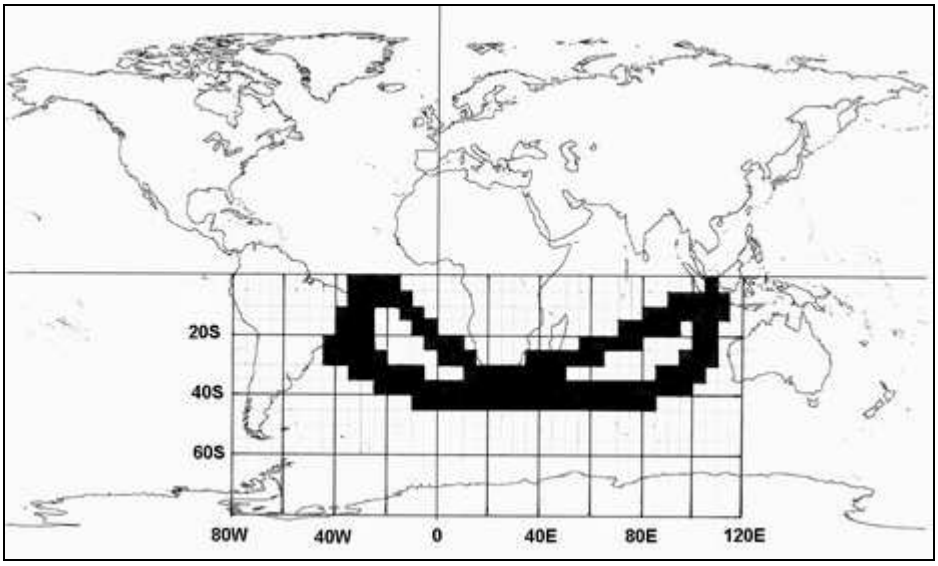


Figure 2.14 Region of coverage in the Southern Hemisphere corresponding to shipping routes between Europe and Indonesia or Far East

Based on the HadCRUT4 method of determining coverage the region shown in Figure 2.14 accounts for 13.95% of the Southern Hemisphere (although the map projection used in Figure 2.14 suggests a greater figure.) Despite this, the annual average percentage contribution of this area to the HadCRUT4 Southern Hemisphere coverage

exceeded 66.6% in every year of 1861 to 1867, peaking at 74.9% in 1866 (Figure 2.15). The percentage was not consistently less than 30% until 1898, which is still more than double 13.95% of the southern hemisphere that this region accounts for.

In 61 months (50.8%) of the period January 1860 to December 1869 the percentage contribution to SH coverage was greater than 66.6%, with three months of 1866 exceeding 80% and a peak of 84.3% in September of that year.

As with the Northern Hemisphere, Southern Hemisphere average temperature anomalies in the late 1800's are very much biased towards weather conditions in only small fractions of the total hemisphere.

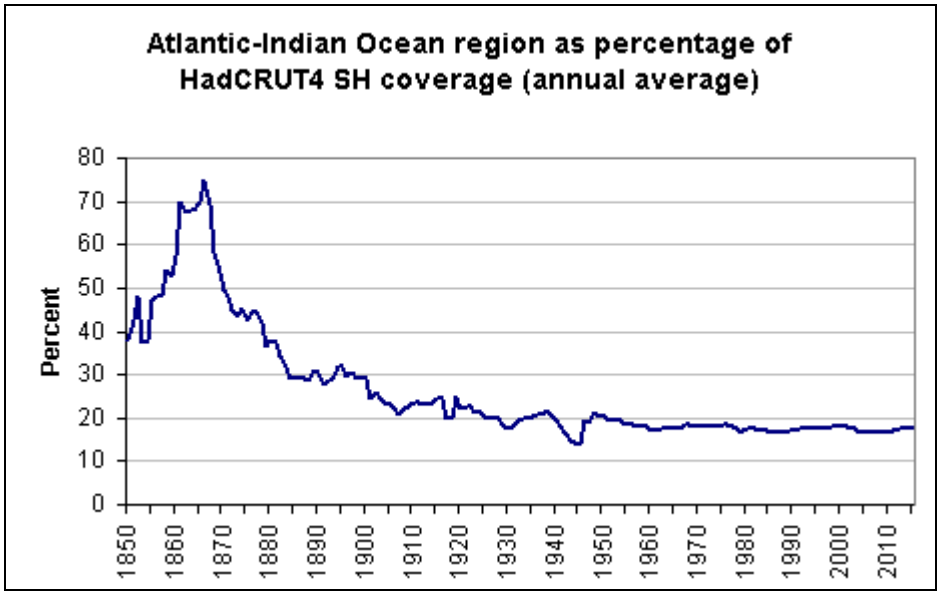


Figure 2.15 Annual average percentage of HadCRUT4 coverage attributable to the area shown in Fig 2.14

2.10 Summary

HadCRUT4 data coverage is related to the grid cell size that the dataset uses and because it has only notional application to the physical world it should be used with caution. Even as a notional concept the global coverage has varied greatly over time,

being less than 50% of the Earth's surface for about one-third of the data record from 1850 to 2015 and only having been greater than 75% for only the last 60 years (since 1956). Even at the end of 2015 there was no coverage of 16.6% (i.e. 1/6th) of the Earth's surface.

Coverage has not only varied annually but also on a month-to-month basis throughout every year, which means that the average HadCRUT4 global and hemispheric temperature anomalies have been calculated from data sources that vary in each month.

There is an implicit assumption with HadCRUT4 global average temperature anomalies calculated with less than 100% coverage that had the missing data been present it would have produced identical averages to those reported from the available data. The HadCRUT4 global average temperature anomaly in May 1861 was -0.761°C when coverage was just 12% but claiming it to be global and the practice of calculating global averages from any coverage whatsoever is to claim that if data was available for the rest of the world (i.e. that coverage was 100%) then the global average would also be -0.716°C. Expressed more generally, claiming that an average is global or even hemispheric when it is derived from less than 100% coverage is to assume that the average temperatures across the grid cells that failed to report data would exactly match the average. This is extremely optimistic because the data that is available differs between grid cells, the hemispheric averages in each month since 1850 differ by up to 1.9°C and, as Figure 1.2 showed, sea surface temperature averages have differed from temperatures from observation stations on land. .

In a similar fashion any of trends in global average temperature anomalies from varying temperature data coverage assumes that the trends apply to everywhere on the Earth's surface, even those that did not report data during some or all of the period over which the trend was calculated. This is also an unsustainable argument because it is easy to illustrate that different locations have in fact had different temperature trends over the same period of time.

It was also noted in this chapter that for both CRUTEM4 and HadSST3 data a general inverse relationship exists between data coverage and month-to-month variation in

average global temperature anomaly, with a greater variation during times of low coverage. This undermines the published global and hemispheric averages for early years of the datasets by implying that averages would likely be different (and month-to-month variations less) if coverage was greater.

The imbalance between the coverage by latitude bands, longitude bands and the regional bias in each hemisphere means that hemispheric and global average temperature anomalies have at times been biased towards areas with substantial coverage and away from those that supplied little data. In the Northern Hemisphere the coverage by latitude bands since 1960 has been almost consistent but in the Southern Hemisphere it continues to be inconsistent even in 2015, although less than it was prior to 1980. Coverage by longitude bands shows a similar pattern to that by latitude bands, which is near-consistency in the Northern Hemisphere but inconsistent even in 2015 in the Southern Hemisphere, albeit less inconsistent in western longitudes since 1980 or in eastern longitudes since 1970.

It has also been shown that in the late nineteenth century and into the twentieth century the supposed hemispheric average temperature anomalies for both CRUTEM4 and HadCRUT4 in the Northern Hemisphere and HadCRUT4 in the Southern Hemisphere drew more than 50% of their data from grid cells that covered only a small percentage of those hemispheres, making them based heavily on weather patterns in those regions.

Ultimately the poor and uneven HadCRUT4 data coverage prior to 1950 means that large scale average temperature anomalies could have a very large error margin associated with them, so large as to render the average values meaningless.

At a smaller scale, i.e. when the focus is on smaller number of grid cells, the data might be acceptable but coverage should be carefully analysed before a decision to use it is made.

The interesting but difficult to resolve question is whether Southern Hemisphere coverage after 1950 is still too low and inhomogeneous to use with confidence when calculating the average temperature anomalies over large areas.

Chapter 3: Variation in data quantity

3.1 Introduction

The previous chapter discussed data coverage, which in the context of HadCRUT4 data is simply whether data is present for a grid cell in a given month. Although coverage and volume of data are clearly related, this chapter will delve deeper into the volume of data from which a grid cell's value in a given month was generated and to look at how the overall quantity of data has changed over time. This examination will also go some way to showing whether a grid cell's value can be regarded as representative of both the area covered by the grid cell and across the entire month.

The HadCRUT4 dataset is constructed from HadSST3 and CRUTEM4 data so the data for these two contributing datasets will be considered separately.

3.2 SST Observations

The HadSST3 data is compiled from a combination of data from the International Comprehensive Ocean Atmosphere Data Set (ICOADS) with, since 2009, data obtained directly from ship's logs. The number of observations made in each grid cell in each month is published in a file available for download from the Hadley Centre, which creates and maintains the HadSST3 data. Figure 3.1 shows the annual averages of the number of sea surface temperature observations per month for the period 1850 to 2015.

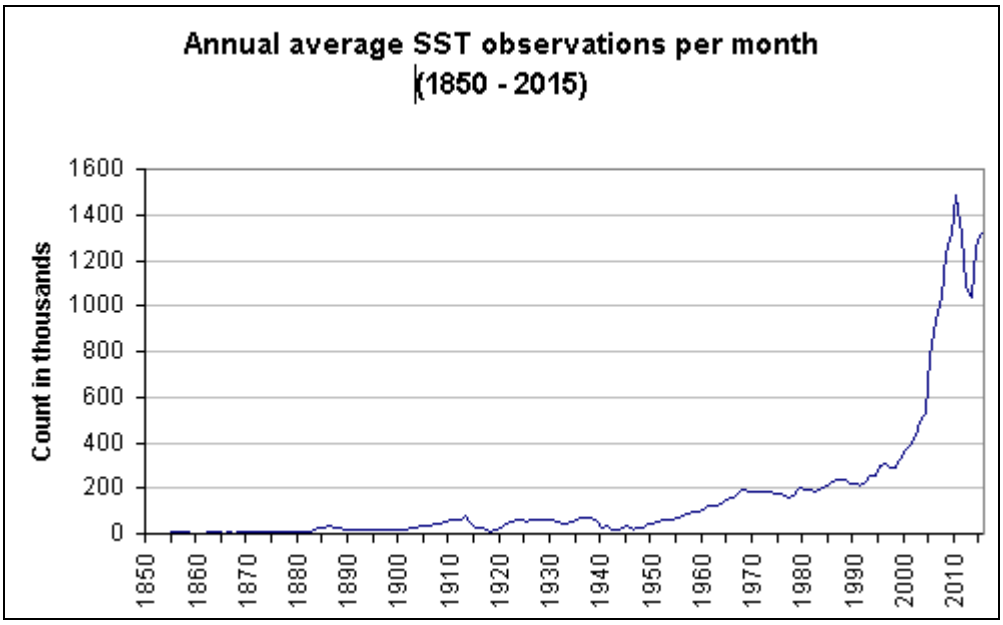


Figure 3.1 Annual average SST observations per month

McKitrick (2010) reports similar total numbers of observations for the period 1936 to 2005, albeit for the HadSST2 data rather than the HadSST3 dataset. The only change of note between HadSST2 and HadSST3 is the inclusion of data transcribed from ship's logs, which is not generally a significant inclusion but might be of value in areas for which no previous data exists.

The number of SST observations in each month has generally increased over time, observations in the northern hemisphere especially being more seasonal. Starring from 1850, the total number of observations per month did not exceed 10,000 until May 1879, having been as low as 1,430 in November 1851. The number reached 89,810 in August 1913 but fell sharply the next year, from 72,390 in February 1914 to 41,470 in July and 28,000 in August due to the outbreak of World War I. As Figure 3.1 indicates, the number of observations per month recovered to pre-war levels in the 1920's and 1930's but fell again during World War II and did not consistently exceed 200,000 until 1985. The surge in observations per month after year 2000 can be attributed to the increasing use of Argo buoys.

The number of observations per grid cell in each month has also varied throughout the period from 1850 to 2015. Since the year 2002 some grid cells have monthly

observation counts of more than 5000 and in some instances in excess of 10,000 because of the use of Argo buoys for SST monitoring, but at the other end of the scale (Figure 3.2) in the same period some grid cells have fewer than 6 observations (and of course some have no observations at all because of sea ice or because no ships took sea surface measurements within the grid cell in the month.)

Even in December 2015, 33 grid cells had only a single observation for the month and a further 67 had from 2 to 5. Only in nine of the 103 years from 1850 to 1952 did the average annual percentage of grid cells with a single observation fall below 5% of all reporting grid cells.

The situation is summarised in Figure 3.2 which shows the global number of grid cells with the specified range of observation counts as a percentage of all reporting grid cells. From this Figure we see that grid cells with from one to five observations per month accounted for an average of about 30% of reporting grid cells from 1870 to 1950 and grid cells with from 6 to 16 observations accounted for only slightly less, meaning that more than 50% of reporting grid cells had an observation rate of one or fewer every two days.

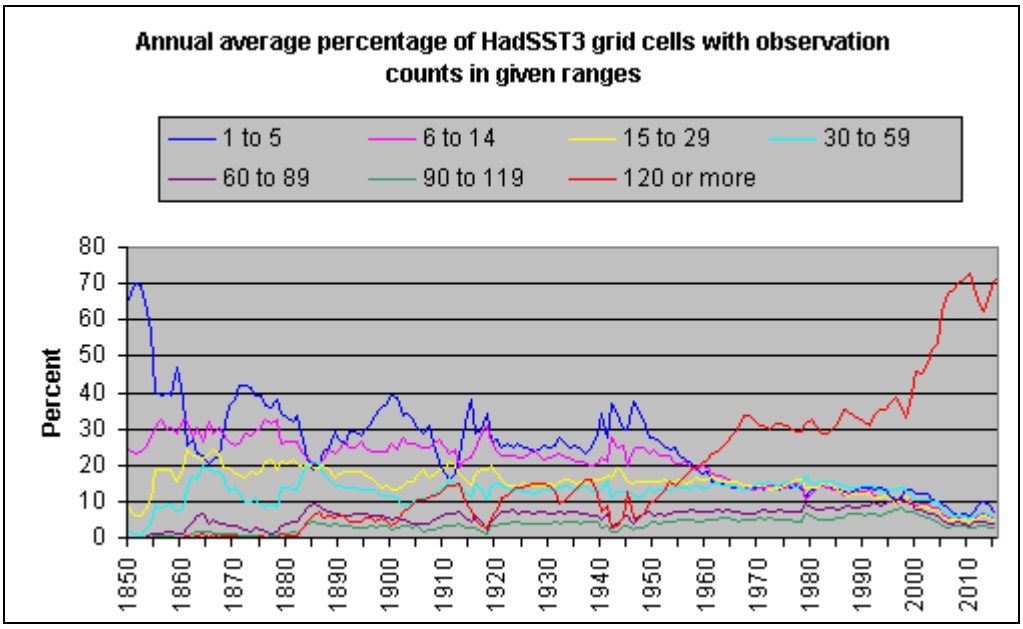
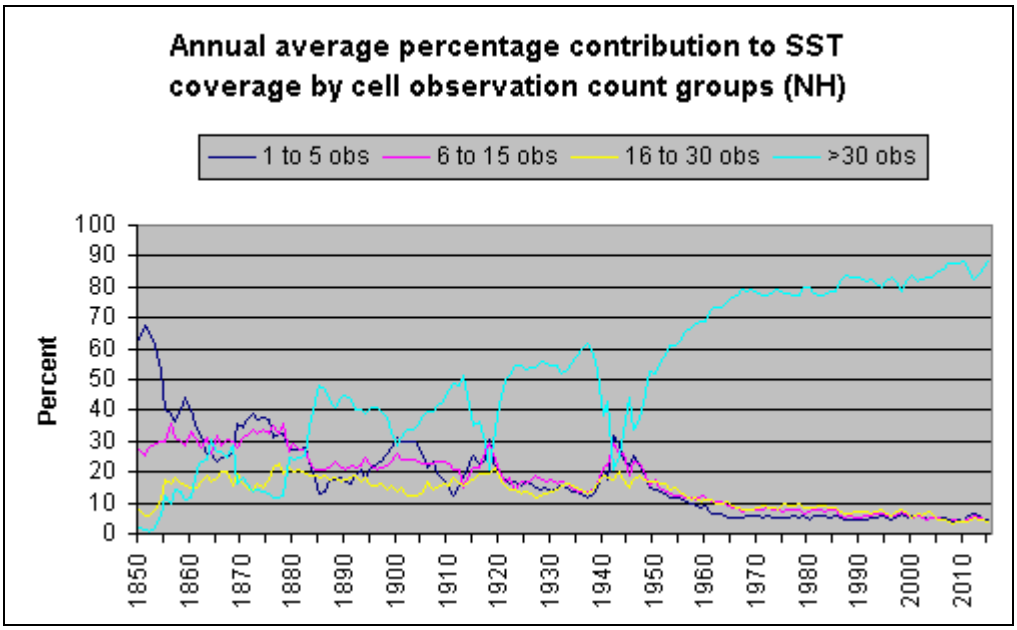


Figure 3.2 Annual average of the monthly number of grid cells with the specified ranges of observation counts expressed as percentage of all SST grid cells with data in the month (e.g. after 1990 the data for more than 30% of reporting grid cells was based on at least 120 observations).

This can also be expressed in terms of the contribution that such grid cells made to the coverage of each hemisphere (Figure 3.3). In January 1850, for example, according to the HadSST3 concept of coverage, grid cells with from 1 to 5 observations accounted for 13.9% of the Northern Hemisphere surface area, cells with 6 to 15 observations accounted for 5.3%, cells with 16 to 30 observations accounted for 1.6% and cells with more than 30 observations accounted for 0.2%, making a total coverage of 21.0% of the hemisphere's surface. Grid cells with 1-5 observations account for 66.1% of that fraction of the total surface area and cells with 6 to 15 account for 25.1%, meaning that the hemispheric average SST in January 1850 was derived from very sparse observations and consequently should have a high error margin.

In the Northern Hemisphere the annual average coverage contribution of grid cells with from 1 to 15 observations per month was 93% in 1852 but fell to approximately 45% by 1885, remained at that point with only some mild fluctuation until 1950 before falling gradually to around 13% by 1970 and since then to about 8.5% by 2008. This number of observations ranges from an average of 1 per month to 1 every second day. The set of lowest number of observations, from 1 to 5, is from one observation per month to one every 6 days and grid cells with this number of observations account for about 5% of the annual average Northern Hemisphere SST coverage, as they have since about 1964 (i.e. the last 50 years).

The situation in the Southern Hemisphere is different because grid cells with from one to five observations accounted for more than 33% of SST coverage from 1850 to 1951 except for three brief periods in the 1860's, the mid 1880's and around 1910. Grid cells with from 6 to 15 observations accounted for only a slightly smaller percentage of the coverage, meaning that about 80% of the coverage from 1850 to 1950 came from grid cells that averaged one observation every two days or fewer. It was not until 1995 that grid cells with more than 30 observations in the month, an average of one per day, consistently accounted for more than half of the SST coverage of the Southern Hemisphere.



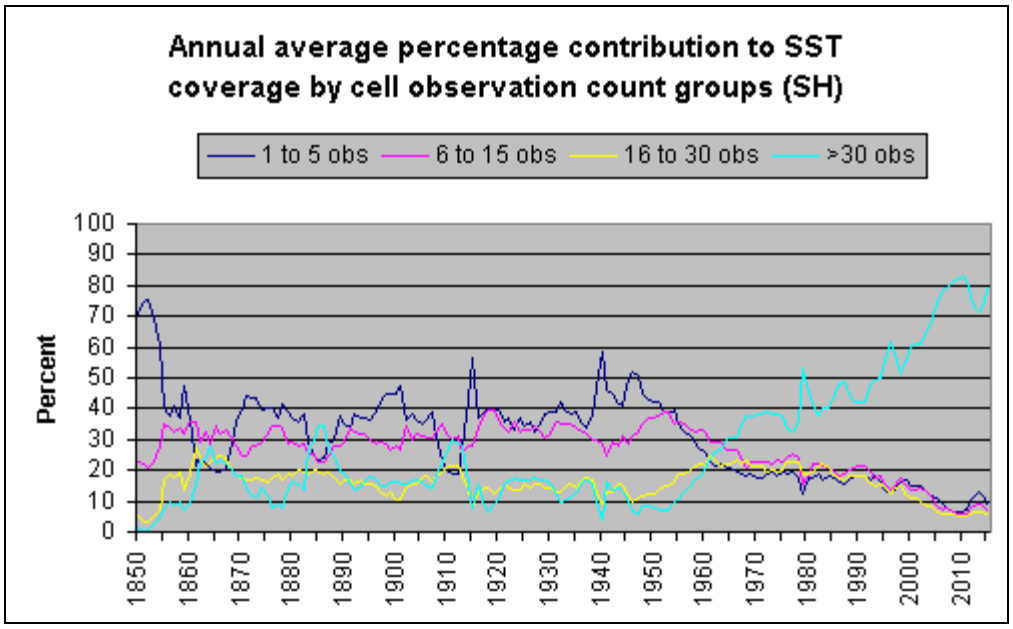


Figure 3.3 Annual average percentage contributions of grid cells with certain ranges of observation counts to the total hemisphere coverage. Top: Northern Hemisphere; bottom: Southern Hemisphere.

To put these observation counts in HadSST3 grid cells into context, the calculation of monthly SST values is based on grid cell sizes of 1° latitude x 1° longitude and according to 5-day periods known as "pentads" (Chapter 1). On this basis a single observation in each grid cell and pentad would require that a minimum of 150

observations (i.e. 5 x 5 x 6) were made in each HadSST3 5° x 5° grid cell in each month. On purely statistical grounds the accuracy of any interpolation to resolve a value for a monthly value for a 5° x 5° grid cell will depend on the temporal and spatial distribution of the available data for the month, the latter particularly relevant if any significant ocean currents occur in part of the 5° x 5° grid cell. The temporal dimension also adds a factor not present in data from land-based observation stations where the monthly mean temperature is calculated from a minimum of 20 days of data, using the daily minimum and maximum temperatures, whereas SST data is from any time of day.

3.3 Variation in number of observation stations

The investigation of the number of land-based observation stations started with the processing of each station data file, made available in a composite archive file by the CRU. Jones et al (2012) describes the criteria for rejecting the data for various stations from the CRUTEM4 calculations. This meant the rejection of:

(a) 16 stations that have no valid Latitude and/or Longitude (typically set to 99.9 and 199.9 respectively),

(b) 852 stations where the long-term averages have been set to -99.99 (i.e. missing data), because no temperature anomalies could be calculated for their data

(c) 75 stations where the standard deviations were fixed at 100, which seemed to indicate that standard deviations could not be calculated.

Analysis of the entire set of station metadata revealed several instances of pairs of stations with the same latitude and longitude (to 1 decimal place) and one instance of 3 stations in that situation. The filtering described above as (a), (b) or (c) removed 27 such entries from consideration, leaving just 21 station pairs that shared the same location. In some cases the paired stations were clearly different (e.g. different elevation or different long-term averages) but in one case, Norwegian station "Kjobli i Snasa", all information for both entries was identical so one instance was rejected.

The data for all remaining stations, including those retained after the filtering described in the paragraph above, were used to determine the number of stations that

reported their monthly mean temperature in each month of each year from 1850 to 2015. The results, expressed as annual averages are shown in Figure 3.4 where the totals excluding those in the USA are also shown separately because of the large number of stations in that country relative to its size (~1.85% of the Earth's surface or ~6.2% of total land area). More than 64% of stations that reported in 1895 were located in the USA, a percentage that fell to about 25% for the period from 1960 to 1975.

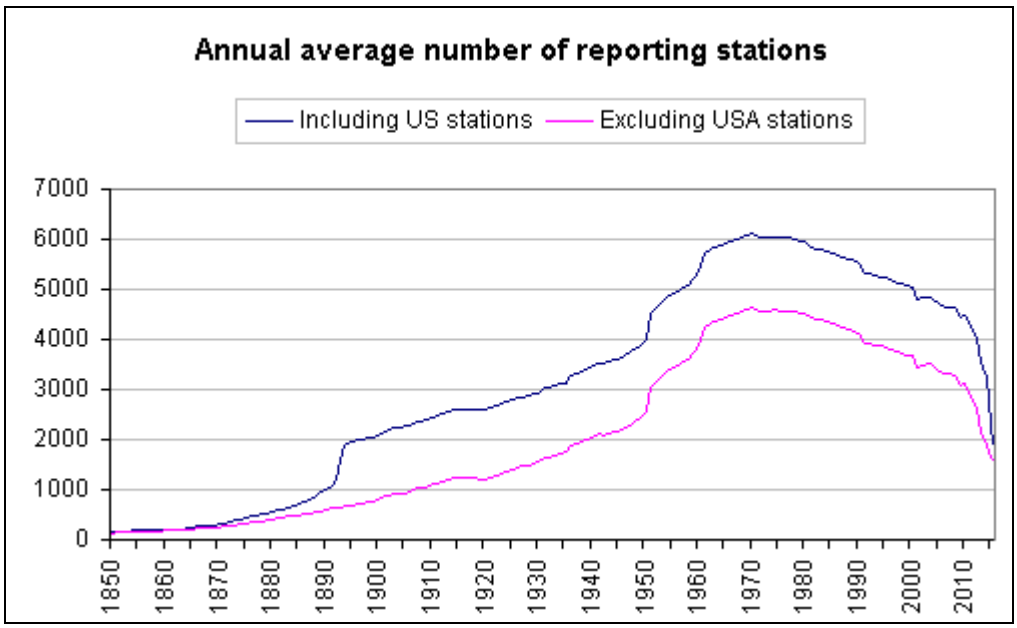


Figure 3.4 Annual average total number of reporting observation stations in each month.

McKitrick (2010) discusses a sharp reduction in station numbers after 1990 and shifts late in the twentieth century that included an increasing number of stations located at airports and reductions in the number of stations at high altitude or latitude. Apart from the matter of McKitrick (2010) working with CRUTEM3 data rather than the newer CRUTEM4 data, according to Jones et al (2012) some national meteorological services supply data to the CRU at decade intervals rather than annually or monthly, which accounts for what McKitrick (2010) described as a shortfall in 1990 now being a reduction after 2010.

The large disparity between land coverage in the two hemispheres prompted the separate totalling of the two hemispheres (Figure 3.5) and attention is drawn to the separate scales for each hemisphere caused by the great difference in total counts.

For the entire Southern Hemisphere only a single station - Padang/Tabing in Indonesia at 0.9°S 100.4°E (Station No. 961630) - reported data in each month from January 1850 to December 1852. The number of reporting stations in the Southern Hemisphere did not exceed 10 until October 1860, exceed 25 until in late 1874, exceed 50 until May 1889, which is almost 40 years after the start of CRUTEM4 data, and did not exceed 100 until January 1904. The amount of data in these early years of the record means that a very substantial error margin should be associated with the CRUTEM4 data of that time, not only for the southern hemisphere average but also for the global average calculated using that SH average.

In contrast, in the Northern Hemisphere 145 stations reported data for the first month of the CRUTEM4 dataset, namely January 1850 and six years later this had reached 200. In May 1889, when the number of reporting Southern Hemisphere stations reached 50 the number in the Northern Hemisphere was 909. With the exception of 2015, when the data might not be complete, the annual averages since the start of data in 1850 show at least 84.6% of reporting stations being located in the Northern Hemisphere, this despite the CRUTEM4 global average being calculated from a weighting of 2:1 for the Northern Hemisphere opposed to the Southern Hemisphere. Some of the bias beyond the ratio or land areas in the hemispheres is due to the number of stations in the USA but even so the Northern Hemisphere station count increased from 145 to 171 before the Southern Hemisphere had its second reporting station.

There might be an inclination to assume the observation stations to be evenly spread across the relevant grid cells but data from the 'observation count' file, published by the CRU and noting the number of reporting stations in each grid cell and month, shows that such an assumption would be false. From this supplementary file the number of grid cells with the same number of reporting stations (e.g. grid cells with 2 reporting stations) was determined on a monthly basis. Figure 3.6 shows the annual

average of those counts, grouped into the described ranges, expressed as a percentage of all CRUTEM4 grid cells that reported data in the month.

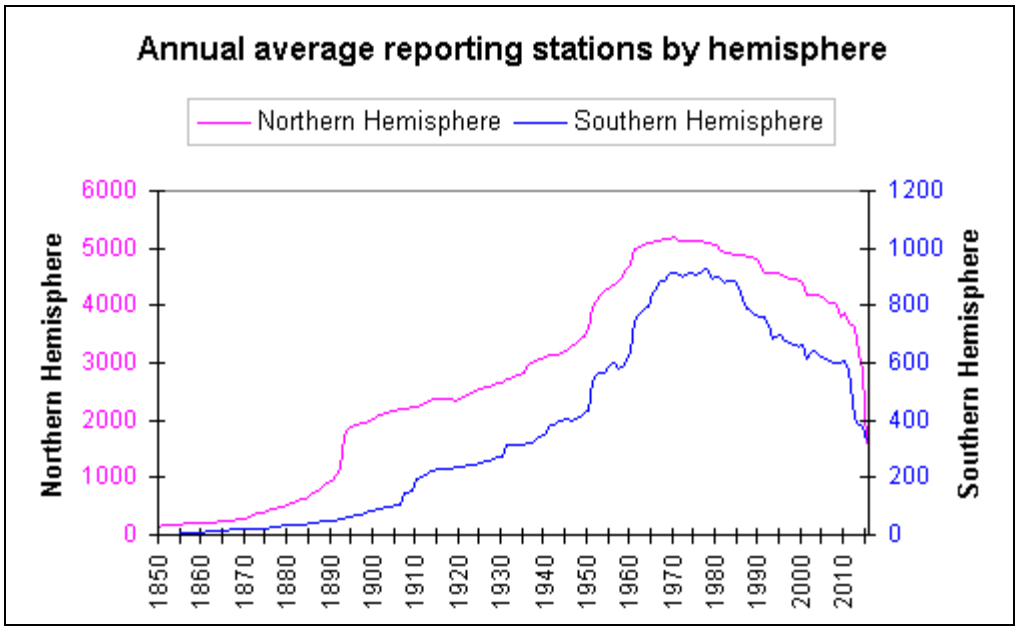


Figure 3.5 Average annual reporting observation stations for each hemisphere. The data is separately scaled on the Y-axes at each side of the graph

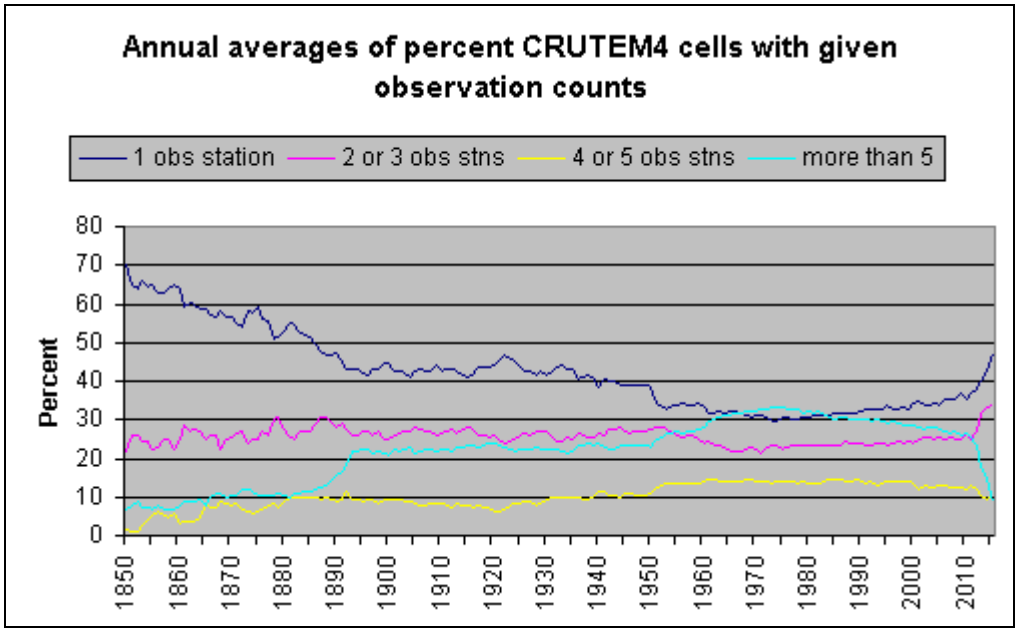


Figure 3.6 Annual average percentages of reporting grid cells with the specified numbers of reporting observation stations

The key point of Figure 3.6 is that until 1945 the data for more than 40% of grid cells was derived from a single observation station and that the data for about 28% of grid cells came from just two or three observation stations. Together these mean that the data for more than two-thirds of reporting grid cells was derived from fewer than four observation stations.

When grid cells have few reporting stations the average temperature anomaly for the grid cell can be greatly influenced by the presence or absence of other stations. If, for example, a grid cell contains two observation stations and one station has an anomaly of +2.0°C and the other of +1.0°C in the same month (which might happen when the latter station is near the coast) the average of the two stations is 1.5°C but the omission of either would mean the value for the grid cell was either +2.0°C or +1.0°C depending on which station was present.

The error margin associated with some data is often calculated by dividing the standard deviation by the square root of the sample size. On this basis grid cells with a large number of reporting stations will produce a lower error margin that a grid cell with the same standard deviation but fewer reporting stations. Instances of more than 40 observation stations were reported, in any month from January 1850 to December 2015, for 20 grid cells, all in either North America or central Europe. The greatest number was when 79 observation stations reported data in the grid cell centred at 42.5°N 72.5°W, which is in the north eastern USA. In this grid cell more than 70 observation stations reported data in each month from February 1893 to December 2015.

3.4 Summary

For HadSST3 sea surface temperature anomalies, on the basis of annual averages, for most of the period from 1850 to 1950 HadSST3 grid cells with from 1 to 5 observations per month comprised about 30% of all reporting grid cells, and these in

combination with cells with from 6 to 15 measurements, accounted for more than 50% of reporting cells.

Based on annual averages, from 1850 to 1950 more than 40% of the NH SST coverage can be attributed to grid cells with from 1 to 15 observations for the month, with about 25% of the total coverage coming from grid cells that had from 1 to 5 observations per month. In the Southern Hemisphere about 70% of the SST coverage can be attributed to grid cells with from 1 to 15 observations per month, of which 40% of the total coverage came from grid cells with from 1 to 5 observations per month.

For CRUTEM4 temperature anomalies from observation stations, for the period from 1850 to 1950 more than two thirds of grid cell data was derived from cells with from 1 to 3 observation stations.

For both of these datasets a very high proportion of the grid cell data is derived from very few observations. It is impossible to determine or even estimate the difference that the inclusion of more data might have made to grid cell values and therefore to the hemispheric and global averages.

In relation to the HadSST3 dataset there can be little confidence that so few observations in a month presents an accurate summary of sea surface temperatures over that month, not even with as many as 15 observations because they might be clustered in particular 1° x 1° sub-cells or in specific pentads (i.e. 5-day periods). As discussed above, a minimum of 150 SST measurements is required in each HadSST3 grid cell in each month in order to avoid any interpolation, so having just 15 observations falls well short of that.

In relation to the CRUTEM4 data the reliance of a single station to provide data for a grid cell risks using data from an atypical location (eg. elevated when much of the area covered by the grid cell is not) and risks having no data whatsoever from that grid cell when the single station fails to supply it. Grid cells that contain two or three observation stations are less likely to fail to have data than a grid cell with a single station but if the observation stations are in very different locations and produce quite

different temperature anomalies then the failure of one station to report data could skew the grid cell's value.

The issues of sample size that have been discussed here are unlikely to skew the HadCRUT4 average temperature anomalies in any particular direction, any errors in one direction being likely to generally counterbalance any errors in the opposite direction, but they do widen the uncertainties, especially when focus is narrowed to smaller numbers of grid cells as might be the case with studies of regional temperature patterns.

Chapter 4: Long-term average temperatures

4.1 Introduction

HadCRUT4 monthly temperature anomalies are calculated relative to a base temperature, which for a given month is defined to be the average of the mean temperatures in the same calendar month over the 30-year period from 1961 to 1990. For land observation station data the long-term average is referred to as the "normal" and for sea surface temperatures it is known as the "climatology".

This chapter will discuss issues of the suitability of this period and some of the relevant statistical issues before looking in more detail at the calculation of long-term averages for CRUTEM4 observation stations and HadSST3 sea surface temperatures.

One of the key aspects of this chapter is whether the data used to calculate long-term averages has normal (i.e. Gaussian) distribution. Certain conclusions and generalisations about the data might not apply if the distribution is not normal.

WMO (2011) states explicitly that the period of 30 years was set as a standard period over which long-term average temperatures are calculated, "mainly because only 30 years of data were available for summarization when the recommendation was first made". Perhaps by good fortune 30 entries is the minimum that some statisticians recommend for the application of the Central Limit Theorem (CLT), the means of drawing general conclusions about a large population (i.e. the full set of data) from a smaller sample that shows normal (i.e. Gaussian) distribution, regardless of whether the large population itself has a normal distribution. Rumsey (2011) indicates a minimum of 30 years is required for the application of CLT, although Witte &Witte (2010) says "depending on the degree of non-normality in the parent population, a sample size of between 25 and 100 is sufficiently large".

WMO (2011) provides guidance on the use of goodness-of-fit tests, by which the data can be tested for normal distribution, in climatology studies involving data from

observation stations, mentioning that the chi-squared and Kolmogorov-Smirmov tests are commonly used. It gives no reason for using these tests over any other test, especially the widely used Shapiro-Wilk test that is also used in this chapter.

The Shapiro-Wilk test determines how well the distribution of data fits the normal Gaussian distribution bell-curve by considering both the symmetry and spread of the data. Ultimately it determines a type of correlation coefficient between the data and the Gaussian distribution pattern that would be expected from the same sample size, then turns the coefficient into a p-value for the null hypothesis that the data is indeed normally distributed. Often misinterpreted, the p-value is a level of confidence rather than a probability and actually indicates whether the data *might be* normally distributed, not to be confused with *is* normally distributed. A low p-value, below a somewhat arbitrary threshold, is regarded as indicating whether the data is not likely to be normally distributed. Opinions vary as to whether a threshold of 0.05 or 0.1 is more appropriate in various circumstances. A high p-value is merely an indication that the data seems more likely to be normally distributed but is no guarantee that it is and is not even a probability that it is.

This chapter includes discussion of the application of the Shapiro-Wilk goodness-of-fit test to both the CRUTEM4 long-term averages and the HadSST3 mean values over the period 1961-90. The p-values shown in these analyses were obtained through bespoke software created according to the Excel code of a source[10] that draws directly on Shapiro and Wilk (1965), the original work that describes the method, and with the software verified by testing it against the given Excel code.

4.2 The suitability of 1961-90 as the base period

The period over which the long-term average temperatures are calculated for each observation station (in the case of CRUTEM4) or 1 x 1 grid cell (in the case of HadSST3) is of questionable suitability given that two of the major drivers of

[10] http://www.real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/shapiro-wilk-test/

temperature, volcanic eruptions and the El Nino-Southern Oscillation, are somewhat unusual across this time.

Major volcanic eruptions, with a volcanic explosivity index (VEI) (Simkin et al., 1981; Newhall & Self, 1982; Simkin and Siebert, 1994) of 5, occurred three times during 1961-1990 - Agung (1963), St Helens (1980, VEI:5), El Chicon (1982 - along with 19 other volcanic eruptions with a VEI of 4. Stothers (2001) shows that these eruptions caused perturbations in the stratospheric optical depth, which means that they blocked a portion of the solar insolation, which in turn meant a reduction in measured temperature. Stothers (2001) also shows that prior to the Agung (Indonesia) eruption the last previous major perturbation was caused by the Santa Maria (Guatemala) eruption in 1902, more than 60 years earlier.

The other major influence on temperature is the El Nino-Southern Oscillation with El Nino events bringing generally warmer weather and La Nina events bringing generally cooler weather. McLean et al (2009) and de Freitas & McLean (2013) respectively show that changes in the Troup Southern Oscillation Index (SOI) (Troup, 1965) correlate well with corresponding changes in lower tropospheric temperatures about 7 months later and near-surface temperatures about five months later.

The trend in the annual average HadCRUT4 global average temperature anomalies was near-flat from 1945 to 1979 (downward trend of ~0.1°C/century) before showing a persistent if slightly irregular increase until 1997 (upward trend ~1.32°C/century). Figure 4.1 shows the monthly values for the HadCRUT4 global average - assuming for the moment that it is accurate - and the Troup SOI. The delayed inverse relationship between the pair, due to by negative values of SOI indicating conditions on the El Nino side of neutral and El Nino events being associated with higher temperatures, is clear, especially prior to 1987.

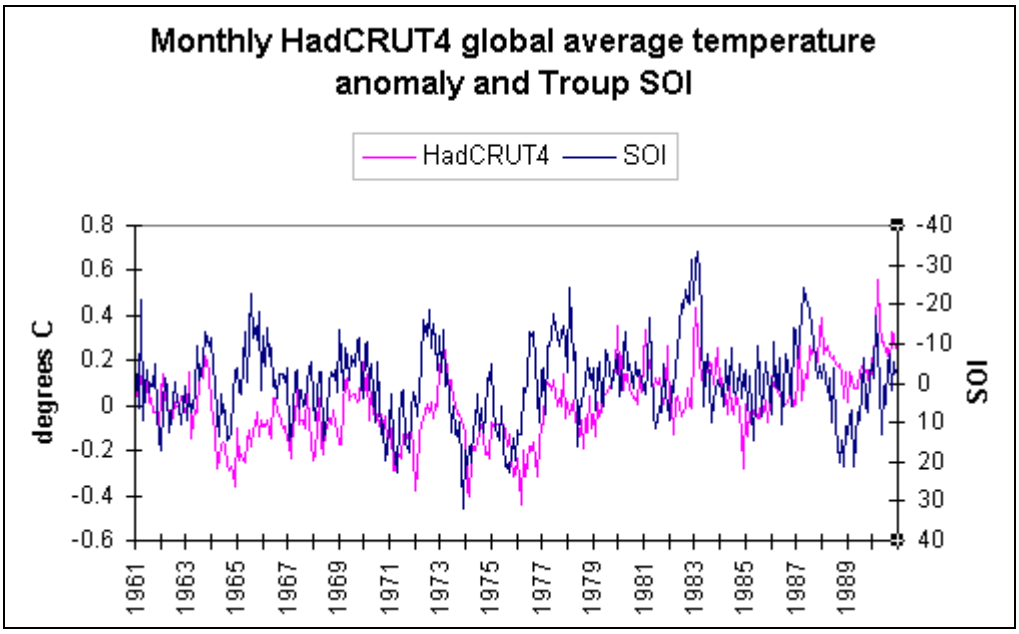


Figure 4.1 Monthly HadCRUT4 temperature anomalies and Troup SOI for the period 1961 to 1990'

Another way to consider this data is in the form of running aggregate values - which is a simple total of the values to date from the start point - (Figure 4.2) where a downward trend indicates a preponderance of negative values and an upward trend a preponderance of positive values. Figure 4.2 shows a clear switch in the SOI in 1976 (April-June), from a predominantly positive state (associated with La Nina events) to a predominantly negative state change (associated with El Nino events). The HadCRUT4 global average temperature anomalies are disturbed somewhat by the cooling due to the volcanic eruptions mentioned above but briefly switched from mainly negative values to positive values in December 1976 or January 1977, reverted to mainly negative values about 12 months later, and then back again to mainly positive values after May 1979.

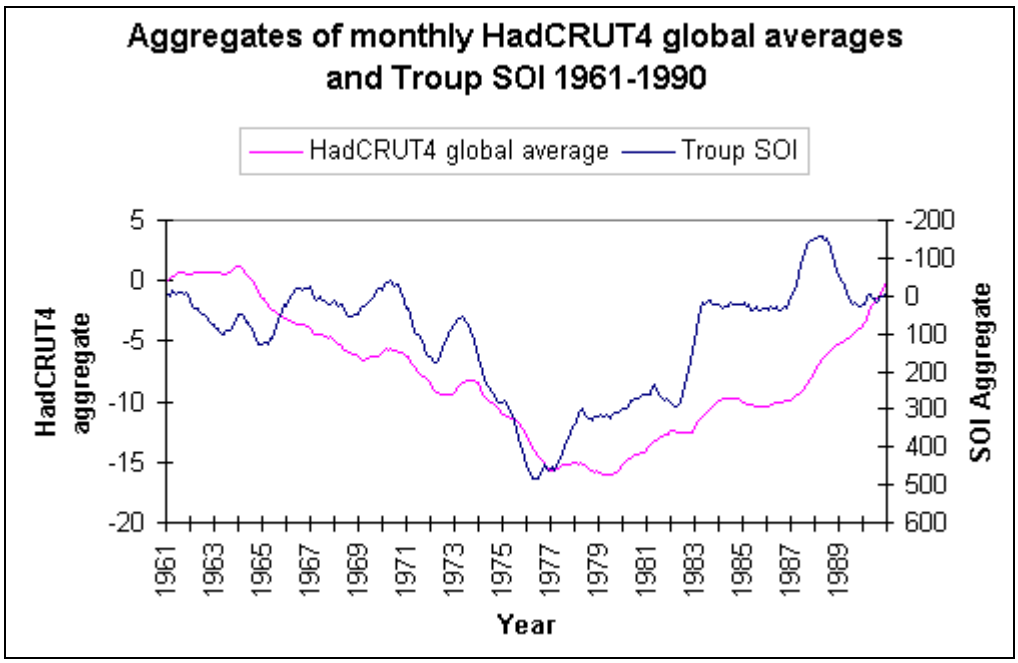


Figure 4.2 Running aggregate of monthly HadCRUT4 temperature anomaly and Troup SOI from January 1961 to December 1990, one with an inflexion in 1976 and the other in 1979.

The changes in temperature trend across the period indicate that if less than 30 years of data for that month are available the mean temperature might be biased above or below the mean that would be calculated if data was available for all years. If the data for a given station follows the pattern of the HadCRUT4 global average for example, a shortfall in the period 1961 to 1976 will likely produce a higher mean value than a shortfall in the period 1977 to 1990.

In terms of anomalies for individual stations (CRUTEM4) or individual grid cells (HadSST3) the values will uniformly decrease or increase with different long-term averages but trends in the anomalies will remain the same. The problem arises when the anomalies for stations or grid cells with long-term averages calculated from less than complete data are averaged with other stations or grid cells that have anomalies based on long-term averages over other periods.

As a simple example, consider two nearby observations in the same grid cell with, for the same calendar month, long-term averages of 17C and 18C respectively, one of which was calculated using data from all such months during 1961-1990 and one calculated from identical monthly data but for just 20 of those years. A mean monthly

temperature of 20C at both stations would mean anomalies of +3°C and +2°C and an average temperature of 2.5C. The failure of one station to report data in that month would result in an average temperature anomaly of +3°C or +2°C depending on which station was missing. Any relative expression such as "two degrees above average" is unsafe because it would depend on the anomaly being used, which in turn would depend on the long-term average being used.

4.3 Issues with observation station long-term average temperatures

4.3.1 Introduction

As noted above, there are issues associated with the long-term averages from observation stations. Firstly there is the method by which the long-term averages are calculated and the minimum sample size used by the CRU. Secondly we have the assumption of normal distribution, the calculations of the standard deviation and the limits used to identify data outliers in the entire data time span, this despite possible data outliers being included when calculating the standard deviation.

It cannot be overstated that errors in long-term averages will be carried through to every temperature anomaly calculated using those averages and that error margins will likewise be carried through. As noted previously, errors in the anomalies for a single observation station will have negligible impact but when anomalies are averaged with those of other stations the impact and implications might be important.

4.3.2. Calculated and estimated long-term averages

The World Meteorological Organization (WMO) provides guidelines about the procedures and practices when determining 'normals' (i.e. long-term average

temperatures[11]) for surface temperatures, guidelines that are not followed by the Climatic Research Unit (CRU) when it derives its normals.

WMO TD-No. 341 (1989) says "If for a given month (e.g. January) 3 consecutive year-month values (e.g. January 1970, 1971, 1972) are missing or more than 5 values in total for the given month are missing, the 30-year standard normal should not be calculated." The same document goes on to discuss "provisional normals", which may be calculated if at least 10 year-month values are available but insufficient to calculate the standard 30-year normals, and emphasises that such normals should be identified by a "provisional" indicator.

WMO (2011) says "As a guide, normals or period averages should be calculated only when values are available for at least 80 per cent of the years of record, with no more than three consecutive missing years."

The CRU ignores the WMO recommendations, first by failing to provide an indicator that long-term averages are provisional when stations have less than 25 years of data for a given calendar month, and then by including station data when less than 80% (24 years) are available.

The minimum amount of data acceptable for inclusion in the CRUTEM4 dataset is 14 years of data across the 30-year period, a minimum that has fallen with each successive version of CRUTEM data. (The first version of CRUTEM required 25 years of data, the second 20 years with at least 4 in each of the three decades from 1961-90 and the third 15 years from anywhere in that period.) The 14-year threshold is not for a number of years of a given month but for years as a whole, each year subject to a minimum permissible number of months without data. The previous version of the dataset, CRUTEM3, had an instance where the observation station had the requisite minimum of 15 years of data (as per the limit then) but the data was such that the station had no more than 14 years of data for any calendar month.

[11] Personally I prefer the term "long-term average temperatures" because as section 4.2 shows, the period over which these figures are calculated is not normal. The term "normal" is however used here for consistency with the cited documentation.

The omission of any monthly mean temperatures for a given calendar month will bias the normal towards whatever data was present. WMO (2011) states that normals calculated from incomplete datasets can be biased and says, rather obviously, that a normal calculated without a particularly cold month would be warmer than a normal with that month included. This can be extended to say that if the omissions are of a number of warmer years then the normal will be biased low and consequently temperature anomalies calculated from these low normals will be higher than they would be with the full set of data.

While temperature data from any given station tends to vary randomly, the global average annual CRUTEM4 temperature anomaly shows a slight downward trend from 1961 to 1978 then an upward trend from 1979 to 1990. If individual stations followed this pattern then the absence of data in the period 1961 to 1978, when temperatures were falling, but present in from 1979 to 1990 will skew the long-term average upwards and the opposite situation skew it downwards. Further, the initial period with the slight downward trend is more than 14 years, meaning that an observation station might meet the criteria for the minimum number of years but with data that comes entirely from the period of cooling.

Before looking specifically at the derivation of CRUTEM4 station long-term average temperatures (i.e. 'normals') it is useful to consider the scope of the potential problem of including data from stations that failed to meet the WMO criteria of a minimum of 25 values out of the 30 for the period from 1961 to 1990 and no more than three consecutive values flagged as 'missing'. Figure 4.3 shows the average annual number of HadCRUT4 and CRUTEM4 grid cells for which data is available, along with the number of grid cells that have data but fail to meet the WMO criteria in each calendar month.

Figure 4.4 shows the average number of grid cells with data from sources failing to meet the criteria expressed as a percentage of the total numbers of HadCRUT4 and CRUTEM4 reporting grid cells. The peak percentage of HadCRUT4 grid cells was 20.57% in 1974 but for the period from 1969 to 1978 inclusive only one year averaged below 20%, which was 19.68% in 1972. For the 73 years from 1909 to 1981, except for 13 years (11 during and immediately after WW II plus 1959 and

1960) did the percentage of the annual average number of reporting CRUTEM4 grids sourced from data that failed to meet WMO criteria fall below 33.33% (i.e. one-third).

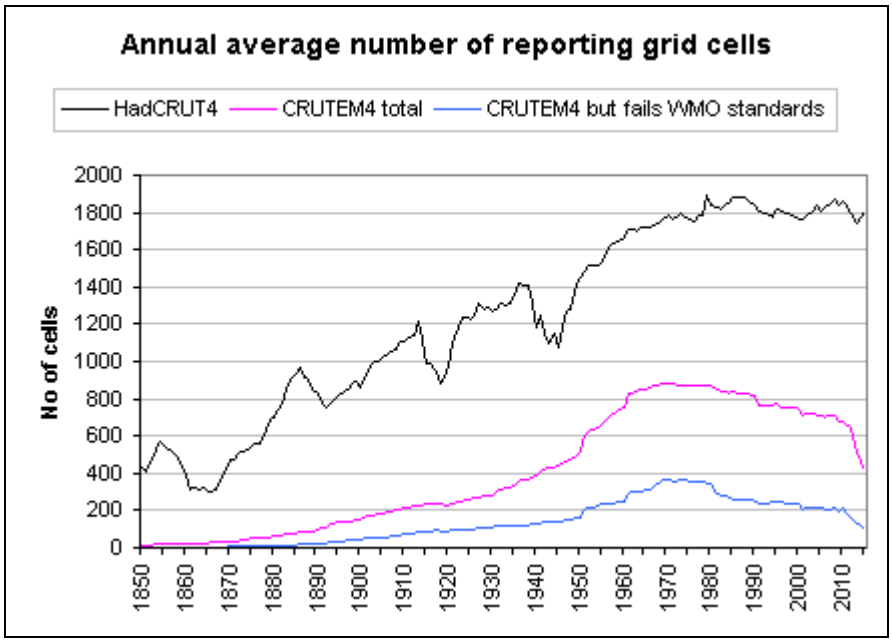


Figure 4.3 Average annual number of reporting grid cells for HadCRUT4 and CRUTEM4 datasets plus the number of grid cells that included data for which long-term average temperatures do not meet WMO standards

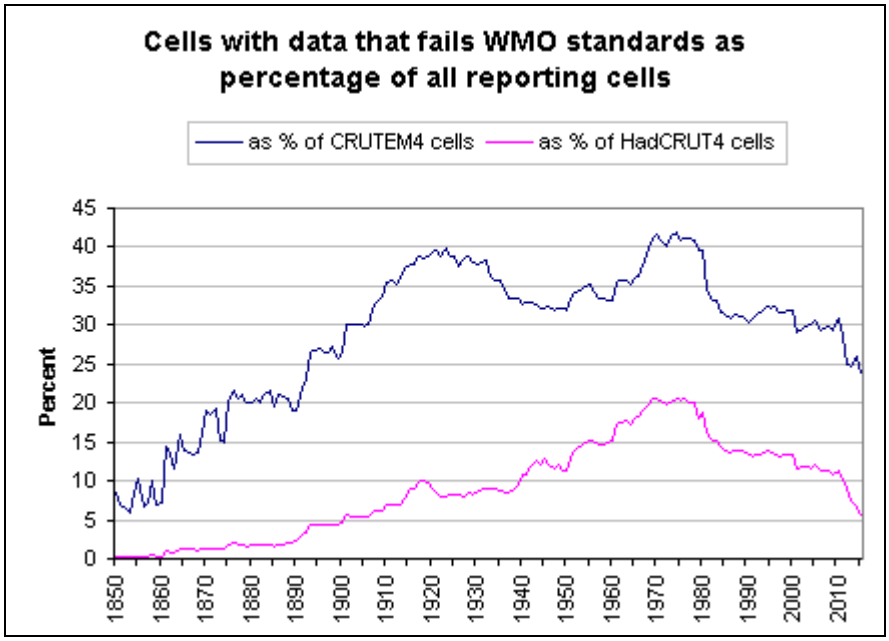


Figure 4.4 The average annual number of grid cells that fail to meet WMO criteria expressed as a percentage of the total number of reporting HadCRUT4 and CRUTEM4 grid cells

The CRUTEM4 normals for a given station might be determined from the station's data during the period 1961-90 or they might be estimated, which is to say calculated by means other than via a minimum of 14 years of data during 1961-1990. Osborn and Jones (2014) describe the alternative methods by which the normals might be estimated, the list in order of priority as follows:

> 1. If the station has sufficient data to estimate a normal for the 1951–1970 period and the grid box from an earlier version of CRUTEM contains data across the longer 1951–1990 period, then we estimate the 1961–1990 normal for the station using its 1951–1970 normal adjusted by the difference between the grid-box averages (in the earlier version) for 1961–1990 and 1951–1970.
>
> 2. If a neighbouring station does have sufficient values to determine its normal, then we calculate the mean difference between the temperatures recorded at this neighbouring station and the temperature recorded at the current station over a different period when they both have data (e.g. 1951–1970), and assume that this mean difference still holds during the reference period. The normal for the current station is then calculated as the sum of the normal for the neighbouring station plus the mean difference between the two stations' temperatures.
>
> 3. If the World Meteorological Organisation (WMO) have published a 1961–1990 normal for the station and the reference period (perhaps because the National Meteorological Service had calculated it from additional data not available to us), then we use that. We rely on a WMO normal for about 150 (~ 2.5%) of the stations.

Each of these methods is not without its problems. Method 1 is flawed because earlier versions of CRUTEM used data that was inconsistent with data supplied by national

meteorological services (NMS)[12]. A comparison of CRUTEM3 and NMS data for various locations in Russia, Norway, Iceland, Switzerland, Australia, New Zealand, Tahiti and the USA revealed differences in data supply (i.e. missing months) and mean monthly temperature between the two sources. The greatest differences in the NMS and CRUTEM3 station data of those examined were from eastern Russia[13], some examples of which are for the following locations:

Kirensk (ID: 302300)	No difference between CRUTEM3 and NMS data in 241 of the 360 months across 1961-90 but 10.0°C in January 1963, 10.0°C in March 1974, 3.4°C in May 1979[14]
Vitim (ID: 300540)	across 1961-90 only 41 of the 360 months had matching data, the differences of the other months almost all within -0.8°C to +1.2°C but 3.8°C in May 1979[15].
Tura (ID: 245070)	Of the 360 months 236 had no difference but others included 11.3°C (Feb 1978), -9.7°C (Jan 1962), 8.6°C (Oct 1964) and -4.5°C (Mar 1964).
Bratsk (ID:30309)	The differences in 1981 in each month are as follows: -9.5°C, -10.2°C, -6.7°C, -0.5°C, -0.3°C, -1.7°C, -0.5°C, -0.2°C, -3.0°C, 0.3°C, -7.1°C and -5.7°C, this despite about 80% of all other months 1961-90 showing no difference at all.
Kyra (ID: 30949)	Of the 360 months across 1961-90, 67 had differences either ≤-1.0°C or ≥+1.0°C, with only one year in which those limits were not exceeded in any month and only the months from

[12] The most significant change from CRUTEM3 to CRUTEM4 was that adjusted data for each station is supplied by the NMS's, which means that CRUTEM4 data is consistent with the data available directly from those NMS's.
[13] The NMS data for locations in eastern Russia can be found via http://neacc.meteoinfo.ru
[14] With few exceptions, the differences between Kirensk HadCRUT3 station data and the NMS data for every February from 1892 to 1940 was -2.7°C and for March over the same period -2.4°C
[15] Differences between Vitim HadCRUT3 station data and the NMS included -11.6°C (Jan 2003), -10.8°C (Mar 2002) and -9.8°C (Jan 1993)

March to July inclusive showing close agreement between CRUTEM3 and the NMS across the period.

The obvious conclusion is that CRUTEM3 station data is unreliable and should not be used to estimate long-term average temperatures.

Estimating using method 2 is questionable because it involves several assumptions, in particular that there is a consistent relationship between the data from the two (or more) locations, which might not be true if exposure to certain weather conditions differs.

Estimating using method 3 makes the false assumption that mean temperatures taken from a period other than 1961-90 will still apply across 1961-90, this despite most (or perhaps all) stations showing temperature anomaly trends that are not flat over time.

As an example, data for Canadian station Cape Dorset (ID: 715750) is available continually since 1927 but the other station in the same grid cell, Nottingham Island (ID: 719080), has data only from 1930 to 1970, which means it has less than the minimum 14 years for the period 1961-90. The source of the normals for Nottingham Island is described as "extrapolated" but this does not distinguish between methods 1 (using an earlier period) and 2 (via relationship with neighbouring station). Figure 4.5 shows the difference in normals for the two stations in each calendar month along with the average difference in mean monthly temperature for the period of overlap (1930 to 1970 with data 22 months missing). In both cases the differences were calculated by subtracting data for Cape Dorset from data for Nottingham Island.

The pattern of the difference in average mean monthly temperature is quite different to the pattern for the difference in normals, which suggests that the normals for Nottingham Island were not estimated from those for Cape Dorset, i.e. method 2 was not used, but at the same time it can be shown that estimation method 1 described above does not produce the normals defined for Nottingham Island.

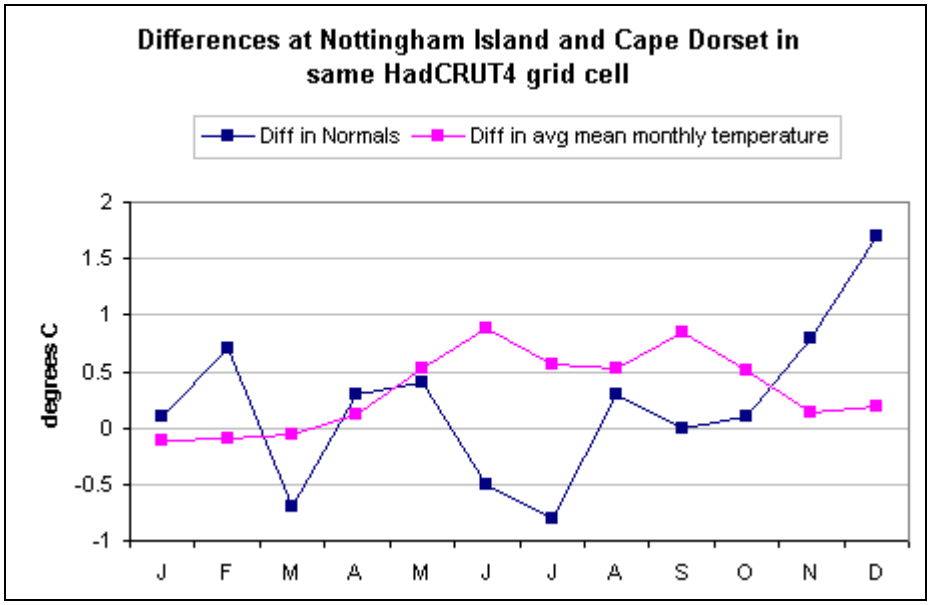


Figure 4.5 Difference in the long-term average temperatures ("Normals") and the average difference in mean monthly temperatures for those calendar months at two locations within the same grid cell.

The CRUTEM4 station metadata indicates the source of the long-term average temperatures as "Data", "WMO" or "Extrapolated" (or "Nil" for rejected stations). With the "Extrapolated" group having no distinction between long-term averages derived from 1951-70 data and those estimated from a neighbouring station the number of stations for which each method was used cannot be determined.

According to the data available in January 2016, of the total number of 7263 stations, 6182 stations have acceptable averages calculated from their own data, 59 have averages provided by the WMO of which 41 were rejected because they failed the minimum data permitted for calculation of the standard deviation, 127 have averages extrapolated from 1951-70 data of which seven failed the same criteria for standard deviations, and 866 stations have no long term average and were therefore rejected.

This mixture of approaches to deriving the long-term averages for a station for a given calendar month has several inherent problems when it comes to data accuracy.

1. Calculating the long-term average temperature over different years will likely produce different results, which would lead to different temperature anomalies being produced. If the monthly averages in the missing years were warmer (cooler) then their inclusion would have meant a higher (lower) long-term average temperature and the calculated anomalies would be lower (higher), albeit the temperature trend at this location for this calendar would be unchanged, the data simply uniformly increasing or decreasing in line with the difference in the average.

2. The assumption that averages extrapolated from earlier years and then adjusted according to grid cell values in earlier versions of CRUTEM (presumably mainly CRUTEM3) is contentious because the observation station temperature data in those earlier versions frequently differed from the data available from the national meteorological services. As mentioned above, for some Siberian stations differences in excess of 5°C were not unusual and in some months the difference was greater than 10°C.

3. Grid cells are likely to contain reporting stations that vary in the amount of data used to determine their long-term averages and therefore the stations have different temperature anomalies that might have been the same if the full sets of data for calculating the long-term averages were available.

4. For some grid cells there are multiple stations that have the entire 30 values or very close to it for every calendar month during the period from 1961 to 1990 and the inclusion of further stations with extrapolated long-term average temperatures decreases the consistency of the grid cell data and arguably adds nothing to the accuracy of the cell value. When the data is sorted into station ID order, the first two observation stations with "extrapolated" averages are respectively in grid cells with 23 stations with an acceptable number of months of data between 1961 and 1990, and in a cell with 54 other stations, 2 with no averages, 4 others with extrapolated averages and 48 having averages from their own 1961-90 data (and 45 of those having the full set of data for each calendar month).

Morice at al (2012) suggests that there is only a small error margin when mixing anomalies derived from varying periods of long-term averages or from estimated long-term averages but the number of inconsistencies in data supply, the uncertainties of estimating 'normals' and the number of assumptions suggest a greater error margin.

4.3.3 Average temperatures and standard deviations 1961-1990

The CRUTEM4 station metadata provides both the long-term average temperatures and standard deviations but the temperatures are notionally for the period from 1961 to 1990 but the standard deviations for the period from 1941 to 1990.

Station-month combinations (i.e.. station ABC/January, station ABC/February etc.) were constructed and the data analysed on this basis. The number of station-month combinations for each long-term average temperature value to one decimal digit - the format in which it was supplied - was counted and graphed (Figure 4.6). Six long-term averages of below -60.0°C are omitted, each mapping to a single value, with the lowest being -67.9°C

One feature of Figure 4.6 is that the distribution of long-term average temperatures is either a skewed Gaussian distribution or an approximation of a reversed Poisson distribution, both distorted by a secondary peak that make neither particularly appropriate.

Another feature is the spike in station counts with long-term average temperature of 0.0°C. The 242 instances of this average compares with the mean of 129.4 (max 145) in each 0.1C interval across the one degree below, i.e. from -0.1°C to -1.0°C, and the mean of 138.1 (max 158) across the one degree above zero. The number of stations with a given long-term mean temperature did not exceed 242 until the 249 at a temperature of 10.8°C.

Neither point has a crucial bearing on accuracy but might be important for the understanding of temperature variations, the second point in particular suggesting that the latent heat transition point for water might need to be considered.

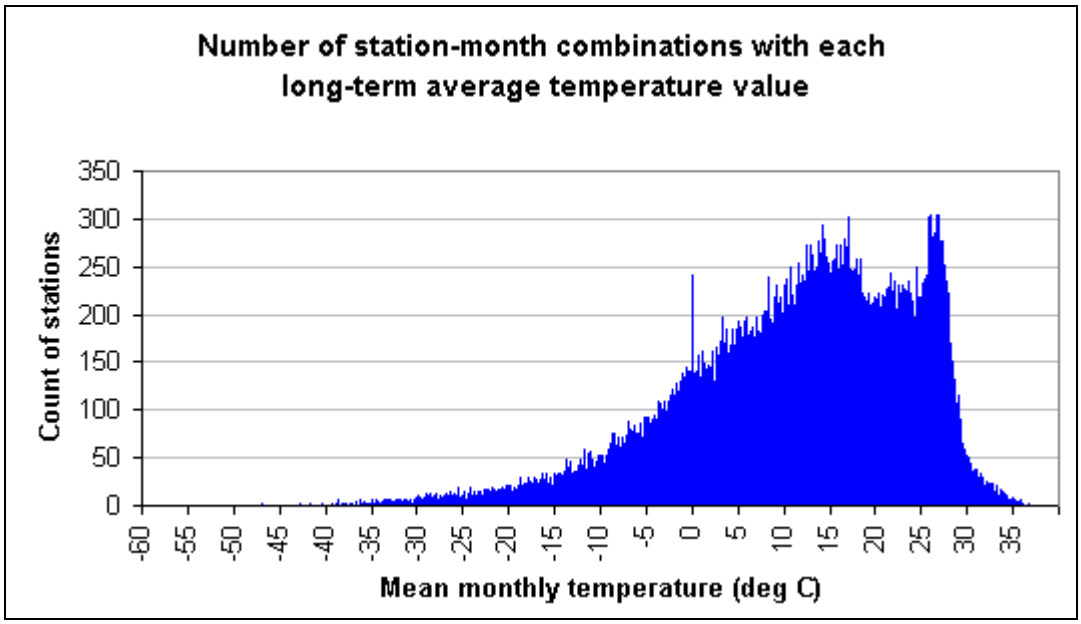


Figure 4.6 Number of station-month combinations for each long-term average temperature, according to CRUTEM4 metadata.

4.3.4 Checking for Gaussian distribution

Both Jones et al (2012) and Osborn & Jones (2014) discuss standard deviation in regard to error margins but neither mentions normal distribution and in particular whether all station data was tested for its approximation to a normal distribution.

To meet the criteria for a standard distribution curve means that the spans of mean ± 1σ (i.e. within 1 standard deviation either side of the mean value), mean ± 2σ and mean ± 3σ should contain 68%, 95% and 99.7% of values respectively.

As mentioned above in section 4.1, the Shapiro-Wilk test is often used to test the goodness-of-fit to normal distribution by considering how well the data distribution

fits the standard curve and the symmetry of the data and returning a p-value between 0 and 1.

Before the Shapiro-Wilk test was applied the 52,444 station-month combinations with 30 entries (i.e. data for the applicable calendar month in every year from 1961 to 1990) were tested for autocorrelation using the Durbin-Watson test. This test produces a score from 0 to 4 with 2 indicating no correlation and scores below 1.0 indicating significant positive correlation. The average score of the station-month combinations was 1.86, which shows only a negligible very slight inclination towards positive correlation. The number of combinations with each D-W score, rounded to one decimal place is shown in Figure 4.7. Just 608 combinations, just 1.16% of the total, show scores below 1.0, so autocorrelation is not a significant issue.

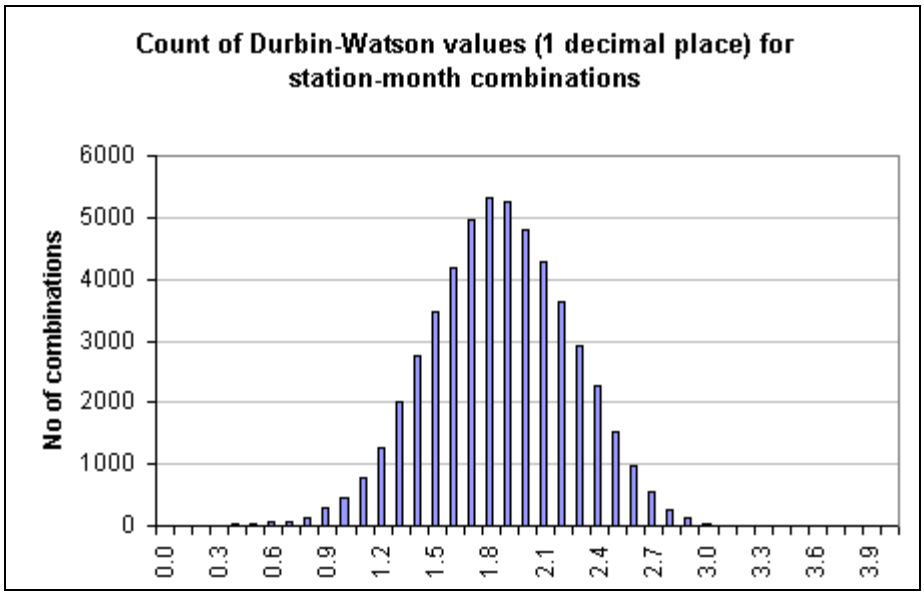


Figure 4.7 Counts of station-month combinations for which the Durbin-Watson test returned various scores, rounded to one decimal place.

The next step was to apply Shapiro-Wilk test to the same 52,444 station-month combinations with 30 entries. The frequency counts for each of the 2 digit p-values is shown in Figure 4.8, with values of 0 assigned for p-values less than 0.01 and 1 for p-values greater than 0.99.

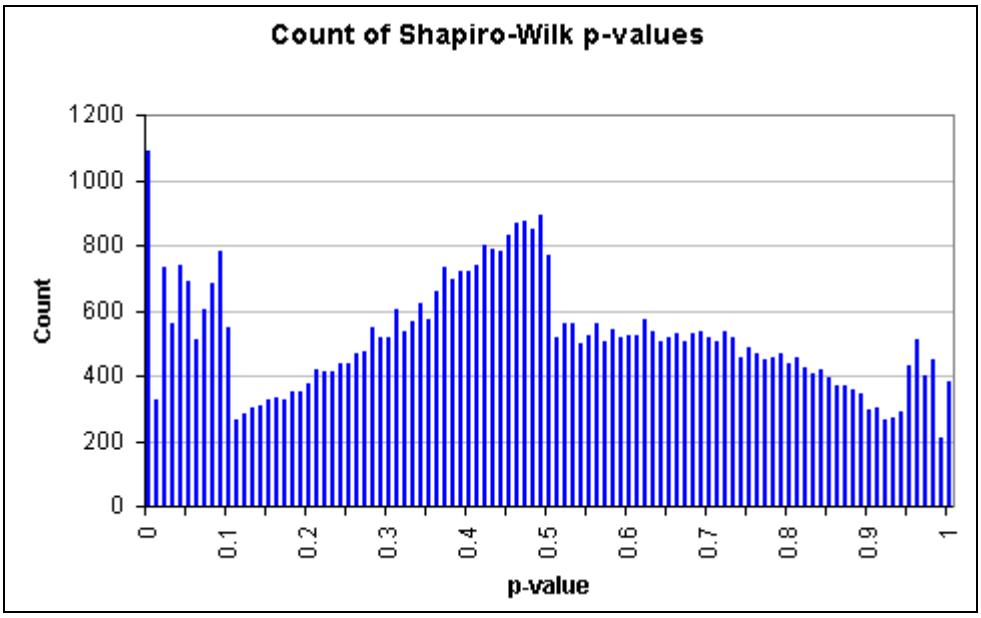
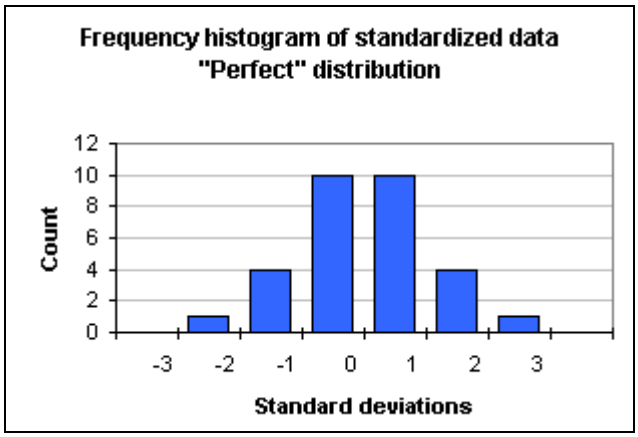
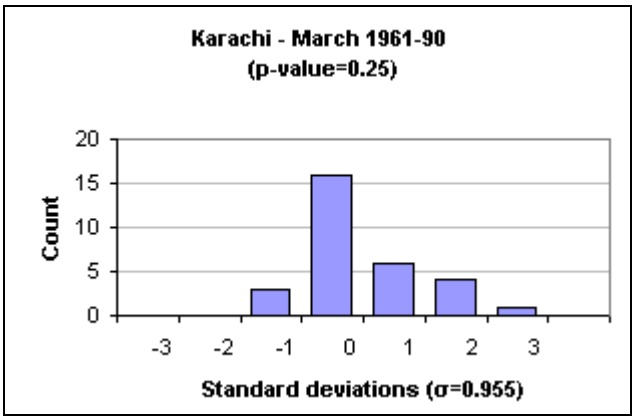
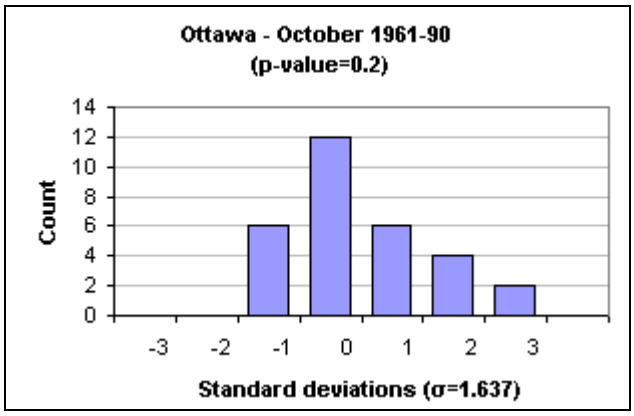


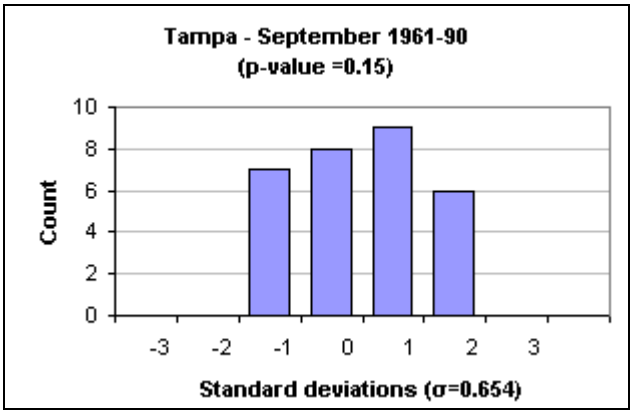
Figure 4.8 The number of station-month datasets, from a total pool of 52,444 with 30 entries, whose data distributions according to the Shapiro-Wilk method had the p-values shown on the X-axis.

Wasserstein & Lazar (2016), stating the position of the American Statistical Society on the use of p-values, says "... a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis" and that p-values do not measure the probability that the studied hypothesis (in this case the fit to the normal distribution curve) is true. This would be true if the 30 entries was a subset of a larger sample and the aim was to determine whether the larger sample was normally distributed but this is not the case here. The distribution of the 30 entries will determine whether conclusions based on Normal distribution can be made for each subset. Figure 4.9 shows firstly the "perfect" distribution, the pattern that complies with the "68-95-99.7 percent" distribution and whose p-value would be >0.99, followed by examples of station-month combinations with p-values of 0.25, 0.2, 0.15 and 0.1, meaning progressively less conformance to Normal distribution.


Frequency histogram of standardized data
"Perfect" distribution
Count
12
10
8
6
4
2
0
-3
-2
-1
0
1
2
3
Standard deviations


Karachi - March 1961-90
(p-value=0.25)
Count
20
15
10
5
0
-3
-2
-1
0
1
2
3
Standard deviations (σ=0.955)


Ottawa - October 1961-90
(p-value=0.2)
Count
14
12
10
8
6
4
2
0
-3
-2
-1
0
1
2
3
Standard deviations (σ=1.637)


Tampa - September 1961-90
(p-value =0.15)
Count
10
8
6
4
2
0
-3
-2
-1
0
1
2
3
Standard deviations (σ=0.654)

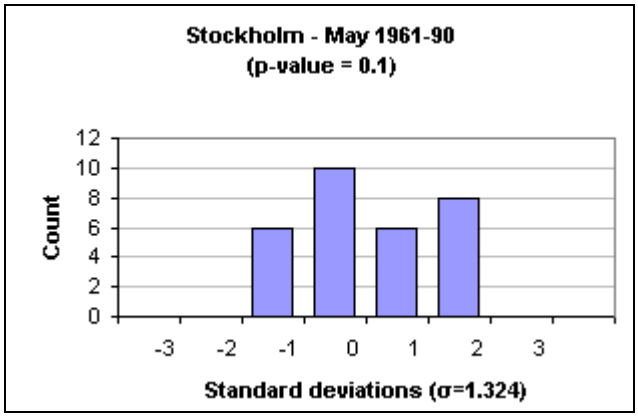


Figure 4.9 The "perfect" normal distribution followed by examples of standardized frequency distributions station-month combinations with 30 entries during the period from 1961 to 1990, with the locations, calendar months and decreasing p-values noted. Lower p-values indicate less conformance to normal distribution.

According to data from which Figure 4.8 was created, of the total 52,444 station-month combinations those with low Shapiro-Wilk p-values are as follows:

7.9% that have p-values ≤0.05
13.9% that have p-values ≤0.10
16.7% that have p-values ≤0.15
and 20.0% that have p-values of ≤0.20.

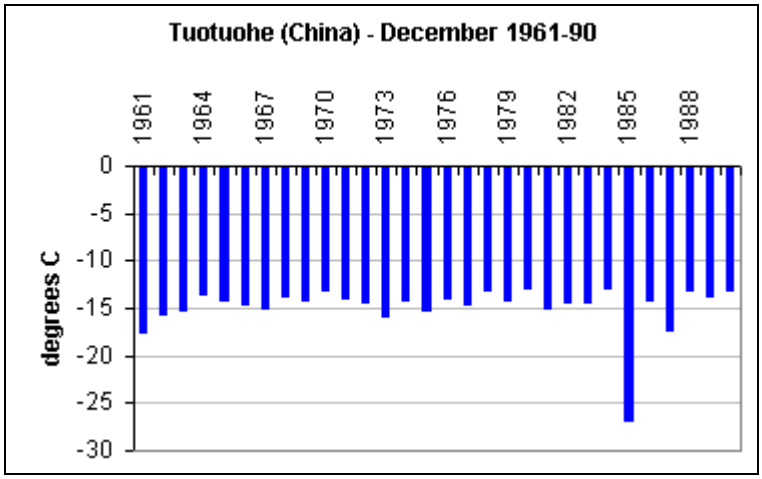
According to the above list, a p-value cut-off of 0.1 would exclude 13.9% of the station month combinations. This is equivalent to a ratio of 1.67:12 so, given that each station has 12 station-month combinations associated with it, an average of 1.66 station months for any given observation station do not meet the criteria for normal distribution.

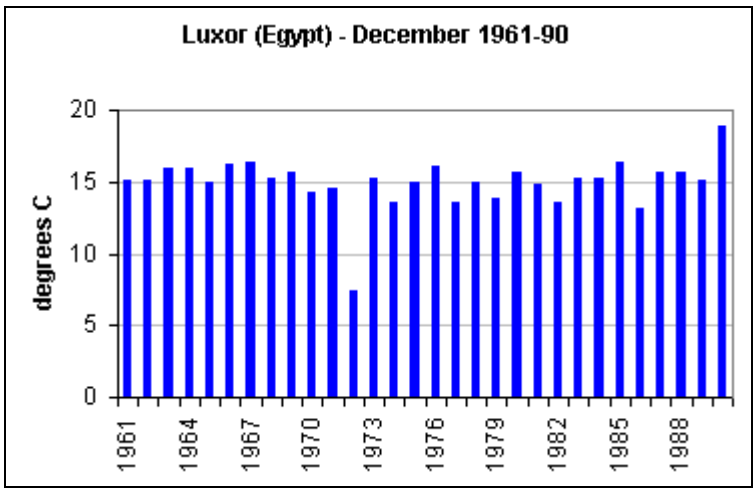
Regarded *en masse* the low p-values of 13.9% of the population is not unusual but when station-months are dealt with individually the situation is different because standard deviations associated with assumed normal distribution are used to identify outlying data values and in the calculation of error margins. If data is not normally distributed both can be distorted.

4.3.5 Inclusion of outliers when calculating long-term average temperatures and standard deviations

Data outliers can be a problem for any data processing. According to Jones et al (2012), outlying mean monthly temperatures of more than five standard deviations from the long-term average temperature for the corresponding month are excluded from CRUTEM4 processing. Five standard deviations is a very relaxed threshold for determining outliers but this and the broad issue of outliers in regard to the entire data record will be discussed in Chapter 5. The concern at this point is the inclusion of outliers in the calculation of the long-term average temperatures or of the standard deviations. Outliers present in this subset of the data will widen error margins in long-term averages, distort temperature anomalies and, for standard deviations, potentially lead to the inclusion of further outlying data in the data record at other times. Figure 4.10 shows some examples of likely outlying data for certain station-months over the 1961-1990 period used to calculate long-term average temperatures.

Figure 4.10 includes just one example of several instances of questionable data from Chinese station Tuotuohe (ID: 560040). Mean monthly temperatures of -28.1°C, -26.9°C, -28.2°C are reported for November 1985 to January 1986 respectively, which unlike the corresponding data from the three other stations in the same grid cell (Naqu, Dangxiong and Suoxian, with IDs 552990, 554930 and 561060) is a marked shift from the long-term average temperatures those months. (The temperatures in October 1985 and February 1986 were also several degrees, and therefore several standard deviations, from the respective means but less so than these three months.)





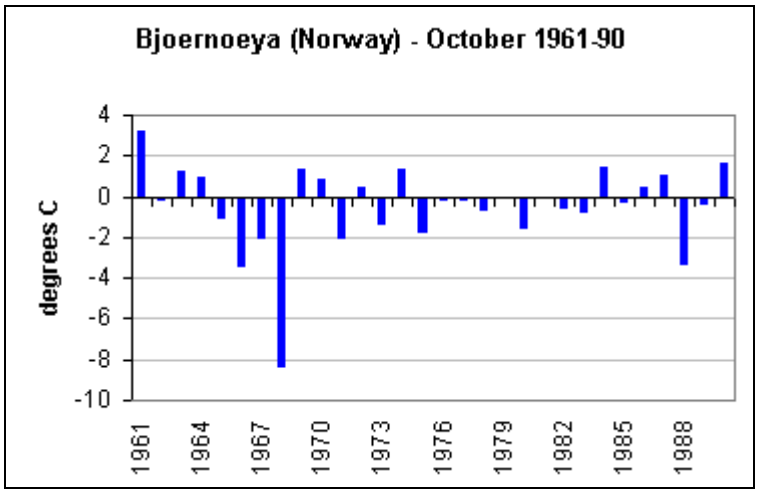


Figure 4.10 Examples of outliers in the 1961-90 data from which long-term averages are calculated with outliers in 1985, 1972 and 1968 top to bottom respectively.

The long-term average temperatures at Tuotuohe over 1961-1990 when the mean monthly temperatures shown above are included in the calculations are -11.3°C, -14.8°C and -15.8°C, with standard deviations calculated over 1956-1990 (almost the same as for temperatures) are 3.5°C, 2.5°C and 2.6°C. This means that the temperatures were at -4.7σ, -6.8σ and -4.8σ from the long-term average temperatures for the relevant calendar months.

When the probably erroneous mean monthly temperature for the three months are excluded from the calculations the long-term average temperatures for the calendar months become -10.7°C, -14.4°C and -15.4°C and the 1956-1990 standard deviations reduced to 2.0°C, 1.2°C and 1.5°C. In particular the "five standard deviation" thresholds for these months are greatly reduced, from 17.5°C to 10.0°C for November, from 12.5°C to 6.0°C for December and from 13.0°C to 7.5°C for January, meaning that more outliers in these months in other years of the record would be rejected.

Another station where obvious outliers are included when calculating long-term average temperatures and standard deviations is Colombian observation station Apto Otu, where the mean monthly temperatures April, June and July 1978 have the implausible values of 81.5°C, 83.4°C and 83.4°C. This data is from the period over which long-term average temperatures and standard deviations are calculated for this location, the first from 1961 to 1990 and the second from 1947 to 1988. As a result of these very high mean temperatures, the long-term average temperatures for the respective months are 27.8°C, 27.9°C and 28.0°C when for all other months of the year the temperatures are consistently between 24.0°C and 24.6°C. The standard deviations for the location are 11.9°C, 11.8°C and 12.0°C in these months, while in the rest of the year no standard deviation exceeds 0.7C. The impact on the CRUTEM4 grid cell values of these three months of high mean temperatures is shown in Figure 4.11.

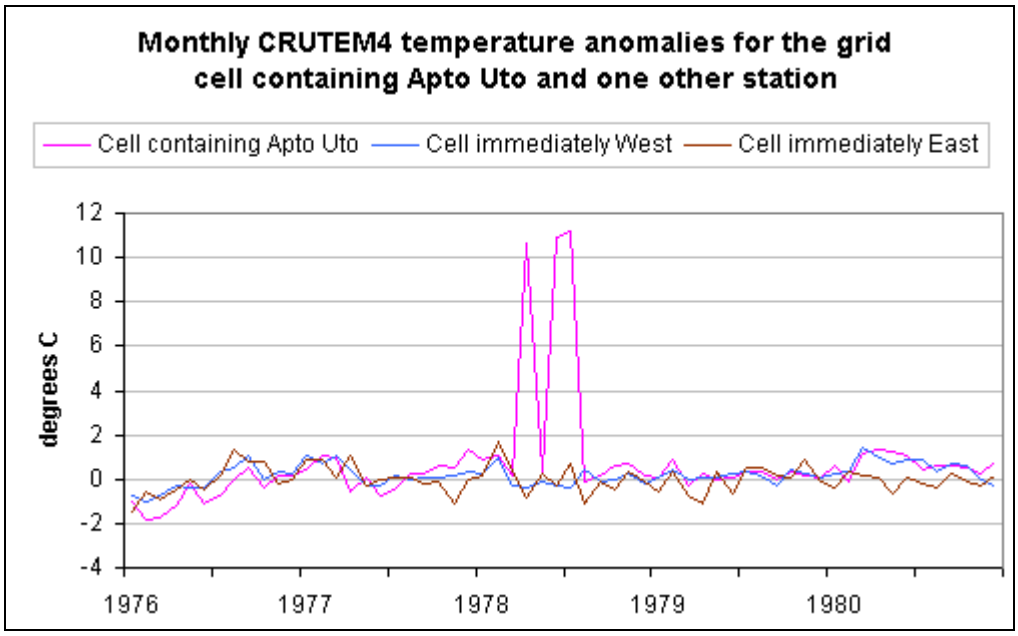


Figure 4.11 Data from the CRUTEM4 gridded dataset for the CRUTEM4 data cell that contains Colombian observation station Apto Uto and one other station, plus the data for the grid cells immediately west and east of it.

4.3.6 Implications of the generous five standard deviations threshold for outliers

The use of the very generous "five standard deviations" threshold for determining outliers means that the entire range of acceptable data for a given calendar month is ten times the standard deviation for that month, centred on the long-term average temperature. For the 75840 acceptable CRUTEM4 station-month combinations Table 4-1 shows the number of combinations with standard deviations above the given thresholds. According to this table, 5414 station month combinations have standard deviations 3.0°C, which means the range of acceptable monthly mean temperatures for that same calendar month will span more than 30.0°C, among these being 292 have a range that exceeds 50.0°C, i.e. have standard deviations of 5.0C or more.

At the top of the list in Table 4-1 are three station-month combinations for observation station Apto Otu discussed in the previous section. According to that station's metadata the "five standard deviations" threshold would amount to an

acceptable range of values (i.e. from -5σ to +5σ) of 119°C, 118°C and 120°C for April, June and July respectively.

Threshold	Combinations
1.0	53968
2.0	15742
3.0	5414
4.0	1813
5.0	292
6.0	35
7.0	10
8.0	4
9.0	3
10.0	3
11.0	3
12.0	0

Table 4-1 The number of station month combinations whose standard deviations exceed the given thresholds.

The "five standard deviations" threshold is extremely generous because the probability associated with having values at each end of the scale is less than one in one million. The data can be trimmed to a 3 standard deviations limit by calculating of long-term average temperature and standard deviation, removing outlying values beyond 3 standard deviations and then repeating the calculation until no values are removed. When this approach was applied to the 71,500 station-months that have from 20 to 30 years of data across 1961-1990, data was removed from 4400 station-months, 6.15% of the total number.

Even an outlier threshold of 3 standard deviations can be considered excessive because it includes 99.7% of data and therefore only excludes an average of 3 in 1000. Climate is rarely stable for long and can vary considerably in 1000 years. A somewhat arbitrary but arguably more reasonable alternative is to trim to 2.81 standard deviations, corresponding to a probability of 99.5% (i.e. 1 in 200 outside the

range) because at most we are dealing with just 30 values. This limit was applied to the 1961-1990 data in order to ascertain its effect.

Table 4-2 shows the consequence of trimming data to a limit of 2.81 standard deviations for the three examples shown in Figure 4.10, as described above, this trimming being a cyclic operation until no values fall beyond that limit. In all three instances the long-term averages increased slightly because the outliers were the lowest values. More importantly the standard deviations decreased, meaning a reduction in the error margins and in the limits for identifying data outliers.

Station		**Tuotuohe**	**Luxor**	**Bjoernoeya**
Country		China	Egypt	Norway
Station ID		560040	624050	10280
Month		December	December	October
Outlier		-4.8σ	-4.23σ	-3.85σ
Before adjustment	*Mean*	-14.80C	14.93C	-0.45C
	Std Deviation	2.52C	1.78C	2.04C
	Std. Error	0.47C	0.33C	0.38C
After Adjustment	*Mean*	-14.38C	15.06C	-0.18C
	Std. Deviation	1.16C	0.89C	1.46C
	Std. Error	0.22C	0.17C	0.27C

Table 4-2 Summary of result of removing outlying entries from the data shown in Fig 4.9 until all values fell within 2.81σ (i.e. standard normal probability 99.5%).

The limit of 2.18 standard deviations described above was also applied to all 74465 station-month combinations where the long-term average temperatures were calculated from at least 14 years of data in the period 1961-90. The cycle of identifying and removing outliers until all data fell within ±2.81 standard deviations from the long-term average temperature was repeated as necessary and caused modification to 8436 (11.8%) combinations.

Table 4-3 summarises the consequences of removing data to meet the 2.18 standard deviations limit, with the reductions in sample size sometimes meaning the station-month combination fell back to a lower sample size grouping, or in the case of the

lowest group, for "19 to 14 years", sometimes meaning that data for fewer than 14 years remained and the combination no longer met the criteria for inclusion in CRUTEM4 data processing.

The average standard deviation of the 8436 modified station-month combinations decreased from 1.59°C to 1.28°C as a result of trimming to the new limit. This meant a reduced average threshold for determining outliers in data outside the period and, with the small reduction in the number of entries a reduction in the average standard error from 0.3°C to 0.25°C.

This experiment of restricting the data to a smaller range of standard deviations when calculating the long-term average excludes some outlying data but reduces error margins and the thresholds for identifying data outliers at other times in the data record.

		Grouping by years of data (i.e. no. of entries)				Totals
		30 years	**25 to 29 years**	**20 to 24 years**	**14 to 19 years**	
	Initial					
(a)	Count	52444	11798	7258	2965	74465
(b)	Percent of total	70.4	15.8	9.7	4.0	
(c)	Group Av Standard Error (95% confid.)	0.61	0.53	0.52	0.59	
	Modified entries from group					
(d)	Count	6258	1392	615	171	8436
(e)	Percent of group total count	11.9	11.8	8.5	5.8	
(f)	Av Std error prior (95% confid.)	0.62	0.51	0.50	0.67	
(g)	Av Std error after (95% confid).*	n/a	0.42	0.37	0.38	
(h)	Change in Max std error (95% confid.)	1.00	0.69	0.69	6.53	
	After modification (for total group)					
(i)	Revised count	46186	17854	7311	3111	
(j)	Group Av Standard Error (95% confid.)	0.61	0.52	0.51	0.57	

Table 4-3 Summary of changes caused by removing data more than 2.81 standard deviations from the mean. Group average standard errors change little (rows c and j) but the reduction in the standard error of the modified group is as high as 43% (in the "14 to 19 years" group).

4.3.7 Summary for observation station long-term averages

The long-term calendar month average temperatures (i.e. normals) calculated for CRUTEM4 stations suffer from a variety of issues. Firstly we have seen that the CRU does not follow WMO recommendations and that any station's long-term averages can be calculated from as few as 14 years across the period from 1961 to 1990. The acceptance criteria are based on the number of years of data in total rather than a minimum for each calendar month. This is unsatisfactory given the variability of temperatures during the thirty years, especially in the increase in El Nino events in the last 14 years of the period.

Secondly, for some stations, data outliers exist in months of the period over which the long-term averages or standard deviations are calculated. In the case of long-term average temperatures this distortion will mean distorted temperature anomalies for every instance of the given month in the entire data record. For standard deviations increased by the inclusion of outliers this will mean expanded error margins and widened thresholds for identifying outlying data that should be corrected or excluded. Trimming the data over which the long-term averages and standard deviations are calculated has been shown to reduce error margins and reject more data outliers.

The impacts of these issues are relatively minor for individual stations, except for the inclusion of data outliers. The greater problem is that temperature anomalies for individual stations are often averaged or, in the case of HadCRUT4 coastal grid cells, merged with other data and there is a very real possibility that the processing will involve anomalies calculated with long-term average temperatures derived from data over different periods and with different standard deviations to determine outliers. Further, stations commencing or ceasing operations could result in false shifts in CRUTEM4 grid cell values.

Text Box 4A - Observations about temperatures and standard deviations

The CRUTEM4 station metadata provides both the long-term average temperatures and standard deviations but the temperatures are notionally for the period from 1961 to 1990 and the standard deviations for the period from 1941 to 1990. Station-month combinations were created and for all instances of 30 values across the period from 1961 to 1990 (i.e. no missing data) the standard deviations were calculated for that period so that the two periods matched and the relationships could be examined.

Figure 4A.1 shows the distribution of long-term average temperatures across the period. Six values are omitted from below -60.0C, the lowest of which is -67.9C. The relationships between the long-term average temperature and the standard deviation for 50% of the station-month combinations, constrained by the volume of data are shown in Figure4A.2. (The omitted 50% showed very small variation from those displayed here except at the extremes where the number of stations is low.)

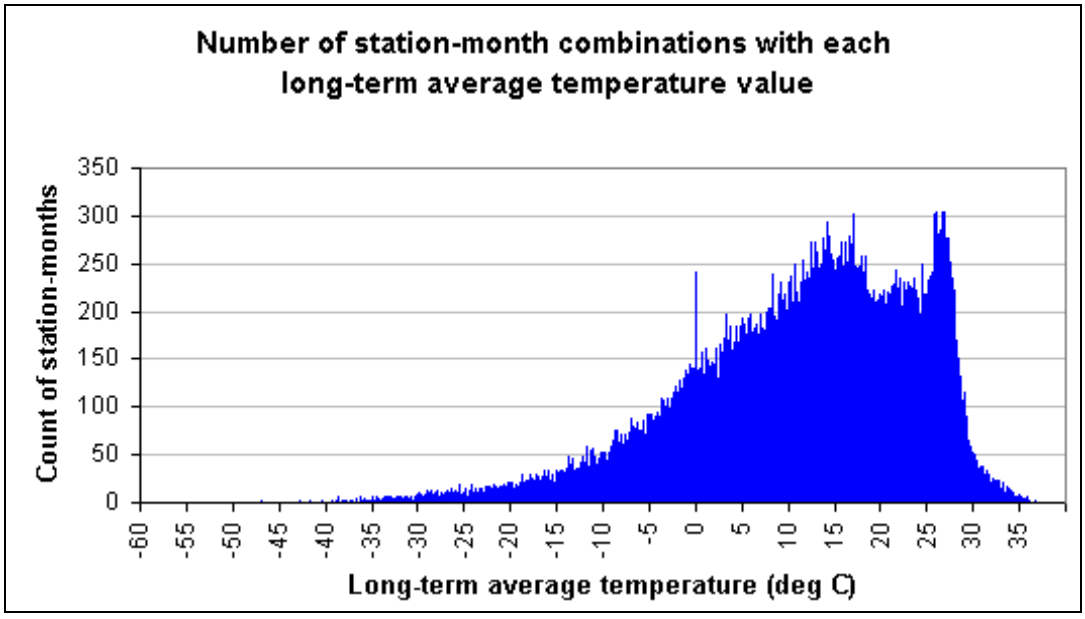


Figure 4A.1 - Number of station-month combinations with the given long-term average temperature.

...

Text Box 4A contd

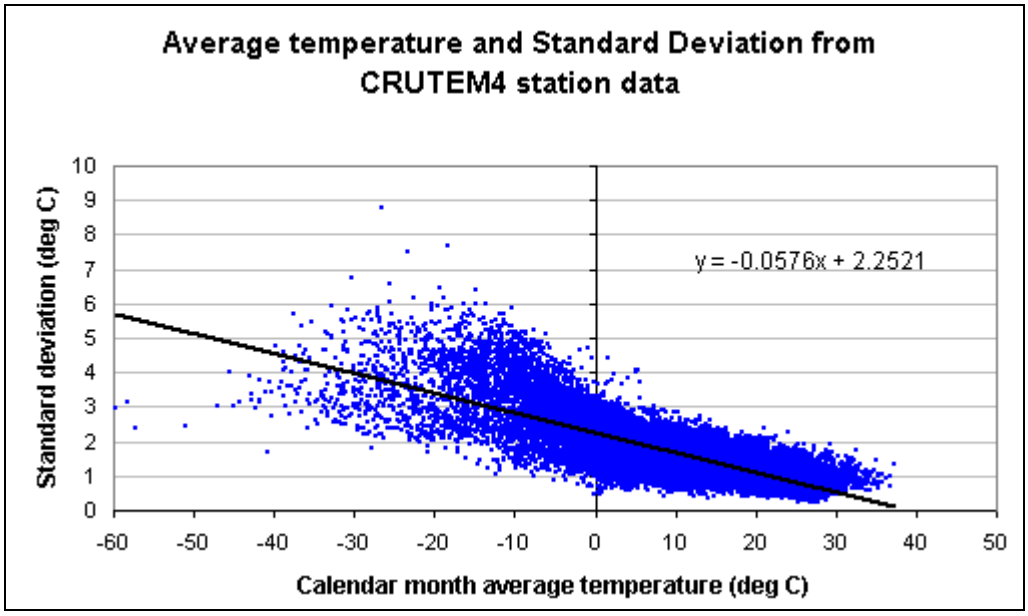


Figure 4A.2 Long-term average temperatures and the corresponding standard deviations for station-month combinations that have data for that month in all 30 years of the period from 1961 to 1990.

Figure 4A.3 shows the mean standard deviation associated with each long-term average temperature. This indicates that greater variation in temperature occurs in colder regions, which by extension means that an increase in temperatures in those regions is of less significance than a similar increase in warmer regions.

...

Text Box 4A contd

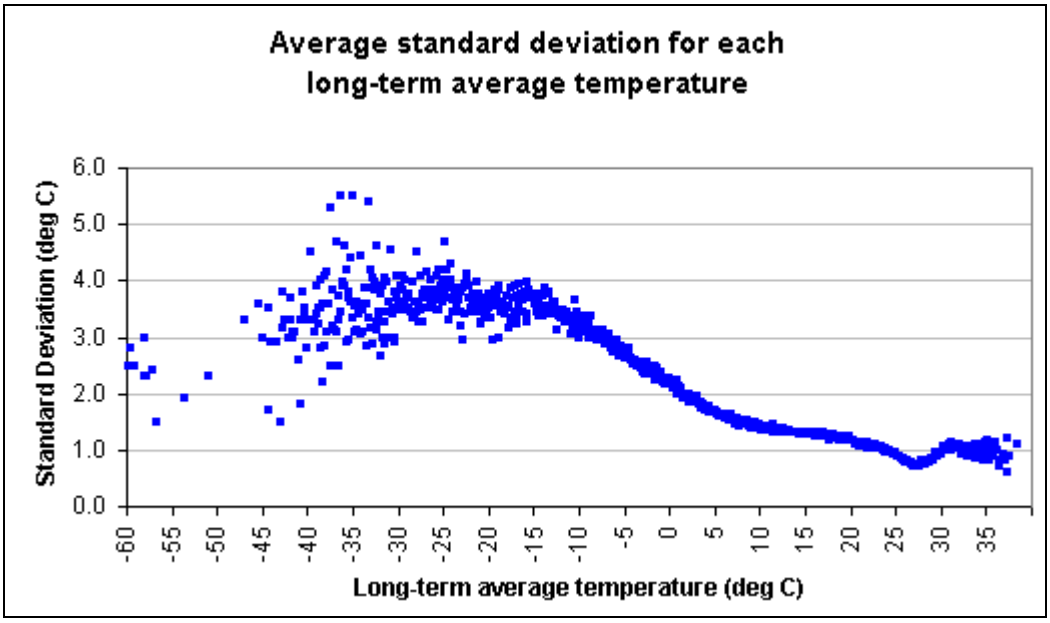


Figure 4A.3 Mean standard deviations for each long-term average temperature.

4.4 Issues with long-term average sea surface temperatures

4.4.1 Introduction

In HadSST3 terminology, long-term average temperatures are known as "climatology" and they are derived quite differently to the "normals" used for CRUTEM4 data from observation stations. This section discusses issues with the derivation of the "climatologies", the standard deviations of HadSST3 data, the presence of outlying data and whether the HadSST3 data has normal distribution.

As before, errors or abnormalities in individual grid cells will not matter greatly. Even though temperature anomalies will be shifted from proper values the shift will

be consistent and the trend remain the same, but combining anomalies from different grid cells might be a very different matter, especially if data for specific grid cells is sometimes unavailable.

4.4.2 The creation of long-term SST averages

Obtaining information about the creation of HadSST3 climatologies is an arduous process. The primary reference for information on the creation of the HadSST3 dataset, Kennedy et al (2011a), refers only to HadSST3 using the same approach as used for the HadSST2 dataset as described by Rayner et al (2006), which provides limited information and refers the reader to Parker et al (1995).

Ultimately we find that HadSST3 data, using grid cells of 5° latitude x 5° longitude grid cells and calendar months, is produced from a method using 1° x 1° grid cells and 5-day periods called "pentads", with 6 pentads making up a pseudomonth (of 30 days), except for August where 7 pentads make up the month, so that a 365-day calendar year corresponds 73 complete pentads.

According to Rayner et al (2006) the establishment of the climatologies is a multi-step process, the first of which is to create the initial SST background dataset that will subsequently have SST observations merged into it.

The initial HadSST2 background dataset consists of data from the Global Ice and Sea Surface Temperature dataset (GISST 2.0), also gridded at 1° x 1°, with quality-controlled SST anomalies added to it, plus data from Polar regions that are sometimes partially or completely covered in sea ice.

The second step is to interpolate the grid cell values using the Laplacian (a partial second order differential equation) to complete the global grid and the third step is to convert the results into calendar months. Rayner et al (2006) says that the new 1961-1990 averages generated by these steps were then used as background fields in iterations of the second and third steps, with six iterations "needed to ensure convergence to a stable result".

This description by Rayner et al (2006) is somewhat unclear. No reason is given for using the Laplacian technique in preference to any other. It also fails to state, in light of the whole task being to create a climatology from which anomalies will be determined, the basis by which the SST anomalies in step one were calculated or how their quality was checked given that quality control involves comparing calculated anomalies. It also fails to describe what is meant by "convergence" and by "a stable result".

In total the climatology appears to be derived from mathematical estimates combined with SST measurements - measured by ships, buoys and by satellite - which are averaged, interpolated and winsorized[16], each step requiring assumptions, the whole process being iterative until some state of stability is produced, which presumably means no large differences between adjacent grid cells.

Given the lack of detail provided in Rayner et al (2006), the apparent complexity of the work and the volume of SST observation records, detailed analysis of the process of determining the "climatology" is too large a task to be attempted here. The focus will instead be on only certain aspects of the HadSST3 data across the period 1961 to 1990, limited somewhat by the available data.

4.4.3 No of years of data per HadSST3 grid cell

The number of years of data between 1961 and 1990 for each HadSST3 grid cell indicates the number of instances in which data for a given calendar month was used to derive the long-term average SST for the grid cell and month.

This can be done using the concept of grid cell-month combinations, with each grid cell paired with each calendar month, similar to the pairing of station and months used earlier for CRUTEM4 data. A total of 2592 grid cells of 5° latitude x 5° longitude

[16] Winsorizing involves setting a certain number or range at the two extremes to certain values (e.g. the 90th percentile is copied to the 91st to 100th percentiles) before averaging, the aim being to reduce the influence of extreme values on the mean.

cover the Earth's surface, meaning 31104 cell-months. Some are entirely over land and will therefore never contain HadSST3 data but all other grid cells either do provide data or could provide data.

For the period 1961 to 1990 no data is available for 10229 cell-month combinations, 4381 combinations have from 1 to 13 values (i.e. data was reported in 1 to 13 years for that calendar month) and 16494 cell-months have data 14 or more years.

A check of the 10229 combinations without data in that period revealed that 2249 combinations had data at other times in the 1850-2015 data record. With no data during 1961-1990 the climatologies for these combinations must be based on estimates and interpolations rather than SST measurements within the area covered by each grid cell. The 2249 cell-month combinations apply to 396 grid cells of which 229 (57.8%) were north of 65°N, just 15 (3.8%) between 65°N and 60°S, and 152 (38.4%) south of 60°S. This distribution is heavily biased towards the Polar regions, which with the weighting system, by the cosine of the latitude of the centre of each grid cell, have very little impact on hemispheric and global average temperature anomalies.

Of the 4381 cell-month combinations with from 1 to 13 values from 1961 to 1990, a total of 1006 (~23%) had a single value (instance of data for that calendar month) across the entire period. The grid cells that these combinations relate to account for an annual average of about 4.5% of the total HadSST3 grid cells containing data but in July 2010 peaked at 15.2%.

The climatologies of these 4381 cell-month combinations seem to be derived from very few SST measurements within the grid cell in the period 1961-1990. Information from the Hadley Centre fails to clarify whether priority was given to these few measurements, which might have been abnormal, over the interpolation of data from neighbouring grid cells.

4.4.4 Number of observations

As noted earlier, the number of observations for each 5° x 5° HadSST3 grid cell in each month are listed in a file available for public download. This data provides further details on the number of SST measurements that revised the estimates of long-term average sea surface temperatures.

Because the monthly data in each 5° x 5° HadSST3 grid cell is derived from data in 1° x 1° grid cells using 5-day pentads it would require a minimum of 150 SST observations, i.e. one for each 1° x 1° cell and pentad, per HadSST3 grid cell to avoid the interpolation of data when creating the underlying set of fine resolution long-term averages.

Figure 4.12 shows the number of 5° x 5° HadSST3 grid cells that have the various specified ranges of observation counts in each month of the period 1961-1990. The group of grid cells with more than 150 observations per month, the minimum number discussed above, increases slowly from 1961 and peaks as the largest of the groups in some months of just a few years after 1985. For most of the 1961-1990 period the largest group is that for 2 to 20 observations, albeit varying by month/season. Figure 4.13 shows the same data expressed as a percentage of all grid cells that have data.

Figure 4.13 shows that for almost all months over the 30-year period more than 70% of grid cells had fewer than 150 observations for the month and about 50% had fewer than 50 observations. About 33% of the total number of grid cells had from 1 to 20 observations meaning that a considerable number of grid cells appear to have had few SST observations to modify the estimated long-term averages.

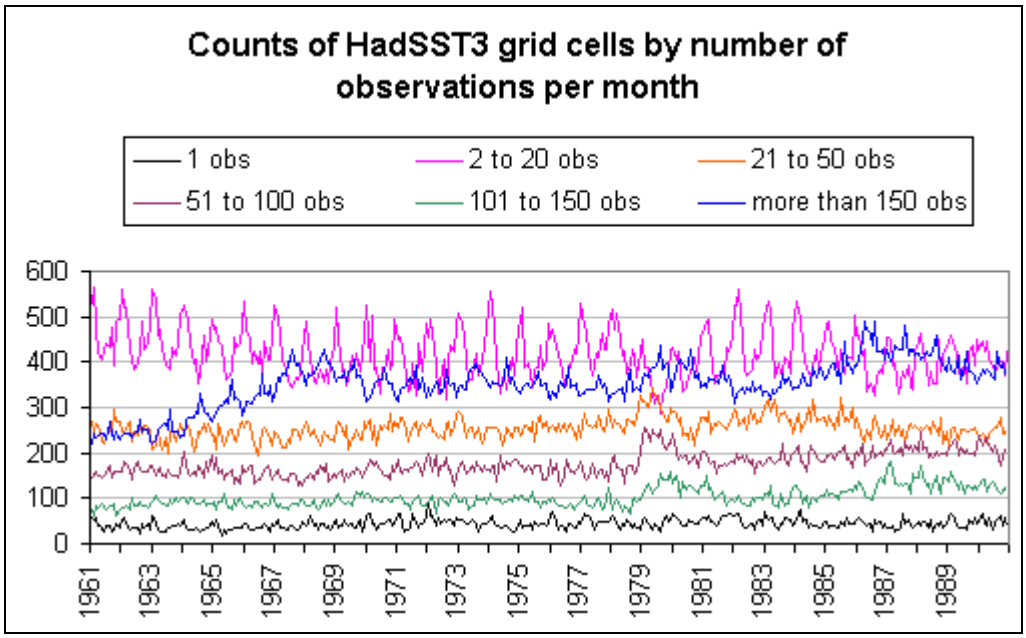


Figure 4.12 Number of grid cells with the specified ranges of observation counts.

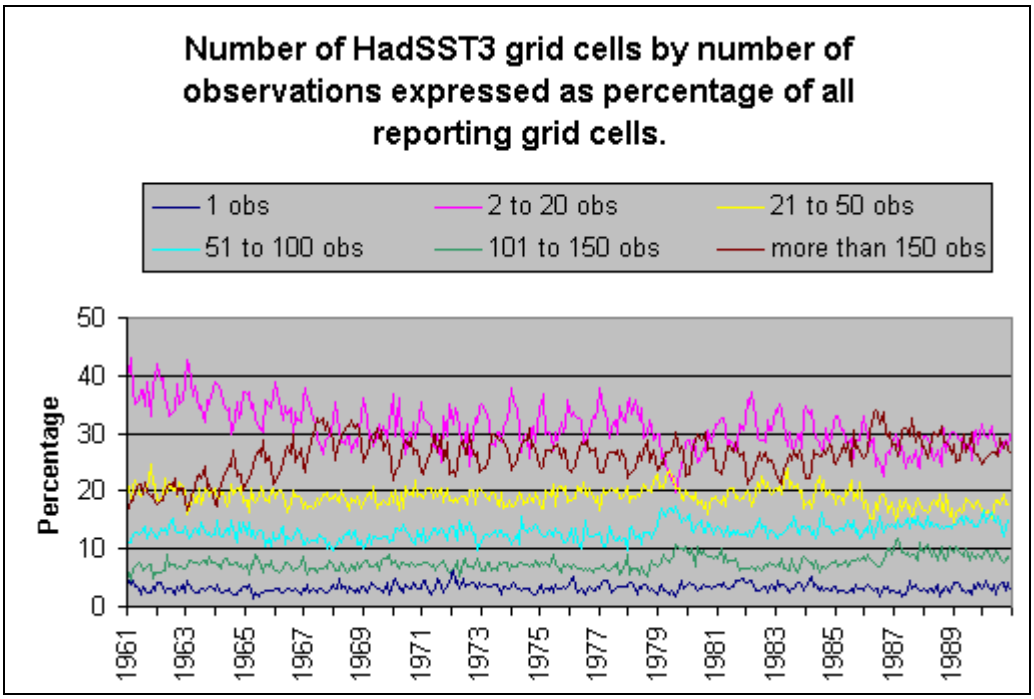


Figure 4.13 As for the previous figure but expressed as a percentage of all grid cells that contain valid data in each month.

4.4.5 Sum of anomalies over the period 1961-90

The average of the values across the period from1961 to 1990 is used as the base value from which anomalies are calculated so if the HadSST3 data is accurate then the global average SST anomalies in each calendar month over that period should sum to zero. Figure 4.14 shows this to not be the case in any calendar month or as an annual average. These figures cannot be attributed to rounding because the data is from HadSST3 summary files in which data is given to three decimal places, which means precision to one-thousandth of a degree despite only 30 values, one for each year, being used to calculate each of the graphed values.

The common failure to sum to zero is likely to be due to the technique of using 1° x 1° grid cells and pentads as the basis for long-term average temperatures and the calculation of anomalies, and then using interpolation and extrapolation to generate HadSST3 dataset monthly values at 5° x 5° grid cell size.

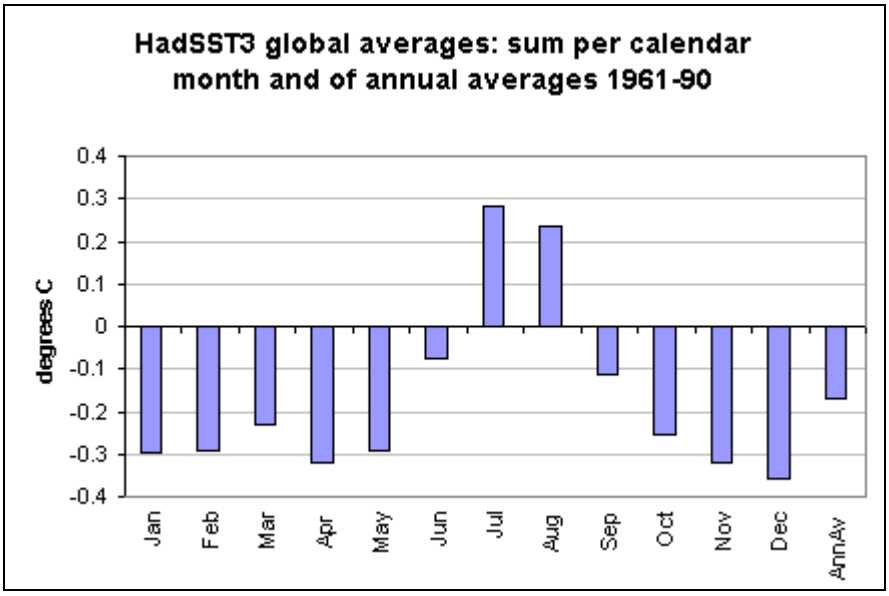


Figure 4.14 The sum of the HadSST3 global average SST anomalies for each calendar month, and on the right the annual averages, across the period 1961-1990.

Another way to consider this is to analyse cell-month combinations with 30 values for the period 1961-1990 (i.e. data for that month in every year). The values are summed then divided by 30 to obtain the average "discrepancy" for that month in each year,

which is to say the amount that should be consistently subtracted from data for that month in each of the years so that the sum of the values over all 30 years is in fact zero. Figure 4.15 shows the distribution of the average discrepancies, rounded to two significant figures, for the 11421 cell-month combinations with 30 entries over the period 1961-1990.

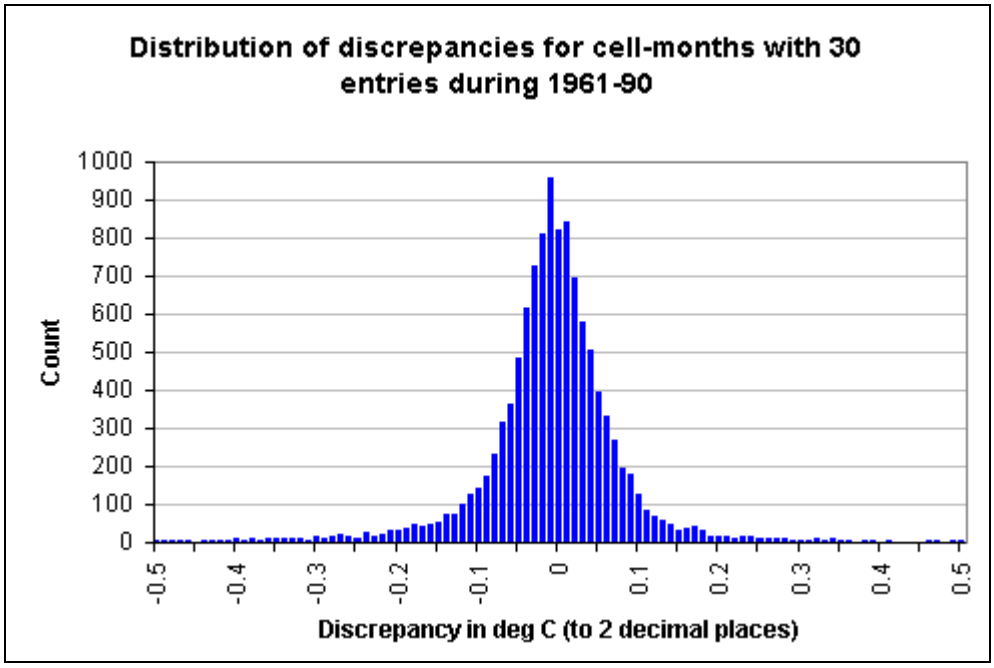


Figure 4.15 Distribution of discrepancies for each cell-month combination that has 30 entries (i.e. one in every year) for the calendar month over the period from 1961 to 1990.

In the worst case the grid cell centred at 32.5°N 67.5°W (off north east USA) has the lowest negative discrepancy of -1.84°C in each February of 1961-90 and two nearby others grid cells have discrepancies of -1.64°C and -1.55°C respectively in December. The grid cell centred at 67.5°N 27.5°W (between Iceland and Greenland) has the highest positive discrepancy of 1.13°C.

If we accept discrepancies of ≤ 0.1°C (i.e. allow a range of ±0.1°C) as being sufficiently accurate in the light of data rounding and the "ensemble" methodology, we are left with the 1661 (14.5% of the total 11421) cell-month combinations that fall outside this range of ±0.1°C, with 658 greater than it and 1003 below it (Figure 4.16). The grid cells with these discrepancies are neither clustered nor distributed evenly as

Figure 4.17 (for January) and Figure 4.18 (for June) indicate (refer Table 4-4 for colour legend).

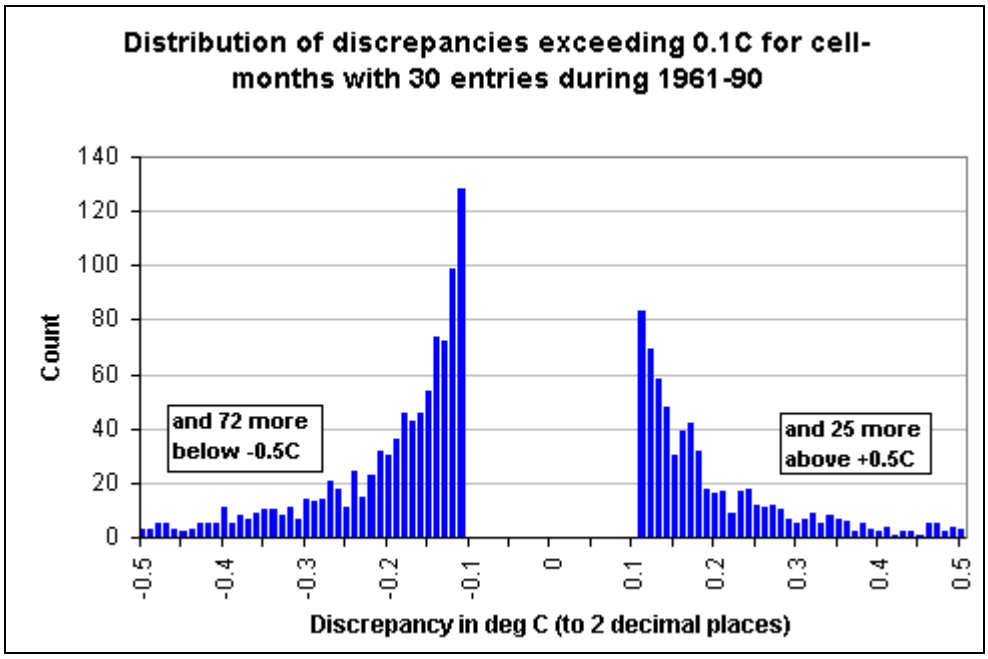


Figure 4.16 As for Figure 4.15 but only where average discrepancy exceeds 0.1°C.

Colour	Meaning
Black	Land grid cells
Grey	Coastal grid cells
White	Seas grid cells
Dark blue	$x < -1.0$℃
Medium blue	-1.0℃ $\leq x < -0.5$℃
Light blue	-0.5℃ $\leq x < -0.1$℃
Light Red	0.1℃ $< x \leq 0.5$℃
Medium red	0.5℃ $< x \leq 1.0$℃
Strong red	$x > 1.0$℃
Where x is the average discrepancy, i.e. the amount to be added to each instance of that cell and calendar month during 1961 to 1990 to make the sum over the period equal zero.	

Table 4-4 Colour legend for Figures 4.17 and 4.18

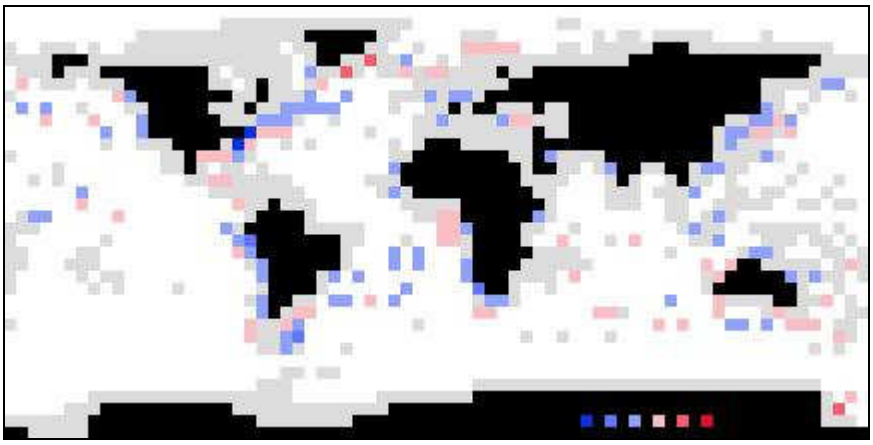


Figure 4.17 Grid cells with 30 years of data for January during 1961 to 1990 but with January discrepancies of more than ±0.1°C (i.e. sum of January values divided by 30 falls outside ±0.1°C indicating an inconsistency). See Table 4-4 for legend.

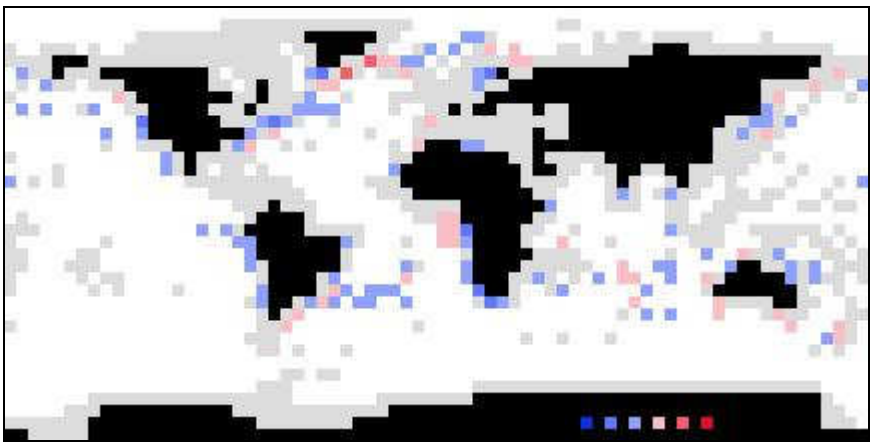


Figure 4.18 As for Figure 4.17 but for the month of June.

While the analysis to this point has focussed on cell-month combinations with 30 values for the period 1961-1990 combinations with less than this number are also relevant. The grid cell centred at 42.5°N 62.5°W (off NE coast of the USA) has 24 years of data for the month of May and has a mean discrepancy for that month of -3.22°C. In total, 42 cell months with from 14 to 24 years of data have mean discrepancies of less than or equal to -1.0°C while 26 have greater than +1.0°C, the greatest of which is 2.7°C.

Correcting these offsets in all months as applicable and recalculating the global average HadSST3 temperature anomaly results in a negligible difference (approx

0.01°C) because of the distribution of the cells needing positive and negative adjustments and the weighting applied to each cell when averaging largely counterbalance the positive and negative influences. As indicated previously, when the region under study is smaller the distortion could well be more significant.

While the above discrepancies are probably a consequence of the original data at 1° x 1° grid cell size and applying to pentads being expanded into 5° x 5° grid cells and months, the discrepancies mean that the HadSST3 data is internally inconsistent.

4.4.6 Testing HadSST3 data over 1961-90 for Gaussian distribution

The SST data for the period 1961-1990 was tested for Gaussian distribution using a similar analysis to that for CRUTEM4 data earlier in this chapter, but this time using the concept of cell-month combinations.

The Shapiro-Wilks test for goodness of fit was again applied, the test examining both the distribution symmetry and the fit to the bell-curve characteristic of Normal Distribution. The HadSST3 dataset is of monthly temperature anomalies rather than mean temperatures but the distribution of the data defines the standard deviation and that should not be affected.

For each cell-month combination with 30 years of data for the period 1961-90 a p-value was calculated according to the Shapiro-Wilks method and the data is summarised in Figure 4.19. Of the total of 11,421 cell-months that were analysed 844 (7.4%) were found to have p-values less than 0.05, 1,538 (13.5%) have p-values less than 0.1, 22,299 (20.1%) have p-values less than 0.2 and 6,240 (54.6%) have p-values below 0.5. The percentage with p-values below 0.1, 13.5%, is almost identical to the 13.9% for observation station data.

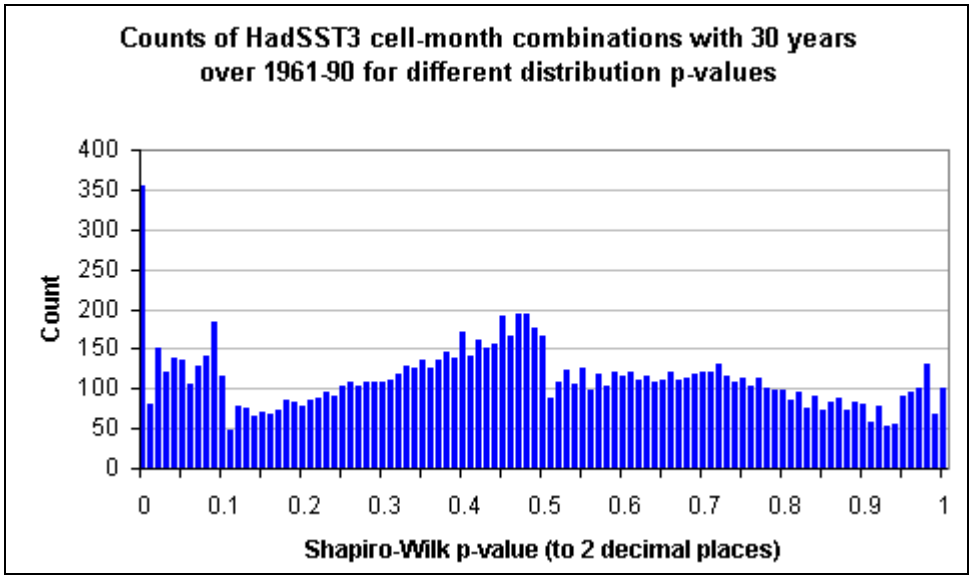


Figure 4.19 Counts of the HadSST3 cell-month combinations with 30 years of data 1961-90 that share a p-value to two decimal places.

As mentioned earlier (sect 4.3.4) Wasserstein & Lazar (2016), says "... a p-value near 0.05 taken by itself offers only weak evidence against the null hypothesis" and that p-values do not measure the probability that the studied hypothesis, which here is the fit to the normal distribution curve, is true. This would apply if the 30 values for each cell-month were analysed to determine if the entire record for the cell-month is likely to be normally distributed but here the analysis is only to consider the data for 1961-1990 and the conclusions that might be drawn from it. With more than 1 in 8 cell-month combinations having p-values <0.1 some caution is called for when drawing conclusions about small numbers of grid cells.

4.4.7 SST Outliers during 1961-1990

Any HadSST3 outliers occurring in the data over the period 1961 to 1990 seem likely to have been included when using SST measurements to refine long-term average temperatures. With this in mind data across the period was examined for any extreme values.

The 1961-1990 average temperatures and standard deviations were calculated for each cell-month combination with 14 entries or more (i.e. 14 or more years with data for

that calendar month), the limit chosen to be comparable to the CRUTEM4 minimum data requirements. The data for each year of all acceptable cell-month combinations was examined for any instances of HadSST3 grid cell values more than three standard deviations from the mean temperature (i.e. the long term average temperature).

After rounding the standard deviations to one decimal place 731 instances were found to be more than 3 standard deviations below the mean, 181 of which were more than 3.5 standard deviations, 51 below 4, and 19 below 4.5. In the opposite direction 1065 instances were found to be more than 3 standard deviations above the mean, 225 more than 3.5, 48 more than 4 and 4 exceeding 4.5 standard deviations.

The lowest extreme was 4.76 standard deviations below the mean, with a temperature anomaly of -7.0°C, but an even lower anomaly of -8.09°C was discovered, it being 4.72 standard deviations below the mean temperature for the grid cell and calendar month combination. At the other end of the scale was a value at 4.9 standard deviations above the mean, with a temperature anomaly of +7.58°C. The maximum temperature anomaly of +7.96°C was 3.81 standard deviations above the mean for the cell and month.

An example of an outlier, at -3.45 standard deviations and a temperature anomaly of -6.9°C, in May 1985 is shown in Figure 4.20 in the central of three grid cells in the mid-Pacific. The smaller anomaly of -4.8°C in an adjacent grid cell later that year is 3.11 standard deviations below the long-term average for that grid cell in that month. Several possible causes for the two aberrations can be found - high winds, unusual upwelling, instrument errors, recording errors - but the impact of most of these would be negligible if numerous SST measurements were made during the month. This is not the case with the anomaly of -6.9°C being from a month with just one observation and the anomaly of -4.8°C having two observations for the entire month.

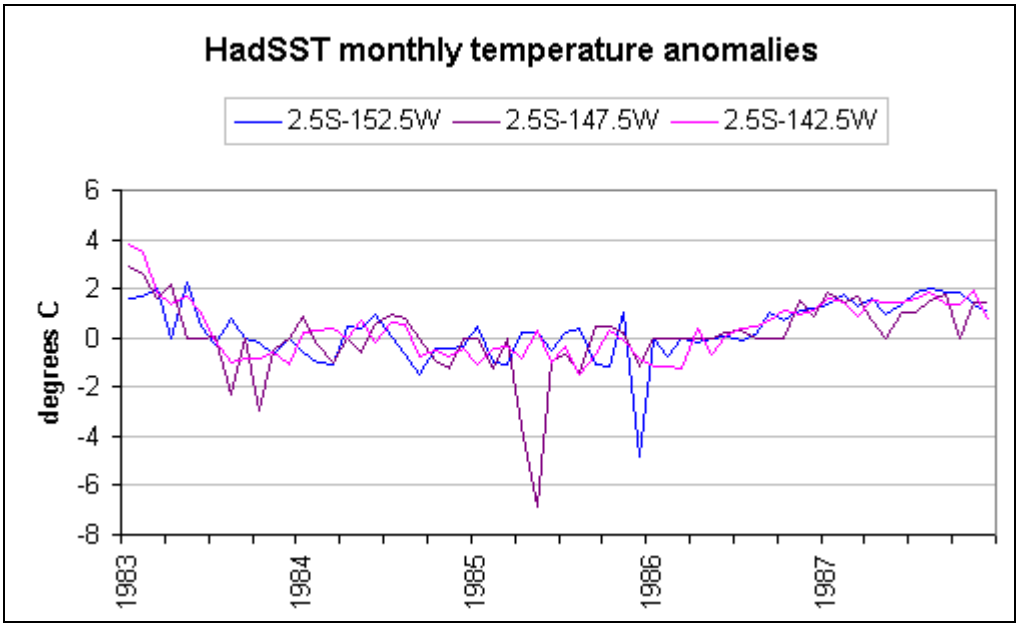


Figure 4.20 An example of two outlying HadSST3 SST anomalies in the mid Pacific.

The presence of outliers in the HadSST3 data during 1961-1990 suggests that they might have been present when long-term average sea surface temperatures were being determined, which would result in distorted averages.

4.4.8 Summary for HadSST3 grid cells

This analysis of HadSST3 grid cells over the period from 1961 to 1990 has revealed three major areas of concern. Firstly, the global average anomalies for each calendar month fail to sum to zero, as is often the case for cell-month combinations, some of which would require average adjustments for the calendar month in excess of 1°C in order to meet the condition.

Secondly, more than 13% of the grid cell - calendar month combinations for the period from 1961 to 1990 do not appear to be normally distributed, which means that caution is required when dealing with or making statements that rely on that distribution (e.g. calculating error margins using standard deviations).

Finally it was shown that outliers beyond three standard deviations are present in the data across the period, suggesting that these might have been included in the process

of determining long-term average temperatures and the associated standard deviations.

TEXT BOX 4B

Mean SSTs are not available but latitude can be used as a crude proxy for temperature when examining the relationship with the magnitude of standard deviations. In a similar manner to that for Text Box 4A grid-cell/month combinations were determined and the standard deviations were calculated for all HadSST3 grid cells that supplied data in every year of the period 1961 to 1990 for the same calendar month. The standard deviations for each grid according to the latitude of the grid cell centre are shown in Figures 4B.1 (for January) and 4B.2 (for July), the arrows indicating particular differences between the two Figures.

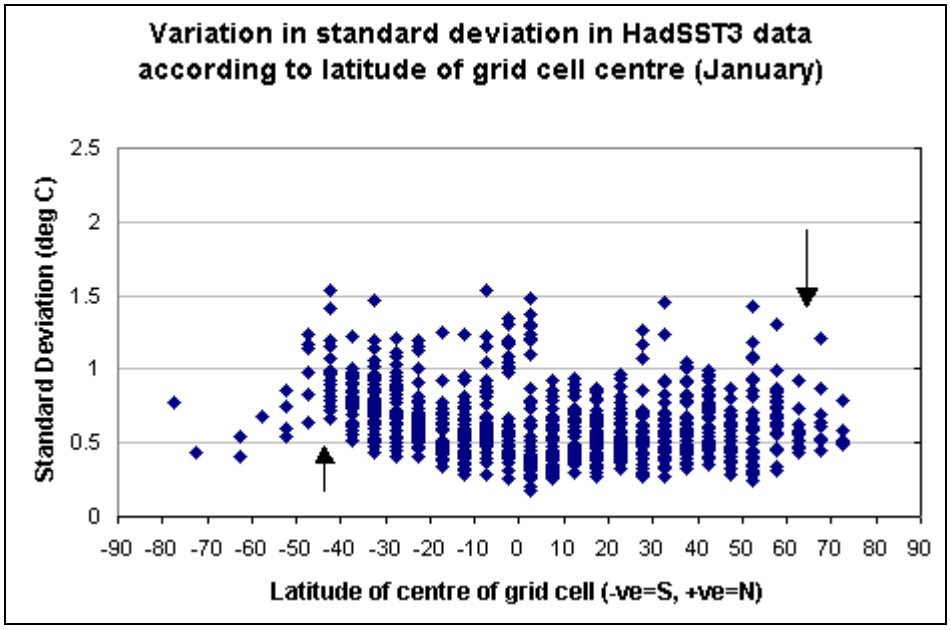


Figure 4B.1 Standard deviations and latitudes for all HadSST3 grid cells with 30 years of data for January 1961-1990. (Arrows indicate notable regions of seasonal variability.)

...

Text Box 4B contd

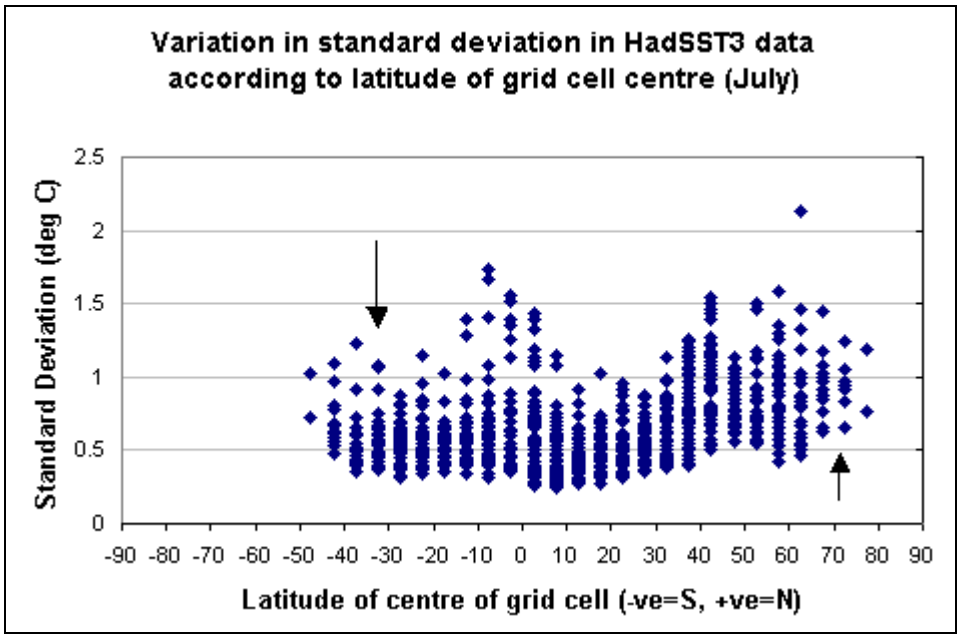


Figure 4B.1 Standard deviations and latitudes for all HadSST3 grid cells with 30 years of data for January 1961-1990. (Arrows indicate notable regions of seasonal variability.)

In a reversal of the situation for station data, standard deviations for SSTs increase as the temperature increases, with both Figures above indicating greatest standard deviations during summer in each hemisphere.

4.5 Summary

The long-term average temperature for a given calendar month and station (for CRUTEM4 data) or grid-cell (HadSST3 data) is imperative for the calculation of temperature anomalies. This chapter has shown that for both temperatures over land and sea surface temperatures there are areas of concern about how these averages are calculated.

While both CRUTEM4 and HadSST3 data have issues specific to those datasets there are also issues common to both. Firstly, it has been shown that 13.5% of such averages are probably not normally distributed (i.e. do not comply with Gaussian distribution patterns). This is a not a significant number in the totality of the CRUTEM4, HadSST3 and HadCRUT4 datasets but could be important when dealing with subsets of that data.

Secondly, both CRUTEM4 and HadSST3 long-term averages are sometimes determined from very few observations. The CRUTEM4 criteria for the minimum amount of data falls well below that recommended by the WMO and sometimes worrying few SST observations are available to refine estimated long-term average sea surface temperatures.

Thirdly, it was noted that both datasets arguably contain outliers in the data used to calculate long-term averages. In the CRUTEM4 dataset these are also included in the calculation of standard deviations and, given that the limit of acceptable data is five standard deviations from the mean, cause the inclusion of data that should realistically be identified as outlying and therefore be excluded. The problems are exacerbated when the outliers are part of a less than complete set of data for the two calculations, as discussed immediately above.

Fourthly, a problem specific to HadSST3 is that the anomalies for a given grid cell and month over the period from 1961 to 1990 often failed to sum to zero even when data is present for all 30 years. This seems likely to be due to the approach of calculating mean temperatures at finer resolution and over shorter intervals, often with few SST measurements, then extrapolating and interpolating that data to produce the HadSST3 dataset. On this basis the technique seems flawed and the error margins quite large.

Much of the above will have only minor impact when only one data source - station for CRUTEM4, grid cell for HadSST3 - is considered, except when outliers are present in the data; anomalies might not be exact but the shift will be constant and trends unchanged. Problems can however arise when the anomalies are averaged with

those from other stations or grid cells that might be to a different base, and particularly when the number of sources varies and the starting or ceasing of sources might cause false steps in the sequence of the composite values.

Chapter 5: Other issues with the HadCRUT4 dataset

5.1 Introduction

This chapter discusses several other issues associated with the accuracy of the HadCRUT4 dataset. It starts by pointing out numerous discrepancies of different types between grid cell values in the HadCRUT4 dataset and the two associated datasets, CRUTEM4 and HadSST3.

It moves on to 'coastal' grid cells, where the data source might be from observation stations on land (CRUTEM4) or from sea surface temperature data (HadSST3) or a combination of both, to show inconsistencies in not only coverage but also the data sources for the same grid cells over time.

Outliers were discussed in the previous chapter in regard to their influence on long-term average temperatures in particular and for CRUTEM4 data, via the calculation of standard deviations, their impact on the identification of outliers at other times. This chapter will look at the inclusion of outliers across the entire 1850-2015 data record in both the SST data and the data from observation stations. While their impact outside the period of calculation of long-term averages and standard deviations is limited, even the existence of outliers is a general concern in itself.

This chapter also points out, for CRUTEM4 data from observation stations, some instances of large ranges of temperature anomalies from stations in the same grid cell. In some cases but not all the removal of outliers would greatly reduce the range of values. This might seem to be a minor issue but the crux of the problem is that the stations that reported data when those wide ranges occurred might not always report data and that shifts in cell values might be due to the failure of stations to report rather than meteorological causes. For example, with two stations within the same grid cell and the range in the anomalies exceed 7.0°C degrees as it sometimes does, the failure of one of those stations to report data would shift the grid cell average up or down as appropriate by 3.5°C.

Errors are also identified in the station data from which the CRUTEM4 and HadCRUT4 datasets are created, in particular that the location for one observation station is obviously incorrect and that the location of another one is doubtful. These are not large errors but they raise the question of quality control and undermine the credibility of the data.

In total it seems unlikely that any of the above issues will have a large influence on the HadCRUT4 dataset in regard to hemispheric or global averages but when studies are focussed on smaller regions there is a risk that if the proportion of erroneous data to good data increases then accuracy of the regional averages could be in doubt. As mentioned above, the very existence of some of these issues undermines confidence in the datasets.

5.2 HadCRUT4 fails to match source dataset

Several discrepancies exist between the grid cell values for the HadCRUT4 data and the two associated datasets CRUTEM4 (derived from observation station data) and HadSST3 (derived from sea surface temperatures).

The HadCRUT4 dataset is created from data used in the CRUTEM4 and HadSST3 datasets so it follows that if only one of the CRUTEM4 or HadSST3 datasets reports data in a given grid cell for a given month then the HadCRUT4 dataset should contain that value or, given that HadCRUT4 uses the "average of the ensemble" approach (see Chapter 1), at least a value that is very close to it. It also follows that if neither CRUTEM4 nor HadSST3 report data for a given cell and month then there should also be no corresponding data in the HadCRUT4 dataset. Instances were identified of both of these situations being untrue.

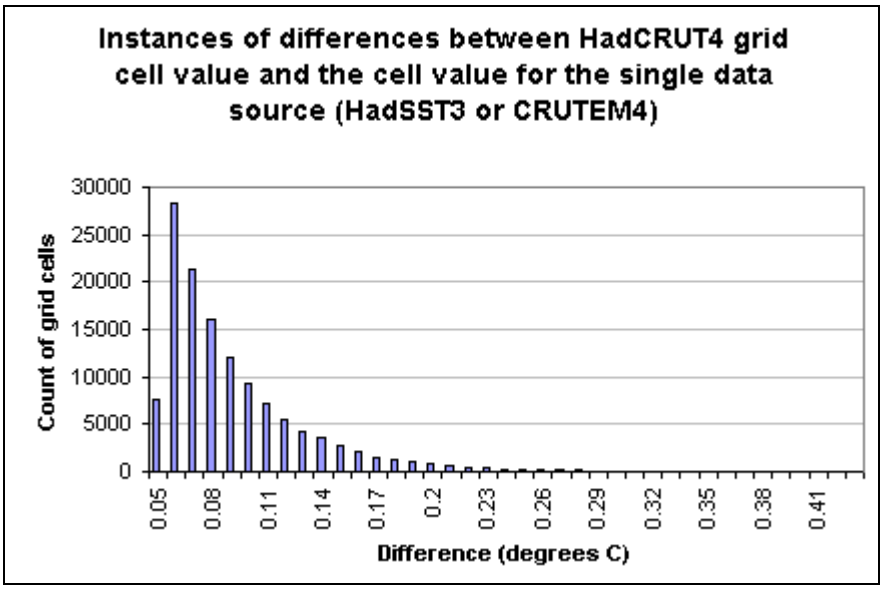
In the latter case it was found that the HadCRUT4 dataset has data for certain grid cells in May and September 2009 when neither the HadSST3 nor CRUTEM4 datasets

have data in those grid cells. The source of the HadCRUT4 data is therefore unknown.

The first situation, of a discrepancy between HadCRUT4 dataset and the single other dataset reporting data is more complicated. An analysis of the entire HadCRUT4 dataset when only one data source is available (2,004,617 grid cell - month combinations) shows 1,634,589 instances (81.5%) of exact matches with the single source, which was either CRUTEM4 or HadSST3. A further 247,538 instances have a difference between corresponding grid cells and months of less than 0.05°C, which is somewhat arbitrary limit intended to take into account the HadCRUT4 and HadSST3 datasets being "ensembles" (see chapter 1).

The problem lies with the 127,490 instances in which the HadCRUT4 dataset shows a difference of 0.05°C or more from the corresponding values in the single other dataset in which they are found. Of these instances 34,435 show a difference of 0.1°C or more. Maximum discrepancies of 0.43°C (in 3 instances), 0.42°C (7 instances) and 0.41°C (16 instances) were discovered. The three instances of the maximum discrepancy, 0.43°C, occur when the HadCRUT4 dataset has values of 0.78°C, 2.53°C and -3.07°C compared to the values in the single other datasets of 1.21°C (in HadSST3), 2.1°C and -3.5°C (both in CRUTEM4) respectively.

Figure 5.1 shows the number of instances of mismatched grid cells where the differences are rounded to two significant digits, the top portion of the figure showing all differences and the bottom showing differences of 0.25°C or greater. Figure 5.2 shows the number of instances where the difference was ≥0.05°C in each year for the full period of the HadCRUT4 data.



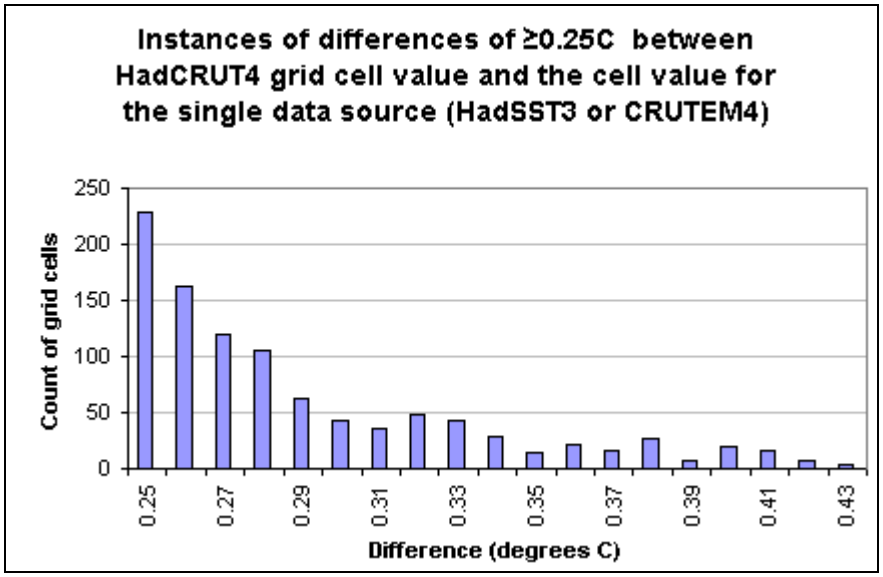


Figure 5.1 Count of instances (i.e. cell-month combinations) when the absolute value of the difference between HadCRUT4 and the single source it was drawn from (either CRUTEM4 or HadSST3) was greater than or equal to 0.05°C. (Top: all data, bottom: data when difference ≥0.25°C)

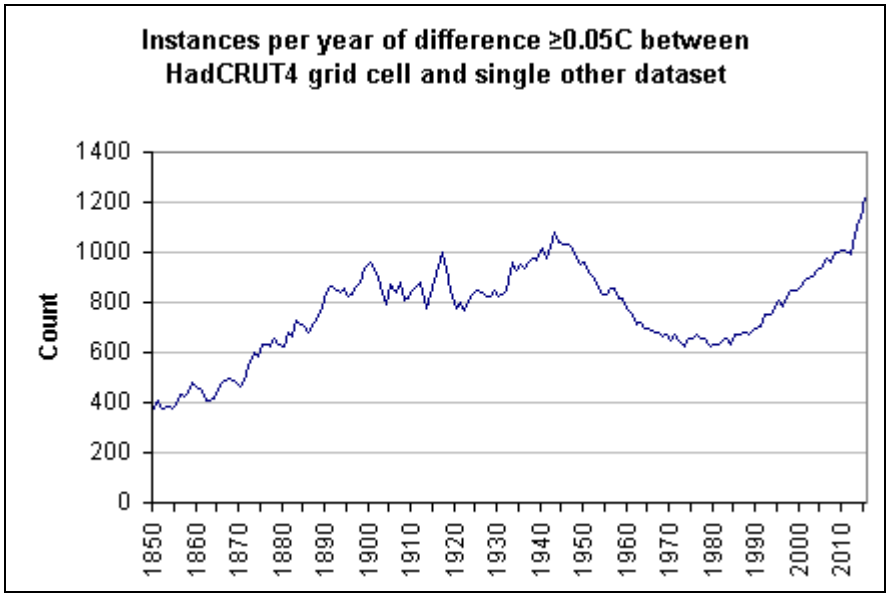


Figure 5.2 Number of instances per year in which the difference between HadCRUT4 and data for a given grid cell and month differed by ≥0.05°C from whichever of CRUTEM4 and HadSST3 contained data in that grid cell and month

Analysis of the instances with differences of more than 0.1°C showed that they were approximately consistent across longitude bands but of the 34,435 instances where that difference was exceeded 30,978 (90%) were for the Northern Hemisphere and only 3,357 for the Southern Hemisphere. The HadCRUT4 data differed by more than 0.1°C from the HadSST3 data in just one instance, leaving 34,434 discrepancies with the CRUTEM4 dataset.

A similar situation was found with the 127,490 instances of discrepancies of more than 0.05°C with over 99% being in relation to CRUTEM4 data and in this case 110,328 (86.5% of the total) being from the Northern Hemisphere and only 17,162 (13.5%) from the Southern Hemisphere.

While some variation between the single data source and HadCRUT4 might be expected due to the ensemble approach, the number of relatively large differences has nor been explained by either the CRU or the Hadley Centre. It appears that, contrary impressions given by the documentation, that the creation of the HadCRUT4 and CRUTEM4 datasets use different processing of the data from land-based observation stations.

5.3 Inconsistent sources for coastal grid cells

"Coastal" grid cells are those cells where part of the cell is over land and part over sea. The data for coastal grid cells might be from observation stations, from sea surface temperatures, or when both are available from a merging of the two according to the fractional areas of land and sea, with a minimum land fraction set to 0.25 (Morice et al, 2012).

To take into account the "ensemble" nature of the HadCRUT4 and HadSST3 datasets when determining the source the data from these, plus the CRUTEM4 dataset, the following principles were adopted in this analysis:

- ***Ignore when data missing from all three datasets***
- ***If HadCRUT4 data present but HadSST3 and CRUTEM4 missing*** assign to special "unsourced" group
- ***If only HadSST3 reported data,***
  (a) If HadSST3 value within ±0.125°C of HadCRUT4 value assign to group "sea"
  (b) else assign to "> 0.125C" group
- ***If only CRUTEM4 reported data,***
  (a) If CRUTEM4 value within ±0.125°C of HadCRUT4 value assign to group "land"
  (b) else assign to ">0.125C" group
- ***If both CRUTEM4 and HadSST3 reported data***
  (a) if HadCRUT4±0.125°C falls outside the range defined by CRUTEM4 and HadSST3 values then list as "outside range"
  (b) else assign to group "merged".

No instances were identified as "unsourced", just four "outside range", which will be discussed later in this section, and only a small number in the ">0.125C", leaving the vast majority as "land", "sea" or "merged".

Figure 5.3 shows average annual "coastal" grid cell coverage from the different sources and Figure 5.4 the average annual percentage contribution of each of those sources to the total "coastal" coverage.

While the group in which HadCRUT4 data differed from CRUTEM4 and HadSST3 by more than 0.125°C was non-zero the coverage of such grid cells peaked at 0.25% of the earth's surface in 1876 and made maximum percentage contribution to coverage of 3.5% in 1868, which are negligible compared to the three main groups and have been omitted from Figures 5.3 and 5.4. Morice et al (2012), the primary reference for HadCRUT4 details, makes no mention of these variations and therefore no explanation as to why they might be this large.

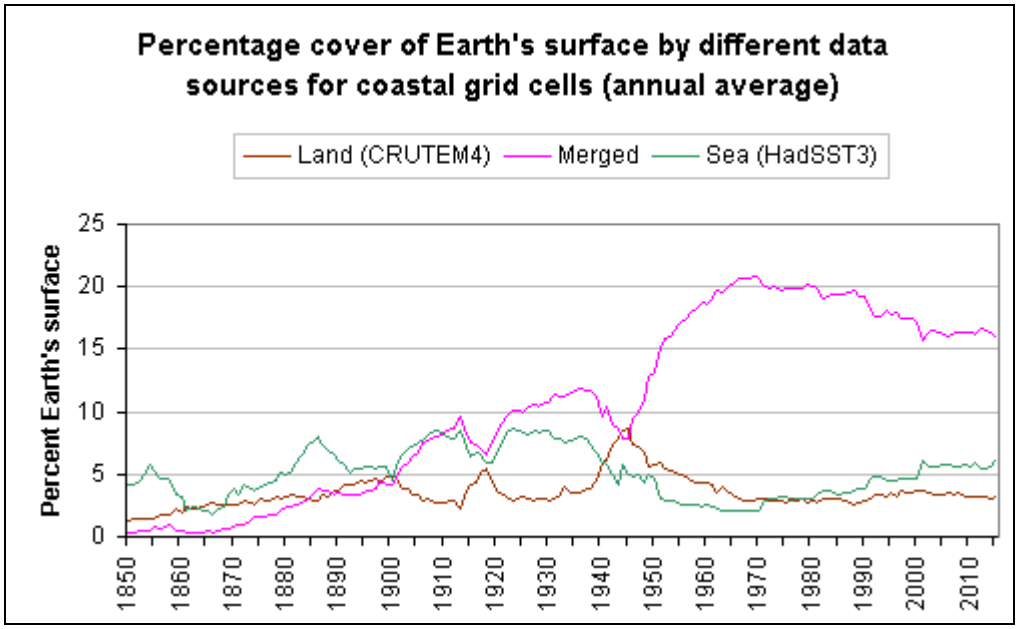


Figure 5.3 Global coverage, expressed as a percentage of the Earth's surface, of coastal grid cells from either one of CRUTEM4 and HadSST3 datasets and when a merging of the values in the two datasets.

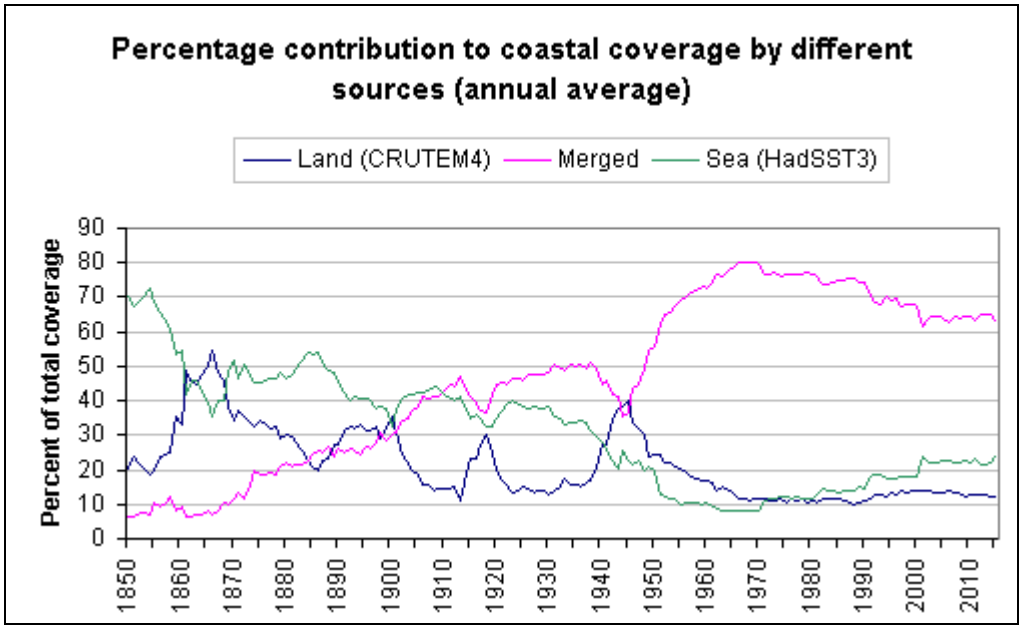


Figure 5.4 Data for Figure 5.3 (above) expressed as a percentage of the total coastal coverage at the time.

The same criteria as above were used to determine the data source for each grid cell in each month and therefore identify those cells in which the source for the calendar month changed over time. Table 5-1 shows that on average the data for a given calendar month was from a consistent source for only about 108 (17.1%) of the 633 coastal grid cells that reported data at any time from 1850 to 2015. In all other cases the data source varied over time between the three categories i.e. exclusively from land, exclusively from sea or a merging of the data from the two sources.

Figure 5.5 shows the average annual percentage of the coverage of each hemisphere attributable to coastal grid cells for the entire period of the record. It was noted in Chapter 1 that coastal grid cells account for 33.9% of the northern hemisphere, which means they were over-represented by as much as 50% in the coverage prior to 1950, with the southern hemisphere 22.2% coverage over-represented only slightly less in magnitude but in general for a shorter period. When considered with Figure 5.4 we see that the reducing over-representation in the northern hemisphere coincided with increasing coastal coverage via the merging of land and sea data, and a decrease in the coverage attributable to data only from land only.

Month	Changed Source	Unchanged Source
Jan	516	123
Feb	529	113
Mar	518	130
Apr	522	121
May	518	110
Jun	523	108
Jul	534	101
Aug	536	94
Sep	537	99
Oct	533	101
Nov	532	93
Dec	529	106

Table 5-1 The number of coastal grid cells that changed and did not change their data source for the same calendar month over time.

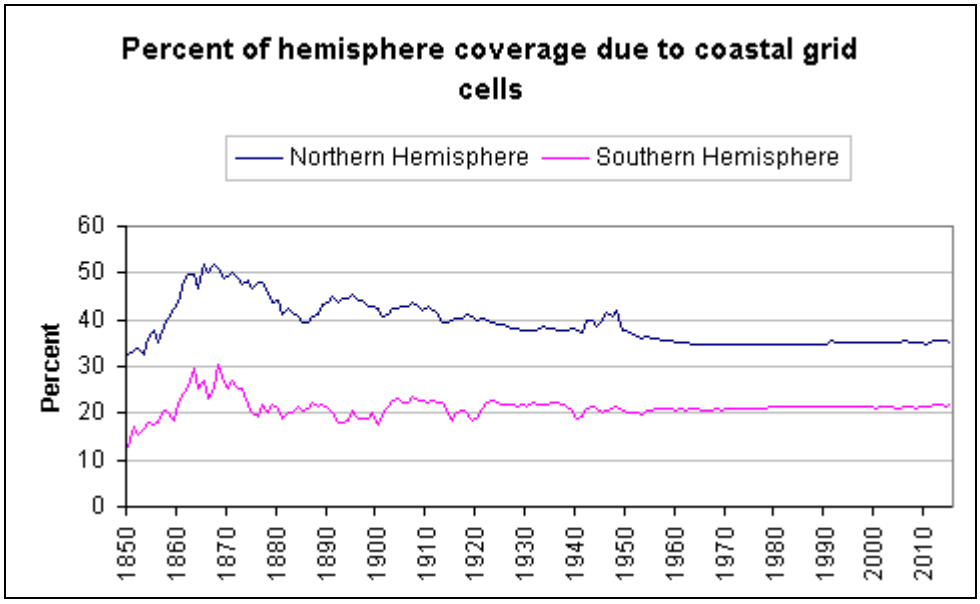


Figure 5.5 Percentage contributions to the reported coverage in both hemispheres from 1850 to 2015

The analysis of the data sources for coastal grid cells also revealed four instances where HadCRUT4 values for a grid cell fell outside the range of HadSST3 and CRUTEM4 data, despite the allowance of ±0.125°C (Table 5-2). As with the case of the HadCRUT4 dataset reporting data when neither CRUTEM4 nor HadSST3 datasets did so, no explanation seems forthcoming for these "out of range": values.

Year	Month	Cell centre Latitude	Cell centre Longitude	HadCRUT4	CRUTEM4	HadSST3	Excess
1869	4	37.5N	117.5E	**-1.25**	-1.4	-1.42	**0.15**
1942	12	42.5N	127.5E	**-0.37**	-0.5	-1.11	**0.13**
1996	5	37.5N	32.5E	**2.14**	2.35	2.62	**-0.19**
2011	10	67.5N	167.5E	**3.93**	3.8	3.79	**0.13**

Table 5-2 Instances where the HadCRUT4 value for a coastal grid cell is more than 0.125C outside the range bounded by HadSST3 and CRUTEM4 values.

The variation in coverage of coastal grid cells (Figure 5.4) is to be expected given that both CRUTEM4 and HadSST3 datasets vary in coverage and the data for coastal grid cells is derived from whatever data from these two sources is available for the given month. Despite this, the problem remains that the source and coverage of this data has been very inconsistent over time. It cannot be ruled out that these inconsistencies might contribute to changes over time in HadCRUT4 hemispheric and global averages, or for that matter cause spurious trends in the temperature anomalies for individual grid cells.

5.4 Data Outliers

5.4.1 Introduction

Chapter 4 discussed the implications of the presence of outliers during the period over which long-term average temperatures are calculated, and in the case of observation station data, during the period over which standard deviations are calculated, because errors or distortions of these two factors will be carried through the entire data record since 1850.

This section will address the presence of outliers during other times in the record, when they will not impact all data for the grid cell since 1850 but have the potential to distort grid cell values in the months in which they occur and distort derived data such

as averages or temperature trends when the outlying data is incorporated into the calculations.

CRUTEM4 data, from observation stations, will be treated separately to HadSST3 data because the analysis is constrained by the available information. This is due to the identification of outliers in CRUTEM4 data being relatively straightforward because monthly mean temperatures are published for each observation station, along with the CRU's calculation of long-term average temperatures and standard deviations.

In contrast, as Chapter 4 discussed, the HadSST3 dataset, with grid size of 5° x 5° latitude and longitude and at monthly intervals is derived from temperature anomalies based on 1° x 1° grid cells and 5-day periods known as pentads. Unlike with CRUTEM4 station data the SST data at smaller resolution and period is not published with the HadSST3 dataset, so this analysis adopts a different approach.

5.4.2 Outliers in CRUTEM4 data from observation stations

Outliers in data, whether due to abnormal meteorological events or to instrument malfunction or errors at some point in data processing, e.g. incorrectly flagged missing data, can distort the final product. The usual practice when dealing with possible outliers is to set a threshold, typically defined as some multiple of standard deviations of the data away from the mean value, and to declare values beyond that point to be outliers and exclude them. When constructing the CRUTEM4 dataset the CRU adopts a threshold for outliers of five standard deviations from the mean temperature and although calculating the long-term average temperatures from data over the 30-year period from 1961 to 1990 the standard deviations used for CRUTEM4 are calculated over a minimum of 15 years of data over the 50-year period from 1941 to 1990.

The use of the long-term averages and standard deviations provided in the metadata of the CRUTEM4 station data files results in 1829 instances of data outliers, i.e. beyond the five-standard deviations limit.

The probability of normally distributed data exceeding five standard deviations is ~1 in 1.25 million. To use this threshold for outliers is exceptionally generous given that for each calendar month there is a maximum of only 166 values (i.e. a value for every instance of the given calendar month in every year from 1850 to 2015).

A threshold of three standard deviations, with probability of exceeding that value of ~26 in 10,000 (i.e. ~1 in 384), would be too low given that the number of years of data (166) is 43% of 384, meaning a high risk of outliers based on probability alone.

The compromise position, adopted for the purposes of this present analysis, is to set the threshold at four standard deviations, where the probability of exceeding that value is ~6 in 100,000 (or ~1 in 16,666), and to use long-term averages and standard deviations based on all available data rather than the CRUTEM4 1961-1990 subset for averages and 1941-90 subset for standard deviations. The advantage of using all available data is that the inclusion of outlying data in the calculation of the long-term average temperature and standard deviation will cause less distortion than it would when calculating from just 30 values for the former and 50 for the latter.

The entire set of station data was analysed, using again the concept of station-month combinations because the long-term average temperature and the standard deviation are determined for each calendar month. This identified 1418 outliers beyond four standard deviations, of which 86 were beyond five standard deviations. The identification of just 86 compared to the 1829 determined by using CRU-supplied long-term average temperatures and standard deviations calculated over a different period (see above in this section) suggests that the CRU might have excessively trimmed the data, especially low values in the early years of station records.

Table 5-3 shows some extreme examples of outliers in the station data. The mean (i.e. long-term average temperature), standard deviation and years of record for the

month in question are given to show the context of the monthly mean ('Tmean') being the number of standard deviations from that mean in the given calendar month.

Stn ID	Station name	Month	Mean	StDev	Yrs	Year	Tmean	SDs
720192	AMITE	8	26.97	0.91	131	1885	32.4	6.0
804100	BARQUISIMETO	8	23.86	0.89	74	2008	29.7	6.6
679640	BULAWAYO/GOETZ-OBS.	5	16.29	3.33	110	2013	-16.3	-9.8
425150	CHERRAPUNJI	5	19.46	1.57	106	1902	33.1	8.7
416240	DERA ISMAIL KHAN	12	13.71	1.12	134	2005	20.4	6.0
670090	DIEGO-SUAREZ	11	27.23	5.13	63	2013	67.3	7.8
434950	GALLE	9	26.65	0.49	111	1901	29.7	6.2
434950	GALLE	8	26.59	0.54	111	1901	29.8	6.0
434950	GALLE	5	27.57	0.64	111	1894	23.2	-6.8
637230	GARISSA	8	27.07	0.81	63	2003	32.0	6.1
417150	JACOBABAD	9	31.81	1.31	138	1938	19.0	-9.8
588340	NANPING	7	29.28	1.68	61	1951	17.8	-6.8
588340	NANPING	8	28.85	1.55	62	1951	18.0	-7.0
852420	ORURO	9	11.55	11.05	52	2011	90.0	7.1
80150	OVIEDO EL CRISTO	6	16.43	2.31	86	1945	0.1	-7.1
483780	PHITSANULOK	8	28.36	0.66	74	2006	32.5	6.3
859310	PUNTA DUNGENES	9	4.84	1.3	89	1907	13.8	6.9
279950	SAMARA (BEZENCUK)	7	21.01	2.59	112	2015	2.0	-7.3
985500	TACLOBAN	7	27.89	1.11	100	1906	18.2	-8.7
14650	TORUNGEN FYR	1	0.3	2.96	148	1937	18.1	6.0
913340	TRUK WSO A	1	27.09	3.22	73	2012	0.0	-8.4
60510	Vestervig	7	15.42	1.6	136	1874	4.2	-7.0
627510	WAD MEDANI	4	32.27	8.61	64	2011	99.9	7.9
480970	YANGON	10	27.76	0.6	116	1993	23.8	-6.6
318290	ZOLOTOJ	10	7.3	1.76	75	2010	19.6	7.0

Table 5-3 Extreme outliers based on mean temperatures and standard deviations calculated across all available data for each station in each calendar month.

Implausible mean monthly temperatures of 67.3°C (at Diego-Suarez), 90.0°C (Oruro) and 99.9°C (Wad Medani) produced implausible standard deviations of 5.13°C, 11.5°C and 8.61°C respectively. (Note that adopting the five standard deviations threshold, as per the CRUTEM4 practice, would mean the inclusion of values within ranges of 51.3°C, 115.0°C and 86.1°C respectively for the calendar months in which the implausible mean temperatures are reported.)

Given that the mean temperature for "Bulawayo/GoetzObs" station is 16.29°C for the month of May a reported mean of -16.3°C appears to have an accidental negative sign ("-"). Further, the mean monthly temperatures of 0.0°C and 0.1°C seem very unlikely at Truk WSO A (an island in the Pacific) and Oviedo El Christo (Spain) when the mean temperatures for the applicable calendar months were 27.09°C and 16.43°C respectively.

It seems very likely that the mean monthly temperatures of 0.0°C, 90.0°C and 99.9°C in Table 5-3 were intended to indicate missing data but failed to use the correct flag value of -99.0°C.

Various outliers in Table 5-3 seem likely to be due to instrument, observer or transcribing error include Cherrapunji where 33.1°C perhaps should have been 23.1°C, Jacobabad where 19.0°C perhaps should have been 29.0°C and 2.0°C for a July monthly mean temperature at Samara/Benzencuk is perhaps missing a digit ("1") when the long-term average for that month is 21.01°C.

The likely causes of certain other outliers are less certain, especially when some are less than 5°C from the mean temperature for the calendar month (e.g. Balasor and Galle). The probability of these mean temperatures occurring might be very low but it is not beyond possibility that they might have been genuine. They might have been caused by a rare meteorological event (e.g. heatwave with hot winds) or, if the temperature gradient through the month is quite steep then the absence of data for a number of consecutive days at one end of the month might have distorted the monthly mean.

The Nanping and Vestervig outliers seem to be part of an extended sequence of errors at both locations. The data for Nanping has seven entries in the same year where the mean monthly temperatures are about 10°C below the long-term averages. At Vestervig four consecutive months in 1874 have mean temperatures more than 5σ below their calendar month long-term average temperatures.

Outside the values listed in Table 5-3 are others that are not as many standard deviations from their long-term averages but appear very likely to be errors. A selection of these is shown in Table 5-4, which is in the same format as the table above.

Stn ID	Station name	Mon	Mean	StDev	Yrs	Year	Tmean	SDs
800890	APTO_OTU	4	27.07	11.89	22	1978	81.5	4.6
800890	APTO_OTU	6	27.07	11.76	24	1978	83.4	4.8
800890	APTO_OTU	7	27.23	11.99	23	1978	83.4	4.7
915920	NOUMEA	11	23.92	1.08	97	1912	29.3	5.0
915920	NOUMEA	4	24.26	1.17	97	1913	29.9	4.8
915920	NOUMEA	5	22.48	1.09	96	1913	28.6	5.6
915920	NOUMEA	8	20.13	1.13	96	1913	26.8	5.9
915920	NOUMEA	9	21.03	1.18	95	1913	27.4	5.4
915920	NOUMEA	10	22.45	1.22	95	1913	28.7	5.1
915920	NOUMEA	11	23.92	1.08	97	1913	29.2	4.9
400610	PALMYRA	6	27.44	1.55	63	1938	18.0	-6.1
400610	PALMYRA	7	29.58	1.4	64	1938	21.6	-5.7
400610	PALMYRA	8	29.51	1.9	63	1938	19.3	-5.4
400610	PALMYRA	9	26.32	1.6	64	1938	16.9	-5.9
913480	PONAPE	1	27.10	0.5	75	1926	24.6	-5.0
913480	PONAPE	2	27.16	0.57	78	1926	24.4	-4.9
913480	PONAPE	3	27.27	0.6	77	1926	24.2	-5.1
913480	PONAPE	4	27.19	0.59	76	1926	24.2	-5.0
672610	TETE	4	26.87	1.44	63	1935	19.7	-5.0
672610	TETE	6	22.17	1.26	64	1935	15.8	-5.0
672610	TETE	7	21.87	1.14	65	1935	15.3	-5.8
672610	TETE	8	23.69	1.39	67	1935	14.9	-6.3
560040	TUOTUOHE	11	-10.43	3.06	60	1985	-28.1	-5.8
560040	TUOTUOHE	12	-14.21	2.29	59	1985	-26.9	-5.5
560040	TUOTUOHE	1	-15.30	2.43	58	1986	-28.2	-5.3
383580	Susamyr	1	-19.9	8.46	32	2013	23.6	5.1
873050	JACHAL	12	23.58	3.86	43	1994	0.0	-6.1
858840	CABO RAPER	12	10.28	1.58	58	1933	0.8	-6.0
722590	DALLAS/FORT WORTH	9	25.32	2.55	55	2009	9.8	-6.1
636300	GULU	7	21.94	1.45	70	1937	13.1	-6.1
404380	RIYADH	2	17.16	2.35	62	2002	30.0	5.5

Table 5-4 Other outliers according to long-term average temperatures and standard deviations calculated across all reported data for each observation station in each calendar month.

The first three instances in Table 5-4 are from "Apto_Uto" (Colombia) with reported monthly mean temperatures implausibly above 80.0°C. The analysis used in this section differs from the approach used to create the CRUTEM4 dataset but as noted in

the previous chapter, these monthly mean temperatures were included when both the long-term average temperatures and standard deviations were calculated, meaning that both are distorted and the outlier threshold based on five standard deviations is so wide as to allow the inclusion of temperatures in a range of over 100.0°C for that calendar month.

Tulcea (Romania) is not listed in either Table 5-3 or 5-4 but for the years 2007, 2008 and 2009 the monthly mean temperatures the data appears to be shifted by one decimal place (e.g. mean temperature for June 2007 is 2.3°C c.f. the June average across 1961-90 of 20.6°C) except for the months from December to March. While the given mean temperatures fall within the range of mean temperatures of the past, the number of degrees C for all months except that for January 2007 could be multiplied by ten and would still fall within the acceptable range.

Both Jones et al (2012) and Osborn & Jones (2014) report that outliers of more than 5 standard deviations are excluded unless they have obvious problems that can be corrected. The data files published in association with the CRUTEM4 dataset do not include a set of station data files that have been corrected or data removed so it is not possible to determine the changes that have been made or to verify that the data has modified as described. What is clear however is that the inclusion of some erroneous values when calculating long-term average temperature and standard deviations has negative repercussions on the CRUTEM4 dataset.

A greater problem to be addressed is why such obvious outliers are included in the station data files. It would seem that not all the national meteorological services that supply the data to the CRU, nor the CRU itself, apply suitable quality control measures to the data in question.

5.4.3 Outliers in HadSST3 data

As discussed in the previous chapter the HadSST3 data, with grid cells of 5° latitude x 5° longitude and intervals of one month are derived from temperature anomaly data

that uses 1° x 1° grid cell size and pentads (i.e. intervals of 5-days). Direct analysis of the finer granularity data is not possible without the complete set of information including the long-term average temperatures used in calculations. Likely outliers can however be identified from the available HadSST3 data.

A first approach is to follow the practice for CRUTEM4 data for observation station data by considering only HadSST3 cell-month combination with at least 14 years of data for the given calendar month during the period from 1961 to 1990. Applying the cell-month combination method described in chapter 4 this identified 14964 cell-months. The long-term average temperatures and standard deviations were calculated for each of these and an outlier threshold of four standard deviations was set to make it consistent with the discussion in chapter 4.

This resulted in 22,729 outliers being identified for the full period of available data (i.e. 1850 to 2015), 14,360 (63.2%) being more than four standard deviations below the long-term average temperature and 8369 (36.8%) more than four standard deviations above. Despite the greater amount of ocean in the Southern Hemisphere only 8,784 (38.6%) of the outliers applied to grid cells located there, leaving 13,945 (61.4%) for the Northern Hemisphere.

Table 5-5 shows details of some extreme outliers, including the month of their occurrence, the number of entries used to calculate the mean (i.e. long-term average temperature) and standard deviation, then the anomaly given by the HadSST3 dataset and the number of standard deviations from the mean that this represents.

From the same analysis the annual total number of outliers per year, i.e. instances of values beyond four standard deviations from the long-term average, was determined (Figure 5.6). Not surprisingly the period over which the averages and standard deviations were calculated appears to contain few outliers, this because the outliers were included in the calculations of the two values.

The problem with this approach is that non-seasonal components of sea surface temperatures seem to change slowly, meaning that, as Table 5.5 suggests, standard deviations calculated over the 30 years from 1961 to 1990 are small. The average

standard deviation for the 16494 cell-months across 1961-1990 is just 0.73°C, during a period when the monthly HadSST3 global average temperature anomalies have a range of just 0.596°C (min = -0.335°C, max = 0.261°C). The low average standard deviation means that quite narrow outlier thresholds are applied across the entire data record (i.e. 1850-2015) where the monthly HadSST3 global average temperature anomalies have a range of 1.469°C (min = -0.744°C, max = 0.725°C).

Cell centre Latitude	Cell centre Longitude	Year	Month	Mean	Std Dev	Av Years	Tanom	SDs
2.5	152.5	1942	4	0.01	0.23	30	-7.11	-31.10
7.5	152.5	1942	4	0.02	0.28	30	-7.49	-26.84
-2.5	107.5	1858	6	-0.01	0.26	30	-6.82	-25.68
-2.5	152.5	1942	4	0.00	0.28	30	-6.74	-24.24
-22.5	-72.5	1918	8	-0.01	0.29	30	-6.91	-23.81
-17.5	-57.5	1854	9	-0.02	0.27	30	-6.19	-23.18
12.5	47.5	1941	7	-0.03	0.41	30	6.59	16.32
17.5	-7.5	1910	10	-0.03	0.42	30	6.92	16.71
62.5	-52.5	2010	12	-0.08	0.42	29	6.91	16.76
-47.5	-62.5	1937	3	-0.52	0.39	30	6.37	17.81
-52.5	162.5	1918	4	-0.07	0.27	30	5.03	19.09
-47.5	-62.5	1952	2	-0.21	0.39	30	7.23	19.23

Table 5-5 Details extreme HadSST3 outliers, where Tanom is the temperature anomaly in the given month and SDs being the number of standard deviations (column 'Std Dev') from the mean temperature anomaly for that month. (Mean and standard deviations based on 1961-90 data, with a minimum of 14 instances of the calendar month.)

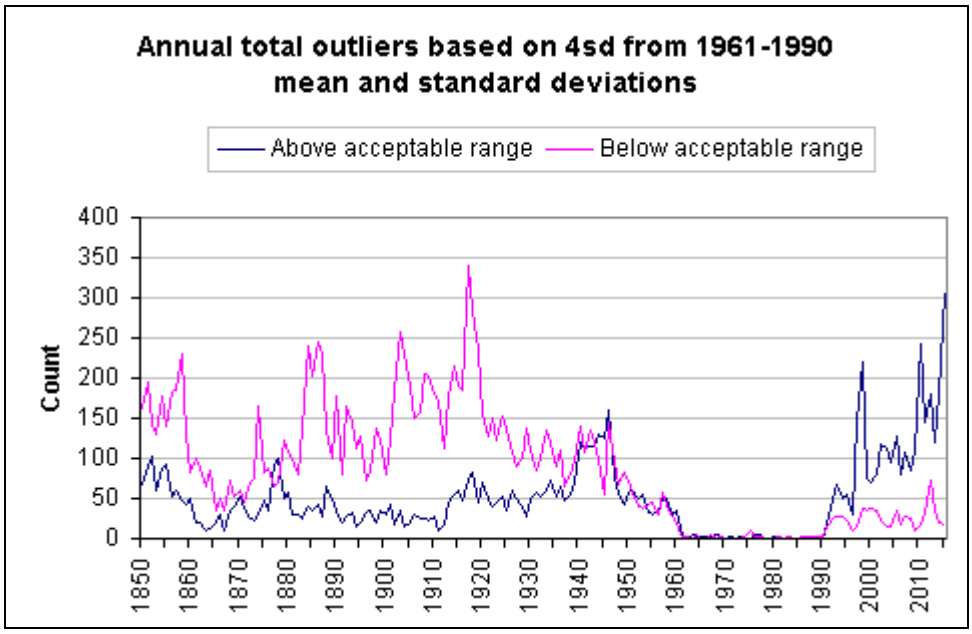


Figure 5.6 Annual total HadSST3 outliers beyond the "four standard deviation" threshold based on long-term average temperatures and standard deviations for each cell-month combination calculated from a minimum of 14 entries for the period 1961-1990.

Figure 5.6 supports this with large numbers of negative (i.e. below-range) outliers and fewer positive outliers in data prior to 1961-1990 and the opposite situation after. (Figure 5.6 also needs to be interpreted in the context of the number of observations increasing over time.)

On this basis a second approach was used, employing the same method as in the previous section for CRUTEM4 station data, viz calculating long-term averages and standard deviations from all available data (i.e. across 1850-2015) for each cell-month combination and then, for all cell-months with at least 14 entries over that time, using the calculated means and standard deviations to identify outliers beyond four standard deviations.

This approach produced 18690 acceptable cell-months combinations, up from the 16494 across 1961-1990 by including more years. The average standard deviation increased to 0.90C (which means the average outlier threshold of four standard deviations increases from 2.92°C to 3.6°C, an increase of more than 20%).

Based on these new criteria 4857 outliers were identified, the most extreme of which are shown in Table 5-6 where none of the SST long-term averages or standard deviations for the cells was calculated with less than 132 entries (i.e. 132 years of data for that calendar month).

Figure 5.7, similar to Figure 5.6, shows the annual total number of outliers beyond four standard deviations from the mean where the mean and standard deviation are based on at least 14 years of data in for the cell-month combination across the entire period from 1850 to 2015. The scale on the Y-axis is one quarter that of Figure 5.6, indicating a smaller maximum annual total.

Cell centre				Long term			
Lat	Long	Year	Mon	SSTav	StDev	SSTval	SDVar
-37.5	77.5	1993	9	-0.28	0.83	6.92	8.67
-17.5	-2.5	1910	10	-0.17	0.83	6.87	8.48
-17.5	-7.5	1910	10	-0.18	0.9	6.92	7.89
52.5	-2.5	1941	6	-0.04	0.97	7.28	7.55
42.5	2.5	1919	2	-0.31	0.57	3.99	7.54
22.5	-77.5	1859	4	0.01	0.72	5.44	7.54
2.5	-32.5	1869	10	-0.09	0.62	4.55	7.48
22.5	-82.5	1859	4	0	0.69	5.16	7.48
27.5	-67.5	1878	10	-0.03	0.72	-6.43	-8.89
22.5	-67.5	1851	8	-0.07	0.7	-6.29	-8.89
22.5	-72.5	1918	8	-0.18	0.75	-6.91	-8.97
2.5	-32.5	1878	12	-0.09	0.67	-6.14	-9.03
7.5	97.5	1851	3	-0.08	0.82	-7.57	-9.13
7.5	92.5	1851	3	-0.15	0.76	-7.34	-9.46
2.5	107.5	1858	6	-0.07	0.71	-6.82	-9.51
17.5	42.5	1940	8	-0.02	0.75	-7.35	-9.77

Table 5-6 Extreme outliers when long-term average SSTs and standard deviations are calculated from data for all cell-months, subject to a minimum of 14 SST values. Column 'SDVar' indicates the number of standard deviations (StDev) that SSTval is from the long-term average (SSTav).

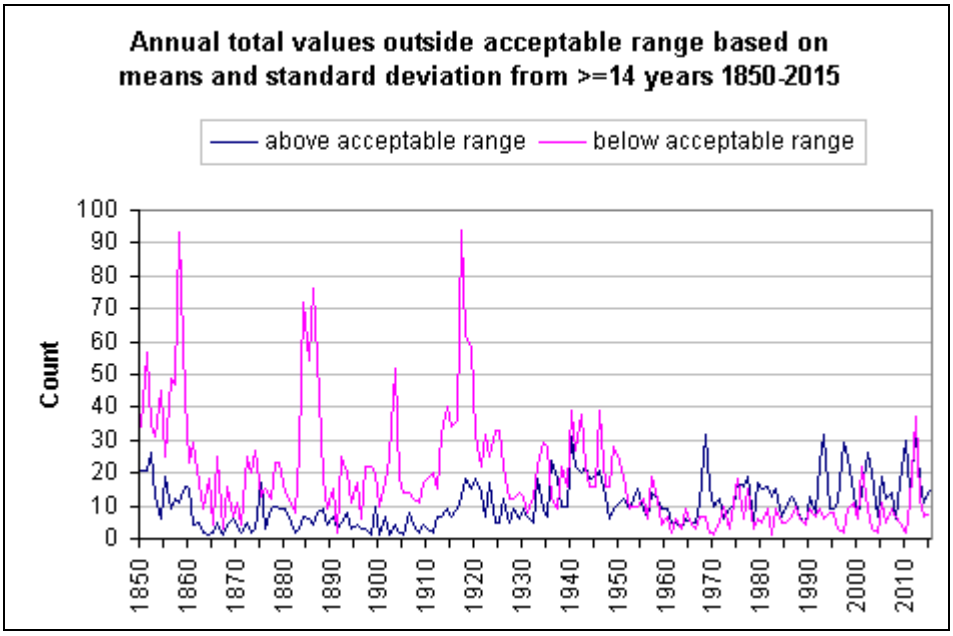


Figure 5.7 Total SST outliers per year when the mean and standard deviation are based on all available data (i.e. 1850 to 2015).

Finally, a critical point about any SST outliers is the magnitude of their variation from the long-term average for the grid-cell and month. Table 5-7 shows some extreme temperature anomalies from the HadSST3 dataset of up to 8.32C from the mean values for the cell and calendar month. In some instances these might be due to abnormal weather and very few SST measurements in the month but in others it might be a problem with the source data (chapter 11).

Cell centre				
Lat	Long	Year	Mon	SSTval
47.5	12.5	2002	1	-8.1
-22.5	-17.5	1957	4	-8.09
-7.5	177.5	1986	10	-8.09
-32.5	132.5	1993	6	-8.01
-47.5	162.5	1976	6	-7.85
-62.5	7.5	1997	4	7.87
-47.5	-12.5	1858	10	7.92
-42.5	42.5	1970	6	7.94
-42.5	127.5	1976	7	7.96
-7.5	-87.5	1919	9	8.32

Table 5-7 Extreme HadSST3 temperature anomalies (calculated according to 1961-1990 averages).

5.5 HadSST3 and CRUTEM4 global averages differ

In was shown in Chapter 1 that the annual average HadSST3 sea surface temperature and the annual average CRUTEM4, from observation stations over land, differ. HadSST3 averages are greater than CRUTEM4 averages until about year 1900, very similar from 1900 to about 1990 and thereafter being less than CRUTEM4 averages. The differences in the annual averages are shown in Figure 5.8 and whether the pattern in differences is correct or a consequence of failings with the data needs to be explored.

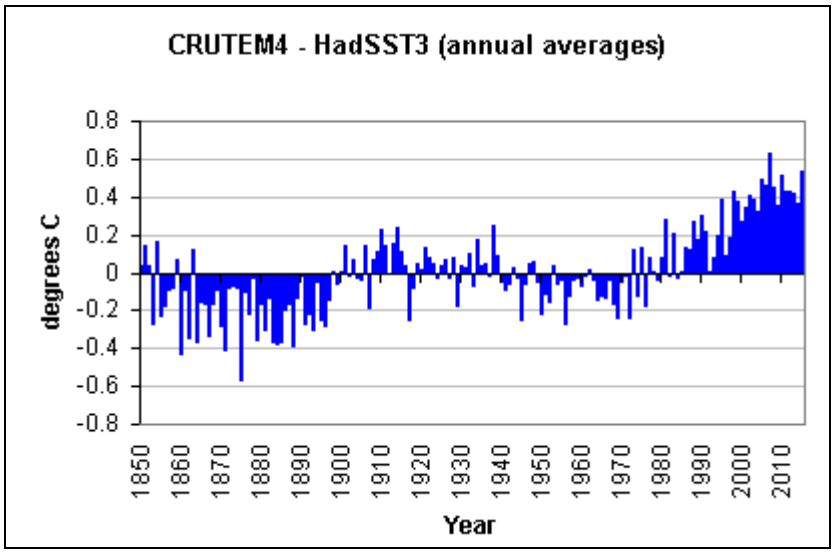


Figure 5.8 Annual average differences between CRUTEM4 and HadSST3 global averages

When viewed by calendar month, with 1850 to 2015 data for each month, a different picture emerges (Figure 5.9). The difference in annual average temperatures changes over time, largely as a consequence of the changing pattern in differences during the six months spanning the Northern Hemisphere winter; the difference in summer temperatures being less.

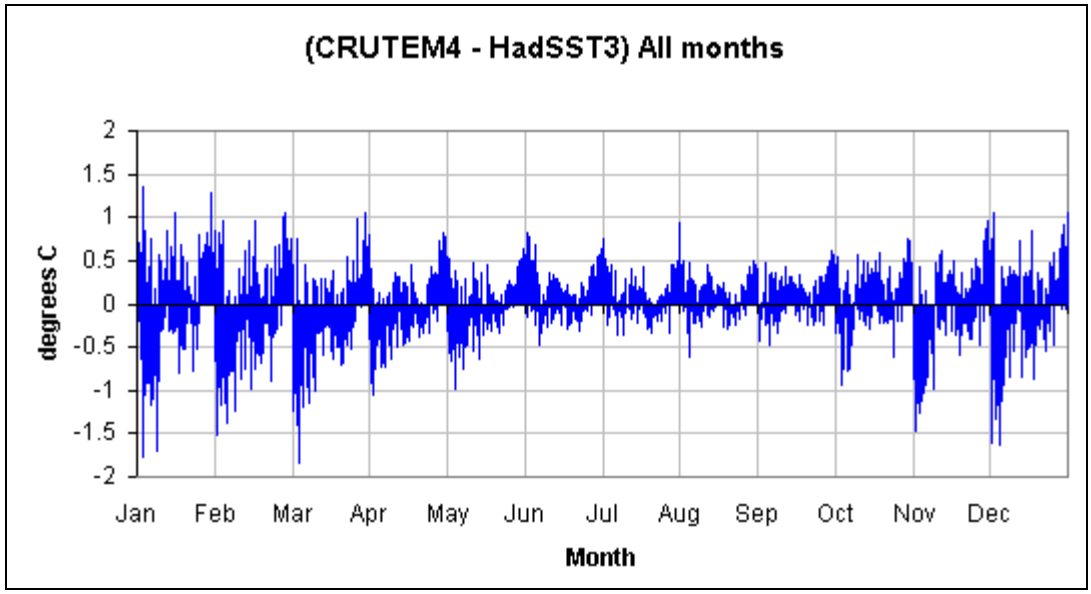


Figure 5.9 Differences between CRUTEM4 and HadSST3 monthly global averages, sorted into calendar months

The situation is made clearer by considering the difference between CRUTEM4 and HadSST3 monthly averages separately for the two hemispheres. (Note that the HadSST3 global average is the average of the two hemispheres but the CRUTEM4 global average uses a 2:1 weighting for the NH over the SH, which means the data is only generally comparable to the global averages.) Figure 5.10 shows the Northern Hemisphere monthly differences for January to March and six months later, for July to September, at the same scale.

The standard deviations for the differences during these months is of limited value because of the obvious trends but the average standard deviations in the Northern Hemisphere for the months of January to March are 0.728°C and for July to September 0.243°C, while in the Southern Hemisphere the differences vary less across the year and the corresponding average standard deviations are 0.342°C and 0.327°C.

The differences in global averages can also be examined in terms of the number of instances where CRUTEM4 averages are greater than or equal to HadSST3 averages and number where it is less, for each month of the year (Figure 5.11). While we might expect approximately equal numbers, varying almost randomly either side of 50% we find that from July to October instances of CRUTEM4 averages being less than or equal to HadSST3 averages range from 33.7% to 40.1%. In contrast only one month of the rest of the year has less than 40% of instances the opposite way, March 36.1%, largely refuting any notion that the variation in months might be due to seasonal differences in temperature and the temperature buffering of the oceans.

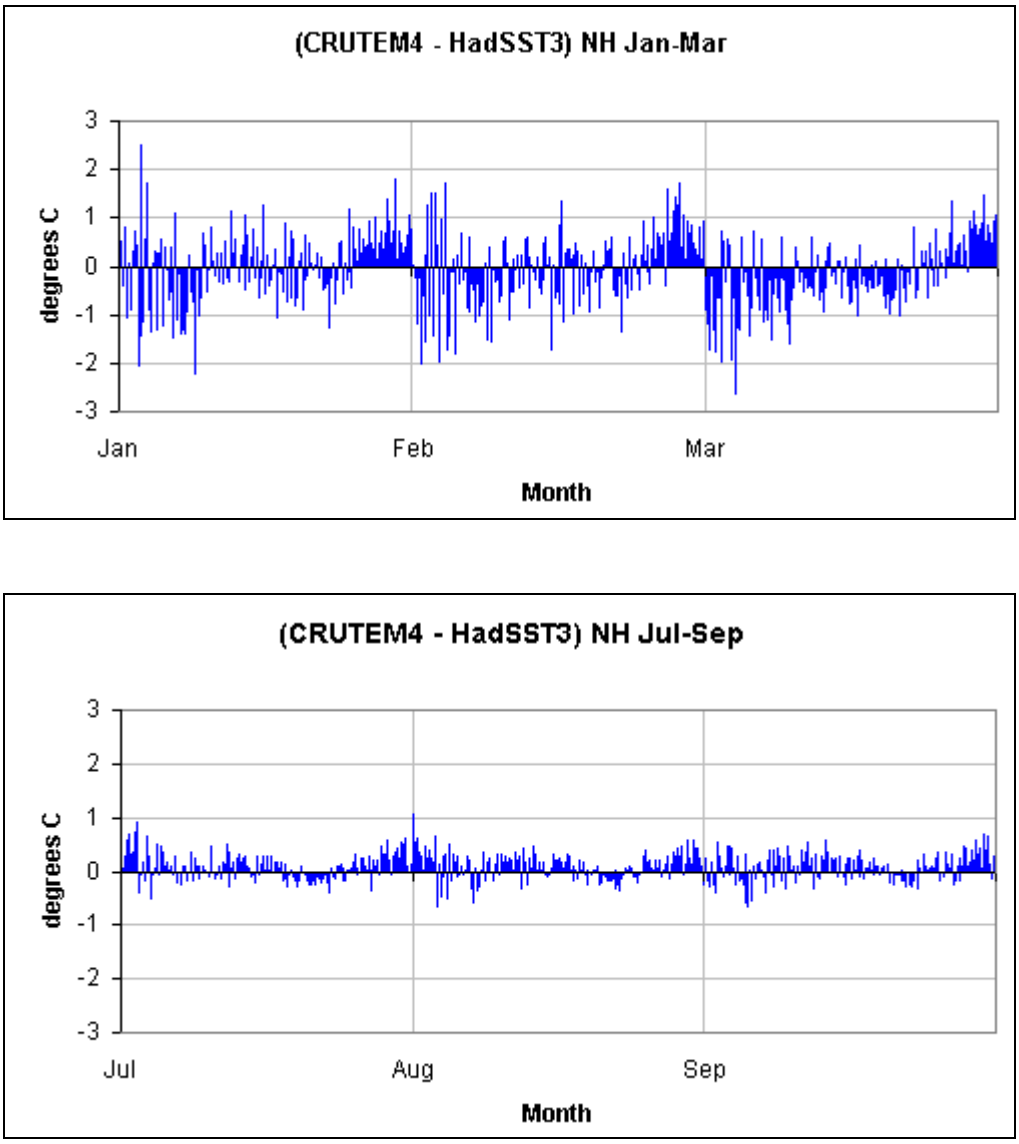


Figure 5.10 Monthly differences in CRUTEM4 and HadSST3 averages in the Northern Hemisphere during two periods of the year (top: January-March, bottom: July-September).

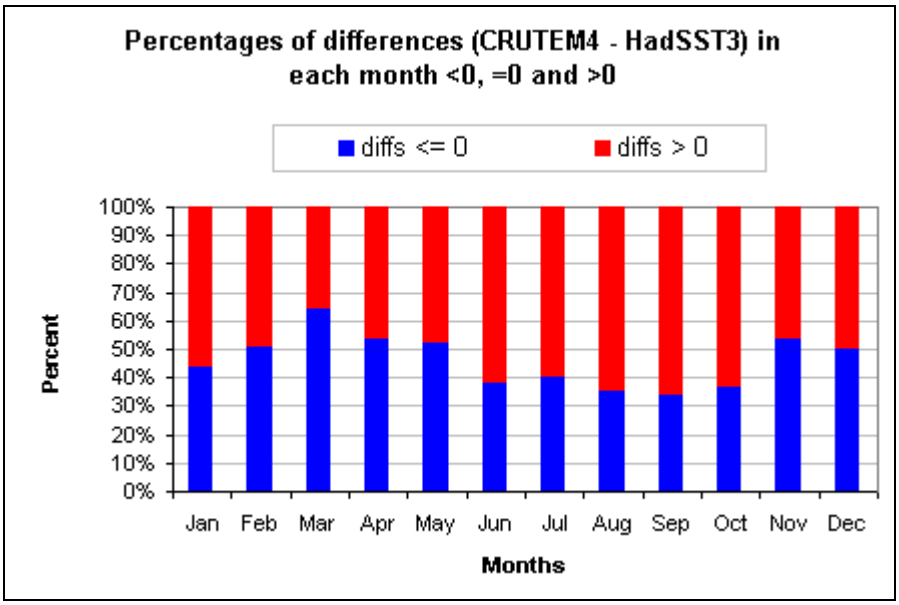


Figure 5.11 Per calendar month percentages of the 166 years in which the CRUTEM4 average was less than or equal to HadSST3 ("diffs <=0") and in which it was greater ("diffs > 0").

Jones (2016) comments more than once in relation to these differences between CRUTEM4 and HadSST3, saying:

> "Related to this, adjustments for land data are estimated completely independently from the marine series, so these two components mutually support each other." [section 3 of that paper]
>
> "If the [SST] adjustments were not applied then century-timescale warming would be greater, and there would be a major discrepancy between the land and marine components prior to about 1940." [section 4.1]
>
> "If the latter [(ie. SST records)] had not been adjusted for the large bias due to the change from bucket measurements, then the agreement with the land record would not have been produced". [section 6].

The second and third comments seem to contradict the first, the both seeming to justify the adjustment of data on the grounds that the CRUTEM4 and HadSST3 values are brought closer together, this despite them being quite different at certain stages of the 166 year record.

The differences in the record during the NH winter months might have more than one plausible explanation and those explanations need not address both the start and end of the record simultaneously. As chapter 2 mentioned, Western Europe contributed more than 50% CRUTEM4 coverage for the Northern Hemisphere from 1850 to 1869, decreasing to 30% (almost twice the land area as a percentage of the total hemisphere land area) by about 1890. Further, at this time Western Europe was still emerging from the Little Ice Age that ended around 1850, which means that CRUTEM4 temperature anomalies at that time will be less than in later years.

One plausible explanation for the CRUTEM4 global averages exceeding those of HadSST3 since around 1990 is that the data from many parts of the NH might be corrupted by urbanisation, particular locally generated heat during the colder months. Another explanation is changes in cloud cover (McLean, 2014). The important difference between these two explanations is that the first is not a meteorological cause but the second is.

Another possible explanation is that the CRUTEM4 data is correct but the SST data in cooler months is too high in the last half of the 1800s and too low since around 1990, but this would require distinct seasonal effects that have changed over time.

It might also be that both CRUTEM4 and HadSST3 are correct and that conditions genuinely have varied as illustrated, but evidence would be needed to support such an assertion. Until such evidence is presented the accuracy of at least one of CRUTEM4 and HadSST3 is in doubt.

5.6 The averaging of station temperature anomalies

The CRUTEM4 practice is to average the temperature anomalies for all reporting stations in the same grid cell in order to obtain the grid cell value for that month. The failure of individual stations to report in some month, either for short-term reasons such as instrument failure or over a long-term (e.g. ceasing operation), will have an

impact on the average for the grid cell according to the range of station anomalies and generally inversely according to the number of stations within the grid cell.

The range of temperature anomalies in each grid cell with from two to 10 reporting observation stations in the month in question was analysed. This revealed 8968 instances with a range of temperature anomalies greater than or equal to 4.0°C. Among these were 4225 instances where the range was greater than or equal to 5.0°C and the number of instances for different numbers of reporting stations is shown in Figure 5.12.

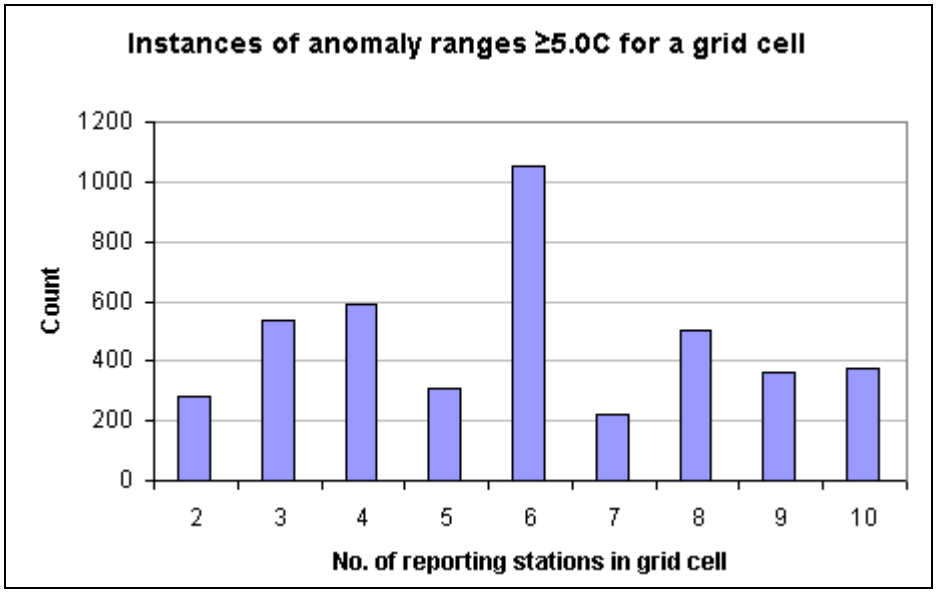


Figure 5.12 Number of instances of grid cell temperature anomaly ranges ≥5.0°C for grid cells containing different numbers of reporting stations

Figure 5.13 maps the grid cells that at some time in the record had instances of the range of temperature anomalies exceeding 5.0°C. Chapter 4 contained discussion of the general inverse relationship between mean temperature and standard deviation so the large range of temperature anomalies in northern Canada and much of Russia is no surprise. The same cannot be said about the mapped grid cells in the tropics where temperatures are consistently warm and the daily and annual range of temperatures is likely to be low.

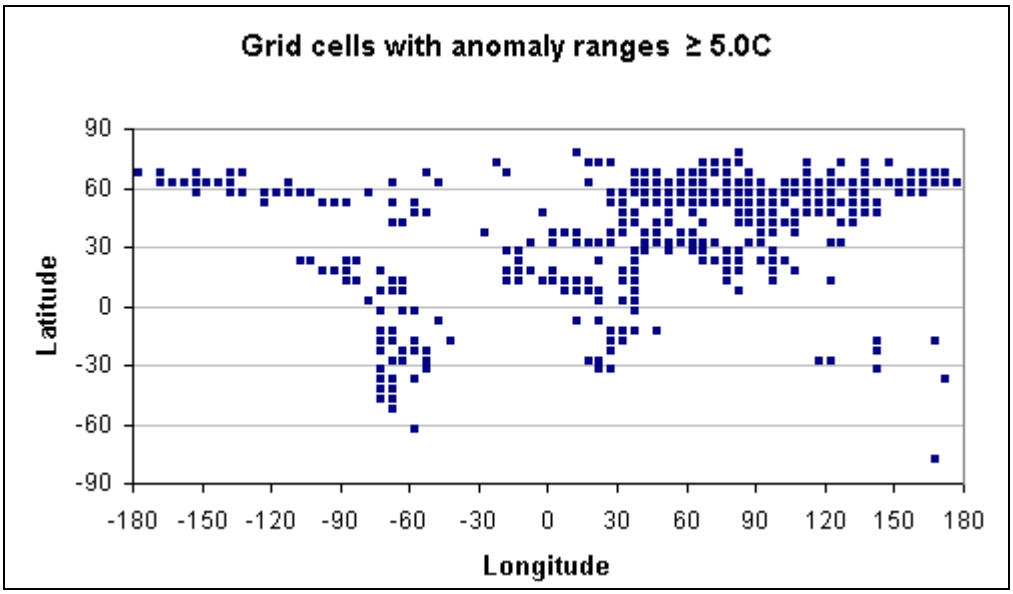


Figure 5.13 Map of the locations of CRUTEM4 grid cells where anomaly ranges in one or more months were ≥5.0°C. (Americas to the left, Africa in the centre and Asia to the right)

The 20 highest anomaly ranges are shown in Table 5-8, with many of the highest values in the list probably the result of the presence of outliers in the temperature data for one or more stations in the given grid cell.

Of particular interest are instances of large temperature anomaly ranges but few observation stations within the grid cell because the omission of a single station from a grid cell in that month could have a large impact on the grid cell value. Table 5-9 lists the 15 greatest ranges of temperature anomalies for grid cells with two or three reporting observation stations. (In the case of just two observations the absence of one station would increase or decrease the grid cell average by 50% of the range of anomalies. With three reporting observation stations the shift depends on the actual temperature anomalies but dividing the range by three gives the average magnitude of the shift.)

The presence of grid cells with large ranges in the temperature anomalies of the stations those cells contain firstly suggests that outlying data has been included, secondly raises questions about the accuracy of the mean temperatures for each station and thirdly suggests that either meteorological conditions are not homogenous

across individual grid cells or if they homogenous then the consequence change in temperature when the conditions change are not.

Extreme ranges of anomalies have been discussed here but any changes in the number of reporting stations in each grid cell will have some impact on the calculated grid cell average that appears in CRUTEM4 and HadCRUT4 datasets. Changes in the number of reporting stations in a grid cell might therefore result in temperature anomaly shifts or even cause false trends in those anomalies.

Cell centre					
Latitude	Longitude	Range	Year	Month	Cell Stns
-17.5	-67.5	80.6	2011	9	9
12.5	32.5	68.1	2011	4	6
7.5	-72.5	55.9	1978	6	5
7.5	-72.5	55.7	1978	7	5
7.5	-72.5	54.2	1978	4	5
-12.5	47.5	40.3	2013	11	4
-22.5	27.5	33.9	2013	5	9
72.5	127.5	23.8	1931	2	3
57.5	27.5	20.4	1970	12	6
32.5	67.5	20.2	2010	5	4
62.5	102.5	20.0	1951	2	4
42.5	47.5	19.9	2012	2	9
72.5	127.5	19.7	1929	12	3
67.5	62.5	19.2	1997	3	8
57.5	62.5	19.2	1927	1	7
52.5	47.5	19.0	2015	7	7
72.5	112.5	18.7	2014	3	3
47.5	97.5	17.8	1994	11	4
62.5	112.5	17.5	1959	3	3
47.5	92.5	17.4	1977	11	6

Table 5-8 Details of the greatest ranges in temperature anomalies for observation station located within the same grid cell

Cell centre					
Latitude	Longitude	Range	Year	Month	Cell Stns
72.5	127.5	23.8	1931	2	3
72.5	127.5	19.7	1929	12	3
72.5	112.5	18.7	2014	3	3
62.5	112.5	17.5	1959	3	3
67.5	-137.5	17.4	2006	2	2
72.5	127.5	16.9	1931	1	3
67.5	157.5	16.6	1953	11	2
67.5	-152.5	15.8	1998	1	3
72.5	127.5	15.5	1930	12	3
62.5	122.5	14.7	1957	1	3
72.5	127.5	14.6	1930	1	3
32.5	12.5	14.5	2015	7	3
27.5	42.5	13.8	2014	2	2
52.5	67.5	13.7	1909	4	2
27.5	-12.5	13.6	2012	12	2

Table 5-9 Greatest ranges in observation station temperature anomaly for grid cells with 2 or 3 observation stations

5.7 Observation stations assigned to incorrect grid cells

During one of the analyses described above it was noticed that the observation station at Ghanzi (Botswana) was, according to the CRU's published data, located within the same grid cell as an observation station in southern Libya. This obvious error was traced to Ghanzi being shown with positive value for its latitude, placing it in the Northern Hemisphere rather than in its correct location in the Southern Hemisphere.

It was also discovered that the given latitude for the observation station at Garissa (Kenya) is 0.5 (i.e. 0.5°N) when -0.5 (i.e. 0.5°S) is much closer to the town of Garissa and would perhaps refer to the observation station at Garissa airport. The distance between the two locations is only 125km but 0.5°N places Garissa in one grid cell and 0.5°S places it in another.

These appear to be the only likely errors with station locations but the fact that they occur at all indicates a failure to check the data.

5.8 Summary

This chapter has discussed several problems with HadCRUT4 data ranging from inconsistencies to outright errors.

Discrepancies were shown to exist between data in the HadCRUT4 dataset and the data in the CRUTEM4 and HadSST3 datasets, sometimes by excessive difference, sometimes with the HadCRUT4 data outside the range defined by the values from the two other datasets and sometimes where the HadCRUT4 dataset contained a value for certain grid cells in certain months but no data existed in either of the other two datasets. An important conclusion in regard to the dominance of discrepancies with the CRUTEM4 dataset is that it appears that different processing is applied to observation station data to create the CRUTEM4 and HadCRUT4 datasets.

It was also shown that data for individual HadCRUT4 "coastal" grid cells often switches between near-surface temperature data from observation stations, sea surface temperatures and a merging of the two. It was also shown that the percentage of coverage from coastal grid cells using each of these sources has varied over time, making the whole situation very inconsistent. It appears likely these inconsistencies could cause shifts in the patterns of anomalies for individual grid cells, or in bulk, shifts in HadCRUT4 hemispheric or global averages, as well as generate false or exaggerated trends.

The presence of data outliers was revisited but this time outside the period over which long-term average temperature are calculated and, in the case of data from observation stations outside the different period over which standard deviations are calculated. It was shown that while some values were likely identified as outliers and either corrected or excluded (no clarification provided by the CRU) it appears that some outliers would not have been identified as such and would have been included in the processing to create the CRUTEM4 and HadCRUT4 datasets. It was also found that a

change in period over which long-term average temperatures and standard deviations are calculated can lead to changes in the data identified as outliers, even more so when the very generous thresholds for outliers in both CRUTEM4 data and the HadSST3 dataset are trimmed to less generous limits.

Outliers were also identified in HadSST3 sea surface temperature data, with changes in the period over which long-term averages and standard deviations again making a significant difference to the data identified as outliers. Admittedly the data used in the analysis is HadSST3 data rather than data, at a smaller grid cell size and shorter time period, from which the HadSST3 data is derived but it seems likely that the problem will be similar at that level.

The presence of any outliers in either CRUTEM4 station data or HadSST3 sea surface temperature anomaly data indicates poor quality control both at the sources of the data and by the CRU and Hadley Centre for the respective datasets. It also suggests that monthly HadSST3 temperature anomalies for some grid cells might be calculated from very few SST measurements and that in abnormal weather this results in abnormal outlying values.

The next section of the chapter showed that global average temperature anomalies for the CRUTEM4 and HadSST3 datasets differ inconsistently from each other. The difference between the annual global averages of the two datasets appears to be due to changes in the relationship between the two datasets during the northern hemisphere winter. It is suggested that the lower CRUTEM4 average temperature anomalies during the late 1800s was due to Europe, the dominant source of such data at the time, coming out of the Little Ice Age and still suffering from very cold winters.

It was also shown that the range of temperature anomalies for observation stations in the same grid cell could at times be very wide. This implies that changes in the number of reporting stations within a grid cell could have a substantial impact on the average of the station temperature anomalies, which is to say the grid cell value. It appears possible that changes in the number of reporting stations in a grid cell might cause sudden shift in the temperature anomalies and therefore cause distorted trends.

Finally it was shown that at least one and possibly two observation stations have incorrect latitudes and have been assigned to incorrect grid cells. As with the presence of data outliers and extreme ranges of temperature anomalies within individual grid cells the errors in location the two stations raises doubts about whether any data quality control is undertaken either by the sources of the data or by the CRU and Hadley Centre before the CRUTEM4, HadSST3 and HadCRUT4 datasets are created.

As with concerns raised in earlier chapters individually these likely errors will probably have little impact on HadCRUT4 hemispheric and global averages save perhaps for inconsistencies with coastal grid cells. The situation could be different when focus is narrowed onto a smaller number of grid cells or when the number of grid cells with errors increases.

It might have been useful to attempt to correct the various errors and inconsistencies shown in this chapter and to recalculate the CRUTEM4, HadSST3 or HadCRUT4 datasets. This task would be very time consuming and perhaps not possible without complete data because it would require

(a) Correcting station locations,
(b) Ensuring that HadCRUT4 data was consistent with its source when only one of CRUTEM4 and HadSST3 provided data for the given grid cell and month
(c) Resolving inconsistencies with data sources for coastal grid cells,
(d) Determining and excluding data outliers for each observation station and for each HadSST3 grid cell,
(e) Determining appropriate limits to the ranges temperature anomalies within the same grid cell then investigating and correcting any cause of extreme ranges,
(f) Resolving the reason for the difference between CRUTEM4 and HadSST3 data and making the appropriate data adjustments

Any corrections would assume that the source data used by the CRU and the Hadley Centre is generally accurate apart from the occasional outlier. The next two chapters will show that such an assumption could easily be false.

Chapter 6: Observation station data prior to CRUTEM4 processing

6.1 Introduction

The Climatic Research Unit (CRU) creates the CRUTEM4 dataset from data submitted to it by national meteorological services (e.g. UK's Met Office, Australia's Bureau of Meteorology) and errors, uncertainties and inconsistencies in the data prior to its submission should not be ignored.

The issues can be broadly classified in to two groups, those related to the measuring of temperature and those related to subsequent explicit modification or other processing of the data. This chapter will discuss these issues with less emphasis on the first group of issues, because such problems are likely to be random and less common, than on the second group of issues.

The term "homogenisation" appears frequently in this chapter. It is the deliberate adjustment of temperature measurement to make them supposedly consistent with having been measured at the same location at the same time of day. It is a two-step process that first requires the identification of inhomogeneities and a second step to adjust them. The aim of the process is to remove all non-meteorological influences from the data record (e.g. the relocation of an observation station). The data is almost always homogenised to correspond to the latest location of the observation station but if the latest location has been in use for less than a few years its data might be tentatively homogenised to the previous location pending greater confidence in understanding the relationship of the data at the old and new locations.

This chapter will discuss issues related to the accuracy of data prior to its use in the construction of the CRUTEM4 and HadCRUT4 dataset. At the outset it has to be said that observation stations were primarily intended only to provide information about local weather and its patterns for purposes such as comparisons and predictions; they

were not intended to provide climate information and to accurately represent conditions over much larger regions.

6.2 How HadCRUT4 creators see the issues

To provide some context to the discussion of the issues upstream of HadCRUT4 it might be useful to understand how the creators of the HadCRUT4 dataset see the problem of data uncertainty. This is a relatively new issue for the CRU and the HadCRUT4 dataset because in previous versions of the dataset either the CRU or the GHCN (Global Historical Climate Network) homogenised the data, perhaps under guidance from national meteorological services, but now the adjustments are the responsibility of those services.

A key reference for the perspective of the creators of the HadCRUT4 dataset is Morice et al (2012), which draws heavily on the discussion by Brohan et al (2006) of the uncertainty of the earlier HadCRUT3 data which states "Uncertainties in the land data can be divided into three groups: (1) station error, the uncertainty of individual station anomalies; (2) sampling error, the uncertainty in a grid box mean caused by estimating the mean from a small number of point values; and (3) bias error, the uncertainty in large-scale temperatures caused by systematic changes in measurement methods."

Brohan et al (2006) went on to describe the station mean monthly temperature at a given stations as follows:

$$T_{actual} = T_{ob} + \varepsilon_{ob} + C_H + \varepsilon_H + \varepsilon_{RC}$$

where T_{actual} is the actual station mean monthly temperature that becomes part of the historical record, T_{ob} is the reported temperature, ε_{ob} is the measurement error, C_H is any homogenisation adjustment that may have been applied to the reported temperature and ε_H is the uncertainty in that adjustment, and ε_{RC} is the uncertainty due to inaccurate reporting or calculation of the station mean.

Morice et al (2012) revised this by talking in terms of a "true monthly average temperature" and described the situation as:

$$T_{obs} + C_H = T_{true} + \varepsilon_{obs} + \varepsilon_H$$

where T_{obs} is the observed temperature, C_H is a homogenisation correction applied to remove inhomogeneities in the station record, T_{true} is the true monthly average temperature, ε_{obs} is a random measurement error and ε_H is the error in the applied homogenisation correction.

Morice et al (2012) includes no recognition of Brohan et al's "inaccurate reporting or calculation of the station mean" but states an important qualifier about its aims and discussion:

> "Ideally, the above equation would also include the effects of urbanization and changing sensor exposure, arising from changes in enclosures used to shield thermometers from the elements. These terms are omitted at this stage as the urbanization and sensor exposure models used here are based on studies of the influence of these factors on regional averages, and the derived biases may not be representative of the influence of these factors on individual station records. These factors are instead applied to gridded temperature anomalies."

This statement appears to be saying that adjustments for urbanisation are not made individually for each station but somehow in a composite form to the datasets' grid cell value. This would be inaccurate and unwise because over time magnitude of the urbanisation influence will differ for each station.

Both Brohan et al (2006) and Morice et al (2012) recognise that temperature normals (i.e. long-term average temperatures calculated across the period 1961-1990) for a given station would themselves have uncertainties and that because temperature anomalies are calculated by subtracting the normal for the calendar month from the

mean value in the current month, the uncertainties in each must be combined. Morice et al (2012) describes the true station temperature anomaly T_a as

$$T_a = T_{obs} + C_H - \varepsilon_{obs} - \varepsilon_H - (T_N + \varepsilon_N)$$

where T_N is the estimated station normal and ε_N is "the error in this estimate arising from measurement error and the computation of normal temperatures from a finite number of years of data" and other values as described above.

Morice et al (2012) goes on to say that because urbanisation and exposure biases "represent the possible influence of these factors on regional averages rather than individual stations" these are also omitted from the published gridded data file of uncertainties.

Both Brohan et al (2006) and Morice et al (2012) have two major flaws. Firstly only the available data can be used to generate any statistical error margin and neither the calculated mean nor the error margin can anticipate the situation if more data were available. Secondly the accurate quantification of the error margins in the equations given above would be very arduous given that monthly data is used and while the given equations might be theoretically correct they do little to advance the accuracy of the data save perhaps for a minor refinement of large scale averages.

6.3 A brief review of issues with observation station data

As the introduction to this chapter noted, the issues with observation station data can be grouped into issues with the measuring of temperature and those related to subsequent explicit modification or other processing of the data. This section presents a brief review of these issues.

The World Meteorological Organization (WMO) is the primary global authority on meteorological practices and publishes standards for the recording of temperature observations and any subsequent processing of the data. WMO 8 (WMO, 2010)

describes the instruments that should be used, the enclosures (i.e. screening) in which they should be housed and where, in general terms, the enclosures should be sited. In recognition that locations might not always be ideal it defines four siting classes, explicitly stating of the less ideal sites that temperatures recorded at a class 3 site have "additional estimated uncertainty added by siting of 1°C", at a class 4 site an uncertainty of 2°C and at a class 5 site an uncertainty of 5°C (see Appendix 1).

On the subject of the enclosure it points out that good airflow is essential because without that airflow the difference between the temperature of the outer walls of the enclosure might be "markedly different from the air temperature", "perhaps reaching 2.5K and -0.5K respectively in extreme cases."

No information is provided as to whether the observation stations fully comply with WMO standards and to which siting class they belong, but as the above suggest, these issues are important as regards the uncertainty of the data.

In discussing potential issues with the measurement of temperature WMO 8 (WMP, 2010) lists seven possible errors with all liquid-in-glass and a major potential error with the spirit thermometers that are used in very cold conditions when mercury might freeze. It also notes that electronic thermometers can produce false recordings when their power supply fluctuates and that they have a faster reaction time to temperature changes, which might mean that they record maximum or minimum temperatures that slower reacting liquid thermometers do not.

Folland et al (2001) say "... estimated the two standard error (2σ) measurement error to be 0.4°C in any single daily [land surface-air temperature] observation". Brohan et al (2006) claimed that this meant the "random error in a single thermometer reading is about 0.2°C (1σ)" having wrongly reduced both the number of observations and the number of standard deviations. Frank (2010) argues instead, on the basis of parallax issues when reading thermometers and issues with instrument manufacture, use and degradation, that the lower limit instrument uncertainty is 0.359°C.

The WMO statements above show the potential for errors in the measurement of temperature and the references given above show that the matter of instrument

uncertainty is unsettled. Further, the duration of the potential errors described above extends from a single instrument reading to the entire set of recordings made by an individual observer or individual instrument, so the impact of these uncertainties is very uncertain.

Another issue with temperature observations concerns the time at which the meteorological day starts and ends. WMO 8 (WMO, 2010) states that the change of day should occur between 7:00am and 9:00am with the maximum temperature in the previous 24 hours assigned to the previous day and the minimum temperature for the previous 24 hours assigned to the current day. Most sites appear to use 9:00am but this can lead to the minimum temperatures on successive meteorological days being made within a few hours of each other (Figure 6.1)

A common practice in the USA until; the 1970s was to end the meteorological day at 3:00pm or 4:00pm local time but doing so at this time risks maximum temperatures recorded for successive meteorological days occurring just a few hours apart.

This "time-of-day" bias is even more likely when summer time (also known as daylight-saving time) is in operation or in the western side of time zones, both situations having local time ahead of true solar time for the location in question.

The monthly mean temperature for a given observation station is the average of the mean daily minimum temperature and the mean daily maximum temperature, so a 9:00am observation time (and therefore change of meteorological day) will risk lower mean daily minimum temperatures and therefore lower monthly mean temperature, while a 4:00pm change of day will risk higher mean daily maximum temperature and therefore higher mean monthly temperature.

Temperature data has often been adjusted from a different observation time to bring it in line with the current practice of a 9:00am observation time. The accuracy of such adjustments is open to question especially when observations were made once per day compared to the later use of electronic thermometers that almost constantly log temperature (although data is usually presented at 30-minute intervals) which should

provide better guidance when resetting minimum and maximum temperatures relative to a 9:00am observation time.

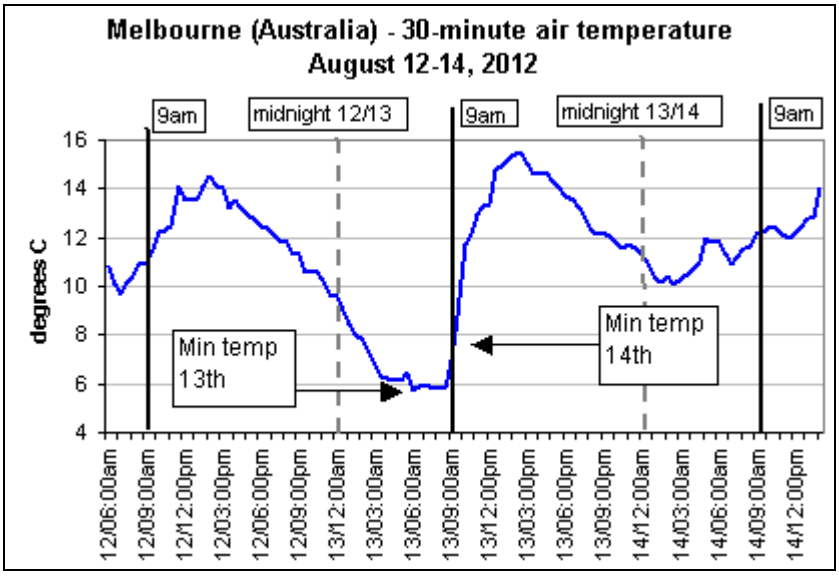


Figure 6.1 Example of where the minimum temperature for the 24 hours ending at 9:00am was recorded only a short time prior to the minimum temperature for the next 24 hours. (Solid vertical lines indicate 9:00am, the meteorological change of day, and the broken vertical lines indicate clock-time midnight.)

Siting and exposure are major concerns for the accurate measurement of temperatures at any given point in time in regard to whether these factors change over time (e.g. annual cycles, vegetation growth.). Local shading caused by natural or man-made obstacles, seasonal changes in vegetation, changes in surface vegetation due to rainfall, variations in prevailing winds and changing exposure to those winds will all influence the recorded temperature. Early morning shading is a significant issue and the WMO standard mentioned above (WMO, 2010) pays particular attention to near horizon shading when defining the five siting classes, three of which have explicit data uncertainties due to siting. Fall et al (2011) reports on the siting of 82.5% of US observation stations and finds that 64.4% of these fall into Climate Reference Network (CRN) class 4 and 6.2% fall into CRN class 5, these classes being very similar to the WMO classes 4 and 5 mentioned above and with uncertainties due to siting of 2°C and 5°C respectively.

Urbanisation influences on observation stations encompass siting and exposure issues and add to those impacts of the properties of the manmade environment that might be different to the natural environment (e.g. heat retention and release, reflection, disposal of surface water) as well as the matter of locally generated heat. Because attempts to adjust data for homogenisation are common it will be discussed in more detail shortly.

Finally in the sequence of measuring, adjusting and preparing data for inclusion in the CRUTEM4 and HadCRUT4 datasets is the calculation of the monthly mean temperature. Neither Jones et al (2012) nor Osborn & Jones (2014) specify the minimum acceptable number of days of data in order that the monthly mean value is acceptable. WMO 100 (2011) says "[i]t is recommended that a monthly value should not be calculated if more than ten daily values are missing or five or more consecutive daily values are missing." In terms of total days this limit is approximately one-third of all days in the month and slightly greater in February. As for compliance with WMO 8 (WMO, 2010) regards siting, enclosure and instruments we have no information regarding the compliance of stations.

6.4 Introduction to temperature data homogenisation

Homogenisation involves the identification and correction of short and long-term errors, or corrections for inconsistencies in how the data was obtained and recorded, in short any inhomogeneities in the data record, and the conversion of all data to be equivalent (at least theoretically) to the data that would have been recorded had the current instruments recorded data at the current observation site for the entire period of the data record (Morice et al, 2011). A more concise definition is that it seeks to remove non-meteorological influences from the temperature record, influences that might be abrupt (e.g. nearby construction, vegetation removed) or gradual over time (e.g. growing vegetation, urbanisation).

Given that the adjustments usually seek to make the data consistent with recording at the current location, screening and instruments the entire set of previous data will be

adjusted as necessary, meaning that early portions of the data record might have been adjusted multiple times.

Any individual errors in homogenisation will therefore be carried forward and any systemic errors be compounded throughout the record. Also important is that data for the period 1961-1990 might be homogenised and therefore impact the long-term average temperatures used to calculate temperature anomalies.

Aguilar et al (2012) adds to earlier work by Peterson et al (1998) to produce a list of 14 methods of identifying potential inhomogeneities in recorded temperature data, some in terms of consistency at a single station and others by comparison with data from other stations or sources (see Appendix 2). Aguilar et al (2012) also notes that one problem with homogenisation is that "... on most occasions the magnitude of the inhomogeneities is the same or even smaller than that of true climate-related variations", meaning that it can be difficult to determine natural variation from those with a non-meteorological cause.

6.5 Issues with homogenisation that uses the Standard Normal Homogeneity Test

6.5.1 Introduction

A popular approach to homogenising data uses the Standard Normal Homogeneity Test (SHNT) that considers the relationship between the data at a target observation station and nearby other stations collectively called for the purposes of the exercise "reference stations".

Menne and Williams (2009) is a key source for information about the technique although it draws heavily on Alexandersson and Moberg (1997), which in turn draws on Alexandersson (1986).

The SNHT technique relies on identifying changes in the relationship between the temperature data at one observation station - commonly called the target station - and the temperature data from one or more neighbouring reference stations (also known as "comparison stations"). The underlying fundamental assumption is that the influence of weather systems (e.g. high and low pressure cells) on temperatures measured at neighbouring stations will show consistent patterns. Disruptions to the pattern of the relationship between stations are assumed to identify points of inhomogeneity, which are then amended by using the relationship to estimate the temperature(s), that according to the relationship, "should have been" recorded at the target station.

Numerous variations on the basic approach of Menne and Williams (2009) can be found. Some use all neighbouring stations as reference stations but others use only neighbouring stations with strong correlations to the target site (e.g. Hausfather et al, 2013). Some use the method pairwise (i.e. compare the data from the target station to the data each individual reference station) as per Menne and Williams (2009) while others merge the data from those neighbouring stations into a composite reference sequence (e.g. Tuomenvirta, 2002).

Some determine the relationship using the actual recorded temperatures, as described in Menne and Williams (2009), but others determine it using the difference between successive recorded temperatures at each site (e.g. Peterson & Easterling, 1994) or it might be based on the normalised difference (e.g. Alexandersson & Moberg, 1997; Toreti et al, 2010) or even on temperature percentiles (Trewin, 2013).

In some cases the adjustment is determined by the median value of all estimated adjustments (Menne & Williams, 2009) but in others a weighted average (based on the distance to a neighbouring comparison station) is used or in the case of Trewin (2013), the weighted average of temperatures mappings based on percentiles.

In some cases the annual mean minimum and maximum temperatures are compared, but in other cases the monthly mean temperatures might be used, or even daily minimum and maximum temperatures (Trewin, 2013).

Tuomenvirta et al, (2002) describes some areas of concern about the technique of Menne and Williams (2009). Those concerns will be discussed in the following sections, along with issues with the adjustments for the Australian Observations Reference Network - Surface Air Temperature (ACORN-SAT) as described by Trewin (2013). The adjustments for ACORN-SAT data are a variation on the use of the SNHT and the creation and use of a "transformation matrix" (i.e. one-to-one mapping) of temperature percentiles at the reference and target stations.

6.5.2 Assumption of similarity of data from target and reference stations

The method described by Menne and Williams (2009) assumes that nearby locations will show "similar variations in climate", which is to say that the impact of passing weather systems cause very similar fluctuations in temperature at the target station and neighbour stations. (For methods described above that use variations of the approach this means relative consistency in whatever data derivative they use.) This makes large assumptions about the spatial coverage and homogeneity of a given weather system, about the consistency of the physical environment surrounding observation station despite irregular changes (e.g. surface moisture, land-use changes, urbanisation) and about meteorological forces at that station (e.g. wind speed and direction, fog, cloud cover) (Stierou & Koutsoyiannis, 2012).

Trewin (2012) states "*Even today, 23 of the 112 ACORN-SAT locations are 100 kilometres or more from their nearest neighbour, and this number has been greater at times in the past, especially prior to 1950.*" This number is more than 20% of all ACORN-SAT stations and is only for the nearest station. ACORN-SAT comparisons require a minimum of ten reference stations, which implies that stations might be separated by much greater distances, which in turn increases the doubt that that the target station and all reference stations would be equally impacted by the same weather system.

Exposure to conditions is an important part of this. Figure 6.2[17] shows an example of variations in the prevailing winds near reference stations used to homogenise data for a target station at Orbost, Australia. The four sites have very different wind patterns and, given that wind can influence temperature, a certain recorded temperature might be a cooler base temperature plus a warm wind or vice versa or perhaps the wind had no influence.

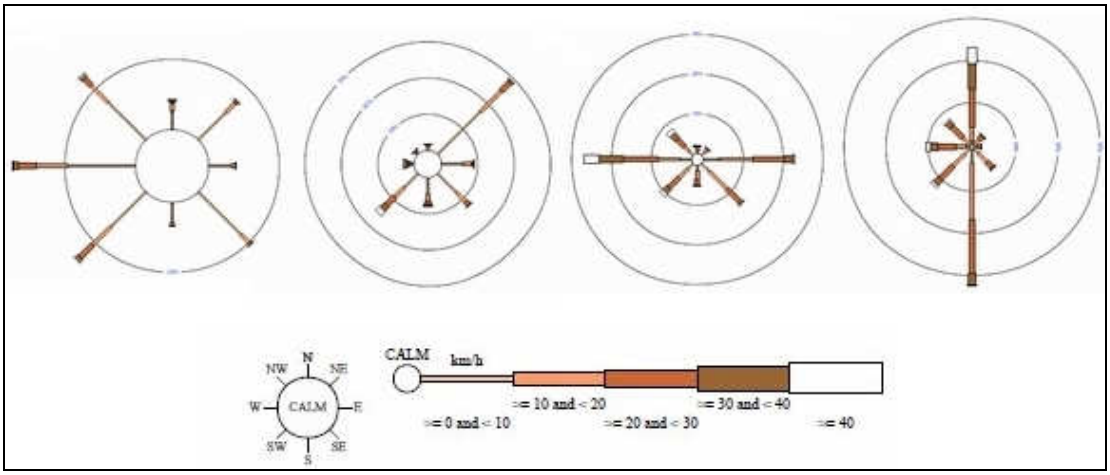


Figure 6.2 Annual average 3pm wind roses at or near weather stations whose data was used to homogenise the daily data at ACORN-SAT station Orbost (Victoria), from left to right Albury, Bega, East Sale and Melbourne. These stations have very different prevailing wind patterns s that might influence the temperature recorded at each of the locations.

The speed at which the weather patterns move through the region covered by the target and comparison stations might also be important. Stierou & Koutsoyiannis (2012) notes that stations located at different longitudes within the same time zone could easily record different minimum temperatures at 9:00am as weather systems move through.

Figure 6.3 shows the difference in time of the recorded minimum temperature across the period 6:00pm to 9:00am at two stations in the ACORN-SAT network, separated by just 20 kilometres. The difference in the time of recording the minimum temperature on the same day varied from over 200 minutes in one direction to almost 400 minutes in the opposite direction, a total range of 10 hours, suggesting that the

[17] Created from data available via http://www.bom.gov.au/climate/averages/wind/selection_map.shtml.

weather was not consistent at the pair of stations despite the separation of only 20 kilometres.

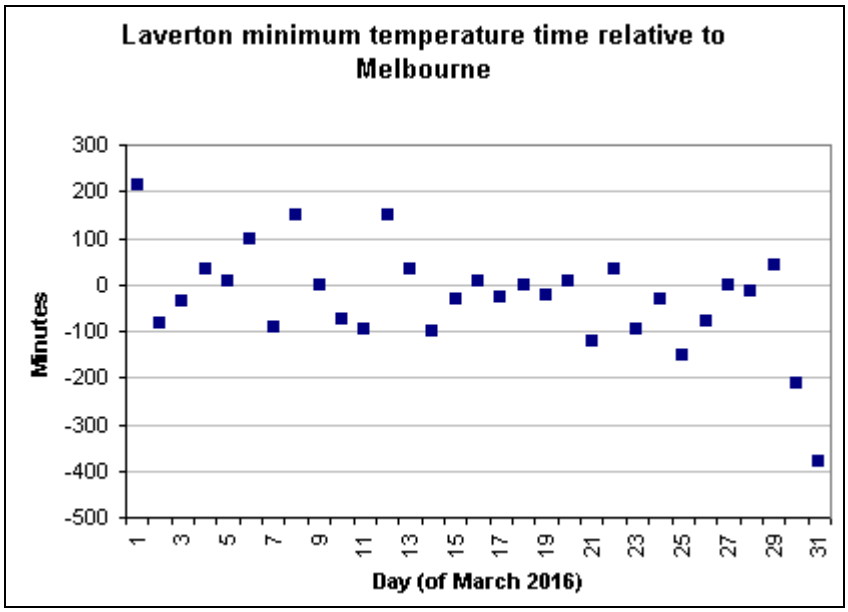


Figure 6.3 Difference in the time at which the minimum temperature was recorded at Laverton and Melbourne (Olympic Park) observations stations, about 20km apart.

6.5.3 Use of data from stations that are poorly sited

It was stated earlier that WMO standards apply five different siting classifications based on a certain set of parameters, with classes 3, 4 and 5 associated with siting uncertainties of 1°C, 2°C and 5°C respectively (see Appendix 1). It would seem unwise to include such observation stations when attempting to homogenise the data from a target site and yet neither Menne and Williams (2009) or Trewin (2012) mention excluding observation stations on these grounds.

One of the observation stations used to homogenise data for the ACORN-SAT station at Orbost (mentioned in 6.5.2) is "Bombala (Therry Street)". The site information provided by the Bureau of Meteorology for this station contains a site map that shows a building "4 to 6m" high located approximately 15 metres from the thermometer enclosure. This station is WMO class 4 site because the enclosure fails to meet the

requirement for class 3 sites that it is "[a]way from all projected shade when the sun is higher than 7°" but does meet the class 4 limit of "[a]way from all projected shade when the sun is higher than 20°". The uncertainty due to siting associated with this site is 2°C, which should have excluded it from consideration.

6.5.4 Data for reference stations likely to already have been adjusted

There is usually no raw temperature sequence from any site that can unequivocally be regarded as homogenous. This results in two issues, the first of which is how the data from the very first station can be adjusted if the data from the available reference stations might contain uncorrected inhomogeneities. The second is that any errors in the adjustment of data at reference stations will potentially impact the identification of inhomogeneities (false positives or false negatives) at the target station and likewise impact the estimated "corrected" value (Tuomenvirta et al, 2002).

6.5.5 Strength of correlation between reference stations and target

The quality of the relationship between the data from the target station and reference stations is measured by its correlation but the recommended threshold of the correlation varies and a very strong correlation cannot be expected if the data at the target site contains inhomogeneities.

In some cases the limits of acceptable correlation are very generous, for example Alexandersson & Moberg (1997) says only that the correlation should be positive, this despite a low correlation indicating a poor match of sites. Tuomenvirta et al (2002) states that a subjective decision will also need to be made as to the priority of using either highly correlated reference stations or stations in the same climatic region that might have lower correlation.

According to Trewin (2012) the ACORN-SAT data homogenisation involved the comparison of daily temperature data at the target location against a temperature estimated from the mean of the 10 nearest neighbouring sites, weighted inversely by distance and by the correlation between the data at the target and comparison sites (excluding neighbouring sites where the correlation coefficient was <= 0). If insufficient stations met the criteria then the distance was expanded with the minimum correlation coefficient set to 0.6. In other words it set a greater priority on correlation than on distance but then weighted the data inversely against the distance (i.e. lower weighting for greater distance), which means that the influence of the reference station with correlation 0.6 or above could be negligible in the weighted mean of the temperatures from multiple neighbours.

6.5.6 Assumption of data showing normal Gaussian distribution

There is a broad assumption that the data from each station - or if used on the basis of data differences, the differences in the data - will conform to a standard normal distribution. Chapter 5 showed that one in eight temperature series for station-month combinations is probably not normally distributed. Even if the data were normally distributed the use of annual, seasonal, monthly or daily data for such comparisons would mean different sample sizes and therefore different error margins and uncertainty (Tuomenvirta et al, 2002).

Related to this is the ACORN-SAT notion of a one-to-one mapping of the data expressed as a transformation matrix and based on temperature percentiles. (No station is going to have a range of temperatures spanning 100°C around the time of the inhomogeneity so it follows that the percentiles are going to be split on the basis of fractions of a degree, this despite data from manual observation stations rarely being recording at intervals of less than one degree because of thermometer scales.)

The fallacy of assuming a normal distribution and one-to-one correspondence will be illustrated shortly in a case study.

6.5.7 The majority of stations rule

As Tuomenvirta et al (2002) expresses it, if the majority of reference stations fall under the same weather influence then the adjustment to data from the target station will have this influence imposed on it whether it be correct or not. These influences might include the extent of urbanisation, exposure to weather from a certain direction, exposure to major weather influences such as ENSO and the extent of soil moisture retention.

Further, if a sufficient number of reference stations undergo a shift (e.g. the painting of an instrument enclosure or a change from manual to automated operation) at about the same time and for an unrelated reason an inhomogeneity occur at the target station the inhomogeneity might not be identified at all if the general relationship is unchanged (Tuomenvirta et al, 2002).

The coincidental changes discussed above need not occur at precisely the same time. It is only necessary that they occur within the same time period being considered and that they be sufficient to cause a shift in the data. For example, when data is averaged across a month the change might occur near the end of the previous month or within the first ten days of the new month (assuming no missing data in the month). The magnitude of the changes might not be the same in both cases but might be sufficiently similar to indicate a shift.

6.5.8 Granularity of input data

Tuomenvirta et al (2002) points out that the use of annual, seasonal or monthly temperature averages would mean different sample sizes for SHNT calculations that might produce different outcomes. A simple example of this is when comparing a coastal observation station, where temperature variation is relatively low, to an inland station that might experience periods of severe cold and severe heat and therefore

have a wider annual range. Comparing annual averages will produce one result but comparing individual months will show an annual cycle that any adjustment should take into account.

It appears that ACORN-SAT adjustments use daily minimum and maximum temperatures when they are available for the target and reference sites. In theory this should reduce error margins but this would require the data to be normally distributed, which might not be true, very similar weather patterns and very similar exposure. (Equally if the daily data is adjusted because it is supposedly correct then there seems to no good reason to continue to operate the station if its data is to be corrected on the basis of data recorded at other stations.)

6.5.9 Possible false positives and negatives with the SNHT

Stierou and Koutsoyiannis (2012) report that the SNHT falsely identified certain data sequences as having shifts and that not all deliberate shifts were identified in other modified sequences. SNHT performance reportedly degrades near the ends of data sequences when the amount of data to analyze is reduced (Toreti et al, 2011). The consequence of false positives is that data is adjusted when it doesn't need to be and of false negatives that adjustments that should have been made are not.

False positives might also arise in situations where the data from comparison stations was weighted, such as by distance, and then conditions at a heavily weighted station underwent major change. The change in the relationship might be falsely flagged as an inhomogeneity at the target site. The usual case is that all comparison stations are checked against each other, the assumption being that the station with the most identified instances of inhomogeneities in the one-to-one comparison at this particular point in time is the station to be corrected (because other stations only record a problem when being compared to the erroneous station). This assumption might however be incorrect because the correlation between stations might be weak or multiple stations might have a simultaneous inhomogeneity (e.g. switch to automatic station).

6.5.10 The subjective interpretation of the SNHT results

The SNHT tests the null hypothesis (that the comparison variable is within the range of natural variation) against the hypothesis that it is not, but the typical test result indicates whether the range *might be* natural rather than *is* natural and therefore either a subjective decision or some form of automated decision is required, but the automated decision might be incorrect for a given situation.

6.5.11 Possible errors in attributing cause of inhomogeneities

Toreti et al (2011) points out that not all inhomogeneities should be regarded as errors that need correcting. It uses the example of a volcanic eruption potentially impacting some observation stations by its downwind plume but not others, and yet such eruptions are recognised natural influence on weather.

In a similar fashion some stations might be more exposed to irregular influences, such as ENSO events, than other stations that their data is compared to. These events might cause inhomogeneities in the data record but they are natural meteorological events.

6.5.12 Errors in determining the magnitude of the adjustment

Menne and Williams (2009) indicate that the adjustment should be calculated by comparing the target location to a minimum of four reference sites, with the adjustment being the median value of the adjustment calculated against each reference station. The median is used so that the influence of outliers in the data will be minimised but there can be no certainty that the media is more accurate than the adjustment to other reference stations, especially those with strong correlation.

As mentioned above, the adjustment system for the ACORN-SAT data applies an inverse weighting by natural log to the distance between stations and puts lower priority on the correlation between data at the target and reference locations. It can be shown that under this system three reference stations at 150km to 200km distance from the target station have greater composite weighting than a single reference station 20km from the target, this despite a closer station being more likely to have similar weather patterns.

6.5.13 Only moderate success with gradual inhomogeneities

The Meene and Williams (2009) reported that their approach to dealing with inhomogeneities that gradually increase over time (e.g. vegetation growth, urbanisation) had a detection rate of about 67% and a false detection rate of 19.7%.

In relation to such inhomogenieties Trewin (2012) says in regard to the ACORN-SAT adjusted data from Australia's Bureau of Meteorology, "*All breakpoints are treated as step changes with no anomalous trend (model M3 of Menne and Williams). This assumption is based on their findings that the method was only moderately effective in reliably identifying more complex breakpoint models (e.g. an anomalous trend superimposed on a breakpoint).*"

Trewin (2012) later discusses a separate check for anomalous trends at locations potentially impacted by urbanisation but adopts a population of 10,000 as the threshold between rural and urban settings. This is unhelpful because the effects of urbanisation have been detected at centres with much lower population and even for roadways and airstrips.

Barrow, Alaska, is one of the better documented instances of urbanisation in a small town, having a population of 4373 (2013 data) and daily average winter temperatures elevated by as much as 6°C (and hourly temperatures sometimes elevated by 9°C) as a result of the Urban Heat Island Effect (Hinkel et al, 2003), this despite its low-density housing, gravel unsealed roads and plentiful open space.

The ACORN-SAT station catalogue specifically mentions that the original site at Deniliquin (Lat: 35.56S, Long: 144.95E) was at the post office, near the centre of the small town, "with the site becoming built up from about 1950 onwards". An adjustment was made from 1 January 1950 with the maximum temperature increased by 0.51°C and no change to the minimum. Ten years later, 1 January 1960, the adjustment to maximum temperature was reversed (i.e. -0.51°C) and again no change to minimum temperature. Following a site move on 13 August 1971, to a site on the outer edge of the town as it was at the time, all previous minimum temperatures were reduced by 1.0°C with an annual average adjustment to maximum temperature of -0.17°C "but with larger adjustments in spring and summer". This site was then replaced with a new station at the local airport, 3.7km away, in September 1984[18].

The three site changes at Deniliquin were due to urbanisation at a town with a population of 7431, which in the terminology of Trewin (2012) means that the site was regarded as "Non-urban".

Another example is the sequence of observation stations at Inverell, NSW, (pop. 9749) about which Trewin (2012) says that the original site, at the post office, "*was very built-up with several buildings within a 10-metre radius*". Despite the "enclosed" (his word) nature of the site the station is described as non-urban in table 8 of Trewin (2012).

The above examples, including that of Barrow (Alaska), show that urbanisation effects can occur in quite small towns and each probably needs to be considered on its own merits if the adjusted data is to be as accurate as it can be.

As noted above, Trewin (2013) says regarding ACORN-SAT adjustments, "*All breakpoints are treated as step changes with no anomalous trend"*. It will be shown later in this chapter that the treating of gradual changes as step changes is a significant error with important consequences for the accuracy and relevance of the temperature record.

[18] Information obtained by the author via interview with local residents, September 2016.

6.6 Homogenisation with site operation overlap

6.6.1 Introduction

It is a relatively common practice that where possible when a station is to be relocated a new station is established at the new site and the old and new stations operate in parallel for some time to collect data to determine the adjustments to be made at the original site to homogenise it with the data from the new.

This section presents two case studies that illustrate homogenisation issues based on data from the parallel operation of old and new sites of stations in the ACORN-SAT network.

6.6.2 Case study 1: Cape Borda, South Australia

A comparison was made of the daily minimum and maximum temperatures (Tmin and Tmax) recorded at two sites at Cape Borda, South Australia, (Lat. 35.75°S, Long. 136.60°E) with Bureau of Meteorology identification numbers 022801 (old site) and 022823 (new site). The ACORN-SAT documentation says regarding the Cape Borda site that the newer site (station 22823) "is further inland [by about 300 metres] and a less exposed location than the former site". The sites operated in parallel during the period 1 March 2003 to 20 July 2007, which leaves about 1175 days of data after the exclusion of data for days when one or both made no recording.

Figure 6.4 shows the distribution of the majority of the differences in minimum and maximum temperatures at the two sites, with 19 outlying differences (1.6%) omitted from the low and high ends of the graph. In order to show both sets of differences above the X-axis the Tmax differences are calculated as data from station 22801 *minus* data from station 22823 but the reverse for Tmin, indicating that in general

station 22801 records higher maximum temperatures but lower minimum temperatures than 22823. Six Tmax differences below the Figure's limit of -1.6°C are not shown, the lowest being -4.2°C and none from the high end. Seven Tmin differences below -1.6°C are not shown, the lowest of which is -7.2°C, and six differences were above the upper limit of the Figure, the highest being +5.7°C.

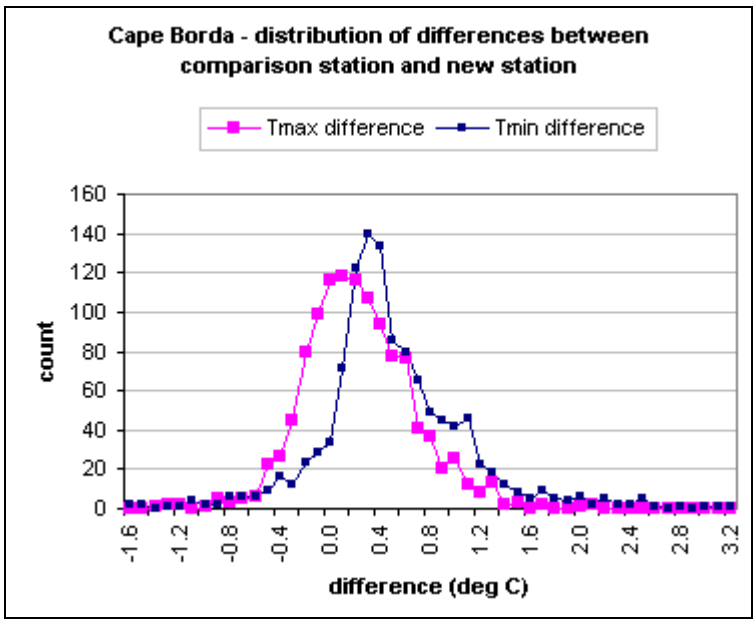


Figure 6.4 Distribution of differences between daily minimum and maximum temperatures at two nearby Cape Borda observation stations with overlapping data, with all data reported and processed to one decimal place.

When all Tmax data are included the average difference is 0.21°C with a population standard deviation of 0.376°C. For Tmax the majority of differences fall between -0.3°C and +0.7°C, with the distribution of differences skewed towards the lower values. The inclusion of all Tmin differences data produces an average difference is 0.475°C with a population standard deviation of 0.735°C. For Tmin the majority of differences fall between +0.1°C and +0.8°C and again the distribution is sharply skewed towards lower values.

The above differences were derived from all of the available data. Narrowing the data for the month of January excluding two outliers at -7.2°C and +3.1°C results in a mean Tmin difference of 0.405°C and a population standard deviation of 0.5°C,

suggesting a seasonal element to the difference between the two sites. This suggests that the temperature differences might be related to specific weather conditions that are more common at some times of the year than at others. On this basis the accurate homogenisation of data would need to take weather factors such as the wind direction and speed into account.

6.6.3 Case study 2: Orbost, Victoria

A similar comparison to the above was made of the daily minimum and maximum temperatures (Tmin and Tmax) recorded at two sites at Orbost, Victoria (Lat. 37.69°S, Long. 148.47°E) with Bureau of Meteorology identification numbers 084030 (older station) and 084145 (intended replacement station)[19]. The sites are separated by about 800 metres and with parallel readings during the period 5 June 2003 to 31 Oct 2011. The old site was at the top of a gentle north-facing slope with a house and several trees between 20 and 30 metres to the southeast, and the new site, according to ACORN-SAT documentation, is a site "adjacent to a reservoir on a southwest-facing slope". The comparison shown here is based on data for the ~3000 days that remained after the exclusion of days when one or both stations made no recordings.

The distributions of differences for Tmin and Tmax (both calculated as the recording at station 084145 *minus* the recording at station 084030) are shown in Figure 6.5 with just a few outliers not shown at each end. Eight Tmax outlying differences are below the lower limit of the graph, three of those being all at the lowest value of -5.0°C, and none from beyond the upper limit. No Tmin outlying differences are below the range shown but four outliers beyond the maximum value of 4.9°C are omitted.

When all Tmax data is included the average Tmax difference is -0.450°C with a population standard deviation of 0.513°C. For all Tmin data the average difference is 0.885°C with a population standard deviation of 0.985°C. (Based on average differences, 84145 generally records lower maximum temperature but higher minimum temperatures than station 084030).

[19] Described in http://www.bom.gov.au/climate/change/acorn-sat/documents/station-adjustment-summary-Orbost.pdf

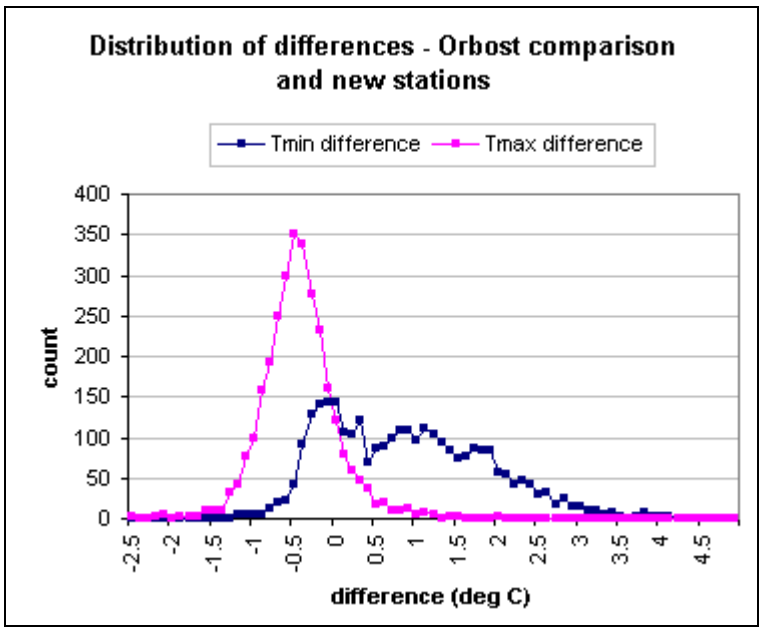


Figure 6.5 Distribution of Tmax and Tmin differences at two observation station locations 800 metres apart at Orbost, Victoria.

The distribution of Tmax differences is a very sharp "bell-curve" while the Tmin difference bears only slight resemblance to such a curve, the difference perhaps being that maximum temperatures on days with mostly clear sky tend to be around 3:00pm when the sun is still high in the sky and shadows across both observation stations are not a problem.

For much of the range of Tmin values most recordings from station 84145 are greater than those from station 84030 but not all. Figure 6.6, cut-off at a maximum of 9.0°C for convenience, shows that this does not hold true for minimum temperatures at station 84030 below 4.0°C where save for very few instances the minimum temperature at station 84145 is always greater.

Analysis on a monthly basis shows no significant change in the spread of the difference, only that the bias indicated above rarely applies in warmer months when minimum temperatures are generally higher.

The bias at low temperatures might be a reflection of different exposure to the conditions that cause the low temperatures (eg. wind) or the influence of other factors in the local environment (e.g. the small reservoir near station 84145). Regardless of the cause, the use of the average difference in minimum temperature to adjust the old data to bring it in line with the new would ignore this low-temperature bias.

Figure 6.6 also refutes the assumption in ACORN-SAT adjustments that there is one-to-one correspondence between the temperatures at two locations because it shows that a single temperature at one location might correspond to several different temperatures at the other location.

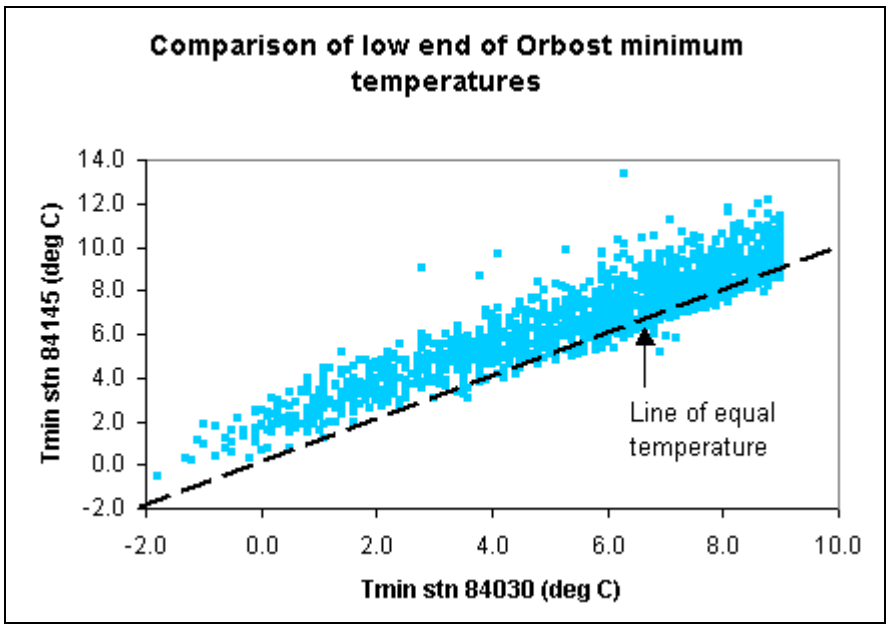


Figure 6.6 Comparison of minimum temperatures at the low end of the range of minimum temperatures recorded at two sites at Orbost.

6.7 Homogenisation using multiple sites

6.7.1 Introduction

This section presents two instances where the homogenisation of temperature involved the use of multiple comparison sites and questionable adjustments have been made. As with any homogenisation it is simply impossible to know whether the adjusted data is correct but it is possible to indicate potential flaws in the approach.

The first case study is an adjustment to Orbost data made prior to the adjustment described above. (Orbost is used because documentation summarising the data adjustments is available[20].) The second case study is of adjustments to observation station data that in composite form is used by the New Zealand government as its preferred national temperature record.

6.7.2 Case study 1: 1966 adjustments to Orbost data

A case study above (section 6.6.3) used an example of parallel operation of two stations over 5 years from 2003 to 2011 at Orbost, Australia, when an observation station was relocated to a new site and replaced with an automatic weather station. According to ACORN-SAT documentation the station at Orbost was also relocated in 1966 and its data adjusted according to a comparison with that from several other stations. This study will consider recorded and adjusted minimum temperatures before addressing the same situation for maximum temperatures.

(a) Adjustment to data for minimum temperatures

Figure 6.7 shows January mean minimum temperatures at the Orbost site and at six of the ten other sites that it was compared to, over the period 1938 to 1980. (Four other sites are omitted here because they commenced operations just prior 1966 and the amount of data is insufficient for comparison.) This plot shows that the data from Orbost was very similar to the data from its two closest comparison stations Bairnsdale (74km) and East Sale (125km), not just prior to 1966 but also after it. The correlation coefficient for Orbost and Bairnsdale data across the period 1938 to 1970

[20] http://www.bom.gov.au/climate/change/acorn-sat/documents/station-adjustment-summary-Orbost.pdf

(when Bairnsdale station closed) is 0.94 and Orbost and East Sale (1938-1980) is 0.88.

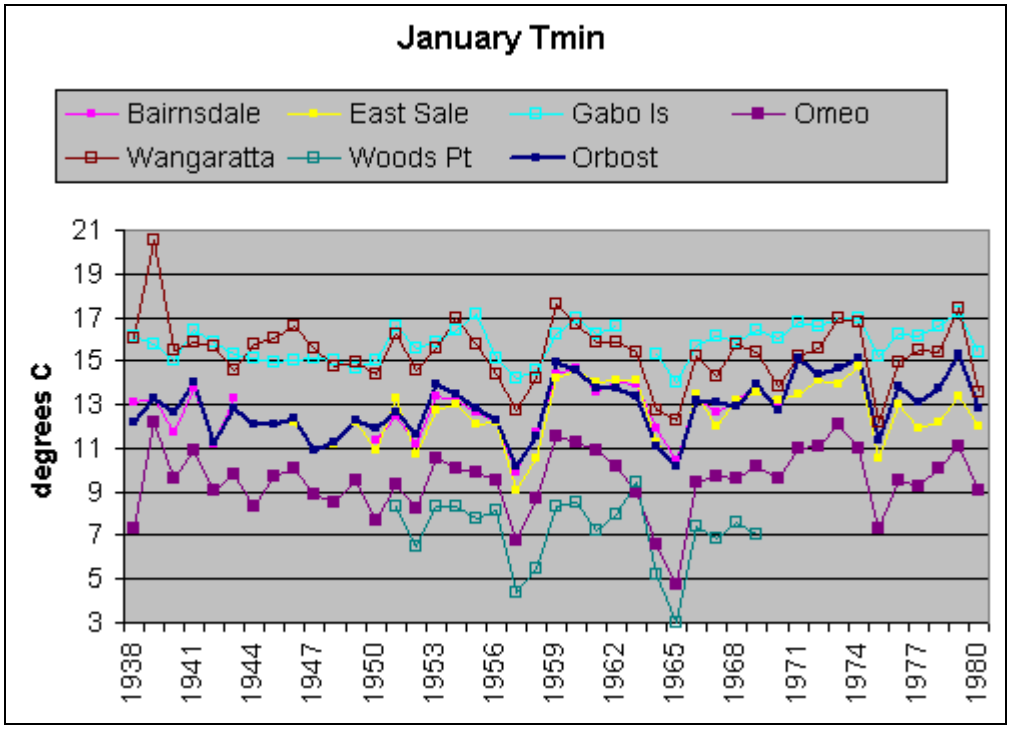


Figure 6.7 January mean minimum temperatures for Orbost and stations that its data was compared with and, starting in 1966, adjusted according to.

Figure 6.8 shows the same situation for July. The mean minimum temperature at Wangaratta (NW, 242km, over a range of mountains) is now similar to Orbost but Wangaratta is well inland, away from the buffering effect of the ocean. Orbost data is again very similar to the data from Bairnsdale, and similar to East Sale until about 1970 when East Sale data starts fluctuating by several degrees each year, as do two of the three stations other than Orbost.

From the data shown in Figures 6.7 and 6.8 it is difficult to see why Orbost minimum temperature data was adjusted at all. If there is any issue it appears to be with the data for East Sale. The East Sale observation station in question closed in 2005 after several years of parallel operation with a new automatic weather station 90 metres south. The metadata for the old site indicates two trees about 20 metres south from

the thermometer screen. The GPS coordinates of the old site location place it what is now a golf course, in a position with a single large tree to the south, large enough to have been present when the observation station was nearby. The open question is whether the loss of the other tree exposed the thermometer to southerly winds, which at this site are usually cold.

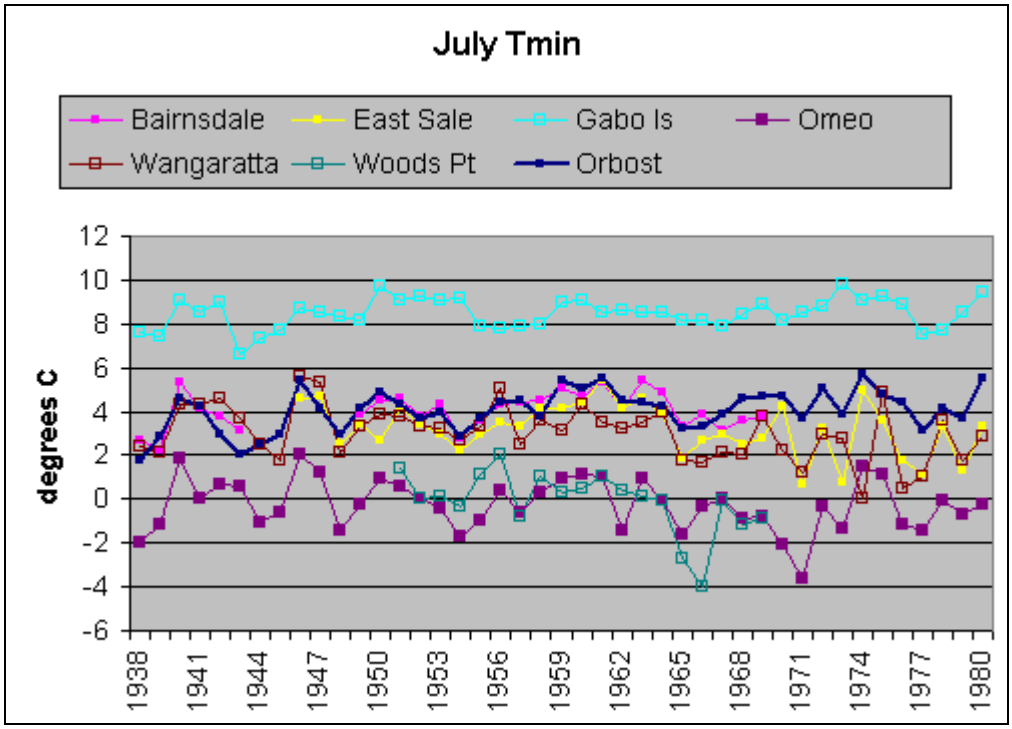


Figure 6.8 July mean minimum temperatures for Orbost and stations that its data was compared with and, starting in 1966 onward, adjusted according to. (i.e. as for Figure 6.6 but for July)

The adjusted data available from the ACORN-SAT system indicates that minimum temperature data for both Orbost and East Sale were homogenised. Figures 6.9 and 6.10 show the adjustments for each calendar month across the periods as stated. The data was sorted into month and then year order to show the changes in each calendar month. A negative (positive) difference indicates that the adjusted data was less than (greater than) the original data.

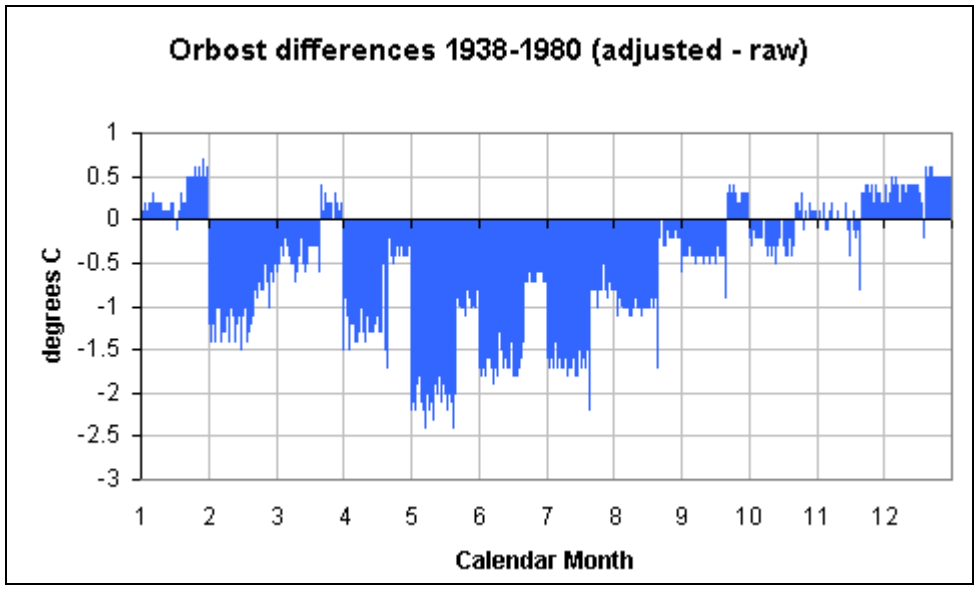


Figure 6.9 Difference between unadjusted and adjusted monthly mean daily minimum temperatures at Orbost 1938-1980.

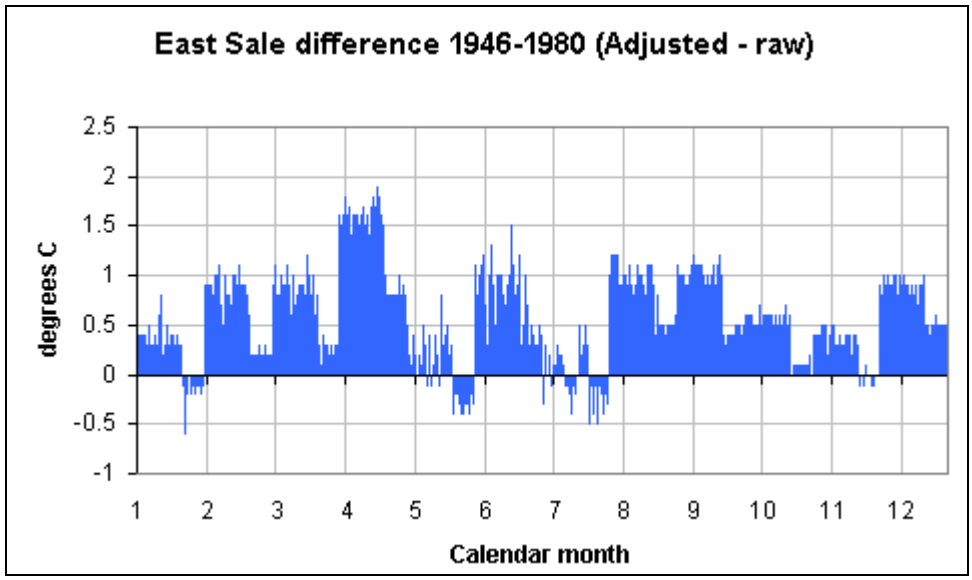


Figure 6.10 Adjustments of East Sale monthly mean daily minimum temperature for each calendar month over the period 1946-1980.

It appears that Orbost minimum temperature data has been adjusted in a way that increases the difference between summer and winter temperatures, this despite the location being less than 20km from the coast and therefore being located where the ocean buffers air temperature variation. It is not possible to determine whether the adjustment is correct or a consequence of the majority of comparison stations being away from the coast and therefore having a greater range of temperatures.

(b) Adjustment to data for maximum temperatures

In similar fashion to the minimum temperatures discussed above, Figures 6.11 and 6.12 show mean monthly maximum temperatures for January and June at Orbost and the six other stations that it was compared to. The striking feature of Figure 6.11 is that the maximum temperature at Orbost in January very closely matches that from four of the comparison sites, the exceptions being the coastal station at Gabo Island and the very inland station at Wangaratta, 243km inland to the north west.

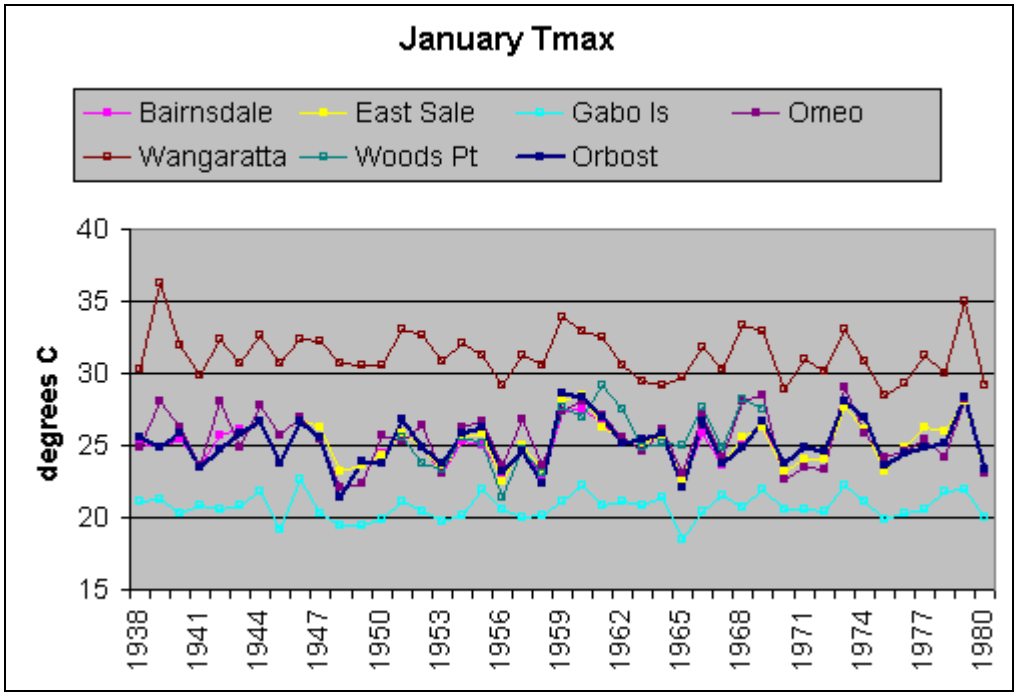


Figure 6.11 January mean maximum temperatures for Orbost and six comparison stations.

Again we have a situation where it is difficult to see any rationale for adjusting Orbost's maximum temperatures, especially in January, but they were adjusted in January and in every other month (Figure 6.13), including a rare upward adjustment of more than 3°C in November 2011.

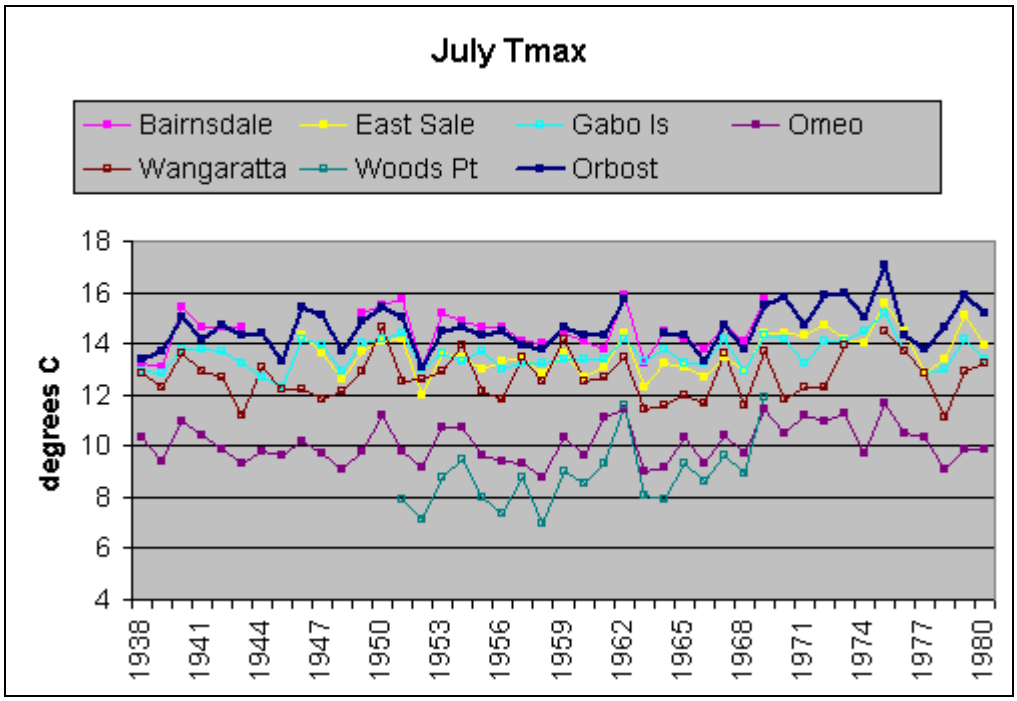


Figure 6.12 July mean maximum temperatures for Orbost and six comparison stations.

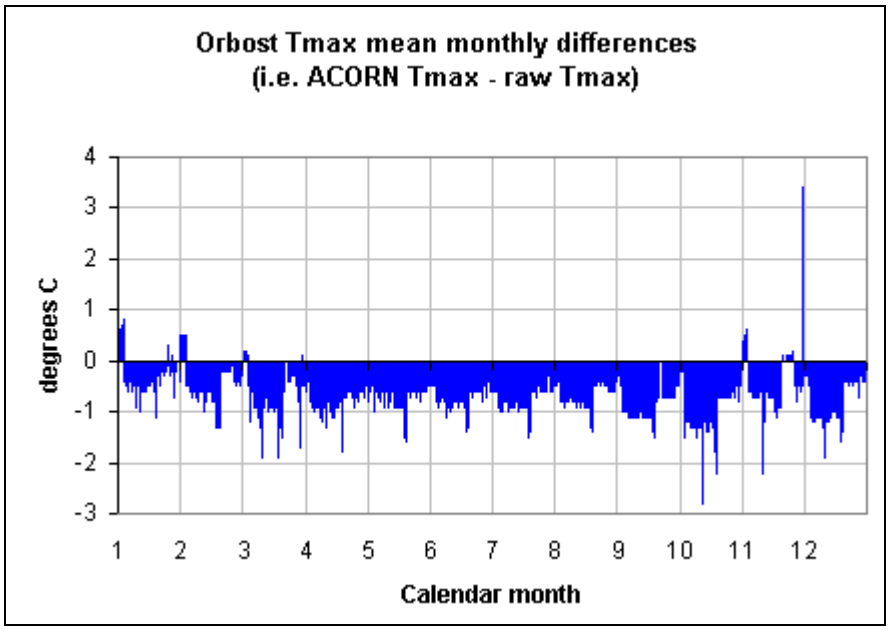


Figure 6.13 Mean monthly adjustments made to maximum temperatures at Orbost (in all cases adjusted values minus original data)

The maximum temperature for the relatively nearby East Sale observation station were also adjusted (Figure 6.14) despite this station being at low elevation and close to large areas of water (a lake and 20km away over countryside with few hills, the

ocean) when the stations that it was compared to are probably not. (Comparisons are often made to data from stations that are not in the ACORN-SAT network and complete information about all data adjustments is not published.)

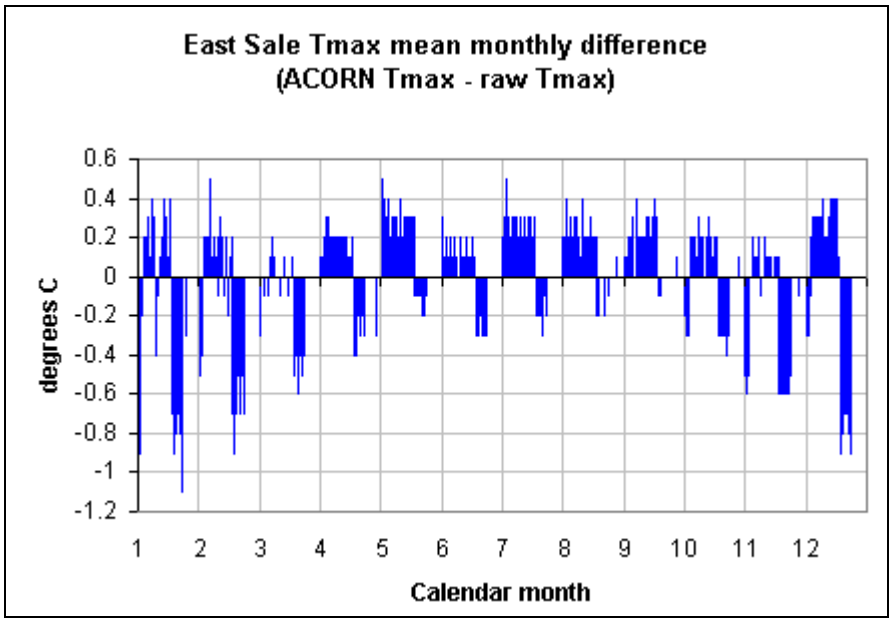


Figure 6.14 Mean monthly adjustments made to maximum temperatures at East Sale (in all cases adjusted values minus original data)

General comments

The adjustment to East Sale minimum temperature data raises certain questions. Its adjustment occurred in 1970 but its metadata record reportedly shows no change from the station's establishment in 1945 until 1996. Coincident with this adjustment of Sale data was the closure of neighbouring station Bairnsdale (59 km away) in April 1970 and Cooma SMHEC (277 km away), plus Woods Point station (93 km away) closed in October 1969.

Winter minimum temperatures (e.g. during July) are probably also subject to orographic airflow that sees temperatures at the top of an incline, as Orbost was, warmer than stations at the bottom of an incline as Woods Point (located in a steeply-sided valley) and possibly other comparison stations were.

The need for adjustment to Orbost minimum temperature data does not appear to be justified and the imposing of a warmer summer- cooler winter pattern might be genuine or might be an artefact of the majority of comparison stations being located further inland from the coast. The Bureau of Meteorology has given no reason for the data adjustment at East Sale but those adjustments occurred at a time when several comparison stations, including two that were likely the closest, ceased operation. Both issues undermine the credibility of such adjustments.

The adjustment to Orbost maximum temperatures is also of concern, especially in January where it shows good agreement with maximum temperatures from comparison stations. Even the adjustments of maximum temperatures in July are questionable because no account seems to be taken of exposure to winter weather.

Trewin (2012) says that identified breakpoints, indicating error or inhomogeneities, were regarded as potentially significant if the number of comparison stations generating such a breakpoint exceeded a threshold - 2 if <5 sites, 3 if 5 to 9, 4 if 10 to 19, 5 if 20 or more - chosen on the basis of less than 5% probability that the significant breakpoints could occur by chance. No detailed data is available for adjustments made to East Sale data but in this case three stations that might have been used for comparison - Woods Point, Bairnsdale and one other - were closed at or about the time the adjustments to East Sale data began.

On top of this Orbost was probably used as a comparison station for East Sale, just as East Sale was for Orbost, which indicates a degree of circularity as well as suggesting that the adjustments at each site might depend on which site was adjusted prior to the other.

Overall the extent of adjustment is worrying. It would seem that an observation station would only need to operate for a few years in order to establish the relationship to data from other stations and beyond that time the data could be synthesised from the data at those other locations. This would be a farce and would fail to recognise that stations have different exposure and that the observation stations were established to record local temperature in a consistent or almost consistent local environment. The

relationship between their data at that at other sites was never regarded as an issue worthy of detailed attention when weather, rather than climate, was the focus.

6.7.3 Case study 2: New Zealand national temperature record

The New Zealand Institute of Water and Atmospheric Research (NIWA) publishes a "7-station" composite and an "11-station" composite for the record of New Zealand temperatures over time. Mullan et al (2010) is the primary documentation for the "7-station" composite series, which is created from the data from three station from large urban environments (Auckland, Wellington and Dunedin), one station in a medium-sized urban area (Nelson), one station in a small urban area (Hokitika) and two stations in rural regions (Masterton and Lincoln). Except for the last mentioned rural stations the stations are situated in towns and cities on the coast but in each case the coast facing in different directions.

Site relocations were rare in the large urban environments. The Auckland observation station relocated once in the 48 years from 1951 to1998, Wellington's location was unchanged from 1928 to 2005 and Dunedin's in the same location for 38 years (1960-1997), meaning they were all susceptible to increasing urbanisation throughout these periods.

In the 101 years described by the documentation, 1909-2009, a total of 32 data adjustments were made, sometimes from overlapping local observations and at other times by comparison with other of the seven stations. In the order of the seven stations listed above in the introduction 3, 3, 5, 4, 5, 4 and 8 adjustments were made. The adjustments at each site were sequential, which means that adjustments were carried forward and the earliest data from the stations was adjusted 3, 4, 5 or 8 times.

A graph of the annual averages of the raw and adjusted composite data is shown in Figure 6.15. It is very notable from Figure 6.15 that the early temperature data for the composite record was adjusted downwards, which has resulted in an increased warming trend. Figure 6.16 shows the difference between the original and adjusted

data. There is no reasonable explanation of why earlier data should need greater adjustment because temperature observations should have the least interference and distortion when the station commences operation at any given site. It is also notable that the pattern indicated in Figure 6.16 is virtually a 180° rotation of rising temperatures that might be expected under increasing urbanisation.

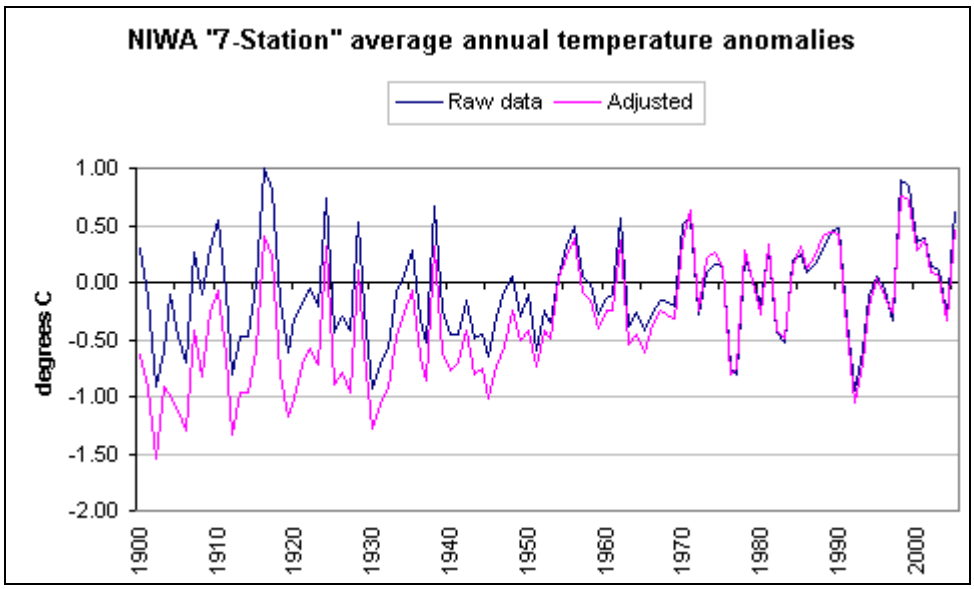


Figure 6.15 Average annual temperature anomalies for the raw and adjusted data used in NIWA's "7-station" composite for New Zealand

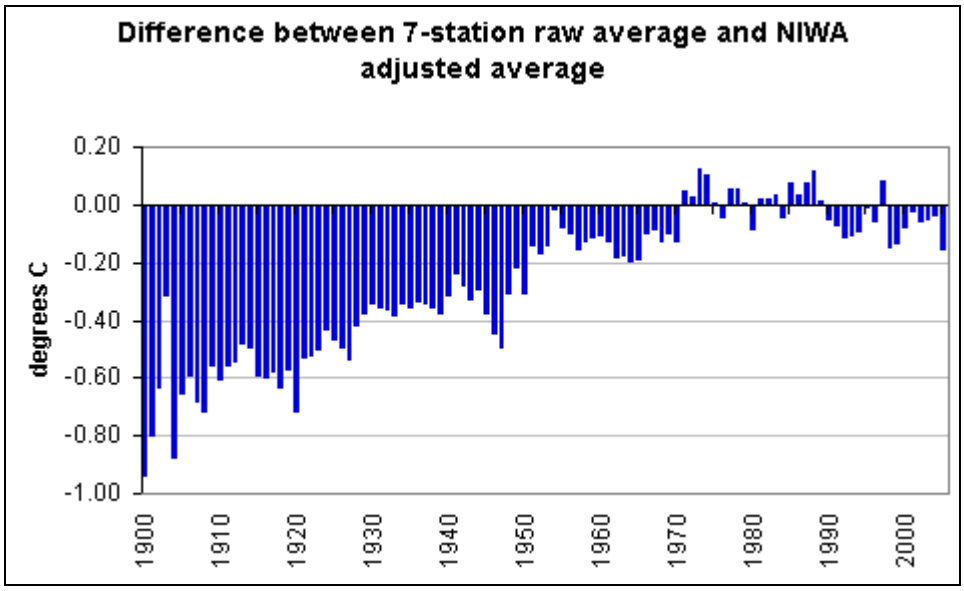


Figure 6.16 Difference between the two sets of data shown in the previous graph

Hessell (1980), the author being from the New Zealand Meteorological Service, finds serious problems with urbanisation at many urban sites and growth of local vegetation to be a problem at many rural sites, especially those run by the New Zealand Forest Service. It states "There are no sites at all which have been unaffected since 1930 by site changes less than 10m, increased sheltering, urbanisation or screen changes." and goes on to say "It is concluded that the warming trends in New Zealand previously claimed are in doubt".

Hessell (1980) predates NIWA's "7-station" composite record but de Freitas et al (2014) does not when it re-examines the data. de Freitas et al (2014) recounts how M.J. Salinger developed a homogenisation technique published in 1980 (Salinger, 1980), elaborated on it in an unpublished doctoral thesis in 1981 and in 1993 co-authored a different technique of homogenisation (Salinger & Rhoades, 1993).

NIWA's adjusted 7-station composite, which uses the 1980/81 homogenisation technique and 35 adjustments, has a warming trend of 1°C/century but de Freitas et al (2014), using the 1993 technique plus two well-documented site changes that did not appear in NIWA's list of adjustments, found an average 0.32°C/century warming for the five urban locations and 0.2°C/century for the two rural stations. This finding illustrates two significant issues, firstly that different methods of homogenisation produce different temperature estimates and secondly that the homogenisation of data to remove the impact of urbanisation might be flawed. The next section of this chapter will discuss the second point in more detail.

6.8 Homogenisation for urbanisation reconsidered

Much of the temperature data used in CRUTEM4 and HadCRUT4 datasets is sourced from observations stations that at some time have been subjected to the influence of urbanisation because people recording the data at manual stations lived in close proximity to the site. The habit in Australia at least was to locate observation stations at post offices where government employees could record the data. The adjustment of such data to compensate for local shielding or the greater problem of urbanisation

(local heat generation, distortion of natural processes) is vital to the accuracy of temperature records when they are used for climate matters.

It was noted above that Menne and Williams (2009) reported a detection rate of gradual inhomogeneities such as urbanisation of about 67% and a false detection rate of 19.7% and that the adjustment of ACORN-SAT data always used a step-wise change (Trewin, 2012).

An alternative view of the situation with trend inhomogeneities is to recognise that local influences on measured temperature will change over time and that the accurate adjustment of temperature needs to be cognisant of those changes. Hansen et al (2001) mentions a similar concept, calling it "undisturbed temperature", and briefly mentions the potential error of incorrect adjustment but fails to fully explore the situation.

"True temperature" can be defined as the exact meteorological temperature for a given location with its fixed geography and baseline physical environment that existed when observations began at the location. The temperature is "true" in the sense that it is consistent with the location, environment and meteorological conditions.

"True temperature" will vary throughout the calendar year because of the changes to the position of sunrise and sunset, the changes in the angle of incidence of insolation, changes in meteorological conditions and the changing influence of the constants or constant cyclic factors (e.g. leaves on trees) in the physical environment. It will however disregards all changes of a non-meteorological nature that might occur over time, such as any local manmade heat, manmade changes to the nearby physical environment and natural changes to nearby vegetation.

On this basis the measured temperature at a given site at any time, Tm, can be described as

$$T_m = T_t + T_{\triangle e} + T_n$$

where T_t is the true temperature, $T_{\triangle e}$ is the temperature contribution from changes to the environment and T_n is any general change in climate. The changes to the

environment might occur gradually and in regular fashion or might be abrupt or a combination of both, and they might come into effect at any time.

When a station is relocated the difference in measured values at the two sites is given by

$$T_{DIFF} = (T1_t + T1_{\triangle e} + T1_n) - (T2_t + T2_{\triangle e} + T2_n)$$

where the factors are as before but identified as being from site 1 or site 2.

A site relocation is usually of just a few kilometres at most, meaning that $T1_n$ and $T2_n$, the general change in climate, will be equal at both sites, assuming no large different in elevation. Just after start-up at site 2 the site is still in its baseline environmental state and therefore changes to the environment are not a factor. The difference in temperatures therefore resolves to

$$T_{DIFF} = (T1_t + T1_{\triangle e}) - T2_t$$

In other words the difference in temperature is the measured temperature at site 2, which at this time is its true temperature, subtracted from the combination of the true temperature at site 1 plus any distortion caused by environment changes at site 1.

Figure 6.17 illustrates this situation, with 'site 1' recording the "true temperature" from start-up (point 'A') for an undefined period of time (until point 'B'), then non-meteorological causes have an increasing influence on local temperature. The skewing from true temperature ('B' to 'C1') is shown as a solid line that in reality might be in steps, some of which might even be downwards, or an almost constant increase or a mix of some steps and steady increases. After some time a new station ('site 2') starts operating in parallel with 'site 1' for a certain period before the average difference between the old and new locations, C1 and C2 respectively, is calculated.

In Figure 6.17 the average difference between the data from the two sites is shown for illustration purposes as 0.3°C and the difference in 'true temperature' at the two sites is 0.2°C caused, for example, by differences in the local geography.

Regardless of whether a simple average difference or some other homogenisation technique (e.g. percentile-matching as described by Trewin, 2012) is used, the principal is the same. All data for the old site, 'site 1' in Figure 6.17, is adjusted according to the homogenisation algorithm devised from the data recorded during the period of overlap with 'site 2' despite the data from 'site1' already being distorted by non-meteorological influences. As the middle portion of Figure 6.17 shows, this would incorrectly leave all data prior to point 'B' 0.5°C below the correct values because the magnitude of the adjustment does not reduce over time to counterbalance the extent of the distortion from true temperature.

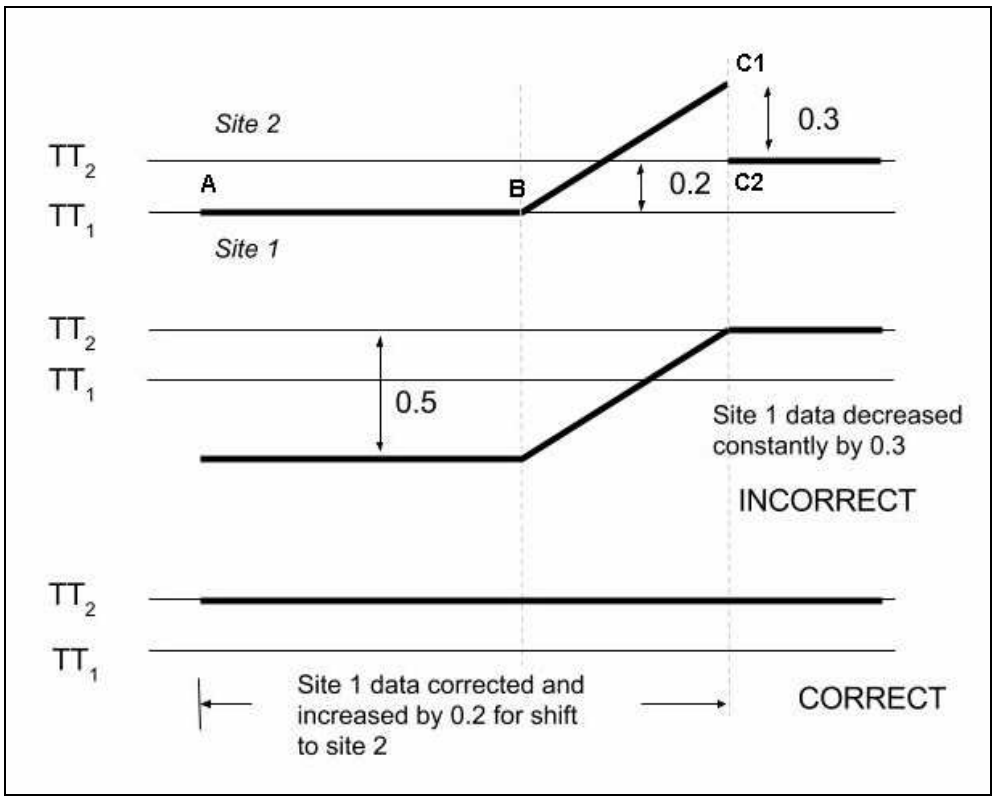


Figure 6.17 Concept of 'true temperature' as applied to an observation site being relocated and the data from the original site being adjusted. The X-axis is of time but cannot be scaled because specific points in time are unknown.

As noted at the bottom of 6.17, the correct adjustment of 'site 1' data and all previous data is a two-step process. The first step is to correct the divergence from true temperature by correcting each point along the path (from 'B' to 'C1' in Figure 6.17)

and the second step is to adjust the modified site 1 data and all earlier data by an amount equal to the difference in true temperatures at each site.

Neither the exact pattern of the divergence from true temperature is known nor the point in time at which it began, but the earliest that the environmental changes could have started influencing temperature is when the first site commenced operations because that was the time of the baseline environment against which true temperature was defined. In the case of an urban environmental those external influences might have started increasing from very soon after site start-up but in non-urban environments any skewing might have occurred much later or even not at all.

One approach would be to assume that all sites began by reporting true temperatures and that the distortion increased in linear fashion, but this might be a poor description of reality.

To correctly adjust the data from the original site to be the equivalent of what would have been recorded at the new site requires that the true temperature at the first site and the point of divergence from that true temperature be known or estimated with good accuracy. The problem is that unless the progressive distortion is recognised and the data adjusted in accordance with the extent of distortion there is a risk that the adjusted data will be incorrect and contain false trends.

6.9 Summary

This chapter has shown that the original data from observation stations, which is the input for the CRUTEM4 dataset and for the HadCRUT4 dataset, is likely to be adversely impacted by errors, uncertainties, assumptions and inconsistencies at many levels, these varying in magnitude according to specific issues at individual stations.

Issues arise at every step of the measurement and subsequent data processing, starting with the instruments, their enclosures (or screening) and the siting of the observation station. The last mentioned is particularly important because World Meteorological

Organization standards indicate up to 5°C uncertainty for individual measurements for some siting situations.

A key issue with the data from observation stations is its homogenisation, which involves attempts to remove non-meteorological distortion of the temperature record. This chapter has shown that the commonly used Standard Normal Homogeneity Test is not without uncertainties and potential errors.

Four case studies have been presented to illustrate various issues with homogenisation. The first and second (Cape Borda and Orbost site relocation), both using nearby stations operating in parallel to determine the relationship between the data, were shown to make assumptions about exposure and in the second case in particular illustrate that there is not necessarily one-to-one correspondence between the temperatures recorded at one location and those at another less that 1km away.

The third and forth case studies (Orbost ACORN-SAT adjustment and New Zealand national temperature trend) involved the use of data from multiple sites, in case three to make questionable adjustments to the data at a target site. The fourth case involved data from multiple sites being merged to create composite records both with and without station data adjustment. The post-adjustment composite showed increasingly lower temperatures than the unadjusted data the further we move back in time. There is no reasonable explanation for this other than errors in the data adjustment.

With this last point in mind the homogenisation of data to remove the effects of urbanisation was reconsidered and a serious systemic error identified, that of adjusting all data for a previous observation station despite the fact that when it commenced operation the station probably accurately recorded the true (i.e. undistorted) temperature.

Case four also showed that different methods of homogenisation can produce quite different temperature estimates, especially when there is a sequence of homogenisation operations.

Homogenisation is nothing more than a process of estimating likely temperature recordings that would have been recorded if the instruments, enclosure and site had been consistent throughout the entire data record. There is no practical method of determining the accuracy of the adjusted data because there is nothing to directly compare it to. One might say that it "looks correct", or that some high level statistics are "consistent with" certain assumptions or estimates, but neither is confirmation that it is in fact correct.

Unlike previous versions of the datasets, HadCRUT4 and CRUTEM4 now require that national meteorological services supply temperature data that has already been adjusted however the metadata associated with each station gives no indication of the station classification or any details of any data adjustments. In these circumstances it is not possible to determine the accuracy of the adjustments in the supplied data. While the magnitude of any problems with the supplied data cannot be determined, the presence of several data outliers (see chapters four and five) is not encouraging.

Chapter 7: Issues with sea surface temperature data

7.1 Introduction

The previous chapter dealt with issues related to data from observation stations on land prior to the national meteorological services supplying the monthly mean temperatures for inclusion in CRUTEM4 and HadCRUT4 datasets. This current chapter will consider similar issues for sea surface temperature measurements but with some important differences from the CRUTEM4 data analysis.

Firstly, the Hadley Centre does not provide the original measurements of sea surface temperature. Such data is available from the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) database but the lack of information about the average temperatures used at the initial 1 x 1 grid cell size used by the Hadley Centre makes the ICOADS data of limited value and the entire ICOADS database is so large as to be unwieldy to download and process.

Secondly, by its nature sea surface temperature measurements are not from a static (or near-static) site as with observation stations and this makes redundant any need to consider site exposure and station relocations. Further, because sea surface temperature measurements are not of minimum and maximum temperatures over time, issues such as time of day bias and the error margins associated with monthly averages are very different. That said there are still many issues about the measurement and processing of sea surface temperatures that cause concern.

While land makes up 29% of the Earth's surface and 71% water, the maximum global coverage of HadSST3 data is 81.9% because it includes all coastal grid cells, each of which is part water and part land. (Chapter 5 discussed the derivation of data for coastal grid cells.) Given the much greater coverage of the earth's surface by oceans than by land the accuracy of sea surface temperature data is crucial to the accuracy of the HadCRUT4 dataset.

7.2 Inconsistencies with measurement methodology

The methods used for observing sea surface temperature have varied greatly over time. One of the most significant changes has been the increasing reliance on buoys for the measurement and reporting of temperatures. Prior to the use of such buoys sea surface temperatures measurements were made on a voluntary basis by a set of ships often referred to as the voluntary observing fleet, or some similar name, with the individual vessels using a methodology defined by the authority that they reported to.

Matthews (2013) and Folland & Parker (1995) provide a detailed discussion of the issues. The following is a summary in point form of issues mentioned by Folland & Parker (1995), with clarifications where appropriate. The aim is not so much to provide details but to illustrate several issues and the general complexity of deriving SST anomalies from sea temperature measurements.

(a) Variations in instruments and tools

- Seawater samples in which to measure temperature have been obtained using a variety of bucket types - canvas, leather, leather and metal, rubber, insulated canvas, wood and iron, although reportedly the last two types were less common because they risked causing damage to the ship's hull.
- Measurements have also been made in the pipes of the seawater inlet for engine cooling, often referred to as Engine Room Inlet (ERI), as well as via hull-mounted sensors.
- Measurements made in any given month are likely to have been made, at least until very recently, by a mixture of the above techniques, each of which had different characteristics that might need to be taken into account.
- More recently than Folland and Parker (1995), measurements have also been made by moored or floating buoys or since about 2001 by Argo buoys.

(b) Variation in water sampling site and depth when using buckets

- The relevant authorities (typically the national meteorological services of the country in which the ship was registered) usually, but not always, instructed

that the sample be taken from forward of the outlet for the engine cooling water.

- Some authorities, but not all, stipulated that it should be taken as far as possible from the side of the ship.
- UK instructions in 1938 said water was to be drawn from the surface.
- US observers were instructed to draw water from a depth of 1 or 2 metres where possible (although how water from closer to the surface was to be prevented from entering the bucket is not mentioned).
- Folland and Parker (1995) note that rough seas probably made it impossible to comply with the above instructions.

(c) Variation in instructions re the temperature of the bucket itself

- Some authorities specified that the bucket was to be stored in the shade so that it didn't warm prior taking the water sample.
- Some said a dry bucket should be used so that evaporation didn't cool the bucket.
- Some specified the time that the bucket was to be left in the sea and these might differ between authorities (e.g. 30 seconds, 1 minute).
- Some said that the first water sample was to be discarded, presumably on the basis that the first sample would bring the bucket to sea temperature.

(d) Variation in placement of the bucket on board the ship

- Some authorities directed that the bucket be placed in the shade when lifted aboard the ship but others were not specific.
- Some said to place the bucket out of any wind, but again not all.

(e) Variation in instructions about mixing the water within the bucket

- No recommendations were made in the nineteenth century about stirring the water sample but instructions in the twentieth century variously advised stirring "little, if at all", "slowly", "slowly, not touching the walls", "continuously", "quickly" or "vigorously".

(f) Variation in thermometer types and their use

- Some authorities required the use of wooden or metal backing or sheaths to protect thermometers from breakage and/or a small water reservoir to briefly preserve the temperature when the thermometer was withdrawn from the bucket to be read, although this was said to introduce problems due to the initial temperature of the thermometer and its housing.
- Instructions to observers sometimes said that the thermometer should be read with mercury bulb still immersed in the water although some observers reportedly removed the thermometer and took it into some light so that it could be read.

(g) Variation in time from ocean to temperature measurement when using bucket

- Some authorities recognised the issue of the time taken to haul the water sample aboard potentially causing cooling (or in rarer instances, warming) but not others.
- Some specified that the thermometer should be inserted immediately the sample is hauled aboard but not all.
- Some said that the thermometer was to be left in the water for 2 to 3 minutes before it was read although some said 'at least 3 minutes', another '3 to 5 minutes', another ' 4 or 5 minutes' and another implied for large buckets that a maximum of 5 minutes was acceptable. At the other end of the scale some recommended for one minute or less and a few authorities specified that the observer should wait until the reading was steady.

(h) Variation in number of measurements

- Some authorities that used a thermometer with a water reservoir to preserve the temperature said that multiple readings should be taken but others did not.

If these instructions were all precisely followed then the data has been obtained using a wide variety of methods and techniques, suggesting a variety of accuracies that might alter with specific circumstances at the time of each measurement.

Folland and Parker (1995) do not mention Expendable Bathythermographs (XBTs), which became popular in the 1970s for determining sea surface temperature although intended for Naval applications at greater depths. Each XBT has a metal nose weight

to force it to sink and near the front of which is fixed a thermistor (Kizu et al, 2001; Geise et al, 2011). XBTs are dropped in a controlled fashion from the sides of ships, the XBT manufacturers referring to it as a launch. The sea temperature is determined from the electrical resistance of the thermistor, which is initially activated by the XBT reaching the water surface, and with the depth measured by cable run-out. The XBT will pass through the notional depth of 500mm below the sea surface, the depth at which bucket samples are notionally taken, about 0.8 seconds after being launched from 5.5 metres above the sea surface, with more than 0.7 seconds of that time spent in the air. At the time of dropping the XBT would be at air temperature, or perhaps warmer if it had been in the sun, so there is uncertainty as to whether the temperature of the metal nose weight of the XBT will match the water temperature after less than 0.1 seconds in the water.

As noted briefly above, SST measurements have also been made by moored or drifting buoys, said by Rayner et al (2006) to account for 65% of observations in 1997, or since about 2001 by Argo buoys. Argo buoys appear to currently contribute the greatest number of SST measurements, although not necessarily the greatest number of "sea" grid cells. Argo and drifting buoys have an advantage of being able to travel through regions of shallow water that ships cannot traverse.

Over time the precision of thermometers used for these observations has improved, especially on engine room intakes. Matthews (2013) mentions Saur (1963) noting that ERI thermometers were sometimes graduated in intervals of 2°F or 5°F, this being of sufficient accuracy for the process of engine cooling. In recent times, with computer-assisted operation of ships engines, electronic thermometers are probably accurate to 0.1°C. Kent and Taylor (2006) make the point that prior to this precision being available readings were often rounded to the nearest whole number, a point supported by analysis of part of the data for Brohan et al (2009) (see later this chapter) which showed that sea temperatures were generally given in whole degrees (Fahrenheit), only rarely to 0.5°F (perhaps 5%) and never at any other fraction of a degree.

The measured temperature might also be influenced by time of day, month of the year, direction and speed of any wind (a moving ship will generate on-board wind and

a natural wind will cause waves and overturning at the surface), presence or absence of cloud, rainfall (latent heat exchange with the ocean) and the influence of any turbulence caused by the passage of the vessel through the water (e.g. bow wave). It will be shown in sections 7.5 and 7.6 that the measured temperature could also be influenced by the recent weather conditions at that point in the ocean causing distinct thermal layering and by the ocean's internal waves. All of these issues are outside the control of the method of measurement, some influences being meteorological and others as a consequence of the ship itself.

7.3 Different methodologies produce different temperatures

Several papers (eg. Folland & Parker, 1995; Kennedy et al, 2011b; Matthews, 2013) report differences in temperatures measured by different methods, with some of the findings likely attributable at least in part to one or more factors mentioned in the section immediately above. This section reviews some of those findings to illustrate how the temperatures vary.

Kennedy et al (2011b) discuss several papers that found that, according to tests or various assumptions, certain biases were present and goes on to describes some of the data adjustments that were made and the reasons for rejecting others. It lists 19 papers that found biases from -0.13°C to +2.3°C but the situations varied as to bucket type, number of ships for which data was analysed and the years of data collection. Most are anecdotal to some extent, and all illustrate the complexity of the situation.

Folland and Parker (1995) found that, depending on the location of the ship, adjustments of up to several tenths of a degree were required to correct for the widespread use of canvas and wooden buckets in the collection of water samples, but only until 1942, after which time it was assumed that the method of data collection was constant.

Matthews (2013) cites Brooks (1926) as reporting an average difference of 0.3°C between a canvas bucket used from one height and a tin bucket from a lower height,

but with the average difference doubling to 0.6°C when another person took measurements using the same canvas bucket. This suggests that SST monitoring might be inconsistent even when the same methodology is used.

Matthews & Matthews (2013) detail a comparison between temperatures measured by wooden, canvas and rubber buckets during a one-month voyage from the Hawaiian Islands to Tahiti during May and June 2008 at the end of a La Nina, reporting mean differences between those bucket types of less than 0.1°C during both day and night and across different weather conditions (eg. wind speed, wind direction, cloud, atmospheric pressure). ERI measurements were also made at 3 metres depth but found to be noisy, the conclusion being that ERI-measurements cannot be corrected for any near-surface gradients in water temperature, a point that we'll return to shortly. Matthews & Matthews (2013) also caution that their findings might not apply in other months or seasons or with different El Nino-Southern Oscillation conditions.

Kent and Taylor (2006) concluded that difference between measurements made via buckets and via engine room intakes varied on a seasonal and even decadal basis. It shows that for the North Atlantic from 20°N to 50°N and 80°W to 0°W over the three decades 1970 to 1999 the SST measured using buckets was usually cooler than that measured via the engine room intake. In the 1970s the mean January and June differences were around 0.4°C and 0.2°C respectively, those for the same months during the 1980s were around 0.34°C and a maximum, in July, of 0.1°C. By the 1990's the temperatures showed less difference with the January average difference (as bucket temperature *minus* ERI temperature) being -0.16°C and the June average difference of +0.15°C, the latter being part of the March-September period when SST measured via buckets was slightly warmer than that by engine room intake.

Kennedy et al (2011a) claims that the SST sensor in floating buoys (c.f. self-submerging Argo buoys) measured temperatures at 250mm below the surface. Emery et al (2001) estimates that temperature measurements recorded by such buoys were 0.15°C cooler than measurements made from ships.

Rayner et al (2005) show graphs of SST anomalies derived from floating buoys and from ships from off the east and west coasts of the USA, the equatorial Pacific and

around the UK. The smoothed differences vary over time, but on average the data from buoys were colder than that from ships by from 0.1°C to 0.2°C in most locations and ~0.4°C off the west coast of the USA. More telling are occasional differences exceeding 1.2°C at various times in all locations except the USA, sometimes in the Pacific for several months, again illustrating the inconsistency of the situation.

Kennedy et al (2011b) make the observation that "during hauling the water sample collected in a bucket can lose heat via evaporation, and can be cooled or, less commonly warmed, by the exchange of sensible heat with the air". This appears unlikely given the very short time in which the operation takes place but Matthews and Matthews (2013) report this occurring with small rubber buckets.

The question of whether cooling or warming might occur during lifting the bucket onto the ship is intriguing given that an analysis of the downloaded data for the 69 of the ships that Brohan et al (2009) discussed (see also 7.7 and 7.8 below) found that of 251,867 recordings of both sea surface temperature and dry bulb air temperature (DBT), 24392 (10%) were equal, SST was greater than DBT in 118,567 (47%) instances and less than DBT in 108,908 (43%). This data shows only a small difference in the number of situations when the water might have been cooled or warmed by the air. Further analysis would be required to determine the relationship of temperature to latitude and month but it does suggest that the situation is more complex than the statement made by Kennedy et al (2011b).

Kent and Taylor (2006) mention that the difference in measurement depth "(surface for the bucket compared with the noted average of 10m for the engine intake) means that the engine intake should, if anything, be colder than the bucket SST". This raises two issues, the first being that the depth of the engine room intake will vary according to the total weight of the ship and its cargo, and hence to the depth of water it displaces. The second is that over time the size of ships has increased and although it is not clear whether depth has increased or only width it does suggest that that engine room intakes might have been lowered. Matthews (2013) says that the depth of the ERI on the observer ships are "typically around 7-10m, although can exceed 15m" and lists 10 earlier papers that report ERI depths from 2m to 10m. These comments about increasing ERI depth conflict with Kent and Taylor (2006) who noted that the

difference between bucket-measured and SSTs and ERI-measured SSTs seemed to be decreasing over time.

The above collection of sometimes-contradictory findings illustrates some fundamental problems associated with attempting to resolve the different methodologies of measuring sea surface temperature. The uncertainties are substantial even when the same method is used, whether that is with different people drawing water from the ocean via a bucket or engine room intake measurements made at different depths. Matthews and Matthews (2013) is so concerned about the problem of deriving sea surface temperatures from subsurface temperatures at a depth of even 2 metres that it calls for ERI and other subsurface temperature data to be removed from the SST records. The removal of that data would not be easy because sea temperature measurement records in the ICOADS database often fail to state the method used, as the next section will discuss.

7.4 Macro adjustments to SST data

With all the above variability in the method of measurement, the uncertainty over the measurement technique being used at the time and the various reports of differences in temperature from one method to another, the adjustment of SST data to make it all equivalent to hypothetical measurements involves numerous assumptions that might not be valid. Apart from these per-measurement or per-ship adjustments two macro-scale adjustments to SST data have been proposed.

Numerous sources (e.g. Folland & Parker, 1995 citing the earlier Folland et al, 1984; Folland et al, 2001; Smith & Reynolds, 2004; Rayner et al, 2006; Kennedy et al, 2011b) report an abrupt shift in average SST's in December 1941 and several papers (e.g. Thompson et al, 2008; Kennedy et al, 2011b) report a similar shift in late 1945, the former particularly in the North Pacific and the latter especially in the Southern Hemisphere. These shifts (Figure 7.1) have been attributed to changes in the most common technique of measuring SST, firstly as a shift from buckets (favoured by

non-US ships) to engine room intake (favoured by US ships) as the USA entered World War II and then the use of buckets increasing after the end of the war.

Folland et al (2001) refer to the switch from buckets to engine room intake for SST measurements when describing how SST measurements were adjusted to more closely match global average temperature anomalies derived from models, the assumption being that SST measurements via bucket recorded lower temperatures than measurements via engine room intakes. It states "*The global annual mean bias correction increases steadily from 0.17°C in 1872 to 0.30°C in 1900 and 0.39°C in 1920, remaining around 0.4°C until 1941*" and later says "*Without bias corrections to SST, ensemble mean LATs [land-surface air temperatures] are significantly too cold between 1872 and 1941 in all major regions except extratropical S. America ...*". This suggests that SST data was adjusted on the assumption that it was incorrect and that it needed to be adjusted to match the output of models. Given the frequent failure of the prediction of climate models the claims by Folland et al (2001) are simply extraordinary.

Folland and Parker (1995) describe adjustments to SST data, first pointing out sudden shifts of ~0.8°C and ~0.5°C in the night marine air temperatures (NMAT) in the Northern and Southern Hemispheres respectively (Figure 7.1). The paper fails to quantify the data shifts so they must be estimated from graphs and the time-scale along the X-axis is unhelpful, on top of which the influence of natural variation must be considered. Folland and Parker (1995) go on to point out that the first jump, in 1941-2, coincides with the entry of the USA into World War II and "is likely to have resulted from a realization of the dangers of hauling sea buckets onto deck in wartime conditions when a light would have been needed for both hauling and reading the thermometer at night".

To illustrate the impact of the adjustments to SST in light of the above Figure 7.2 shows the monthly anomalies according to the adjusted HadSST3 data.

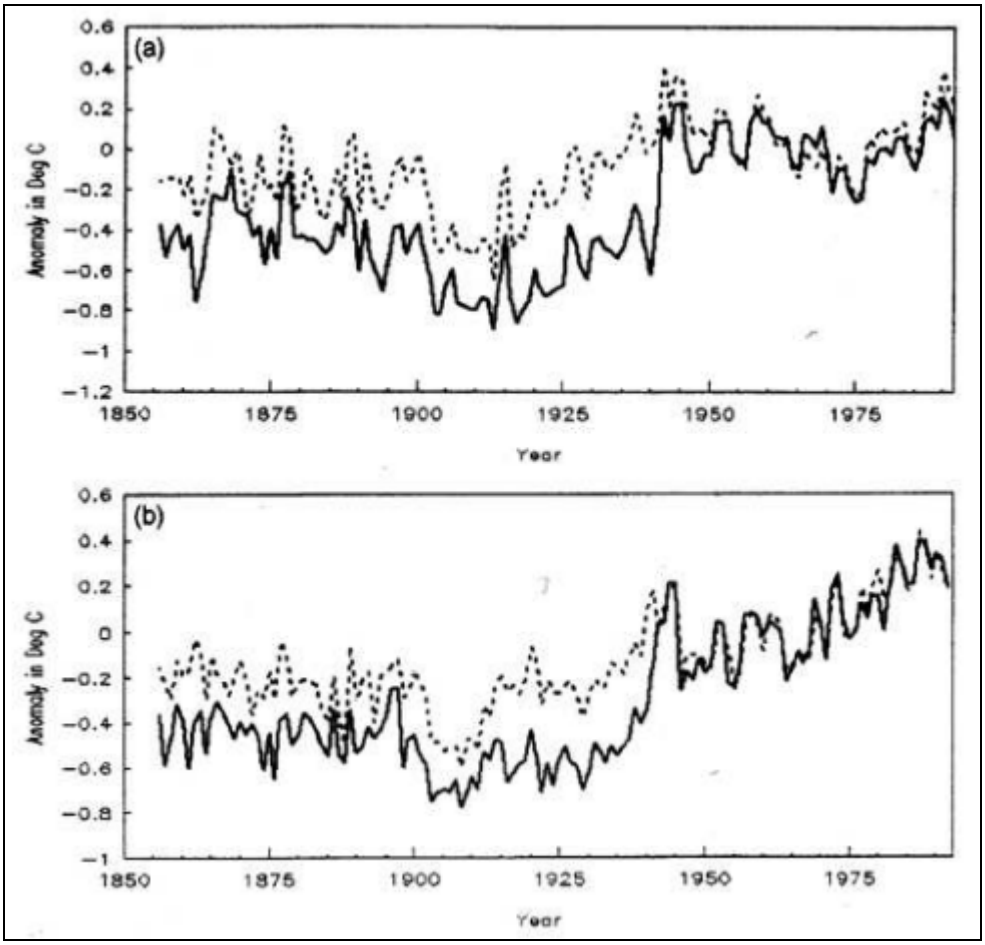


Figure 7.1 Copy of Figure 3 in Folland and Parker (1995) showing the jump in measured SST (solid line) in the mid 1940's and the adjusted - stated as "corrected" - night marine air temperature NMAT (dashed line) for (a) northern (b) southern hemisphere, (1856-1992)

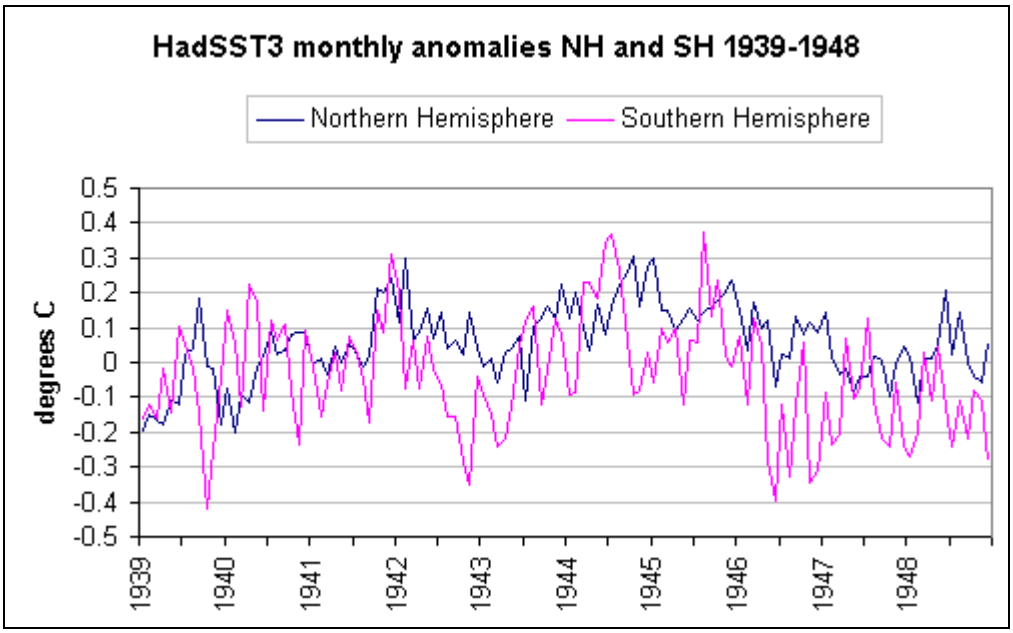


Figure 7.2 Monthly HadSST3 temperature anomalies for both hemispheres (1939-1948)

Rayner et al (2005) say of this period "It was clear that the FP95 [Folland and Parker (1995)] corrections needed modification in this period [1939-42] for the new analysis, so they were altered to gradually decrease to zero between January 1939 and January 1942, rather than the previous continued increase to December 1941 and then sudden end at December 1941... This was achieved by decreasing the annual proportion of canvas buckets from 100% in 1939 to 50% in 1940 to 25% in 1941. Smith and Reynolds (2005) linearly decreased their bucket corrections to zero over this period for the same reason."

McKitrick (2010) has a different take on the matter, pointing out that according to WMO data, buckets were still widely used well into the 1980s. McKitrick et al (2010) cites Kent et al (2007) and replicates a figure from that paper but the earlier Kent and Taylor (2006) has similar if slightly different information (Figure 7.3). According this figure, ≥ 40,000 SST measurements per month were taken using buckets from 1979 to 1992, the ratio of that to all SST measurements varying largely according to the measurements via engine room intake.

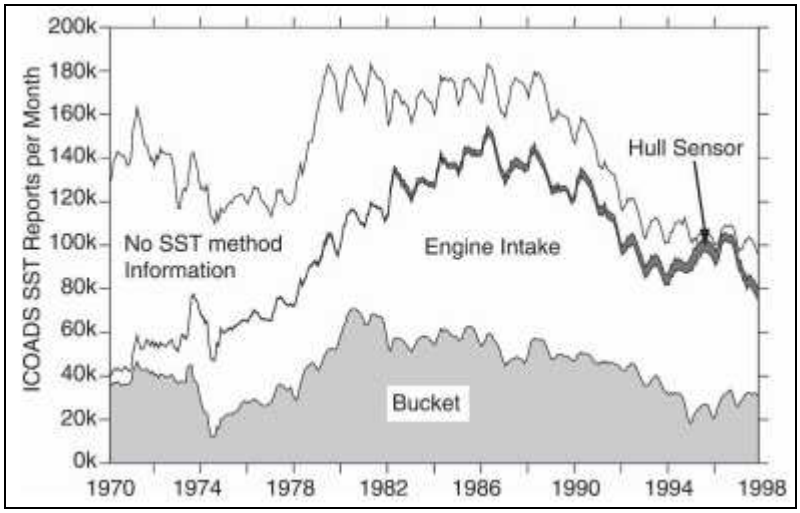


Figure 7.3 From figure 3(b) of Kent and Taylor (2006), breakdown of the average number of ship SST reports by measurement method using information from the ICOADS SI flag (i.e. method indicator) for the period of 1970–97, supplemented with metadata information from the WMO, all smoothed with a 3-month running mean filter.

McKitrick (2010) makes the point that average hemisphere and global SSTs seem very sensitive to the relative proportions of measurements by buckets and engine room intakes. That might have been a concern with HadSST2 but is less of an issue with the HadSST3 dataset, the version that is usually used being the average of 100 datasets, each with slightly different assumptions about the percentages of different techniques for measuring SST when no method was logged when recording the temperature.

The above papers seem to ignore the fact that while the entry of the US into World War II coincided with a change in average global SST it also constrained shipping and reduced data coverage. Global SST coverage altered significantly in December 1941 and again during 1946, the first being a contraction and the latter being an expansion. SST coverage of the Northern Hemisphere in December 1941 is 40.09% of the hemisphere, down from 53.9% in the previous month, a decrease of 25.6% of the original figure. The Southern Hemisphere the fall was greater with coverage of 31.5% of the hemisphere in November 1941 falling to just 18.2% in December. Figures 7.3 and 7.4 show the coverage by 10-degree latitude bands, excluding those bands that failed to ever exceed 2.0% during the period.

The abrupt changes in SST, based on Folland and Parker (1995), coincide closely with an abrupt change in coverage. Figure 7.4 shows the Northern Hemisphere coverage in 10-degree latitude bands from 1937 to 1948, with the December 1941 decrease in coverage very obvious.

The shift in SST from November to December 1941 is most noticeable in the northern Pacific Ocean (0-60N and bounded by coastlines to the east and west) and corresponds to the coverage in December 1941 of 15.47% of the hemisphere, down from 24.02% in November and 27.6% in October (Figure 7.5). In particular coverage of Pacific Ocean grid cells in the band 0°N to 5°N fell sharply from 8.1% of the total of the North Pacific's 24.02% hemisphere coverage (i.e. 1.96% hemisphere coverage for the band), to 3.9% of the North Pacific's 15.47% hemisphere coverage (i.e. just 0.6% hemisphere coverage for the band) in December. In band north (i.e. 5°N to 10°N) coverage in December was half what it had been in November.

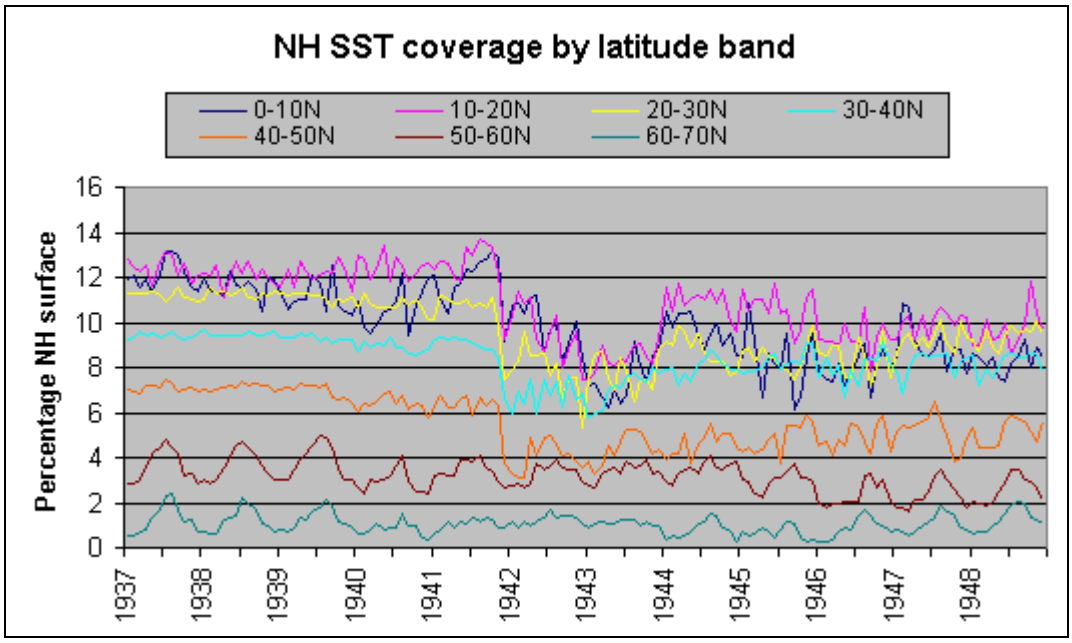


Figure 7.4 Northern Hemisphere coverage by latitude band 1935-1946

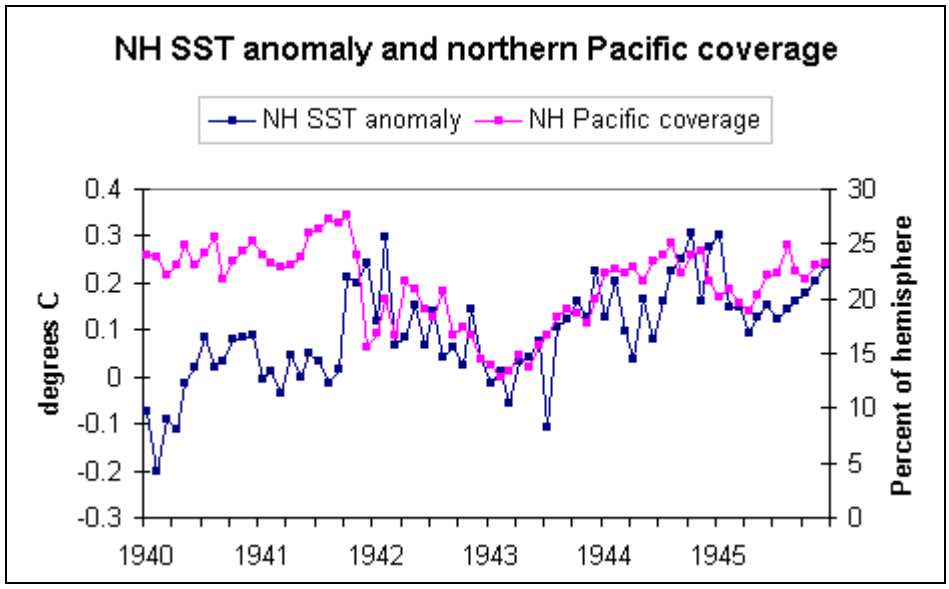


Figure 7.5 Northern Hemisphere SST coverage of the Pacific Ocean, expressed as percentage of total hemisphere surface area, from 1940 to 1945 (labels at January position) showing sharp decrease in coverage in December 1941 and almost simultaneous change in NH SST.

The situation in the Southern Hemisphere is similar but this time with a shift in SST coverage in late 1939 and again in 1946 (Figure 7.6). The coverage in August 1945 was 13.68% of the hemisphere and in September just 11.60% but 12 months later

these were in 41.58% and 32.99% respectively as shipping resumed after the war. The increases occurred in all latitude bands (Figure 7.6), with substantial increase in coverage in bands centred at 22.5S and beyond. During 1945 but prior to August the coverage in the latitude band 40-50S was less than 1%, and virtually nil for some months up to August, after which it expanded rapidly, exceeding 2% by the end of the year.

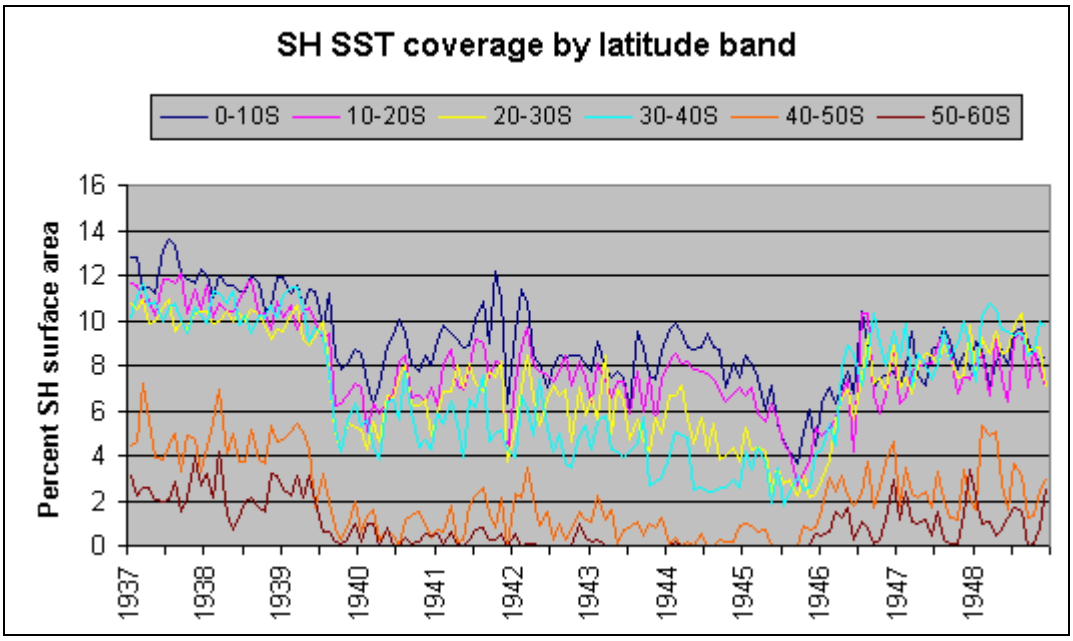


Figure 7.6 Southern hemisphere coverage by latitude band 1937-1948

Not surprisingly this expansion of SH coverage can be attributed to the war in the Pacific ending on September 2, 1945 and shipping routes reopening. During 1945 shipping was very much along the coasts (Australia, South Africa, South America) but after the war ended, shipping on the route from South Africa to Australia recommenced as did shipping from Australian and New Zealand to Cape Horn, these routes traversing the ocean where warmer waters meet Antarctic waters and SSTs are variable.

The sources cited earlier in this section focus on only the two methods of measuring sea surface temperature and the estimated relative proportions of the methods. This discussion of coverage has shown that changes in coverage and the locations of those

changes might go a long way towards accounting for the reported abrupt shifts in sea surface temperatures in both hemispheres. Further, if SSTs prior to 1941 have been adjusted upwards excessively it might make quite a difference to the earlier SST pattern.

Kennedy et al (2011b) talk at length about the assumptions involved with assigning SST measurements to either buckets or engine room intakes. It then goes on to describe how 100 different datasets, each based on different assumptions about uncertainties, were developed to explore the impacts of different assumptions and how the "ensemble mean" of these datasets is commonly used. This has two problems. Firstly the "ensemble mean" is derived from datasets that are constructed according to assumptions that vary between the datasets, which means that if any one set of assumptions was correct its data is averaged not only with incorrect data but even with the data based on assumptions that might differ quite substantially from it. Secondly, as indicated above in the discussion about 1942 and 1946 data, the magnitude of the SST adjustments might be quite incorrect.

7.5 The inherent problem of thermal layering in the ocean

Few, if any, papers related to HadSST3 data seem to take into account some significant problems with estimating the sea surface temperature when the thermal layers of the ocean vary over time, both in depth and according to the temperature of each layer.

Figure 7.7 is copied from Donlon et al (2005). Its legend, focussed on sea surface temperatures, is summarised as follows:
STT_{int} is the interface of atmosphere and ocean,
SST_{skin} is the skin layer,
$SST_{subskin}$ is below the level at which heat is lost by evaporation,
SST_{depth} is "an in situ measurement near the surface of the ocean that is typically reported simply as SST " (Donlon et al, 2005), and

SST_{fnd} is the "foundation" temperature obtained from below the layer that varies diurnally.

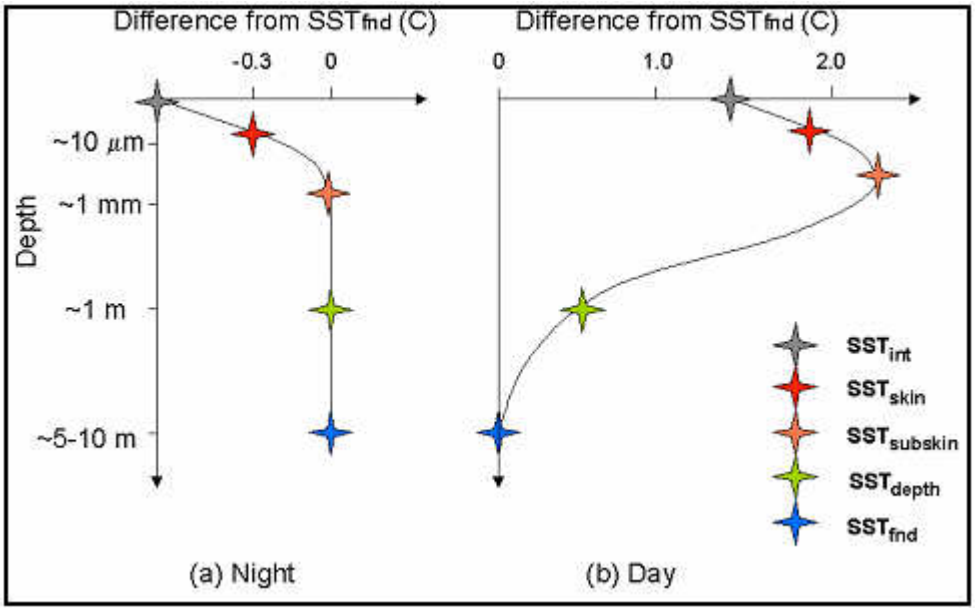


Figure 7.7 (Copied from Donlon et al, 2005) Schematic of the variation in temperature with water depth. See text for explanation of the legend.

Figure 7.7 is a schematic but it indicates the typically SST measurement, presumably taken by bucket, as being at 1m depth whereas other sources (e.g. Webster et al, 1996) give the depth as 500mm. It is also greatly generalised and gives no indication that surface wind or other turbulence, plus the amount of solar radiation will vary the relationship between depth and temperature on a diurnal basis and residual heat might be carried over from one day to the next.

The top layer is the sea surface skin layer (Woodcock, 1941; Ewing & MacAlister, 1960; Hasse, 1963; Fedorov & Ginsburg, 1992; Mobasheri, 1995; Fairall et al, 1996; Wells et al, 2009). This is the layer where heat is lost to the atmosphere (evaporation, sensible heat loss and infrared radiation) and in some cases gained (eg. radiation in some bands is absorbed within microns of the surface). The layer is typically 1 or 2 millimetres deep but exceeding 3mm is not unusual and instances of 6mm have been reported (Mobasheri, 1995). Webster et al (1996) and Fairall et al (1996) report instances in the tropics where the temperature difference between the surface and the

layer beneath the skin was as much as 3.0°C, Fairall et al (1996) indicating that this was during the daytime but at night 0.1°C to 0.3°C was more likely.

The direction and rate of heat exchange are according to the current temperature difference and the magnitude of incoming or outgoing radiation, meaning that the heat transfer differs during sunlight and darkness, seasons/latitude and the absence or presence of clouds. Windspeed is also a significant factor because up to a certain point it will enhance evaporation and conduction, but beyond that will cause over-turning of the ocean surface and therefore mixing with the waters below the surface skin.

Measurement of sea surface temperature can involve passing a bucket through this layer, which might collect a substantial amount of this layer water especially if the angle of the bucket is approaching horizontal as it enters the water, as could easily be the case with moving ships. Matthews and Matthews (2013) reported that when drawing water samples using different types of buckets it was found that rubber buckets would enter the water "*near-vertically and did not need to be dragged like the wood and canvas buckets. The canvas buckets tended to close flat when dragged and so not fill while the wood buckets would bounce along the surface when under-motor. ... Retrieval of the wood and canvas buckets became difficult if too much line was released and they drifted far back towards the stern [of the ship]*."

This skin layer is the only one "seen" by satellite-based SST measurement, which means an adjustment is necessary to bring them into line with the temperature that might have been measured below that surface skin but within 500mm of the surface. A web page from the Hadley Centre[21] discusses this skin in the context of converting satellite-based measurements of sea surface temperature to values ostensibly equivalent to what would have been measured by bucket.

The layer beneath the subskin layer of Figure 7.7 is sometimes referred to the "warm layer", which some papers (e.g. Fairall et al, 1996; Webster et al, 1996; Soloviev & Lukas, 1997; Kawai & Wada, 2007) refer to as a "diurnal warm layer". Fairall et al

[21] http://research.metoffice.gov.uk/research/nwp/satellite/infrared/sst/conversion.html

(1996) say that roughly half of the insolation received from the sun is absorbed in the first 2 metres of water, which on a daily average of 500 w/m^2 is sufficient to heat the 2m deep layer uniformly by 2.0°C. Soloviev & Lukas (1997) shows a composite of 5 temperature profiles taken during calm weather during the TOGA COARE project, with a temperature difference of almost 3°C between depths of about 200mm down to 1 metre. Kawai & Wada (2007) show an example of a diurnal temperature variation of 4.5°C in a warm layer 200mm to 450mm below the surface.

Matthews and Matthews (2013) report that during a voyage from Hawaii to Tahiti the temperature gradients from depths of 0 to 3 metres averaged 0.4°C±0.2°C and were 0.5°C±0.2°C during the day but 0.3°C±0.1°C at night, and goes on to say "*Evidently the near surface thermocline did not breakdown overnight.*" The temperature gradient was greatest in the early afternoon and weakest overnight. The gradient was also greater in regions where the south Equatorial Current was stronger.

Farrar et al (2007) show a figure indicating temperatures at different depths off Martha's Vineyard in the USA, with negligible difference until about 2:00pm but then an abrupt temperature increase of about 2.0°C at 1.0m depth while the temperature at 2.2m remained unchanged until about 22:00 hours (Figure 7.7).

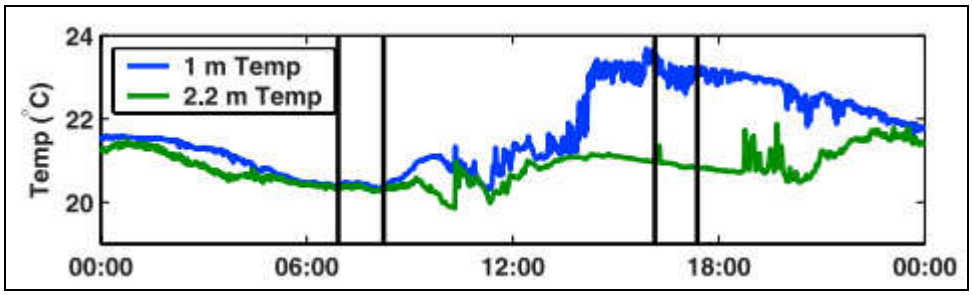


Figure 7.7 copy of Figure 5 from Farrar et al (2007) showing temperatures at two depths off Martha's Vineyard on 15 August 2003. The divergence between the temperatures after about 2:00pm is very obvious and the whole shows that the relationship between SSTs measured at the two depths is inconsistent.

If calm conditions with high insolation persist the warm layer becomes hotter and it deepens, with not all the heat appreciably dissipating at night. The development of El Nino conditions can be described in similar terms and Wyrtki (1989) points out that in

the Pacific warm pool, under calm conditions with high insolation, a warm layer can easily extend to a depth of 80 metres. From a sea surface measurement perspective this warm layer is a challenge, Kawai & Wada (2007) says "The large temperature difference between the sea surface and about 1m depth, where ships and buoys usually measure the seawater temperature as 'SST', has been recognized as one of the major sources of error in satellite-derived SST." As noted above, satellite-based observations are of the sea surface and the data are often adjusted in an attempt to bring them into line with bucket-based measurements that draw water from a depth of about 0.5 metres. The variable temperature gradient between the surface and that depth is the result of insolation over time and water mixing due to turbulence and bow wave, mean that estimating a 0.5 metre temperature from surface measurements is far from simple.

The above description applies to calm weather conditions and high insolation. Changes in the wind speed or insolation (due to the diurnal cycle, changes in cloud cover or in latitude) will modify the layering. Under thick cloud the warm layer will be shallower and of lower temperature, meaning that the thermal layering will differ from that under clear skies. Under windy conditions the surface turbulence will cause mixing of the upper layers, homogenising their temperatures, and the removal of heat from the sea surface, meaning a temperature profile shifted towards a vertical line. Ultimately in the high latitudes, where insolation is low, the schematic described in Figure 7.4 will become virtually a straight line with the temperature at the sea surface being almost identical to deeper temperatures.

El Nino conditions in the Pacific are widely regarded as occurring due to a pause in the easterly trade winds across the ocean, the calm ocean causing the thermal layering described above. HadSST3 global average temperature anomalies increased after 1976 (see chapter 1, figure 1.2), which coincided with a switch in the El Nino-Southern Oscillation (ENSO), the average Troup[22] SOI for the 25 years[23] prior (i.e. 1952-1976) being 2.04 and for the 25 years after -2.88, with similar averages over 15,

[22] The alternative index, the multivariate ENSO index, or MEI, includes sea surface temperature in the Nino 3.4 region so to use the MEI here would be circular reasoning.
[23] Under the Troup SOI, values greater than 7 for 3 months or more generally indicate La Nina conditions and less than -7 for a similar time indicate El Nino conditions.

20 and 20 years. While neither of these averages is beyond the La Nina or El Nino event threshold they respectively indicate a propensity for those conditions. The open question is therefore whether the shift towards more frequent El Nino events coupled with incorrect assumptions about the relationship between SST measurements made at different depths artificially increased sea surface temperatures, particularly in the tropical Pacific where the ENSO influence is greatest.

In summary, thermal layering of the ocean and the establishment of various temperature gradients across the depths at which sea surface temperatures are measured make it difficult to accurately convert temperatures measured, using any specific methodology, at a particular depth to equivalent temperatures that, conceptually at least, would have been measured by another methodology at a different depth.

7.6 Ships in port when recording sea surface temperature

Brohan et al (2009) discusses the transcription of temperature observations in the log books of the Royal Navy into the ICOADS database, the source of HadSST3 and HadCRUT4 sea surface temperature data.

Samples of that data were available from Brohan's web page[24] and data for the 69 ships whose names commenced with 'A', 'B' or 'C', was downloaded for analysis. Of the 253,600 recordings of SST data a total of 127,846 recordings (50.4%) were made while the ship was in port (i.e. a non-blank 'port' field in the record) leaving only 125,754 recordings made (49.6%) with the ship at sea (assuming that the field would not be left blank while the ship was in port which might not be true if it remained there for many days).

As an example, ICOADS data shows that ship "ABERDARE", one of those downloaded as described just above, made 32 SST observations between 24th and

[24] See http://brohan.org/hadobs/digitised_obs/docs/

30th of September 1939 (inclusive) at latitude 1.4°N longitude 103.8°E, which according to the downloaded data file for ABERDARE is "RN Base Singapore". The ICOADS search, covering a box of latitudes 0°N to 2°N and longitudes 103°E to 105°E over the period from 24 to 30 September 1939, revealed 11 ships of the Royal Navy making observations at or very close to the same location, with just 86 of the 445 observations made by those ships over the 7 days apparently outside the port but within the given HadSST3 grid cell.

In ports the water is shallower, likely fed by local natural and manmade discharges, sheltered from currents and often from wind, with far less mixing of surface water with deeper water; and on the basis of these factors the thermal layering will be different. Sea surface temperatures are therefore likely to be higher in ports than in the open sea and are therefore likely not representative of the grid cell.

The data in question was from ships of the Royal Navy during the years of World War II, which might spend considerable time in port depending on requirements for their deployment or the action they were undertaking. Commercial shipping is likely to spend less time in port on the basis that it is paid to move cargo. Without accessing a very large part of the International Comprehensive Ocean-Atmosphere Data Set (ICOADS) database, on which the HadSST3 dataset is based, it is not possible to determine the magnitude of any "in-port" temperature bias in the HadSST3 dataset but the recent additions from the logs of Royal Navy ships certainly have a problem.

7.7 Errors in transcription of data from ships' logs

Data from Brohan et al (2009) was also found to have flaws in its transcription from handwritten logs from Royal Navy ships for the periods 1938-39 and 1941-47 into digital form for inclusion in the ICOADS and HadSST dataset. It appears that at times the automatic digitising (apparently by character recognition software) failed to correctly distinguish between similar digits, such as '1' and '7', and '4' and '9'. Figure 4 of Brohan et al (2009) provides an illustration of the latter when it shows details from ship HMS Warspite on January 12 1941 (Figure 7.8 and Table 7-1).

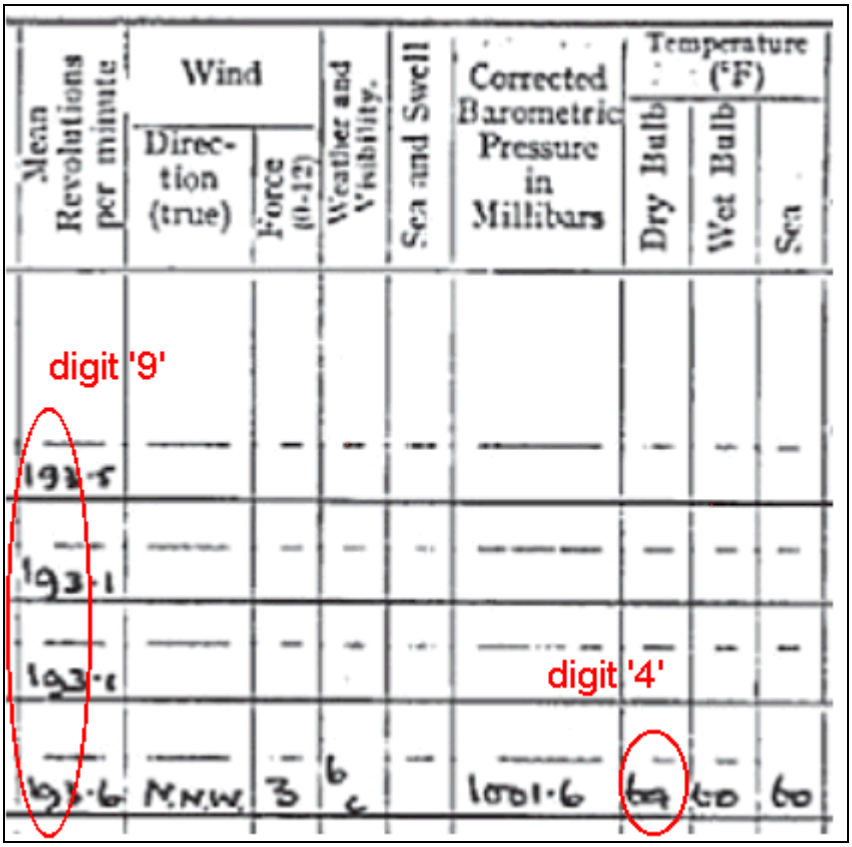


Figure 7.8 Enlarged extract of Figure 4 from Brohan et al (2009), with several instances of the digit '9' indicated and one of the digit '4', the latter shown in the table below to be misinterpreted as a '9'.

YearMnDyTime	Lat	Long		Location	WDir		DB	WB	SS
1941 1111800	3442	1948	0204 310WARSPITE		20SE	04bc 061110038	064	060	60
1941 1112200			020 310WARSPITE		24SE	03c 061110025	065	059	60
1941 112 200			020 310WARSPITE		04NNW	03c 06 10016	**069**	060	60
1941 112 600	3472	2317	0204 310WARSPITE		08NNW	06c 062110023	063	061	61
1941 1121000	3410	2427	0204 310WARSPITE		12WNW	06bc073110053	065	060	60
1941 1121400			020 310WARSPITE		16NW	07bc 06221008	069	061	61
1941 1121800	3288	2647	0204 310WARSPITE		20W	06b 062210104	063	060	63
1941 1122200			020 310WARSPITE		24WNW	06bc072210113	069	060	60

Table 7-1 The digitised data (downloaded from ICOADS) corresponding to the data in Figure 7.9, with the relevant data from the last line of the extract of the ship's log underlined and, in bold, the dry bulb temperature with the '4' incorrectly identified as a '9'.

The above problem also occurred with the latitude and longitude of the ships, in some cases placing ships well inland from the sea. Brohan has subsequently stated that

various errors with the latitude and longitude have been corrected by quality control processing but it seems that the less obvious errors in temperature data persist.

The error described here is a single instance of an ICOADS error and while it means little in itself it raises the question of the accuracy of the transcription of written records into the ICOADS database.

7.8 Other possible discrepancies in ICOADS

As mentioned above, the ship "Warspite" was in the south-east of the Mediterranean Sea on 12 January 1941. Data from the ICOADS database shows that other ships were nearby. According to data from "Warspite", at 2pm the air temperature was 20.0°C whereas ship "Barham", just 0.07° latitude north and 0.07° longitude west away (approximately 3km across open sea), recorded the air temperature as 16.1°C. A few days later, on 14 January 1941, in the same area of the Mediterranean ships "Eagles" and "Barham" reported 6pm air temperatures of 21.1°C and 16.7°C respectively when "Eagles" location was given as 34.85°N 20.83°E and "Barham" at 34.87°N 20.90°E, which puts the vessels even closer than 3km.

Other simultaneous observations made by "Eagle" and "Barham" are shown in Table 7-2, where pairs of readings at the same time are grouped. Some of the differences in recordings made from ships just a short distance apart[25] are surprising:

- Sea level pressures differ greatly when the ships are very close (e.g. 1005.8 hPa v. 1017.1hPa at 16:00hrs)
- Air temperatures are sometimes more than 3°C different when the ships are close but with wind supposedly from different directions (at 10:00) but even with wind from the same direction one is 16.7°C while the other 18.3°C (at 16:00hrs)
- Wet bulb temperatures are consistently different at 16:00 and 18:00 hrs

[25] At this latitude and longitude a difference of 1° latitude is 100km and 1° longitude is 92 km.

- The sea surface temperature from ship "Barham" fell by almost 3.5°C between 16:00 and 18:00 hrs but for ship "Eagle" only by 0.5°C, with a difference of 0.6°C at 16:00hrs becoming a difference of 2.3°C at 18:00. (Both SST's at 18:00 hrs appear to be incorrect repeats of wet-bulb temperatures.)

On the basis of this data the accuracy of the ICOADS database is open to question.

Day	Hour	Ship	Latitude (N)	Longitude (E)	Wind direction	Wind speed (m/s)	Sea Level Pressure (hPa)	Air Temp (℃)	Wet Bulb Temp (℃)	Sea Surface Temp (℃)
12	6	BARHAM	34.60	23.20	225	6.7	1004.3	16.7	15	17.8
12	6	EAGLE	34.65	23.13	248	6.7	1003.6	18.3	15.6	17.2
12	10	BARHAM	35.15	23.22	270	6.7	1004.9	15.6	13.3	18.3
12	10	EAGLE	35.12	23.12	315	6.7	1001.2	18.9	17.2	17.2
12	14	BARHAM	35.29	23.32	293	6.7	1005.6	16.7	14.4	17.8
12	14	EAGLE	35.26	23.24	293	6.7	1017.1	16.7	16.1	17.2
12	16	BARHAM	35.35	23.36	293	6.7	1005.8	16.7	14.4	17.8
12	16	EAGLE	35.33	23.30	293	6.7	1017.1	18.3	16.7	17.2
12	18	BARHAM	35.42	23.41	293	4.6	1008.9	15.6	14.4	14.4
12	18	EAGLE	35.40	23.36	293	4.6	1017.2	17.2	16.7	16.7
12	22	BARHAM	35.56	23.51	315	2.6	1009.2	16.1	15	14.4
12	22	EAGLE	35.55	23.49	248	2.6	1007.6	16.7	16.1	15.6
13	2	BARHAM	35.69	23.60	315	4.6		17.2	15.6	16.7
13	2	EAGLE	35.69	23.61	293	2.6		15.6	14.4	17.2
13	6	BARHAM	35.83	23.70	293	9.3		12.8	11.1	16.7
13	6	EAGLE	35.83	23.73				15.6	13.9	16.1

Table 7-2 Extract from ICOADS database for two ships in close proximity on 12 and 13 January 1941. The differences in data for two ships less than 100km apart are sometimes quite great (e.g. sea level air pressure 1005.8hPa to 1017.1hPa at the same time of day).

Further, ICOADS entries often provide latitude and longitude co-ordinates to one decimal place. This is insufficient to precisely pinpoint the location in which the observations were made and it requires some allowance when attempting to pinpoint locations near coasts. Despite this, when ICOADS data was accessed for a study of SSTs along the Great Barrier Reef (see chapter 11) it was found that in several cases the co-ordinates were further inland than can be explained by rounding to one decimal place. Some these erroneous locations are shown in Table 7-3.

Latitude	Longitude	Distance inland	Coast longitude
-12.9	142.5	100 km	143.3
-13.5	142.5	125 km	143.5
-14.6	143.2	75 km	143.8
-15.6	144.0	300 km	145.3
-15.7	144.3	125 km	145.3
-18.5	145.4	125 km	146.3
-19.8	146.0	100km (SW)	146.8 (Lat -19.3)
-21.8	148.4	250km	149.4

Table 7-3 Example instances of ICOADS co-ordinates defining inland locations rather than locations at sea. (Distances to sea are approximate via scaling of distances.)

The examples shown in this section of ICOADS meteorological data varying considerably for ships only a short distance away, poor transcriptions of ship logs and finally ship's co-ordinates mapping to land areas rather than sea indicate that the quality control of ICOADS data is far from ideal. When that data is incorrect so too, to some extent, will be the HadSST3 and HadCRUT4 datasets. Whether the errors are sufficient to shift the hemispheric or global average temperature anomalies is unknowable without a very large amount of work and even then some assumptions will probably be required.

7.9 Summary

In December 2015 SST data was available from 69% of the Earth's surface and data from observation stations accounted for 31%, with coastal grid cells that have data from both sources covering for about 15% of the Earth's surface (see section 5.3). HadCRUT4 coverage in the same month totalled 84% of the Earth's surface, which means that SST data accounts for 82% of the coverage - more than three-quarters - in that month either directly or when merged with data from observation stations as is the case with the coastal grid cells.

Accurate sea surface temperature data appears unlikely given the variety of methods of measurement, assumptions about adjustments to those temperatures and uncertainties associated with the physical environment. These are listed in point form below.

(a) Methodological issues

Over the time of the HadSST3 dataset the methods of measuring sea surface temperature have been the following:

Canvas bucket	Leather bucket
Wooden bucket	Leather and metal bucket
Rubber bucket	Insulated canvas bucket
Iron/tin buckets	Floating buoy (moored or drifting)
Hull-mounted sensor	Engine room intake
Argo Buoy;	

- The methodology used for each measurement was not always recorded;

- There is no consistency in the instructions given by different agencies for the use of basically the same equipment and method;

- Hull-mounted sensors and thermometers on engine room cooling intakes record temperatures at a greater depth (to 10m or slightly more) than bucket-based measurements and buoys;

- A variety of thermometer types with different scale graduations have been used, meaning differing degrees of rounding or interpolation and therefore not necessarily accurate for a single measurement or consistently accurate over multiple measurments;

- The types of buckets reportedly differ in the kind of sample they take, some more prone to sampling more water from the sea skin than others.

(b) Adjustment issues

- Various papers indicate that SST data should be adjusted to take into account a variety of factors including depth, bucket type, local weather conditions while lifting samples onto the ship and the initial temperature of the bucket, but there is little consistency in the magnitude of the indicated adjustments.

- Broad-scale adjustments have been recommended in various papers on the basis that global or hemispheric average SSTs are supposedly incorrect and that false assumptions have been made about the relative proportions of the use of certain methodologies or the adjustments associated with each.

It has been shown in this chapter that broad-scale adjustments based on SST shifts in December 1941 in particular might be incorrect because the assumption has been that the shifts were due solely to switch in the dominant methodology for obtaining sea surface temperatures rather than the significant change in coverage that was a consequence of changes to shipping patterns during World War II.

(c) Environmental issues

- The ocean has a general series of layers – skin, warm layer, semi-permanent deeper layer - with gradients between them. The warm layer can often be divided into an

upper diurnal warm layer that varies over 24 hours and a lower more permanent warm layer that loses less heat at night and might therefore accumulate heat over some period.

- The layers are dynamic and vary in temperature and depth.

- Taken as a whole the layer structure varies in amplitude (i.e. temperature range) and total depth according to latitude, season, time of day, current weather (insolation, wind, rainfall) and weather in the recent past (varies from days to months). The amplitude is typically very small in high latitudes but often much greater at low latitudes in calm, clear-sky conditions.

- The layering is likely to be different in ports because they are sheltered and have relatively shallow water, which means that SST measurements taken while a ship is in port could be quite different to a temperature made in the open sea.

- Layering in shallow water, such as on a continental shelf might also differ from layering in deep water.

With the above uncertainties the SST data provided to the Hadley Centre for the HadSST3 and HadCRUT4 datasets seems likely to be inaccurate by an amount that cannot be determined because the magnitude of the errors in the various assumptions is both unknown and highly variable.

The HadSST3 dataset attempts to counteract these uncertainties by creating 100 different realizations of the data (i.e. multiple variants of the dataset), each based on different assumptions about measurement techniques when they were not recorded. It seems inevitable that any conflicting assumptions in those realizations will make some realizations redundant. It is also impossible to know if particular realization is in fact correct, this even before it is averaged with all other realizations to produce the "ensemble mean" dataset that has become a de facto standard.

With 82% of the HadCRUT4 gridded dataset for December 2015 derived from SST measurements the many uncertainties about SST data are transferred into the HadCRUT4 dataset with little dilution.

Chapter 8: Summary of Part 1

8.1 Introduction

This thesis has discussed many issues regarding the accuracy of the data in the HadCRUT4 dataset and therefore on information derived from it. It has only been a "first pass" audit of the data and its processing but has identified some obvious errors and numerous uncertainties that together question the accuracy of the dataset. Jones (2016) (see Appendix 2) concurs with several points that have been raised here but makes no mention of many others.

The scope of the uncertainties spans the entire range of activities and processes from individual temperature measurements on land or at sea, to the calculation of global average temperature anomalies when global coverage is variable and has never been 100%.

Some of the uncertainties translate into errors that are likely to be distributed relatively evenly above and below the given values but some suggest bias in one direction for some portion of the record (e.g. European dominance of early coverage) or even all of the record when the long-term average temperatures, from which anomalies are calculated, are incorrect.

Adjustments to individual values for most of the uncertainties will have an indistinguishable impact on global or hemispheric average temperature anomalies because the change will occur many decimal places below the precision of those averages. This won't necessarily be true as the number of corrected errors increases or with an increasing proportion of corrected data relative to the data pool being studied (e.g. to data for individual grid cells or observation stations).

This chapter will reiterate the many findings made in earlier chapters and propose an alternative approach that would use as much of the historical data as practicable to create a revised temperature dataset with fewer uncertainties.

8.2 Major findings

Many issues uncovered by this analysis will only directly impact the data from individual observation stations (in the case of CRUTEM4 station data) or individual grid cells (in the case of HadSST3 data), and might impact a single month, a series of months but also impact the temperature anomaly trend incorporating the specific data. Some issues are however large scale and impact the accuracy of a larger data pool (e.g. global averages) or at least extend the error margin. This section (8.1) will discuss these large scale issues or issues with an implied potential for greater influence.

8.2.1 Coverage Issues

According to the HadCRUT4 method of calculating coverage that coverage has varied greatly over time, generally increasing except in the times of the two World Wars but particularly notable for its lack of homogeneity.

1. Coverage Southern Hemisphere coverage was not consistently above 50% until the early 1950's and global coverage not consistently above 75% until about 1960.

2. In the late nineteenth century grid cells in certain latitude and longitude bands in each hemisphere made a much greater percentage contribution to the hemispheric coverage at that time than their proportionate area (i.e. more data available from these grid cells than from grid cells in other bands). In the Southern Hemisphere for example, the 20% of the hemisphere that lies between latitudes 30S and 50S accounted for about 45% the hemisphere's

coverage from 1850 to 1914. By about 1950 the Northern Hemisphere coverage became homogenous and while the Southern Hemisphere is approaching that state it has still not done so completely.

3. Also in the late nineteenth century but on regional rather than band basis, HadCRUT4 grid cells for western European and the nearby Atlantic Ocean, accounted for more than 60% of Northern Hemisphere coverage at that time despite being only 12.5% of the surface area of the hemisphere. (For CRUTEM4, using only data from observation stations, the continent of Europe, which covers ~9.3% of the hemisphere land area at times accounted for more than 50% of the CRUTEM4 Northern Hemisphere coverage.)

4. In general terms, late nineteenth century data was heavily focussed on Europe, its colonies and trading centres in other parts of the world, and the shipping routes between those points. Some HadCRUT4 grid cells in the central Pacific for example did not consistently report data until the 1970s.

5. The variation in coverage over time is a potential source of systemic (rather than random) errors that the process of averaging cannot remove. If temperature variation trends are not uniform across the globe changes in coverage will potentially cause a misrepresentation of the global average trend. If, for example, the Northern Hemisphere recovery from the Little Ice Age was not homogenous in the late 1800's in particular then the inclusion, if it was possible, of data from where data for that time is currently unavailable might alter the Northern Hemisphere and global average temperature anomalies. HadCRUT4 global averages rest on the implicit assumption that if data was available from other locations it would not alter those averages regardless of how many locations supplied data and, in terms of trends, that temperature anomalies varied across the globe in a manner consistent with the available data.

8.2.2 Sample size

The sample size affects the amount of interpolation that might need to be done and the square root of the sample size is often used as the denominator when calculating the error margin, which means the greater the sample size the less the error margin. The following findings are relevant in this regard:

6. While the number of reporting stations in the Northern Hemisphere in 1850 was 145 in the Southern Hemisphere only a single station reported data until the start of 1853, and the number only grew to 9 in the first decade of data.

7. In 1850 grid cells with a single observation station accounted for an average of 70% of all reporting CRUTEM4 grid cells. This fell below 45% in 1891 but continued to be above 38% until 1951. Grid cells with 1, 2 or 3 observation stations accounted for 92% of reporting grid cells in 1850 and reached its minimum of 52.7% in 1974. Unless the geography and prevailing weather conditions are very consistent over the entire grid cell the data from a single observation station seems unlikely to be a good representation of the entire cell. In the case of 2 or 3 stations any failure to report could alter the grid cell average value so the inconsistency of the number of stations is also important.

8. HadSST3 and HadCRUT4 datasets are based on a grid cell size of 5° latitude x 5° longitude and intervals of one month but HadSST3 data is developed from a dataset at a grid size of 1° x 1° using pentads (5-day periods). A minimum of 150 observations in the month in the 5° x 5° grid cell would be required in order to provide each 1° x 1° grid cell in each pentad with data from a single observation, but many of the HadSST3 5° x 5° grid cells have fewer than 15 observations per month.

9. Grid cells with from 1 to 5 SST observations in the month account for at least 33% of all grid cells with SST observation from 1850 to 1955 except for three brief periods, none exceeding a decade. Grid cells with from 1 to 15 observations (i.e. group 1 to 5 and group 6 to 15) accounted for 93.8% of reporting SST grid cells in 1850 and did not fall below 33% until 1995.

8.2.3 Long-term average temperatures

Long-term average temperatures are used to determine temperature anomalies. Trends and differences in anomalies for any observation station (CRUTEM4) or grid cell (HadSST3) will be the same for any consistent long-term average temperature even if it is erroneous but problems arise when merging or averaging anomalies calculated from different bases. Findings in relation to the calculation of long-term average temperatures are as follows:

10. The period from 1961 to 1990 inclusive over which long-term averages are calculated is climatically abnormal because of various volcanic eruptions in and around the Pacific Ocean and because of the Pacific Climate Shift of 1976. The absence of data across part of this period could easily impact the long-term average temperature more than a similar absence of data from a more climatically stable period.

11. The acceptance of as few as 14 years of data from observation stations for the calculation of long-term (i.e. 1961-1990) average temperatures - 14 years as a whole not 14 years for the given calendar month - is well below the WMO recommendation of a minimum of 80% be present and that no data sequence of more than three values be missing. In fact the minimum sample size set by the CRU is less than 50% of the maximum possible data.

12. The standard deviations associated with long-term average temperatures for observation stations appears to be inversely related to that temperature, and when considered in bulk the range of standard deviations for a given mean temperature widens as the mean temperature decreases. This warrants further study because it suggests the temperature range at observations stations in the same grid cell but with different average temperatures might have a different natural range of temperatures and therefore a different range of temperature anomalies.

13. HadSST3 long-term averages are first estimated on a 1° x 1° grid cell size and using pentads (5-day periods) and then modified according to SST observations in those 1 x 1 grid cells and pentads. When considered as grid cell and month combinations (where each cell is 5° x 5° and the combination with calendar months means 12 entries per cell), 4381 combinations (21% of the total with any data in any year 1850-2015) have less than 14 years of data for the given month over the period 1961-1990 and 1006 of these had a single instance of data in that month. These low sample sizes, at the 5° x 5° level, make any modified estimates of long-term average SSTs at the 1° x 1° and pentad level quite uncertain.

8.2.4 Outliers present in data

Data outliers will cause distortion at some scale, perhaps not with global or hemispheric averages but they will when the focus is on a small number of grid cells and/or a smaller period of time. Several issues with outliers are identified:

14. Data for CRUTEM4 is regarded as an outlier if it is more than five standard deviations from the long-term average temperature (the two factors strangely being calculated over a different time periods). The "five standard deviation" threshold is excessively generous because on the basis of Normal Distribution the probability of data exceeding this range is 1 in more than 1 million.

15. Outliers, some of which are obvious errors, are present in the CRUTEM4 station data across the period used to calculate the mean temperature and across the different period used to calculate standard deviations. When outliers increase standard deviations they automatically extend the threshold for identifying other data outliers, at times producing ridiculous acceptable ranges of values.

16. Outliers were also found in the SST data. When using HadSST3 data for the period from 1961 to 1990 to calculate mean temperatures and standard

deviations for each grid cell and month one outlier was identified at more than 31 standard deviations from the mean temperature. Changing the period to cover the entire range (i.e. 1850 to 2015) when calculating the two factors reduced the range of standard deviations from the mean but still revealed instances of more than 7 standard deviations from the mean.

8.2.5 Errors in ancillary data files

During this study errors were discovered in three ancillary files published with the HadCRUT4 dataset. None of these errors directly impacted the main file of gridded temperature anomalies but they could be an issue when using the data in these files.

17. Data processing errors identified in ancillary files are as follows:

    (i) Two summary data files, one file for each hemisphere, contain monthly average sea surface temperature data and hemispheric coverage and while each contained the correct temperature data, they contained the coverage for the other hemisphere.

    (ii) The file containing the number of observations in each cell of a global grid had its monthly data in reverse order to the main temperature file (i.e. listed in order from latitude 90°S to 90°N rather than 90°N to 90°S).

    (iii) The same file as for (ii) above had fields where the data was too large for the number of bytes allocated, these fields containing '*******', which is the "field overflow" indicator for the Fortran programming language.

In all likelihood these problems have existed since 2012, when HadCRUT4 replaced the earlier version, HadCRUT3 and perhaps even before then. In February 2016 these errors were reported by the author to the Climatic Research Unit and the Hadley

Centre, both of which promptly corrected the files and advised users of the changes[26]. Whether more subtle data processing errors exist in other files is unknown but attention to detail appears to be missing.

8.2.6 Other issues

Several other inconsistencies, unexplained differences and uncertainties were identified in the HadCRUT4, CRUTEM4 and HadSST3 datasets. Some are few in number of instances but cast doubt on the overall quality of the processing of data and documentation associated with it.

18. Despite the HadCRUT4 dataset being created from the same data as the CRUTEM4 and HadSST3 datasets the HadCRUT4 grid cell data sometimes differs inexplicably from CRUTEM4 and HadSST3 data even after allowing ±0.1°C for the ensemble nature of HadCRUT4.
    a. Of just over 2 million instances of either CRUTEM4 or HadSST3 but not both having data for a given grid cell and month the HadCRUT4 value differs by more than 1.0C from the single other dataset in 34,435 instances, all but one being differences from CRUTEM4 data.
    b. In two instances the HadCRUT4 dataset contains data for a given grid cell and month and yet neither CRUTEM4 nor HadSST3 have data for that cell and month.
    c. When an even wider HadCRUT4 allowance of ±0.125°C was applied four instances of HadCRUT4 values were found to be outside the range of the two values given by the CRUTEM4 and HadSST3 datasets for the same grid cell and month.

19. Annual average CRUTEM4 and HadSST3 global average temperature anomalies differ with SST averages generally being greater than CRUTEM4 averages in the late nineteenth century but generally less than CRUTEM4 in the last 30 years (i.e. 1986-2015). As discussed earlier, the difference in the

[26] See above subheading "File Formats" on web page https://crudata.uea.ac.uk/cru/data/temperature/ and at top of http://hadobs.metoffice.com/hadsst3/

early period might be due either to the bias towards European data (see above in this section), to incorrect downward adjustment of observation station data (see below this section) or to the incorrect bulk upward adjustment of SST data (ditto) but the later (1986-2015) lacks a good explanation.

20. The data sources for coastal grid cells, i.e. observation stations or sea surface temperature measurements, have changed over time for most coastal grid cells with only about 17% of HadCRUT4 data being from a consistent source. This lack of consistency can cause temperature anomaly trends that might not be present if the data was from a consistent source and, given that coastal grid cells account for 28% of the Earth's surface, the impact on the global average temperature anomaly might be significant.

21. If step-wise adjustments are made to observation station data to remove the influence of urbanisation there is a very real danger that earlier data will be excessively adjusted downwards, this causing falsely exaggerated upward trends. Given that adjustments are far more likely to be decreases than increases this would mean a systemic bias.

22. The CRUTEM4 station metadata supplied by the CRU gives no indication of any corrections or adjustments that have been made to the data either by the national meteorological service (NMS) supplying the data or by the CRU itself. This severely limits any thorough audit but the presence of outliers and obvious errors in the data indicate failings in quality control both by the NMSs and the CRU.

23. The different methodologies for obtaining sea surface temperatures are well known but this thesis has identified that the bulk adjustments to SST data to take into account those changes in methodology, derived from sudden shifts in SST averages, could be excessive because changes in coverage can account for part or all of those shifts in SST.

24. The adjustments to SST recordings to take into account the depth into the water are questionable given the thermal layering that irregularly occurs particularly in calm tropical waters.

25. The ICOADS database, from which SST data is obtained for the creation of HadSST3 and HadCRUT4, contains instances of temperatures measured while ships were in port, which typically means in shallow, calm water, sheltered from winds and ocean currents, and potentially in areas where rivers, stream and drainage will deposit local water that might be at different temperature to the ocean. The number of such measurements across the entire ICOADS database is unknown. This problem could give rise to systematic errors (offset biases), as the amount of data recorded while ships were in port is likely to have changed over time.

26. The presence of erroneous transcriptions of hand-written data, questionable differences in simultaneous observations made by ships less than 100km apart in the Mediterranean Sea and ship's locations that are on land suggest that ICOADS data quality control is poor. Whether the Hadley Centre has taken steps to correct or exclude such errors and inconsistencies is unknown.

8.2.7 General conclusions

Data prior to 1950 is unsatisfactory for the calculation of global or hemispheric averages due to (a) low coverage of the Earth's surface, (b) lack of homogeneity of coverage, (c) the high percentage of SST grid cells with data based on from 1 to 5 observations in entire months and (d) the likelihood that observation station data prior to this time has been adjusted multiple times meaning a risk of compounded errors. Data at grid cell or even regional level might be satisfactory for localised studies but should be checked prior to its use.

On the basis of improved coverage, increased SST sample sizes and less likelihood of adjustments to the data the post-1950 HadCRUT4 data is better but, as shown earlier, not completely free from errors.

Climate assessment reports by the Intergovernmental Panel on Climate Change focus on post-1950 data but give no explicit reason for doing so. This concurs with the finding shown here.

8.3 Less significant issues

A number of lesser issues were identified and while each might have very limited influence at a larger scale they might be relevant when making use of the HadCRUT4 data and associated files:

1. Decreasing the grid cell size will decrease the number of grid cells in which observations were made and therefore reduce coverage, while increasing the grid cell size will have the opposite effect. On this basis the hemispheric and global coverage determined from the HadCRUT4, CRUTEM4 and HadSST3 datasets is a consequence of the grid-based system being used and the size of each grid cell.

2. The number of reporting observation stations has varied over time and therefore also the number of grid cells that report that data. At times the number of stations from the 48 contiguous states of the USA, accounting for only 36 of the 2592 HadCRUT4 grid cells that cover the Earth's surface, exceeded 50% of the total number of stations. While the use of grid-cell averages negates most of that disproportionate representation discussion of the number of stations needs to be taken with caution.

3. The annual average of monthly SST observations did not exceed 10,000/month until 1880 and did not exceed 100,000 until 1960. The number of monthly SST observations increased to about 328,000 by 1999 and grew rapidly to over 1,200,000 by 2015, largely due to the number of observations from Argo buoys especially near the coastline of the USA.

8.4 Statistical issues

A number of statistical issues were identified in earlier chapters of this thesis and they are as follows:

The analysis of CRUTEM4 observation station-month combinations for the period 1961-1990 showed that even when data is present for all 30 instances of a given calendar month and station 13.9% (more than 1 in 8) of all such combinations have Shapiro-Wilks p-values of ≤ 0.1, indicating that the complete data record for the station-month is unlikely to be normally distributed. Care therefore needs to be taken when attempting to draw conclusions that assume normal distribution when discussing a small number of stations.

Error margins should be expressed at every step in the sequence of monitoring, adjusting and averaging data. With data from observation stations this means from error margins associated with the method by which temperatures are measured, through the averaging of daily minimum and maximum temperatures across the month, any adjustments of the recorded data, the calculation of temperature anomalies using long-term average temperatures, the averaging of anomalies for each station within the grid cell and finally the calculation of hemispheric and global averages.

In a similar fashion the deriving of sea surface temperature anomalies also involves steps that should be accompanied by error margins, from measuring the water temperature, the averaging of it at grid cell sizes of 1° x 1° over pentads (5-day intervals) and then its interpolation, extrapolation and averaging to convert it to monthly values for 5° x 5° grid cells.

As mentioned earlier in this chapter, the sample size changes from one month to the next either as changing numbers of reporting observation stations, sea surface temperature observations per grid cell, or at a later stage in the process, the number of grid cells that contain data when calculating the average for the hemispheric. This means that if the error margin is calculated using as the denominator the square root

of the sample size that error margin is constantly changing at each level through the processing sequence. The calculated error margin amounts to nothing more than an estimate based on assumptions about the likely distribution had more data been available. These assumptions might be reasonable when little data is missing but they are less reasonable when large amounts of data are absent and of course error margins are subject to skewing when outlying data is included.

The true error margin for HadCRUT4 would be extremely difficult to calculate because every step in the process of measuring each temperature through to the production of hemispheric or global averages, including all data adjustments and any merging with any other data, would need to be examined in detail and the context (coverage and number of stations) constantly varying. The number of error margins to take into account would be enormous, many of them prior to the submission of data either directly to the CRU (for observations) or to the inserting of the data into the ICOADS database (from where the SSTs are accessed by the Hadley Centre).

8.5 Some general issues with observation station data

The approach to measuring temperature, even according to WMO standards, is more aligned to making comparative weather observations than to considering long-term climate variations. As a consequence of this there are several areas of concern when the data is used in climatological studies.

Firstly, temperate data from observation stations is measured between 1.5 metres and 2 metres in the atmosphere above the Earth's surface and sea surface temperature measurements are made just below the sea surface. In many respects the earth and sea surfaces are the worst locations to measure long-term temperature trends even though they are of most interest to humans. This is because larger temperature fluctuations, both spatially and temporally, occur at these boundaries with the Earth's atmosphere than occur at higher levels in the atmosphere or deeper in the ocean. Heat exchanges between the two mediums will take place, sometimes producing steep temperature gradients, and the conditions and therefore the measured temperature will by

influenced by varying meteorological forcings (e.g. wind speed and direction, surface moisture, ground cover, ocean turbulence).

Secondly, data from observation stations has many potential flaws:

- The mean temperature is determined from just two measurements over a 24-hour period, the minimum temperature and the maximum temperature, both susceptible to short-term weather influences such as cloud, rain or wind.

- Especially in mid-latitudes, minimum temperatures on clear-sky mornings are typically recorded within ~5 minutes of sunrise, which means that the sun is low in the sky and therefore under conditions where the solar radiation could easily be impeded by landscape, vegetation or manmade structures. .

- Urbanisation is a significant issue and its influence on recorded temperature can change over time either abruptly with the construction of buildings near observation stations or more gradually as the pattern of generation of manmade heat changes (e.g. changing traffic patterns).

- The WMO standard of recording minimum and maximum temperature for the last 24 hours as they stand at 9:00am is far from ideal given the chances in mid to high latitudes of the minimum temperature for two 24-hour periods being recorded within just a few hours of each other (see Chapter 6).

- Adjustments to recorded temperatures are very common but rely heavily on possibly invalid assumptions about the relationship between the temperatures measured at different location. It has also been shown that several different methods of homogenisation have been attempted and that different methods produce different outcomes for the same data.

8.6 Comments about the discussion of HadCRUT4 reliability in Jones (2016)

Jones (2016) discusses the reliability of the HadCRUT4 dataset and associated datasets. Earlier chapters of this thesis have addressed several issues that Jones (2016) mentions.

Jones (2016) claims that large-scale area averages are reliable even when the data is sparse, even citing an earlier study, by the same author (Jones, 1994), that found that the sparse network of observation stations in the second half of the 19th century was a reliable indicator of global averages on decadal time scales. Such a statement is unsustainable because there is no means by which it can be judged. Moreover it has been shown that the contribution of certain latitude bands, longitude bands and regions such as Europe was well in excess of their true proportionate area of the hemisphere (Chapter 2), which means that hemispheric and global averages were skewed towards the temperatures recorded in these areas.

The paper (Jones, 2016) also argues for an upward adjustment of sea surface temperatures measured by drawing water samples in buckets between 1900 and 1941, presumably based on the change to temperature measurement via engine room intake and the corrections claimed to be required. Chapter 7 of this thesis has shown that variation in global coverage of SST data, brought about by a contraction of shipping during war-time, might account for much of the "spike" in SST global averages on which the adjustments are based. Jones (2016) goes on to say, "If the adjustments were not applied then century-timescale warming would be greater, and there would be a major discrepancy between the land and marine components prior to about 1940." This argument is questionable given that many historical temperature observations over land have been adjusted at least once for a change from manual to electronic instruments and many have been adjusted for urbanisation. The accuracy of land temperatures prior to 1940 is very debatable so comparing them to sea surface temperatures is of questionable value when apparently Jones is unconcerned about the discrepancy between the two sets of temperatures over the last ~25 years.

Jones (2016) also says that the relative correspondence between the HadCRUT4 dataset and the other global temperature datasets (eg. NASA-GISS, GHCN) is "a testimony to the robustness and accuracy of the resulting homogenized data". It is no such thing because national meteorological services adjust their own data prior to supplying it to the CRU and are hardly likely to supply completely different data to the GHCN database. In other words the datasets are similar not because of the homogenized data but because the datasets share the same data.

Jones (2016) also attempts to dismiss the urbanization effect on measured temperature, claiming that differences between urban and rural areas might have meteorological explanations, but such a claim conflicts with the widespread attempts to mathematically remove the effects of urbanization from temperature records in the belief that the distortions have non-meteorological causes. In section 6.3.2 we saw how alternative adjustments to raw data from seven observation stations in New Zealand resulted in an average trend of 0.32°C/century for the five stations in urban areas but just 0.2°C/century for the two rural stations.

The treatment of urbanization is somewhat simplistic in Jones (2016) because, as McKitrick (2013) shows, a variety of situations can cause an apparent temperature trend and a trend in a rural area can, for a variety of reasons, be greater than a trend in an already urban area. McKitrick (2013) also points out that apparent conflicts in the results of different studies is likely due to the absence of common evaluation frameworks.

Jones (2016) goes on to claim that urbanization is not an issue because the data over land is similar to SST data, having argued (see just above) that the SST data was adjusted to match data from land-based observations. Later it claims that adjustments for urbanisation have no net effect, which is an admission that adjustments were made and a dubious assertion of the correctness of those adjustments given the discussion in chapter 6 about urbanization adjustments.

In total the claims of Jones (2016) are weak. They also fail to include many of the findings mentioned earlier in this chapter, especially the presence of outliers and their distorting influence, and that their presence implies poor quality control.

8.7 Towards a more accurate global temperature dataset

So far this chapter has noted several of the failings or at least areas of concern with the HadCRUT4 dataset. It is now time to consider some possible improvements that would make the best use of the available historical temperature data rather than reject it completely.

Satellite-based observations would seem to be the future of temperature monitoring because their coverage is more complete (Christy et al, 2000; Spencer & Christy, 1992). Being usually from a single satellite, on the rare occasions when data adjustment is required (e.g. a change of satellite) the conversion is applied generally rather than, as is the practice with HadCRUT4 data, adjustments being made to individual sea surface temperature measurements or specific to part of the data record from an individual observation station.

Satellite-based measurements, via microwave sensor units (MSUs), are available only since 1978, which means they are of limited historical value. Radiosonde atmospheric temperature measurements offer some advantage because the atmosphere even at 1000m is more laterally homogenous than very close to the Earth's surface. Radiosonde measurements are available since 1958, albeit not gridded but either for the 87 individual stations that launch balloons with these instruments or for mean values over large scales such as global, hemispheric, tropical and extratropical. Likewise Argo buoys seem to offer the most accurate sea surface temperature data but they have been in widespread use for less than two decades. Temperature data from satellites and Argo buoys would be recommended if the aim was only to create a more accurate future set of temperature data but to do so would be to reject virtually all of the data collected since 1850 and have no historical temperature record of any length.

The alternative dataset proposed here aims to make best use of the available historical temperature data and make minimal, if any, adjustments, and with fewer uncertainties

than the HadCRUT4 dataset. The utilising of available data is key if the new dataset is to have historical value.

The following are some recommendation for such an alternative dataset. They are listed in point form with each followed by the reasoning behind the point.

General

1) A new climate reference network system to be established comprised of non-urban observation stations and a separate specific set of grid cells for sea surface temperatures, with the reference network ideally evenly distributed across all latitudes and longitudes.

   The use of a reference network of non-urban station avoids any contamination (UHI effects, manmade shielding) from urban environments, and for SSTs reduces several uncertainties (e.g. quantity and continuance of data, estimations, data processing issues such as interpolation). Effort will be needed to maintain or improve the quality of data from this network because its utility relies on that quality, which might mean ongoing constraints to any changes in the local environment.

2) The observation stations and SST grid cells should be distributed as evenly as possible across latitudes and longitudes

   The aim of an even distribution across latitudes and longitudes is to ensure a reasonably representative coverage of the Earth's surface, albeit with coverage being a lesser concern than high quality data from a set of consistent environments.

3) In order to minimise the uncertainties due to missing data the start year of the data needs to be the point at which very little data is missing (e.g. 90% of the network providing data).

This starting point cannot be locked to a specific year and month without first knowing the availability of the data because data quality is a primary concern.

4) No reference locations to include coastal grid cells

No coastal data should be used because the local environment inconsistently distorts it. Observation station data is impacted by the buffering effect of the ocean and offshore winds, and sea surface temperatures are potentially coming from areas of shallower water, coastal currents or impacted by onshore winds. Data for the reference network is therefore only from observation stations and SST measurements.

5) The two datasets, one land and one sea, to be referred to as reference networks to clearly define their composition.

The terminology "reference network" also makes it clear that global coverage is irrelevant and that the average of the network values are not a global average but from temperature anomalies at a consistent set of representative locations.

Observation station data

6) The selected stations to be non-urban but also with further constraints of consistency of the local environment subjected to ensure a minimum of non-meteorological temperature forcings (very low if any nearby human population, no natural or manmade shielding, no changes of land use etc.)

Only non-urban observation stations should be used in order to avoid both any Urban Heat Island effect and any adjustments that might have been made to the data when stations were relocated as a consequence of that effect. The

constraints on the local environment are to try to limit temperature variations to only meteorological causes.

7) The mean monthly temperature, as the average of the mean daily minimum and mean daily maximum temperature, to be replaced by the median daily maximum temperature across the month

The maximum temperature to be used because the minimum is frequently recorded shortly after sunrise when the sun is low in the sky and can easily be blocked by landscape, vegetation or manmade objects, the maximum usually occurring before the sun is similarly low in the western sky. Further, the maximum temperature is usually of greater importance and the focus on one rather than two temperatures removes one factor with uncertainties. The use of the median value for the maximum temperature rather than mean is to avoid the mean temperature potentially being skewed by short periods of abnormal weather. Using the median value also avoids any implication that the data precision is at greater than one decimal place.

8) In line with the immediately above, the long-term average for a given calendar month to be replaced with the long-term median maximum temperature over a time period of at least 30 years with minimal weather or climate abnormalities.

The use of a long-term median temperature again minimises any distortion caused by abnormal weather conditions.

9) Anomalies to be derived from the monthly daily median and the long-term median maximum for the corresponding month

This point is consistent with the two points immediately above it and has the side-effect that the anomalies will be at similar precision to both the values used to calculate it and the original temperature measurements. This is in

contrast to the current system that records to one decimal place but the subsequent calculation of monthly mean temperatures, long-term average temperatures and monthly anomalies are at multiple decimal places.

Sea surface temperature data

10) A reference set of sea surface grid cells of size 2° latitude x 2° longitude to be established with those grid cells distributed as evenly as possible across the latitudes and longitudes but free of inconsistent external influences.

    The current SST grid cells at 5° x 5° are too large both for data collection and accurate representation with a single value. As shown earlier, the current methodology of 1° x 1° and 5-day pentads requires a minimum of 150 observations in a calendar month if interpolation is to be avoided when converting to the 5° x 5° grid cell size. Reducing the size to 2° x 2° will reduce the current uncertainties. (Moored Argo buoys offer quality SST data at fine granularity but they have been used for very little of the temperature record.) These grid cells should be located away from influence of significant ocean currents (e.g. Gulf Stream, Humboldt and Kuro-Shio), subsea shelving, sea mounts, coastal continental shelfs, river deltas, extended shallow areas and the influence of ocean gyres because all can distort measured sea temperatures.

As well as the above, the complete set of observation station data should be carefully reviewed and any data adjustments reconsidered at a level that takes into account variations in exposure at comparison sites. All temperature adjustments should be fully detailed so users of this data might make informed decisions about data quality.

8.8 Concluding remarks

This analysis and audit of the HadCRUT4 temperature anomaly dataset has revealed many serious concerns about its accuracy. It has not been possible to establish the error margins because the quantity of data changes every month, either as observations report or fail to report data or as the number of observations of sea surface temperature changes.

HadCRUT4 data prior to 1950 is unsatisfactory, particularly due to low coverage and poor distribution of what little coverage there is. Overall coverage even by the HadCRUT4 method of calculation is low and only exceeds 50% in few of the years prior to 1950. Even in the mid 1940's coverage of the southern hemisphere sea surface temperatures is often below 20%. Pre-1950 data from observation stations is not greatly better. Stations that did report data are likely to be in urban areas or areas that since 1950 have developed into urban areas, which means either the data has been distorted by the urban environment or adjusted to try to remove the influence of urbanisation with unknown and unknowable degrees of success. Temperature data adjustments are sequential so all things being equal older data is more likely to have been adjusted multiple times.

Data since 1950are more complete than earlier data but still not without concerns. Even at the end of 2015 data are unavailable for about 12% of the Earth's surface. The majority of HadCRUT4 data is SST data and it suffers from assumptions about the methodology being used and the associated data adjustments, the homogenisation of observation station data remains questionable and the accuracy of the vital long-term average temperatures on both land and sea remain under a cloud.

At the end of the day the HadCRUT4 dataset is comprised of estimates at almost every level, covering the rounding of instrument measurements, the adjustment of recorded data and through to the calculation of global average temperature anomalies (which implicitly estimate that had coverage been 100% the average would be the same), but they are not estimates in which one can have great confidence.

*****

Part 2

Papers, both published and drafted, dealing with three contemporary issues in climate science.

INTRODUCTION TO PART 2

Part 1 of this thesis was an initial analysis of the HadCRUT4 dataset that highlighted many problems with its creation. Part 2 of this thesis contains three papers, one already published and two in draft form.

The first of these chapters is a reformatted version of a paper (McLean, 2014) written and published in the course of my candidature. It uses the HadCRUT4 temperature dataset that was the subject of Part 1 and presents some plausible reasons for the pattern in HadCRUT4 global average temperature anomalies from 1950 onwards. It accordance with Part 1 it explicitly ignores earlier HadCRUT4 data.

The second chapter is a draft of a forthcoming paper that investigates the Troup Southern Oscillation Index (SOI), a key measure of the state of the El Niño-Southern Oscillation (ENSO). Short term fluctuations, caused by the manner in which it derived, result in a "noisy" data sequence, with implicit uncertainty about the relevance of short term interruptions to SOI sequences and when those sequences commence and end. The derivation of the Troup SOI is discussed and an alternative ensemble index, the average of multiple indices calculated in the same manner as the Troup SOI, is proposed.

The third chapter is also a draft of a forthcoming paper for publication. It uses four different approaches to indicate whether severe and widespread bleaching of coral on the Great Barrier Reef was likely to have occurred prior to 1998, which some people claim or imply was the first such outbreak in the last 100 years.

Chapter 9: Paper - Late Twentieth-Century Warming and Variations in Cloud Cover

9.1 Introduction

This chapter consists of a reformatting of a paper (McLean, 2014) that draws on some of the findings about the HadCRUT4 dataset made by that time (see Part 1). In this reformatting table numbers and figure numbers have had the chapter prefix added to them for consistency with other chapters, but the caption wording and main text are unchanged, save for the inserting of reference to an earlier chapter.

9.2 Published Paper

ABSTRACT

From 1950 to 1987 a strong relationship existed between the El Nino-Southern Oscillation (ENSO) and HadCRUT4 global average temperature anomaly, interrupted occasionally by volcanic eruptions. After 1987 the relationship diverged, with temperature anomaly increasing more than expected, but was re-established after 1997 at an offset of ~0.48°C higher.

The period of increased warming from 1987 to 1997 loosely coincided with the divergence of the global average temperature anomalies over land, which are derived from observation station recordings, and the global average anomalies in sea surface temperatures. Land-based temperatures averaged 0.04°C below sea temperatures for the period 1950 to 1987 but after 1997 averaged 0.41°C above sea temperatures.

The increase in the global average temperature anomaly and the divergence of land and sea surface temperatures also coincided with two significant changes in global average cloud cover. Total cloud cover decreased during the period from 1987 to 1997 and, for most of the remainder of the period from 1984 to 2009, decreases in low-level cloud were accompanied by increases in middle and upper level cloud. These changes can be found in both global average cloud cover and in each of the six 30°-latitude bands. The impact of these changes in

cloud cover can account for the variations in HadCRUT4 global average temperature anomalies and the divergence between land and sea temperatures.

Key words: climate change, temperature shift, insolation, ENSO, ISCCP, HadCRUT4

1. Introduction

The latest report by the Intergovernmental Panel on Climate Change (IPCC, 2013, [1]) reports that some climate models overestimate the climate system's response to increasing greenhouse gases since 1998, in other words they predicted higher temperatures than those observed.

Given that models overestimate the influence of carbon dioxide it follows that the relative accuracy of the models for the period 1950 to 1997, as reported in IPCC's 4AR [2], could only occur if the models under-estimated the influence of other forcings.

One forcing that might have been under-estimated is cloud cover. Variations in total solar irradiance are often discussed but not variations in cloud cover, but cloud cover impedes the flow of radiation, which in general means that it controls the amount of radiation reaching the Earth’s surface during the day, and how much heat is lost during overnight cooling.

The effect of cloud cover on the net radiation budget of the earth has received much attention (e.g. Hartmann et al.[3], and Keihl, [4]), especially in the context of the zone of tropical convection, which is essentially the engine driving atmospheric circulation. In the tropical convection zone the cooling effect of increased albedo due to clouds is largely cancelled by the reduction in outgoing long wave radiation, however there is much that is not yet understood about the role of clouds.

Relevant papers include Goode and Pallé [5], who briefly discussed variations in cloud cover as part of larger paper focusing largely on variations in solar radiation. Herman et al [6] also discussed global cloud cover but dealt mainly with surface reflectivity at 340nm. Kauppinen et al [7] discussed the impact of both humidity and cloud cover on the global mean surface temperature. Eastman and Warren discuss long-term trends in cloud over sea [8] and land [9]. The former focuses on marine stratus and stratocumulus cloud cover when discussing variations in sea surface temperature and the later mentions temperature only very briefly.

This paper will attempt to address the wider issue of variation in cloud cover at all levels and in total, and whether this relates to variations in the HadCRUT4 global average temperature anomaly. Because temperature is expressed in terms of average monthly anomalies, the cloud cover data will likewise be converted to anomalies. Volcanic eruptions and the El Nino-Southern Oscillation (ENSO) are recognised influences on temperature so those influences will be identified and removed before the relationship between the residual temperature and variations cloud cover is considered.

2. Data Sources

This paper draws on cloud cover data from the International Satellite Cloud Climatology Project (ISCCP), at http://isccp.giss.nasa.gov/, and described in Rossow and Schiffer [10]. This data includes total cloud cover as well as low, mid and upper level coverage, all of which will are used in this paper, but is only available, at the current time, for the period from 1984 to 2009.

The primary temperature data is the HadCRUT4 dataset, available via http://www.cru.uea.ac.uk/cru/data/temperature/.

Data for the El Nino-Southern Oscillation is the Troup Southern Oscillation Index [11] published by the Australian Bureau of Meteorology and available at http://www.bom.gov.au/climate/current/soihtm1.shtml. Under the Troup system,

sustained values above +8 for at least three months usually indicate La Nina conditions and sustained values below -8 usually indicate El Nino conditions. The Troup SOI is open-ended and for the period 1950 to 2013 inclusive it averages 0.19 with a standard deviation of 10.53. The Troup SOI data is preferred to the 3-month average of sea surface temperatures in the "Nino 3.4" region (bounded by latitudes 5N-5S and longitudes 170W-120W) because prior to 1957 data coverage of this region rarely exceeded 50% and was less than 75% for many years after, the shortfall being mainly in the western half of this region, which might have been the only portion impacted by mild ENSO events.

Data for volcanic eruptions, particular for the periods of eruption, was sourced from the Smithsonian Institution (see http://volcano.si.edu/search_eruption.cfm).

This paper uses data only since January 1950 for two reasons. The first is that the selected period corresponds to the period for which the IPCC claims that warming was largely due to human activity, meaning that the findings of this paper therefore apply to the same period.

The second reason is that prior to 1950 the coverage of HadCRUT4 temperature data for the Southern Hemisphere was below 50% in both World Wars, and fell to just 23% in 1945. After World War II Southern Hemisphere data coverage increased, with only one month during the 1950s below 50% and after the start of 1960 it was consistently above 60%[27].

The reliability of data during periods of low coverage is one problem but the other is that the month-to-month variation during low coverage is usually greater, meaning more data "noise" than during periods of greater coverage. For the decade January 1990 to December 1999, with average global coverage 83.7%, the average absolute month-to-month variation was 0.084°C (σ = 0.075°C), whereas for the decade 1940-1949, with average coverage lower at 56.62%, the variation was almost 50% higher at 0.122°C (σ = 0.092°C).

[27] This paragraph refers to the issue of coverage that is more completely dealt with in chapter 2 of this thesis.

3. Analysis

3.1 Resolving a Residual Temperature

The first step towards investigating the possible influence of cloud cover on the HadCRUT4 global average temperature anomaly is to establish a residual temperature anomaly from which the impacts of the ENSO and volcanic eruptions have been removed.

Large volcanic eruptions, particularly in the western Pacific, appear to have an impact on the global average temperature anomaly. Table 9-1 shows four volcanoes that are regarded as having a noticeable impact on the global average temperature since 1950. The displayed metrics of Volcanic Explosive Index (Lamb [12], [13] and [14]; Newhall and Self [15]) and the Dust Veil Index (Global) indicate their relative strengths.

Volcanic eruption	Duration	VEI	DVIG
Agung	Feb 1963 - Jan 1964	5	800
Awu	Aug 1966 - Oct 1966	4	200
El Chichón	Mar 1982 - Sep 1982	5	800
Pinatubo	Apr 1991 - Sep 1991	6	Not available

Table 9-1 – Four major volcanic eruptions of the late twentieth century.

Volcanic eruptions are a challenge when working with climate statistics. It is difficult to compare these in any simple way because the VEI scale is logarithmic. A VEI of 4 means in the range 0.1 km^3 to 1km^3 of ejected tephra; a VEI of 5 is from 1km^3 to 10km3 and so on. If an eruption rated as VEI 4 is near the lower end of that band and an eruption rated 5 is near the upper limit of that band the difference in ejected tephra, and presumably sulphides that cause cooling, could be a factor close to 100.

In passing it is noted that the first three of the eruptions listed in Table 9-1 occurred in the period from 1961 to1990. The HadCRUT4 temperature dataset is derived from the variation (or "anomaly") from the long-term average temperatures for each calendar month over this same period, which means that monthly "normals" are lower than what they would be had no volcanic eruptions occurred and that the calculated anomalies are higher.

Determining the influence of volcanic eruptions on temperature is problematic. Various previous attempts have attempted to use either the Dust Veil Index (DVI) or estimates of cooling. The use the DVI is flawed because it is derived from the temperature, thus making the argument circular (Robock [16]; Bradley and Jones, [17]). Using the aerosol optical depth (AOD) to estimate cooling, such as the technique used by Sato et al [18], is uncertain because there appears to be no accepted factor for converting AOD into a temperature change. The other approach to estimating cooling relies on models of uncertain accuracy, which is probably inevitable when trying to calibrate them against temperature data with its short-term fluctuations, and consequently the error margins are large.

A further complexity is the potential link between volcanic eruptions and ENSO conditions (Gu and Adler [19]). El Nino events often follow volcanic eruptions in the western Pacific (Wigley [20]; Emile-Geay et al. [21]). Empirical data indicates that the cooling due to the eruption substantially counterbalanced the warming influence of the El Nino events.

Because both the cooling and warming attributable to the two phenomenon are less than expected had they occurred without the other this situation distorts any calculation of the "average" impact of such events. In turn, this impacts on the calculation of a "residual temperature" that remains after removing the influence of volcanic eruptions and the ENSO. Accordingly, the simplest interpretation of the long-term temperature records is achieved by omitting the data for the period over which major eruption caused cooling (cf. Figure. 8.2). This is unlikely to have much impact on residual temperature because the influence of volcanic eruptions is transitory with a maximum of about three years and there is no evident trend in volcanic eruptions during the period of this study (1950-2013).

3.2 The El Nino-Southern Oscillation (ENSO)

The ENSO is widely recognised as having a significant impact on weather patterns both in and around the Pacific Ocean and as far away as the continents of Africa (e.g. Korecha and Barnston [22]; Ogutu et al.[23]) and Europe (e.g. Shaman and Tziperman [24]; Brönnimann et al. [25]).

The ENSO has also been found to influence global average near-surface temperature and Trenberth et al. [26] and Jones [27] found an optimum correlation between ENSO and global average temperature when using the time-lagged ENSO data from six months earlier but de Freitas and McLean [28] found the lag time varied between four and five months when data from 1950 to 1995, 1950 to 2000, 1950 to 2005 and 1950 to 2012 was considered.

Figure 9.1 shows the Troup SOI and HadCRUT4 global average temperature anomaly for the period from1950 to 2013 inclusive with no time lag to the SOI data. (As with other graphs in this paper a 5-month centred average is used when plotting, but monthly data is used in analyses.)

The vertical axes in Figure 9.1 are scaled for optimum correspondence across the period 1950-1985. The graphs show a generally close relationship, with the SOI leading the HadCRUT4 temperature data slightly, from 1950 until the late 1980s and an offset of ~0.5°C after about 1995.

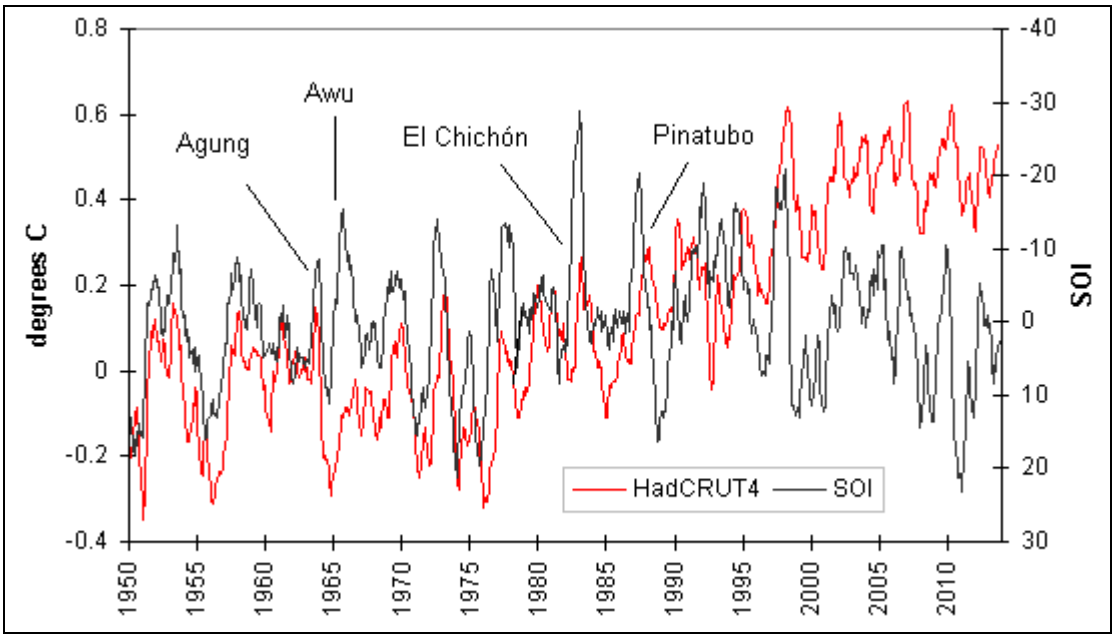


Figure 9.1 - HadCRUT4 global average temperature anomalies and Troup SOI, both using 5-month centred averaging (i.e. from -2 months to +2 months), with major volcanic eruptions indicated

Both signals have significant short-term fluctuations (i.e. "noise"). The SOI is more susceptible to short-term weather effects because it is derived from air pressure monitoring at just two locations, Darwin, Australia and Tahiti. The month-to-month variation in the Troup SOI averages 6.87 (σ =8.74) from 1950 to 2013 but the variation differs across the calendar year with a maximum of 9.88 (σ =7.33) in May and then falls to a September low of 4.46 (σ = 3.45).

The month-to-month variation in HadCRUT4 data averages 0.076°C (σ = 0.103°C) across the entire period, with December to March all exceeding 0.08°C and June to September averaging between 0.04 and 0.05°C. Temperatures over land show a greater month-to-month variation than sea surface temperatures. The CRUTEM4 average across all months is 0.182°C (σ =0.164°C) compared to the HadSST3 average of 0.039°C (σ =0.031°C)

The temporary periods of divergence in Figure 9.1, when temperature are lower than expected from the SOI value (i.e. 1964-66, 1983 and 1991-1993) have been attributed to cooling caused by volcanic eruptions near the tropical Pacific (Agung 1963, Awu 1966, El Chichón 1982 and Pinatubo 1991) (Hansen et al [29]; Dutton and Christy [30]; Douglass and Knox [31]). The two later periods illustrate the issues discussed

earlier. Cooling due to El Chichón appears to be less than expected, with Angell (1988, 1990) attributing blame on the El Nino event that followed, whereas the Pinatubo eruption appears to have caused cooling and reduced the impact of the El Nino event that followed it.

The first step in determining the relationship between the Troup SOI and global average temperature anomaly until the late 1980's, the period of good correspondence in Figure 9.1, was the exclusion of the data for the periods of cooling due to volcanic eruptions. These periods are February 1963 to June 1967 (Awu eruption closely followed by Agung) and from March 1982 to September 1983 (El Chichón), which are extensions on the periods of eruptions shown in Table 9-1.

The second step was to find the optimum correlation between the HadCRUT4 and Troup SOI data when the latter was lagged by different numbers of months. This optimum was found to be -0.551 when using the SOI from three months earlier. The correlation appears to be reduced by two main factors, firstly the month-to-month variation (or data "noise") discussed above and secondly possibly minor volcanoes in 1955, 1972 and 1975 (Figure 9.2).

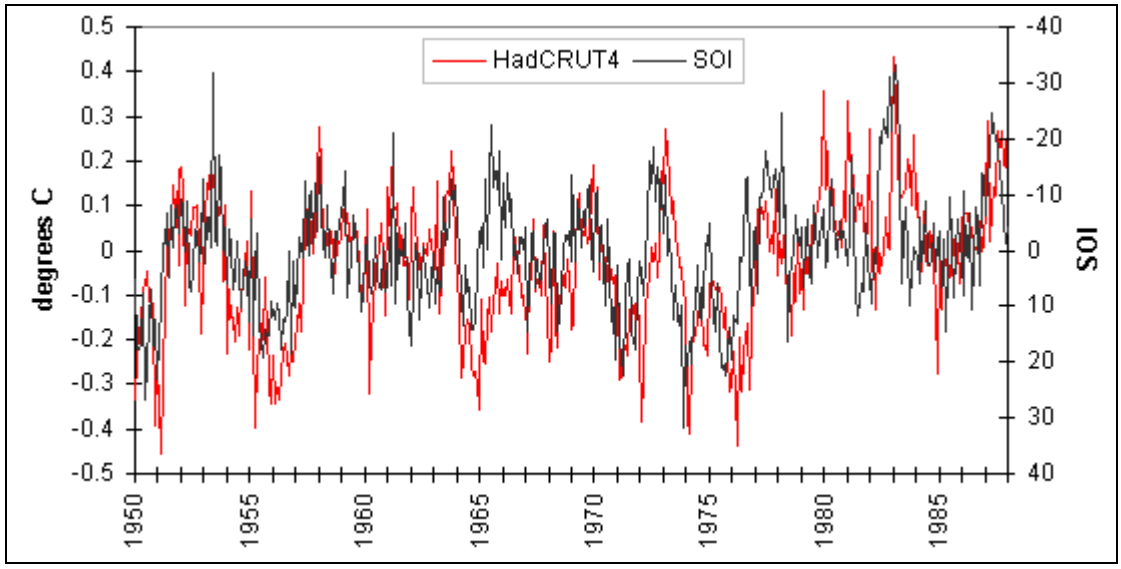


Figure 9.2 - HadCRUT4 and Troup SOI for the period 1950-1987(monthly averages).

After applying the three-month time lag in SOI to the 1950-1987 data with periods of eruption-driven cooling omitted, the linear equation of best fit was found to be.

$$GATA = (-0.0083 * SOI) - 0.021 \qquad \text{(equation 1)}$$

where GATA is the HadCRUT4 global average temperature anomaly and SOI is the Troup SOI (from three months earlier).

Applying the same time lag to the 1998-2012 data, which on visual inspection appears to be a time of relative consistency in relationship, resulted in a different equation of best fit.

$$GATA = (-0.0049 * SOI) + 0.465 \qquad \text{(equation 2)}$$

The R^2 values for the two equations are 0.304 and 0.196 respectively; the weaker correlation in the second period suggesting that temperature variability during that period is somehow different to the earlier period.

The above calculations assume a linear relationship between the SOI and HadCRUT4 global average temperature anomaly. It is however possible that the relationship varies on a seasonal basis, as do temperature and prevailing winds, or in the case of regions where rainfall correlates well with the ENSO, a sustained El Nino or La Nina event could influence the amount of surface moisture, which in turn might impact near-surface temperature until such time as the surface moisture returns to normal levels.

The 95% percentile of the magnitude of the influence of the ENSO on global average temperature anomalies can be approximately determined by applying two standard deviations of the monthly Troup SOI data to equation one. For the period for which equation 1 was calculated, the standard deviation is 9.68 and therefore two standard deviations implies a temperature anomaly contribution of 0.16°C.

The addition of the constants in equations 1 and 2 total almost 0.49°C, indicating a temperature shift of approximately that amount occurred between 1988 and 1997, i.e. the intervening period between the two for which calculations were made.

Figure 9.3 shows the derived residual global average temperature anomaly for the period 1950-2012 after the ENSO influence, as derived from the 1950-1987 ENSO-temperature relationship, and with the periods of volcanic eruption-driven cooling omitted, i.e. the two periods as above (February 1963 to June 1967 and March 1982 to September 1983) and additionally Pinatubo (June 1991 to December 1994).

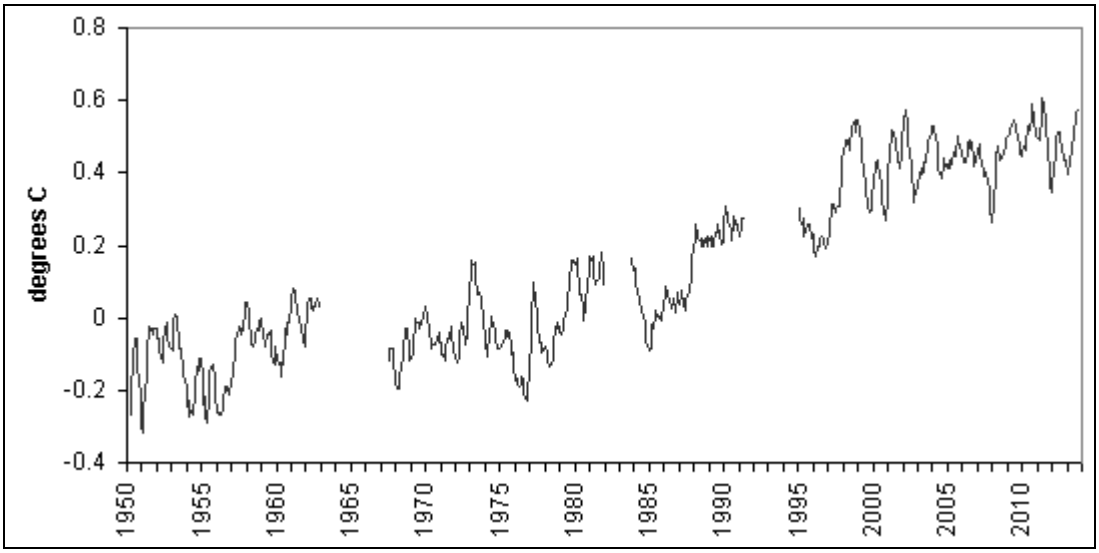


Figure 9.3 - Residual global average temperature anomaly after removing the ENSO influence derived from the Troup SOI (using equation 1) and the data for periods of cooling due to volcanic eruptions (5-month centred average).

Figure 9.4 expands on Figure 9.2 by showing the number of months where the residual temperature anomaly was less than -0.2°C or greater than +0.2°C in each year. To put this range into context, it is slightly less than two standard deviations of month-to-month variation HadCRUT4 global average temperature anomaly described above (σ = 0.103°C).

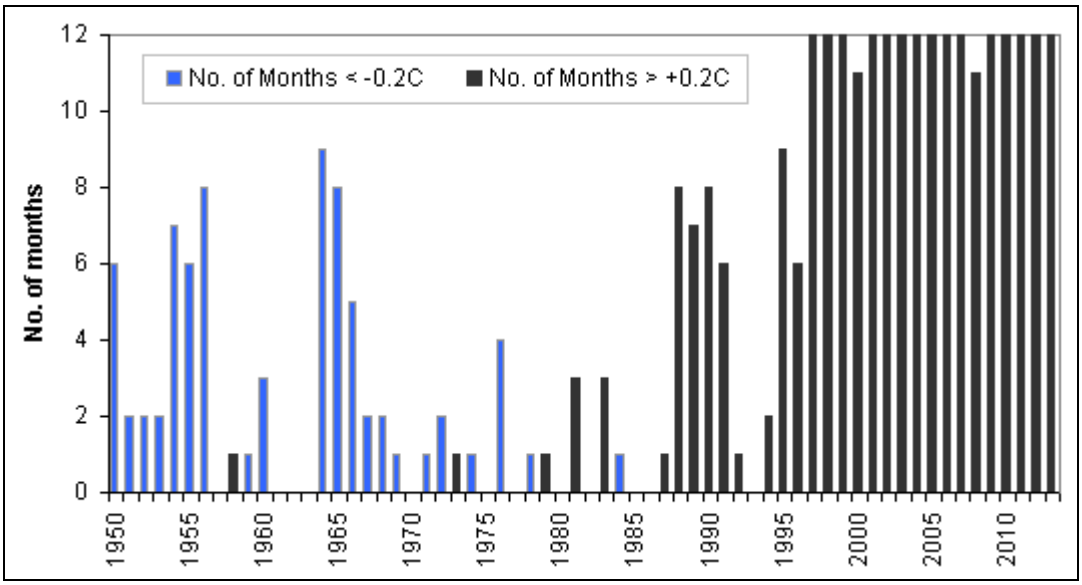


Figure 9.4 - The number of months in each year that the residual temperature anomaly was outside the range 0.0 ± 0.2°C.

The presence of the temperature shift in multiple temperature datasets, including near surface (HadCRUT, GISS and NCDC) and the lower tropospheric temperatures obtained from balloon-mounted instruments (RATPAC-A) and satellite-mounted instruments (UAH, and RSS) indicates that the shift is not an artefact of monitoring (e.g. technique, instrument error, number of stations, coverage) or the data processing methodology (e.g. station data adjustments). Similar shifts in the temperature-ENSO relationship are also found when the Nino 3.4 index is used, which means that the shift is not specific to the Troup SOI.

The residual global average temperature anomaly shown in Figure 9.3 differs from those calculated in other studies such as Trenberth et al. [26], Thompson et al. [32] and Foster and Rahmstorf [33]. Two of these three studies used temperature data prior to 1950, when hemispheric (and global) coverage was poor, and none identified the combination of post-1987 increase in temperature and post-1997 plateauing.

Further, Trenberth et al [26] attempts to split the ENSO-temperature relationship into two periods, 1950-1978 and 1979-1998, and because the change in the relationship after 1987 wasn't identified concludes that only 0.06°C of the warming between 1950 and 1998 could be attributed to the ENSO. Thompson et al [32] attempts to use sea

surface temperature of the eastern Pacific cold tongue as an ENSO proxy across the period from 1900 to 2009 but data in that region is sparse prior to 1950 and particularly unreliable for determining mild ENSO conditions. Further, the ENSO relationship is derived prior to addressing the irregular cooling caused by volcanic eruptions, which were shown earlier to suppress ENSO-driven warming. Foster and Rahmstorf [33] considers only the period since 1979 and therefore largely omitted the period from 1950 to 1987 during which the global average temperature anomaly varied little.

The pattern in residual temperature anomaly of Figure 9.3 indicates warming from 1988 to 1997, the details of which are masked by cooling caused by the Pinatubo eruption, followed by a flattening.

3.3 Divergence of Land and Sea Temperature anomalies

Perhaps associated with this pattern is that 1988 was the beginning of the divergence between the global average temperature anomalies derived from observation stations (CRUTEM4) and those derived from sea surface temperatures (HadSST3). Figure 9.5 shows these two anomalies using five-month centred averages but the averaging spreads the effect of brief peaks in single months.

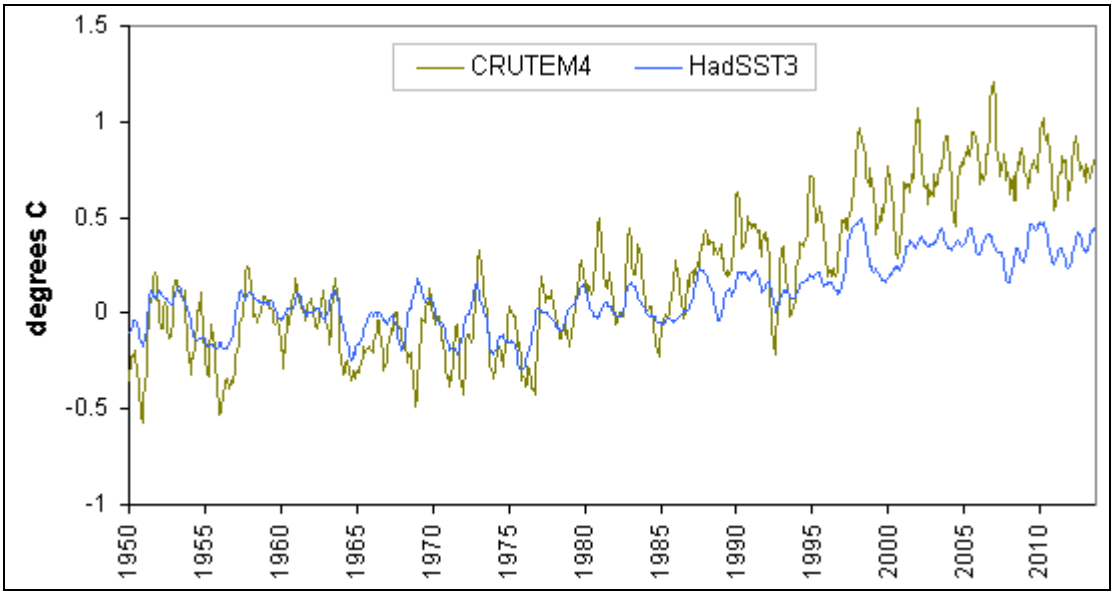


Figure 9.5 - CRUTEM4 and HadSST2 global average temperature anomalies (5-month centred averages)

For the period 1950 to 1986, CRUTEM4 values exceeded HadSTT3 data by more than 0.1°C in 24.1% of all months and only three years had more than six months that met this condition. In contrast, since 1988 CRUTEM4 values exceeded HadSST3 data by more than 0.1°C in 84.3% of months, and only one year had more than six months where the condition was not met. (Unlike above, in these calculations no exclusions have been made for periods of cooling due to volcanic eruptions because both land and sea temperatures could be expected to reduce.)

The average difference between CRUTEM4 and HadSST3 global averages (the former minus the latter) from January 1950 to December 1987 is -0.04°C (μ=0.23°C) but for the period 1998 to 2013 the average is +0.41°C (μ=0.23°C), which together indicate a shift of 0.45°C.

The pattern of a divergence and a later offset stabilisation between these two datasets echoes the pattern in the residual temperature anomaly and the ENSO/HadCRUT4 relationship shown in Figure 1.

3.4 Coincidental Variations in Cloud Cover

The pattern of the residual temperature anomaly does not correspond to changes in atmospheric carbon dioxide, which has been increasing almost linearly from 1958, when monitoring began and certainly has not stabilised since 2000. The pattern is also inconsistent with other greenhouse gases, including methane, whose concentration rose from 1984 to 1999, and CFC-12, which increased from 1979 to year 2000.

The residual temperature anomaly is however consistent with two variations in cloud cover. The first is a reduction in total cloud cover, which would allow more solar insolation to strike the Earth's surface and in particular more radiation in the UV-B range; the second is a decrease in low-level cloud and increase in mid- and upper-level cloud.

As noted earlier, cloud cover data is available from the ISCCP only for the period from 1984 to 2009, which is far less than the period of available HadCRUT4 temperature data.

To compare cloud cover data with the temperature anomalies from HadCRUT4 dataset it was necessary to convert to cloud cover data first to monthly long-term averages calculated from the full span of available data, and from those averages calculate the cloud cover anomalies for each month. These anomalies were calculated for "total cloud" cover as well as for low, mid and upper level cloud.

Figure 9.6 shows the anomaly in global average total cloud cover and HadCRUT4 global average temperature anomaly with cloud cover inverted and both graph lines as three-month centred averages. The HadCRUT4 temperature anomalies fell after the eruption of Mount Pinatubo and this cooling continued for the next few years. The fall in HadCRUT4 anomalies around year 2000 was largely ENSO driven (see Figure 9.3). When these factors are taken into account there is general consistency of a reduction in total cloud cover as temperature anomaly increases, with cloud cover

decreasing from about 1984 until year 2000 followed by a flattening out to 2009, which is the end of the available cloud cover data.

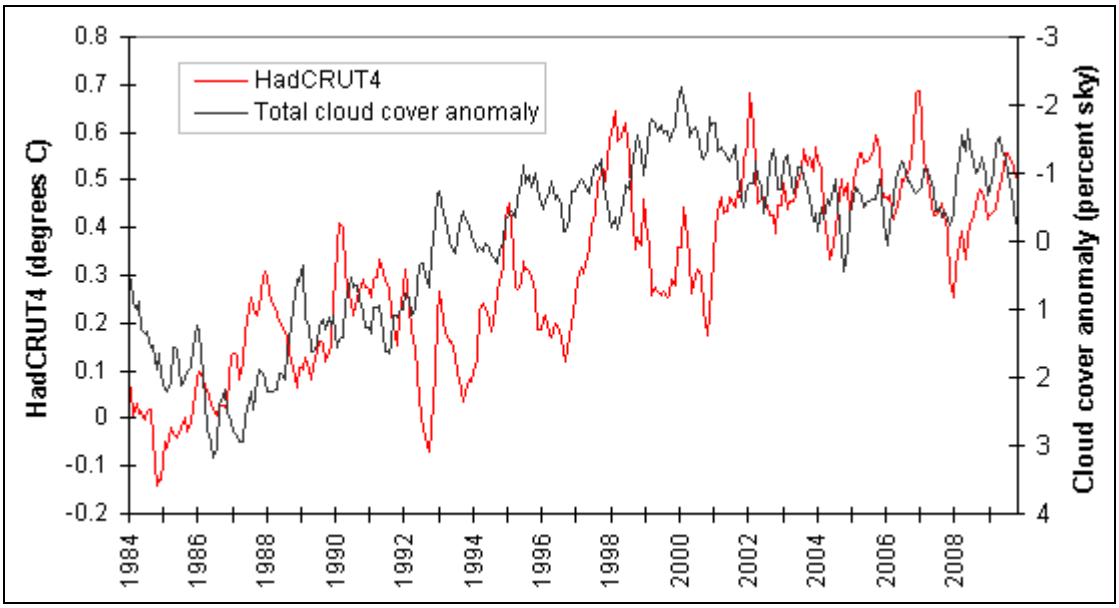


Figure 9.6 - HadCRUT4 global average temperature anomaly and the inverted anomaly in total cloud cover (3-month centred averages).

The decrease in total cloud cover anomaly is approximately 4.5 percent of sky, against the long-term average (all months 1984-2009 inclusive) of 66.4 percent of sky, which means a reduction of 6.8%.

The reduction in total cloud cover is significant in the context of the energy budget described by Trenberth et al [34], which indicates that cloud reflect 23% of the 341 Wm^{-2} (i.e. 79 Wm^{-2}) of incoming solar radiation. The reduction in total cloud cover of 6.8% means that 5.4 Wm^{-2} (6.8% of 79) is no longer being reflected but acts instead as an extra forcing into the atmosphere, some of which will be lost when it adds to the longwave radiation to space. Of course clouds have many other affects on the earth's radiation budget many of which are not fully understood, but a change of 5.4 W/m^2 is potentially of considerable significance.

To put this into context, the IPCC Fifth Assessment Report [1], section 8.5.2, states that the total anthropogenic radiative forcing for 2011 relative to 1750 is 2.29 [1.13 to

3.33] Wm^{-2} for all greenhouse gases and for carbon dioxide alone is 1.68 [1.33 to 2.03] Wm^{-2}.

The increase in radiative forcing caused by the reduction in total cloud cover over 10 years is therefore more than double the IPCC's estimated radiative forcing for all greenhouse gases and more than three times greater than the forcing by carbon dioxide alone. Even the upper limits of the IPCC's estimates fall well short of the increase in radiative forcing caused by the reduction in total cloud cover.

Goode and Pallé [5] examined variations in Earth's albedo from 1984 to 2000 and concluded that the decrease in albedo from the late 1980s to the late 1990s caused additional shortwave forcing of $6.8Wm^{-2}$**.**

Herman et al [6] determined 340nm Lambertian equivalent reflectivity of the Earth from 1979 to 2011 and found a 3.6% ±0.2% decrease in cloud reflectivity over that period, which after applying the shortwave energy balance in Trenberth et al [34] concluded an increase of 2.7 Wm^{-2} insolation, of which 2.3 Wm^{-2} was absorbed by the Earth's offset slightly by increased longwave cooling. This is less than the 5.4 Wm^{-2} described above, possibly because it focuses on 340nm reflectivity, but it likewise finds a reduction in cloud cover and a consequent increase in surface temperature.

The second form of variation in cloud cover is cloud height, which is not evident from total cloud cover. Increasing cloud height is associated with an increase in the transparency of cloud and how much radiant heat passes through it. Figure 9.7 shows the global cloud cover anomaly at low, mid and upper levels. A general decrease in low-level cloud is evident, except for the period from 1992 to 1998, which is when firstly the Pinatubo eruption caused widespread cooling and then HadCRUT4 global average temperature recovered, accompanied by the further warming of an El Nino event.

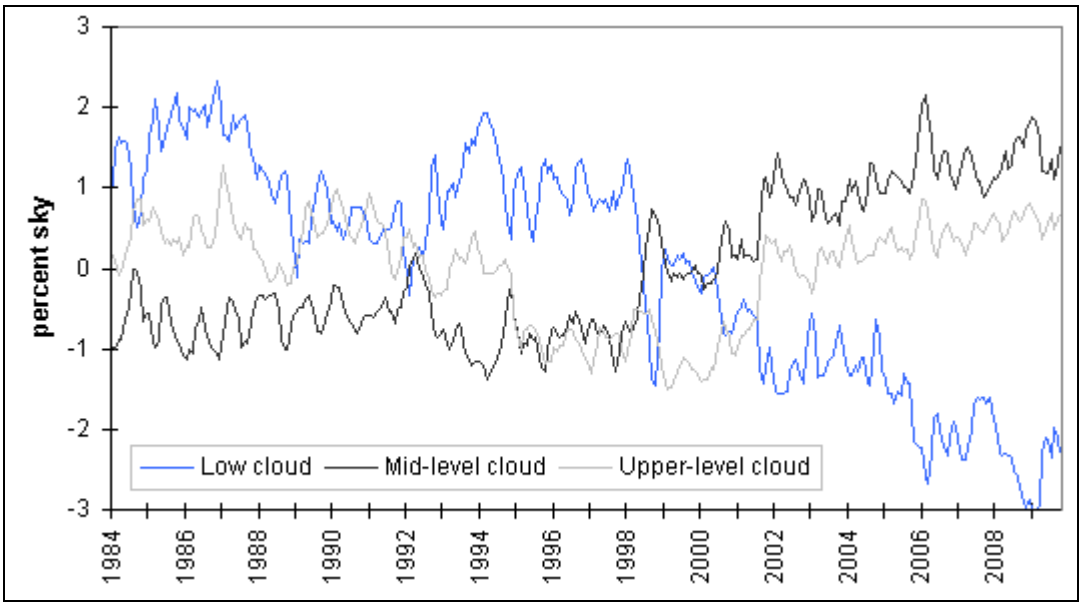


Figure 9.7 - Global average cloud cover anomalies at low, mid and upper levels

The global average percentage sky coverages for low, mid and upper-level cloud are 27.3, 20.6 and 13.1 respectively. The reduction in low-level cloud cover, from 2.0 percent sky above average to 2.5 below, amounts to a change of 15%. The increase in mid-level cloud, calculated in similar fashion, is 10%. Considering the affect of cloud cover on both long and short wave radiation fluxes, these and the 6.8% reduction in total cloud cover are likely to be variations of considerable significance.

Figure 9.8 shows the anomaly in low level cloud cover with the (inverted) anomalies of the sum of mid and upper-level cloud. Across the entire period the average low cloud cover is 27 percent of sky and for the sum of mid and upper-level cloud 33.8 percent of sky. Across the calendar year low-level cloud cover varies from 23.2 percent of sky in February to 26.5 in June, and the combined mid and upper cloud varies from 38.8 in December to 33.7 in August.

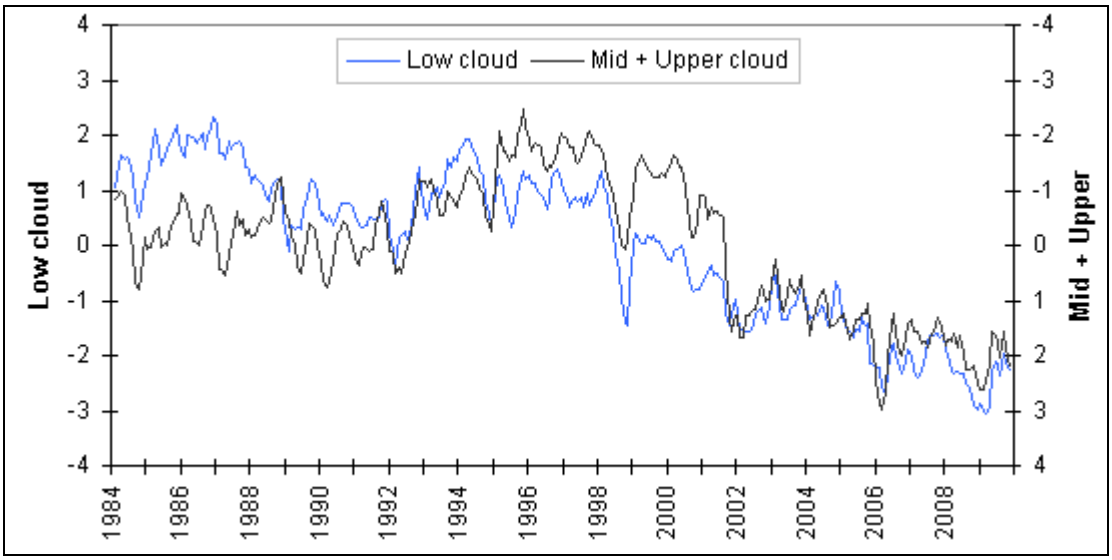


Figure 9.8 - Anomalies in global average low-level cloud cover and the (inverted) sum of mid and upper-level cloud cover (3-month centred average)

The important feature of Figure 9.8 is that from 1988 any decrease in low-level cloud was reflected in an increase in mid and upper-level cloud (and vice versa). Between 1992 and 2002 mid and upper-level cloud reduced slightly and low-level cloud increased. From 2002 to 2009 (end of data) a reduction in low-level cloud corresponded to an increase in mid and upper-level cloud. Further, the absence of any clear shifts in the relationship between the cloud cover at each level indicates a genuine shift rather than an instrumentation issue.

3.5 Analysis by latitude bands

The findings described in section 3.4 (above) were derived from global average temperature anomalies and cloud cover anomalies. A further analysis of each of six 30°-degree latitude bands was undertaken to determine whether the variations were widely distributed or influenced by very large variations at few locations. The contribution each band to the global averages is that the region from the equator to 30° latitude covers 50% of the hemisphere, the region from 30° to 60° covers 36.6% and the region from 60° to 90° covers 13.4%, so global contributions are half of each of these figures.

The mean HadCRUT4 temperature anomalies for each region are shown in Figure 9.9 as 7-month centred averages. The Polar latitude bands, from 60° to 90° in each hemisphere, show the greatest variation in temperature both since 1950 and on a short-term basis, but neither repeats the global pattern of an obvious temperature increase from 1988 to 1997 with relative flat trends before and after. The mid-latitude band for the northern hemisphere shows greater variability than the corresponding band in the southern hemisphere, as well as a clearer warming period from 1988 to 1997.

The two latitude bands nearest the tropics exhibit very similar patterns in temperature anomalies, with a correlation of 0.85 on monthly data, and with both having a distinct step in 1977 when ENSO conditions swung from favouring La Nina conditions to favouring El Nino conditions.

The average of the monthly temperature anomalies for these two bands (i.e. 30°N to 30°S) shows persistently close correlations with the Troup SOI of between -0.614 and -0.674 for ENSO time lags of zero to five months. Residual temperature anomalies can be calculated in similar fashion to that shown in Section 3.2 and the pattern is almost identical to those shown in section 3.2 above.

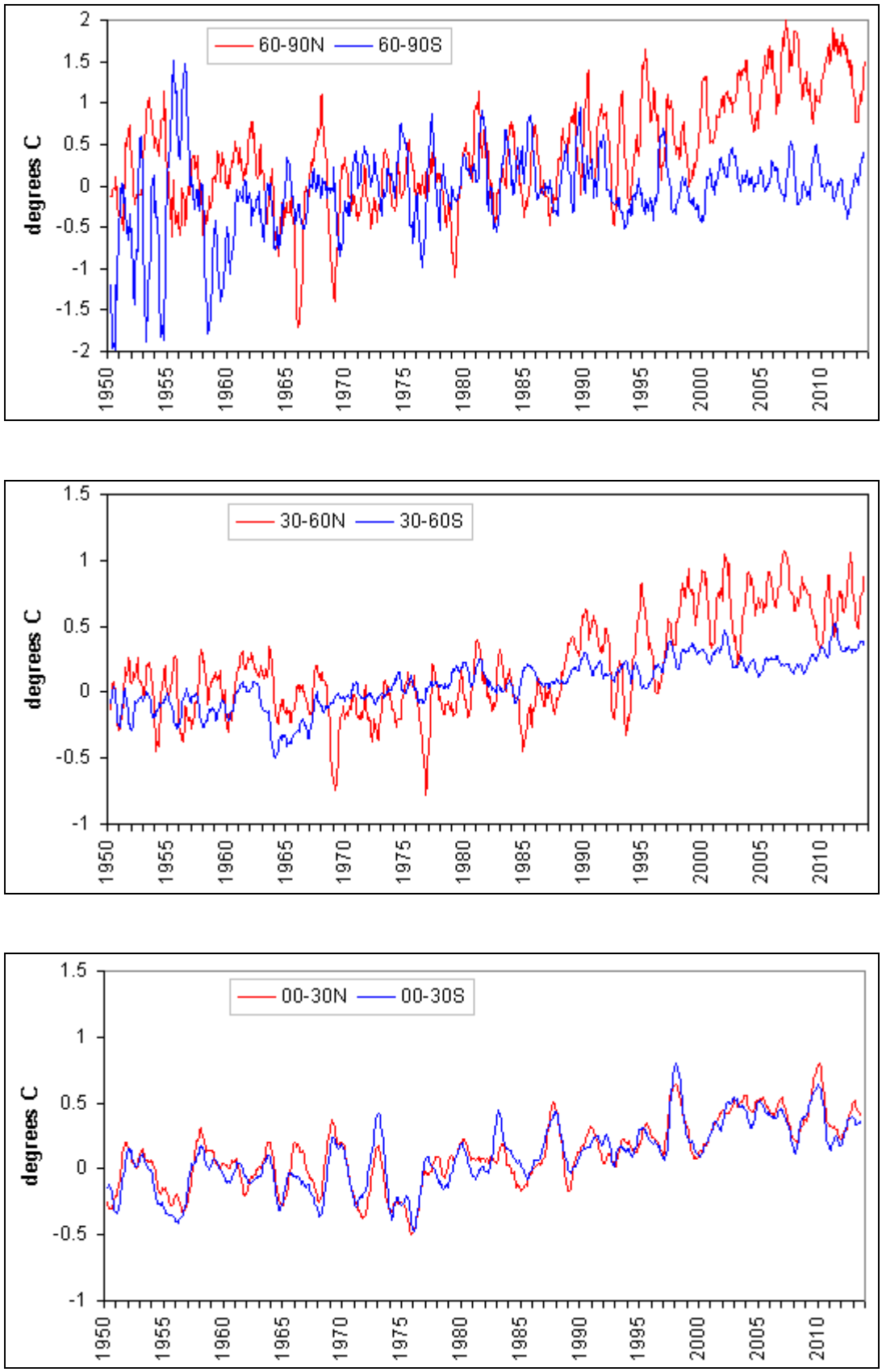


Figure 9.9 - Average HadCRUT4 temperature anomalies for six latitude bands.

The anomalies in total cloud cover for each band are shown in Figure 9.10. In the northern hemisphere, the polar band (60°N-90°N) shows a general decrease of about 3% sky across the period from 1987 to 2009, while the other two bands, 00°-30°N and 30°-60°N, show decreases from about 1987 to year 2000 of about 4% sky and 6% sky respectively, and a subsequent recovery of slightly less than half those figures from 2000 to 2009. In the Southern Hemisphere only the band from 00° to 30°S shows that pattern, the cloud cover in the other bands fluctuating but showing little trend from 1984 to 2009.

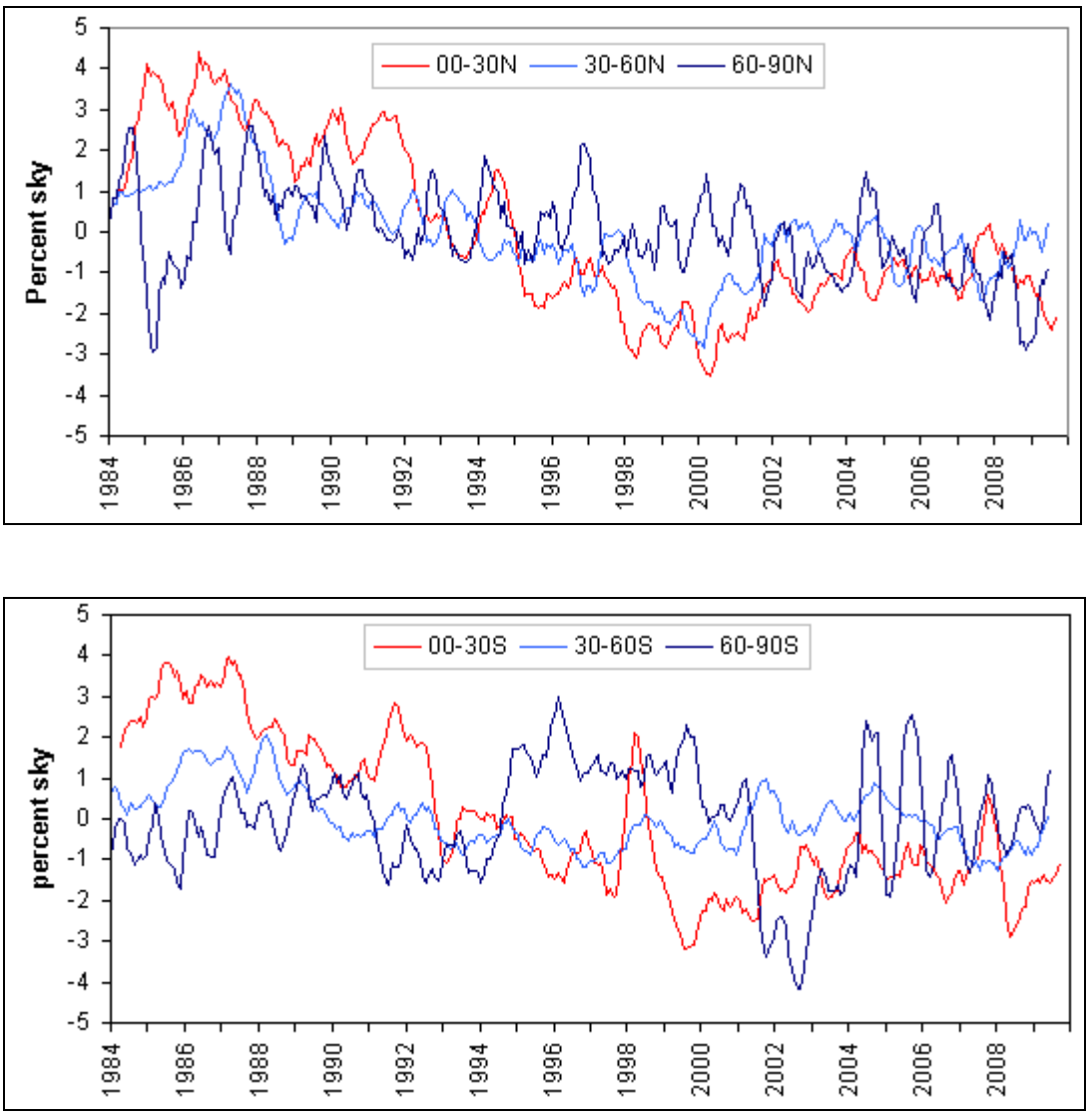


Figure 9.10 - Anomaly in total cloud cover for each latitude band.

The average of the monthly temperature anomalies for these two tropical bands (i.e. 30°N to 30°SN) shows persistently close correlations with the Troup SOI of between -0.614 and -0.674 for ENSO time lags of zero to five months. Residual temperature anomalies can be calculated in similar fashion to that shown in Section 3.2 and the pattern in those anomalies is almost identical to the global residual temperature anomalies shown in section.

The anomalies in total cloud cover for each band are shown in Figure 9.10. In the northern hemisphere, the polar band (60°N-90°N) shows a general decrease of about 3% sky across the period from 1987 to 2009, while the other two bands, 00°-30°N and 30°-60°N, show decreases from about 1987 to year 2000 of about 4% sky and 6% sky respectively, and a subsequent recovery of slightly less than half those figures from 2000 to 2009. In the Southern Hemisphere only the band from 00° to 30°S shows that pattern, the cloud cover in the other bands fluctuating but showing little trend from 1984 to 2009.

Figure 9.11 shows the anomalies in cloud cover at low, mid and upper levels. As for the global analysis, these show a reduction in low level cloud but an increase in mid and upper level cloud across the period for which data is available.

The correlation, for each latitude band, between low-level cloud cover and the sum of mid and upper level cloud is shown in Table 9-2. The correlation for the tropical bands is weak but for other bands is much stronger, where the decrease in low cloud cover is largely offset by increases at mid and upper levels.

Since 1950, temperatures have risen in all latitude bands except for the Antarctic band at 60°-90°S. Warming has been greatest in the Arctic band (60°-90°N), followed by 30°-60°N, the two tropical bands (00°-30°N and 00°-30°S) and then the mid-latitude band at 30°-60°S.

The anomaly in total cloud cover decreased from 1987 to year 2000 in 30°-60°N, 00°-30°N and 00°-30°S. The Arctic band (60°-90°N) suffered a general decrease over the

entire period for which data is available, 1984 to 2009, but for the two remaining bands, 30°-60°S and 60°-90°S, the trend in total cloud cover has been virtually flat. Changes in total cloud cover correspond well to temperature changes in each of these six latitude bands band.

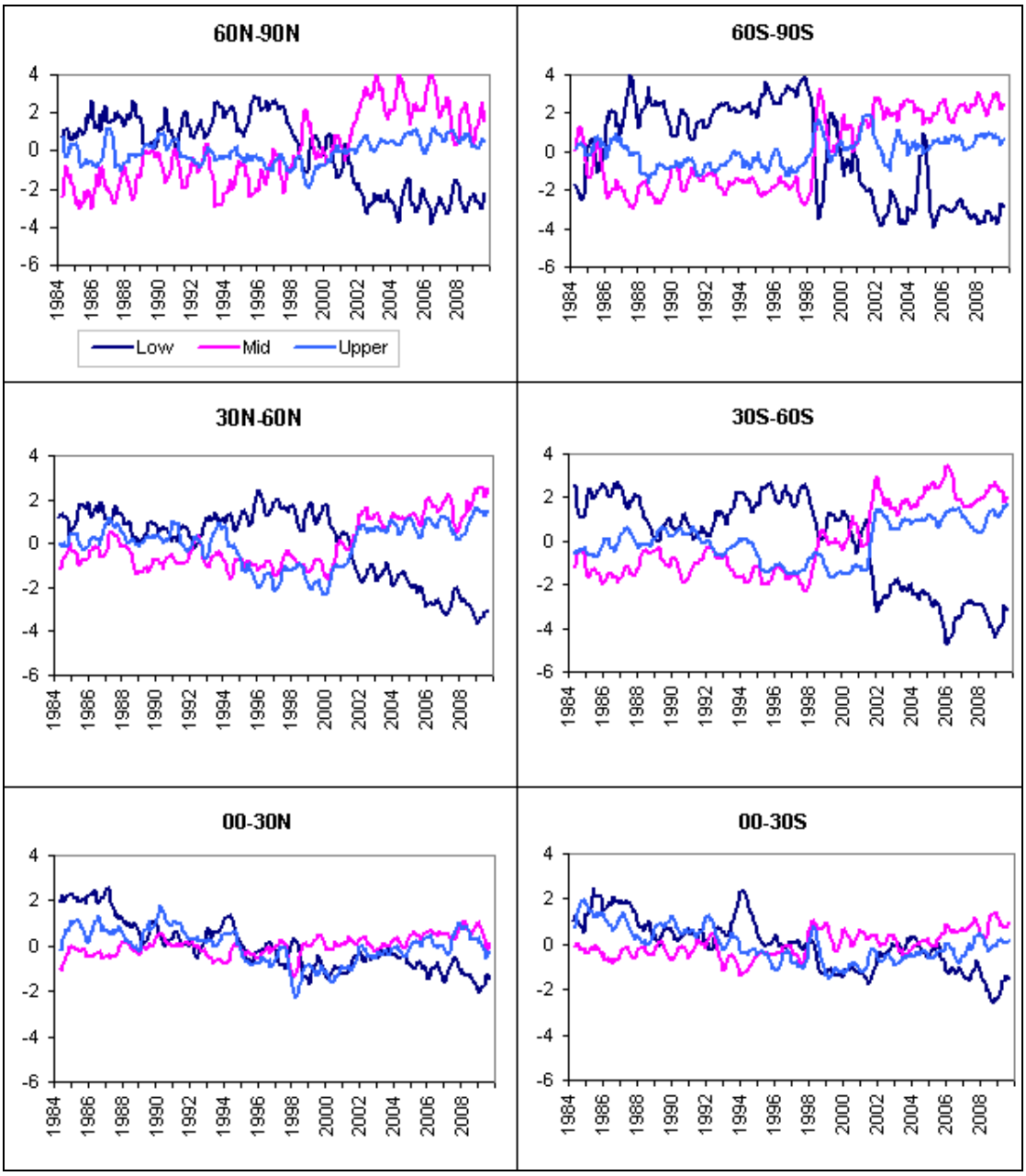


Figure 9.11 - Anomalies in cloud cover at low, mid and upper levels for each latitude band.

Latitude Band	Mean Total Cloud Cover	Mean Low Cloud Cover	Mean Mid Cloud Cover	Mean Upper Cloud Cover	Correlation L to (M + U)
60 - 90 N	68.429	18.283	36.537	6.888	−0.754
30 - 60N	69.707	25.336	24.986	14.465	−0.807
00 - 30N	58.527	24.495	14.596	14.305	−0.032
00 - 30S	59.615	29.459	13.145	12.345	−0.087
30 - 60S	80.852	36.741	24.930	13.388	−0.951
60 - 90S	69.210	17.433	31.371	14.322	−0.875

Table 9-2 - Mean cloud cover at each level for each latitude band and correlation between low level cloud and combined mid and upper level cloud.

Both tropical bands (00°-30°N and 00°-30°S) experienced a decrease in low-level cloud, offset rather imperfectly by an increase in mid and upper level cloud. This shift was more gradual and smaller than the similar changes in cloud cover at mid latitudes and high latitudes where the negative correlation between low cloud cover and the sum of the mid and upper level cloud cover is strong. In the polar bands most of the shift occurred between 1997 and 2002, after which time cloud cover trends flattened, but in the mid-latitudes the shift began later and low-level cloud was still decreasing at the end of 2009 when ISCCP data currently ends. The shift from low cloud to mid and upper level cloud will likely mean an increase in temperature but low or zero solar insolation at high latitudes in winter mean that the consequences of changes in cloud cover are more complex.

This regional analysis shows that the variations in temperature and cloud cover described by the global analysis were widespread across the planet and not due to localised extreme variations.

4. Discussion and Conclusions

This paper has shown that the variation in the global average temperature anomaly since 1950 has three distinct phases, the periods for which are 1950-1987, 1988-1997 and post 1997 respectively.

(a) Phase 1 (1950-1987)

This phase was dominated by variations driven largely by the El Nino-Southern Oscillation, interrupted at times by irregular periods of cooling caused by volcanic eruptions. It can be further split into the La Nina dominated period 1950-1976 and the El Nino dominated 1977-1987. Two La Nina events occurred during the 1950's, one of sixteen months, and no El Nino events. The first El Nino event after 1950 occurred in the latter half of 196 and the next in1972. Three El Nino events, two of which continued for almost 12 months, occurred during the period from 1977 to 1987, but no La Nina events.

It is often implied that the ENSO is a tri-state phenomena but it is a continuum with arbitrary thresholds for La Nina and El Nino events. A better analysis is achieved by examining the Southern Oscillation Index rather than the three defined states.

Averaging the Troup SOI over approximately equal periods across the time span shows the dominance of conditions on the La Nina side of absolutely neutral (i.e. index = 0) during 1950-1976 and the dominance of conditions on the El Nino side after 1976. From 1950 to 1962 the average Troup SOI was 2.61 (σ =9.1), for 1963-75 was 2.37 (σ =10.3) and for 1977-1987 was -4.62 (σ= 9.8). Over the period 1950 to 1987 therefore, the average Troup SOI shifted from the La Nina side of absolutely neutral to a figure almost double and on the El Nino side of absolutely neutral.

The trends in the monthly SOI data over this time appear to be related to temperature trends. From 1950 to 1976 the trend in the Troup SOI data was 0.066/year (i.e. towards La Nina) and from 1977 to 1987 was -0.0111/year, with an overall trend from 1950 to 1987 of -0.21/year. The trend in monthly HadCRUT4 global average temperature anomalies for the corresponding periods were -0.0024°/year, +0.006°/year and +0.0036°/year. Note that the temperature data used for these calculations includes data from periods when volcanic eruptions caused cooling and the precision of the figures is debatable, but an ENSO shift towards El Nino conditions and a corresponding increase in temperatures after 1977 are indicated (see also Figure 1).

(b) Phase II (1988-1997)

During this second phase the ENSO and volcanic eruptions continued to influence the temperature but after excluding the influence of the ENSO, the residual average global temperature anomaly was found to rise across the decade and by 1997 had reached ~0.48°C above the 1961-90 average.

Across this decade the global average temperature anomalies from the CRUTEM3 dataset, based on data from observation stations, increasingly diverged from corresponding averages from the HadSST3 dataset whose data is derived from sea surface temperatures.

The temperature pattern for the period 1988-1997 appears to be generally consistent with the 7% reduction in total cloud cover that occurred across the period 1987 to 1999. Applying that reduction to the influence of clouds in the energy budget described by Trenberth et al [34] results in an increased average solar forcing at the Earth's surface of about 5 Wm^{-2}. This increase is more than double the IPCC's estimated radiative forcing from all anthropogenic emissions of greenhouse gases.

The analysis by six 30° latitude bands shows a loss in total cloud cover in four of the bands but not occurring in the two southern-most bands, 30°-60°S and 60°-90°S.

The reduction in total cloud cover during the period from 1987 to 1999 could also account for the divergence of CRUTEM4 and HadSST3 temperatures. An increase in solar radiation will pass deeper into the ocean, to about 100 metres in clear calm tropical water, and the heat disperse far more than when the same amount of radiation strikes the ground surface.

(c) Phase III (since 1997)

The third phase of the post-1950 pattern in the average global temperature anomaly is the plateauing of the anomaly after 1997.

From 1999 to 2009 total cloud cover increased slightly from its low point but at the same time there was a widespread reduction in low-level cloud that was almost matched by increases in cloud cover at middle and upper levels. This finding applies to the global average cloud cover and to latitudes 30° to 90° in each hemisphere.

Less opaque low-level cloud and more translucent mid and upper-level cloud meant greater solar radiation, albeit diffuse radiation, reaching the Earth's surface. This approximately coincides with both the flat trend in average global temperature anomaly and the a stabilising of the difference between CRUTEM4 and HadSST3 temperature datasets although data "noise", particularly in the former, makes confident conclusions difficult.

5. Conclusions

Since 1950 global average temperature anomalies have been driven firstly, from 1950 to 1987, by a sustained shift in ENSO conditions, by reductions in total cloud cover (1987 to late 1990s) and then a shift from low cloud to mid and high-level cloud, with both changes in cloud cover being very widespread.

According to the energy balance described by Trenberth et al (2009) [34], the reduction in total cloud cover more than accounts for the increase in temperature since 1987, leaving little, if any, of temperature change to be attributed to other forcings.

With ISCCP cloud cover data available only for the period from 1984 to 2009 this hypothesis should be regarded as tentative.

References

[1] IPCC. (2013) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, eds. Stocker, T.F., D. Qin, G-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley,. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp.

[2] IPCC. (2007) Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. eds. Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 996 pp.

[3] Hartmann, D.L., Moy L.A. and Fu Q. (2001) Tropical Convection and the energy balance at the top of the atmosphere. *J. Climate*, **14**, 4495-4511. http://dx.doi.org/10.1175/1520-0442(2001)014<4495:TCATEB>2.0.CO;2

[4] Keihl, J.T. (1994). On the observed near cancellation between longwave and shortwave cloud forcing in tropical regions. *J. Climate*, **7**, 559-565. http://dx.doi.org/10.1175/1520-0442(1994)007<0559:OTONCB>2.0.CO;2

[5] Goode, P.R. and Pallé E. (2007) Shortwave forcing of the Earth's climate: Modern and historical variations in the Sun's irradiance and the Earth's reflectance. *Journal of Atmospheric and Solar-Terrestrial Physics*, **69**, 1556-1568. http://dx.doi.org/10.1016/j.jastp.2007.06.011

[6] Herman, J., DeLand, M. T., Huang, L.-K., Labow, G., Larko, D., Lloyd, S. A., Mao, J., Qin, W., and Weaver, C. (2012) A net decrease in the Earth's cloud, aerosol, and surface 340 nm reflectivity during the past 33 yr (1979–2011). *Atmos. Chem. Phys*, **13**, 8505-8524. http://dx.doi.org/10.5194/acp-13-8505-2013

[7] Kauppinen, J., Heinonen, J. and Malmi P. (2014) Influence of relative humidity and clouds on the global mean surface temperature. *Energy and Environment*, **25**, 389-400. http://dx.doi.org/10.1260/0958-305X.25.2.389

[8] Eastman, R. and Warren, S.G. (2011) Variation in Cloud Cover and Cloud Types over the Ocean from Surface Observations, 1954-2008. *Journal of Climate*, **24**, 5914-5934. http://dx.doi.org/10.1175/2011JCLI3972.1

[9] Eastman, R., Warren S.G. and Hahn, C.J. (2013) A 39-yr Survey of Cloud Changes from Land Stations Worldwide 1971-2009: Long-Term Trends, Relation to Aerosols, and Expansion of the Tropical Belt, *Journal of Climate*, **26**, 1286-1303 http://dx.doi.org/10.1175/JCLI-D-12-00280.1

[10] Rossow, W.B. and Schiffer R.A. (1991). ISCCP Cloud Data Products. *BAMS*, **72** No. 1, 2-20. http://dx.doi.org/10.1175/1520-0477(1991)072<0002:ICDP>2.0.CO;2

[11] Troup, A.J. (1965). The Southern Oscillation. *Quarterly Journal of the Royal Meteorological Society*. **91,** 390, 490-506 http://dx.doi.org/10.1002/qj.49709139009

[12] Lamb, H.H. (1970) Volcanic Dust in the Atmosphere; With a Chronology and Assessment of its Meteorological Significance. *Philosophical Transactions of the Royal Society of London, Series A*, **266**, 425-533. http://dx.doi.org/10.1098/rsta.1970.0010

[13] Lamb, H.H. (1977) Supplementary Volcanic Dust Veil Assessments. *Climate Monitor*, **6**, 57-67.

[14] Lamb, H. H. (1983) Update of the Chronology of Assessment of the Volcanic Dust Veil Index. *Climate Monitor*, **12**, 79-90.

[15] Newhall, C.G. and Self, S. (1982) The volcanic explosivity index (VEI): An estimate of explosive magnitude for historical volcanism. *Journal of Geophysical Research*, **87**, 1231–1238. http://dx.doi.org/10.1029/JC087iC02p01231

[16] Robock, A. (1981) A latitudinally dependent volcanic dust veil index and its effect on climate simulations. *J. Volcanol. Geotherm. Res.*, **11**, 67-80. http://dx.doi.org/10.1016/0377-0273(81)90076-7

[17] Bradley, R.S. and Jones, P.D. (1992) Records of explosive volcanic eruptions over the last 500 years. In: R.S. Bradley and P.D. Jones, eds Climate Since A.D. 1500, Routledge, London, chap 31.

[18] Sato, M., Hansen J.E., McCormick, M.P. and Pollak J.B. (1993) Stratospheric aerosol optical depths, 1850–1990. *J. Geophys. Res*. **98**, 22987–22994 http://dx.doi.org/10.1029/93JD02553

[19] Gu, G. and Adler, R.F. (2011) Precipitation and Temperature Variations on the Interannual Time Scale: Assessing the Impact of ENSO and Volcanic Eruptions. *J. Climate*, **24**, 2258-2270. http://dx.doi.org/10.1175/2010JCLI3727.1

[20] Wigley, T.M.L. (2000) ENSO, volcanoes and record-breaking temperatures. *Geophys. Res. Lett*. **24**, 4101-4104. http://dx.doi.org/10.1029/2000GL012159

[21] Emile-Geay, J., Seager, R., Cane, M.A. and Cook, E.R. (2008) Volcanoes and ENSO over the Past Millennium. *J. Climate*, **21**, 3134-3148. http://dx.doi.org/10.1175/2007JCLI1884.1

[22] Korecha, D and Barnston, A.G. (2007) Predictability of June–September Rainfall in Ethiopia. *Monthly Weather Review*, **135**, 628-650. http://dx.doi.org/10.1175/MWR3304.1

[23] Ogutu, J.O., Piepho, H.-P, Dublin, H.T., Bhola, N. and Reid, R.S. (2008) El Niño-Southern Oscillation, rainfall, temperature and Normalized Difference Vegetation Index fluctuations in the Mara-Serengeti ecosystem. *S. African Journal of Ecology*, **46**, 132-143. http://dx.doi.org/10.1111/j.1365-2028.2007.00821.x

[24] Shaman J and Tziperman, E. (2011) An Atmospheric Teleconnection Linking ENSO and Southwestern European Precipitation. *J. Climate*, **24**, 124-139. http://dx.doi.org/10.1175/2010JCLI3590.1

[25] Brönnimann, S., Xoplaki, E., Casty C., Pauling, A. and Luterbacher, J. (2007) ENSO influence on Europe during the last centuries. *Climate Dynamics*, **28**, 181–197. http://dx.doi.org/10.1007/s00382-006-0175-z

[26] Trenberth, K.E., Caron, J.M., Stepaniak, D.P., and Worley, S. (2002) Evolution of El Niño–Southern Oscillation and global atmospheric surface temperatures. *J. Geophys. Res*, 107(D8), 4065, http://dx.doi.org/10.1029/2000JD000298

[27] Jones, P.D. (1989) The influence of ENSO on global temperatures. *Climate Monitor*, **17**, 80–89.

[28] de Freitas, C.R. and McLean J.D. (2013) Update of the Chronology of Natural Signals in the Near-Surface Mean Global Temperature Record and the Southern Oscillation Index. *International Journal of Geosciences*, **4**, 234-239. http://dx.doi.org/10.4236/ijg.2013.41A020

[29] Hansen, J.E., Wang W.-C. and Lacis, A.A. (1978) Mount Agung eruption provides test of global climate Perturbation. *Science*, **199**, 1065-1068. http://dx.doi.org/10.1126/science.199.4333.1065

[30] Dutton, E.G. and Christy, J.R. (1992) Solar radiative forcing at selected locations and evidence for global lower tropospheric cooling following the eruptions of El Chichon and Pinatubo. *Geophys. Res. Lett.* **19**, 2313-2316. http://dx.doi.org/10.1029/92GL02495

[31] Douglass, D.H. and Knox, R.S. (2005) Climate forcing by the volcanic eruption of Mount Pinatubo. *Geophys. Res. Lett*., **32**, L05710. http://dx.doi.org/10.1029/2004GL022119

[32] Thompson, D.W.J., Wallace, J.M., Jones, P.D. and Kennedy, J.J. (2009) Identifying signatures of natural climate variability in time series of global-mean surface temperature: Methodology and insights. *J. Climate*, **22**, 6120–6141. http://dx.doi.org/10.1175/2009JCLI3089.1

[33] Foster, G. and Rahmstorf, S. (2011) Global Temperature evolution. *Environ. Res. Lett.* **6**, 044022 (8 pp) http://dx.doi.org/10.1088/1748-9326/6/4/044022

[34] Trenberth, K.E., Fasulo, J.T. and Keihl, J. 2009. Earth's Global Energy Budget. *BAMS*, **90**, 311-323. http://dx.doi.org/10.1175/2008BAMS2634.1

Chapter 10: Improving the Troup SOI

10.1 Introduction

The El Niño-Southern Oscillation (ENSO) is an important influence on the weather patterns of the Pacific Ocean and has a lesser but still significant affect on global weather. Three commonly used indices of the ENSO state are Oceanic Nino Index (ONI), the Multi-variate ENSO Index (MEI) and the Troup Southern Oscillation Index (Troup SOI or simply SOI).

All three indices use threshold values to determine which of the three states the ENSO is in at the time - La Niña (widespread cooler and wetter conditions), El Niño (widespread warmer and drier conditions) and neutral (a middle state between these two) - but this is misleading because the range of values is continuous in each index and conditions that fall just short of some threshold are very similar to conditions that just over that threshold.

The ONI used by the US National Oceanographic and Atmospheric Administration (NOAA) is based on average sea surface temperature (SST) in the eastern Pacific, specifically an area covering latitudes 5°N to 5°S and longitudes 120°W to 170°W, known as region "Nino 3.4", with El Niño and La Nina thresholds of +0.5°C and -0.5°C respectively for the three-month running mean SST.

The ONI commences in 1950 but data coverage of the designated area is far from complete in the early years. The Nino 3.4 region covers 20 HadSST3 grid cells but from 1950 to 1959 the annual average number of reporting cells did not exceed 15 and in 57 of the 120 months have less than half of those grid cells report SST data. The cells without data were typically in the western end of the region meaning that milder El Nino conditions, which even with complete coverage might have had little impact on SST at the region's eastern end, are likely to have been unrecorded.

One disadvantage of the ONI is that it is focussed on a single region of the tropical Pacific and therefore related to conditions only within this area. Another problem is that data for five successive overlapping three-month "seasons" (i.e. centred 3-month averages) must be collected before either El Niño or La Niña episodes are identified, with the threshold for El Niño being SST increases of more than 0.5°C in the five seasons and that for La Niña being decreases of more than 0.5°C. The use of 3-month averages means that data for seven months must be collected before the ENSO thresholds are reached, but according to other indices the duration of such events often falls fall short of seven months.

The MEI (Wolter and Timlin, 1993) incorporates sea-level pressure (SLP) with five other environmental factors - zonal and meridional wind, SST, surface air temperature and cloudiness - when it generates a single value to describe the ENSO state. These factors are not independent e.g. a warmer ocean will cause convection that causes a decrease in SLP and an increase in cloud cover. Meridional wind will also cause overturning of ocean water and therefore mixing of surface water with lower layers, and the absence of wind in the tropics can cause thermal layers to form in the water, both issues making it more difficult to accurately determine sea surface temperature [see chapter 7]. Like the ONI the MEI is an imperfect index.

The third index is the Troup SOI, based on the difference in monthly mean seal level air pressure (MSLP) at Tahiti and Darwin. This SOI will be discussed in greater detail below but for now the key point is that according to the Australian Bureau of Meteorology, sustained (three months or more) Troup SOI values of less than -7 "typically indicate" El Niño episodes and sustained values of greater than +7 "typically indicate" La Niña episodes[28]

The three indices for the period 1991 to 2000 are shown in Figure 10.1. The MEI and ONI use average values over two and three months respectively and here the Troup SOI is shown as a 3-month centred average for comparison. Small, short-term variations between the indices are to be expected but sustained differences can be seen

[28] From http://www.bom.gov.au/climate/enso/#tabs=SOI but other BoM web pages say -8 and + 8 (eg. http://www.bom.gov.au/climate/enso/history/ln-2010-12/SOI-what.shtml). Until just a few years ago the term "sustained for at least 3 months" was used by the BoM but now it is simply saying "sustained".

during strong El Niño episodes (eg. 1983 and 1997) and milder El Niño episodes (1992-1994).

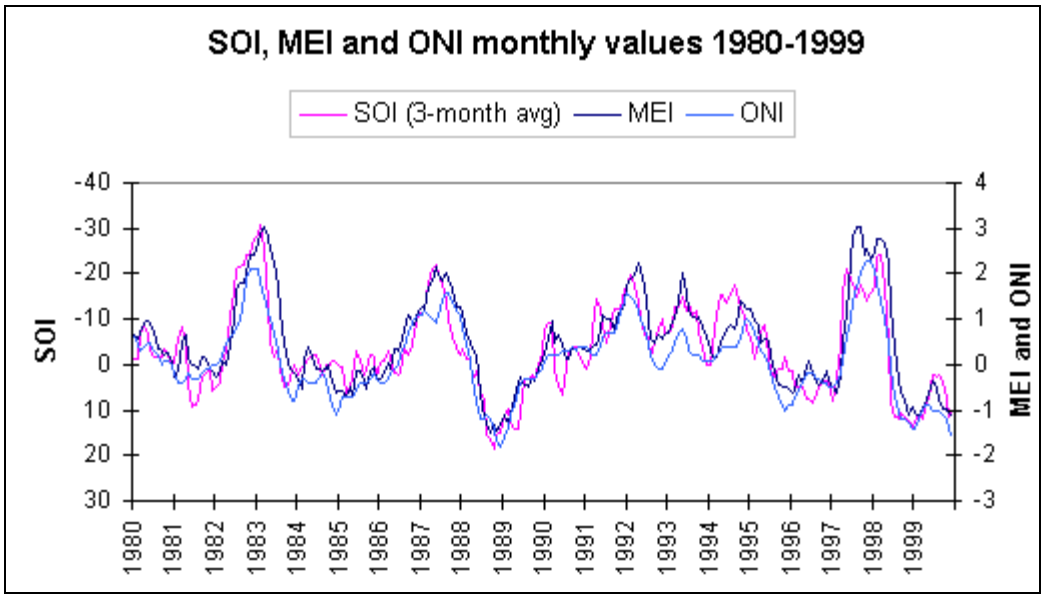


Figure 10.1 Monthly values of three ENSO indices 1991-2000

The Troup SOI is an MSLP-based index to ENSO conditions. The use of the MSLP as an ENSO index has the advantage that many of the characteristics of ENSO events are associated with the state of, or changes in, air pressure and therefore the SLP is a direct measure of an important part of the phenomenon.

Relatively few papers seem to recognise the MSLP-ENSO relationship. Kousky et al (1984) reviews the state of knowledge about the ENSO to that time and discusses what it terms a pressure "seesaw" between the Indian and Pacific Oceans, the former actually meaning the waters near Indonesia. It cites Berlage and de Boer (1959) as identifying the two principal regions of MSLP anomaly correlation, in this case from 1949 to 1957, using MSLP at Easter Island as the base against which other locations were compared. It goes on to talk about indices being created from MSLP at Easter Island and Djakarta but it follows through with the findings of Berlage and de Boer (1959) and when it describes the ENSO it makes no mention of the implications of the changes in MSLP at the two locations.

Reason et al (2000) discussed the relationship of SST and MSLP anomaly in the Indian Ocean and show that both could be linked to rainfall patterns. Their analysis of MSLP data is broken into seasons and it deals only with MSLP anomalies, not standardized anomalies based on long term averages and standard deviations. The aim of the paper appears to be the clarification of certain Indian Ocean characteristics of El Niño and La Niña events rather than to further understanding of the structure of the ENSO system as a whole.

Cai et al (2011) discuss MSLP but only in the context of its variation and how the ENSO affects Australia's climate. Dijkstra (2006) mentions that the Southern Oscillation is "related to changes in surface pressure", discussing MSLP, as do Wolter and Timlin (2011), in terms of characteristics of ENSO. Clement et al (2011) supports this by showing that according to modelling, the Southern Oscillation is the dominant mode of tropical Pacific variability, without needing dynamic coupling to the ocean.

Sea level pressure is conventionally linked to certain weather patterns and its influence is usually expressed in relative rather than absolute terms (e.g. "warmer" rather than "warm") because basic weather at any location depends largely on the latitude (and hence month or season) and on the altitude. The same can be said about how the mean seal level pressure (MSLP) is related to characteristics of El Niño events (Table 10-1(a)) and La Niña events (Table 10-1(b)). (For general context in diagrammatic form see Appendix 3.)

One example of the association with MSLP can be seen in the manner in which the ENSO influences the Indian monsoon, which occurs during the period from July to September. India's monsoonal rains develop when the very hot summer air over the land rises and is replaced with cooler air from over the ocean, laden with moisture picked up by westerly winds. La Niña events, associated with reduced MSLP in Southern India and over the nearby part of the Indian Ocean, mean stronger westerly winds and therefore above-average rainfall is likely on the subcontinent. In contrast El Niño events, associated with increased MSLP in the regions just described, impede or limit those westerly winds and therefore below-average rainfall is likely in India (Reason et al, 2000; Kug and Kang, 2006; Kumar et al, 2006).

The remainder of this chapter will discuss the nature of the MSLP-based Troup SOI, its shortcomings as an ENSO index and how an alternative but associated "ensemble" index might resolve some of its problems.

Factor under El Niño conditions	Relationship to SLP
Surface air pressure across the Pacific	**Increased MSLP in the west, decreased MSLP in the east**
Weather in the western Pacific and nearby - **Warmer and/or drier**	Increased MSLP in the west of the Pacific will typically mean finer weather and less rainfall.
Weather in the eastern Pacific and on the nearby coast - **Cooler and wetter**	Decreased MSLP will likely mean increased evaporation and increased cloud cover and therefore greater chance of rainfall.
Pacific trade winds (easterlies) - **Weaker, wind might even reverse**.	Decreased MSLP means a reduction in wind and increased MSLP in the western Pacific acts against these winds.
Upwelling of cold water in the eastern Pacific - **Weaker**	Less wind means less ocean overturning and less opportunity for cold water to reach the surface.
Sea level - **Flatter across the Pacific**	Elevated pressure and reduced winds mean less "piling up" of the water in the west, which is to say, flatter sea surface
East-west thermoclines - **Flatter across Pacific**	Reduced winds means reduced current and less surface overturning, meaning calmer seas and greater opportunity for the sea to warm in transit along the equatorial current
Walker Circulation (west to east across the Pacific well above the surface) - **Weaker, might even reverse**	The elevated pressure in the west reduces the easterly winds so less air will circulate.
Hadley Circulation (poleward from the equator) - **Stronger**	The elevated pressure in the west pushes air outwards, including to the north. .
Tropical cyclones, typhoons, hurricanes (different names but same thing) in Indian and Pacific Oceans - **Fewer**	Elevated pressure in the west is forcing air outwards, also the difference in surface pressure across the Pacific is reduced and therefore the key driver of these events is reduced.
Summer monsoon rainfall in India and Australia - **Start later and less rainfall**	Increased MSLP acts somewhat like a blocking High to impede the monsoon process.
Cloud near the intersection of the equator and Date Line - **Increased**	Convection increased when air from the east strikes the leading edge of the elevated MSLP.

Table 10-1(a) The characteristics of an El Niño ENSO state described with reference to MSLP

Factor under La Niña conditions	Relationship to SLP
Sea level pressure	**Decreased MSLP in the west and increased MSLP in the east**
Weather in the western Pacific and nearby - **Cooler and/or wetter**	Reduced MSLP in the west means more evaporation and convection, so more cloud and likely increased rainfall.
Weather in the eastern Pacific and on the nearby coast - **Warmer and drier**	Increase in MSLP near South America brings fine weather and impedes on-shore winds carrying moisture.
Pacific trade winds (easterlies) - **Stronger**	Increased MSLP in east pushing winds across Pacific.
Upwelling of cold water in the eastern Pacific - **Stronger**	More upwelling of cold water and more mixing of surface layer due to the increased winds.
Sea level - **Higher in the west**	The combination of wind-driven water, reduced MSLP and thermal expansion of water cause higher sea level.
Thermoclines (eg. for 24C) - **Deeper in the west**	More upwelling in the east means colder water at or near the surface. In the west, cloud impedes ocean cooling.
Walker Circulation (west to east across the Pacific well above the surface) - **Stronger**	Easterly winds mean an enhanced Walker Circulation cycle.
Hadley Circulation (poleward from the equator) - **Weaker**	No impediment to easterly winds across the Pacific so convection occurring in the western rather than central Pacific.
Tropical cyclones, typhoons, hurricanes (different names but same thing) in Indian and Pacific Oceans - **Increase**	Tropical cyclones are more easily generated in the narrow central Pacific transition zone between the regions of increased and decreased MSLP.
Summer monsoon rainfall in India and Australia - **Starts earlier and more rainfall**	Increase in the amount of moisture-laden air reaching land.
Cloud near the intersection of the equator and Date Line - **Decreased**	Convection occurring further west therefore clouds developing further west.

Table 10-1(b) The characteristics of a La Niña ENSO state described with reference to MSLP.

10.2 The background and calculation of the Troup SOI

The Troup SOI is based on the oscillation of sea level pressure across the Pacific that typically sees the MSLP at Tahiti rising as the MSLP at Darwin falls and *vice versa* (Figure 10.2).

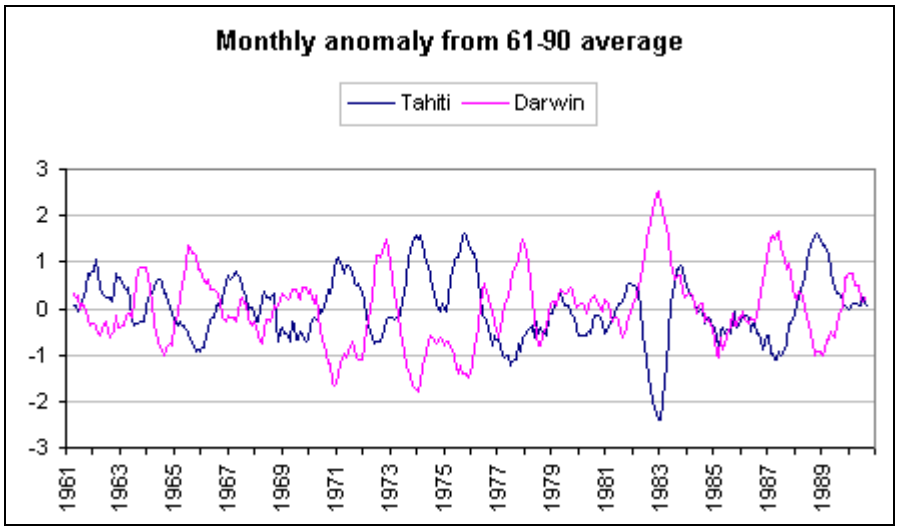


Figure 10.2 Monthly anomalies in MSLP from the 1961-1990 average at Tahiti and Darwin, with the general pattern of MSLP at one location rising while it falls at the other. (Seven-month centred averages)

The Troup SOI is derived from the difference in monthly mean sea level pressure (MSLP) at Tahiti, French Polynesia, and Darwin, Australia, (i.e. MSLP at Tahiti *minus* MSLP at Darwin). Figure 10.3 shows the monthly MSLP at the two locations in each year from 1961 to 1990, sorted into calendar month. The annual cycle in MSLP varies at the two locations, with the range at Darwin being ~6hPa and at Tahiti ~4hpa, leading to average differences in MSLP that range from ~1 hPa in July to ~4 hPa in January.

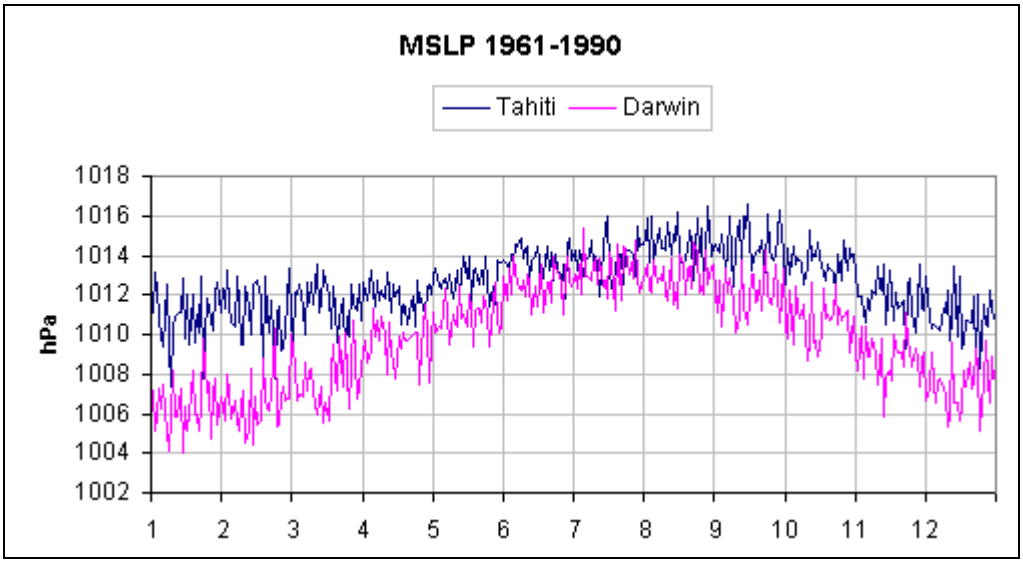


Figure 10.3 Monthly MSLP at Darwin and Tahiti from 1961 to 1990 sorted by calendar month

The Australian Bureau of Meteorology defines the Troup Southern Oscillation Index (SOI) as ten times the standardised anomaly of the Mean Sea Level Pressure (MSLP) difference between Tahiti and Darwin. It is calculated according to

$$\text{Troup SOI} = 10 \frac{[\ \text{Pdiff} - \text{avgPdiff}\ ]}{\text{SD(Pdiff)}}$$

where:

`Pdiff` = (Tahiti MSLP for the month) *minus* (Darwin MSLP for the month),

`avgPdiff` = long term average[29] of `Pdiff` for the month in question, and

`SD(Pdiff)` = standard deviation associated with `avgPdiff`

The multiplication by 10 is a convention to make it possible to express the range as a whole number without any significant loss of accuracy.

Based on the probability associated with Normal distribution and ignoring for the moment the need for thresholds to be exceeded for at least three months, the

[29] In the above calculations 'long term' refers to the period from 1933 to 1992 inclusive according to the Bureau of Meteorology at http:://www.bom.gov.au/climate/current/soihtm1.shtml

thresholds of -7 (for El Niño) and +7 (for La Niña) mean an approximate probability of 0.516 for "neutral" conditions and 0.242 for each of El Niño and La Niña.

The primary variable factor in the above equation is Pdiff, which is the monthly MSLP at Tahiti *minus* the monthly MSLP at Darwin, the whole top line of the equation, the numerator, being the anomaly in Pdiff (i.e. difference from the long-term average). When the Tahiti MSLP increases the SOI will increase and when it falls the SOI will decrease. Figure 10.4 shows the monthly Troup SOI from September 1935 to December 2015 in the form of a 7-month centred average, a compromise between "noisy" data with frequent very short-term fluctuations and over-smoothing that might minimise critical information. (Subsequent Figures will use similar averages for consistency and clarity.).

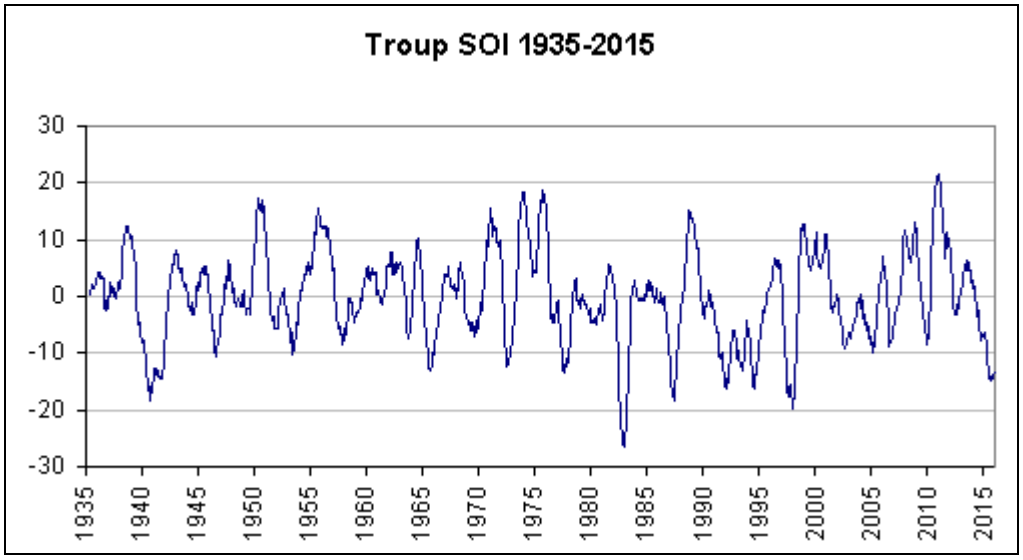


Figure 10.4 Troup SOI from 1935 to 2015. Positive values greater than +7 sustained for more than 3 months correspond to La Niña events and negative values of less than -7 sustained for the same length of time indicate El Niño events.

In its discussion of the Troup SOI the Australian Bureau of Meteorology states "Daily or weekly values of the SOI do not convey much in the way of useful information about the current state of the climate, and accordingly the Bureau of Meteorology does not issue them. Daily values in particular can fluctuate markedly because of

daily weather patterns, and should not be used for climate purposes."[30] The underlying problem is that the data from which the Troup SOI is calculated is measured at just two locations. Local weather events at these locations can cause data "noise" (i.e. very short-term irregular fluctuations in the calculated value), which potentially means positives or false negatives regards meeting thresholds for El Niño and La Niña conditions.

This data noise and a serious problem with the Troup SOI prior to 1935 warrant discussion prior to considering the development of an alternative but similar SOI index.

10.3 Analysis

10.3.1 Investigation of the data "noise"

The Troup SOI is a particularly noisy signal with the average month-to-month variation (i.e. absolute value of difference) from January 1940 to December 2015 being 6.86 (σ = 5.36). These variations are very large given that the difference between El Niño and La Niña thresholds is 14 units - one has a value of +7 and the other of -7. The reason for this large variation lies in the local MSLP. Table 10-2 shows the monthly average mean difference in MSLP at Tahiti and Darwin, along with the standard deviations, according to all MSLP data from 1933 to 1992. Based on the method by which the Troup SOI is calculated (see above) if a particular SLP state continues for more than a few days into the next month it can have a big impact on the Troup SOI because of the different mean differences and standard deviations (e.g. the shift in the mean difference and standard deviation from March to April are falls of more than 33%).

[30] See http://poama.bom.gov.au/climate/glossary/soi.shtml

Month	Mean (hPa)	Std Dev (hPa)
1	4.43	2.10
2	4.57	2.12
3	3.96	1.84
4	2.36	1.21
5	1.64	1.27
6	1.18	1.23
7	0.96	1.53
8	1.68	1.56
9	2.35	1.66
10	2.91	1.64
11	2.91	1.52
12	3.48	1.96

Table 10-2 Long-term (1933-92) mean differences in MSLP between Tahiti and Darwin, with associated standard deviation, for each calendar month.

As mentioned above, the thresholds for El Niño and La Niña events are Troup values of -7 and + 7, which means a variation of ±0.7σ from the mean value. From Table 10-1 this variation is equivalent to differences in the MSLP at the two sites that, depending on the calendar month, vary between 1.47hPa and 0.85hPa from the long-term mean difference for that month.

Based on data from Australia's Bureau of Meteorology this range of differences is equivalent to ~25% of the range in MSLP at Darwin in any month, this conclusion derived from 9:00am and 3:00pm recordings at Darwin, from March 2016 to February 2017, showing SLP monthly range that varied from 4.5hPa during May, June and July to 9.32 hPa for the period from December to February[31]. On this basis a small shift in the Darwin MSLP alone, i.e. without considering the MSLP shifts at Tahiti, could easily result in substantial short-term changes in the Troup SOI.

These short-term shifts cause data "noise" and make it difficult to interpret Troup SOI values. It is not only a question of whether specific values are a consequence of short-

[31] From daily details of the last 13 months provided by the Bureau of Meteorology at http://www.bom.gov.au/climate/dwo/IDCJDW8014.latest.shtml

term weather but the start and end dates of El Niño, La Niña and neutral episodes are unclear, as are apparent interruptions to those sequences.

Accurate probabilistic analysis is not possible but if it is assumed that 25% of the MSLP is distorted by short term weather, in keeping with the figure mentioned above for Darwin, the probability of having just two successive months without that distortion is 0.563 (i.e. 0.75 x 0.75) and three successive months is 0.422.

With the Troup SOI so seriously impacted by data "noise", caused by weather events at the two locations, a reduction in this noise is highly desirable.

10.3.2 Accuracy of the Troup SOI prior to 1935

The Troup SOI data is available from 1877 onwards but there are several reasons why the data prior to 1935 should be treated with caution.

Firstly, the meteorological data recorded at Tahiti prior to 1935 is already known to be suspect. CRUTEM4 temperature data from Tahiti (station ID 919380) commences in August 1876, has six months missing from 1877, more data missing from April 1891 to January 1898 and data missing again from January 1908 to January 1935. Ropelewski and Jones (1987) say that from 1875 to 1933 the meteorological data for Tahiti were recorded at Papeete hospital rather than at an observation station. It is an open question as to whether the measuring and recording of the data at that time conformed to standard meteorological practices.

Secondly the data from Tahiti since 1877 can be split into two periods, the first, from 1877 to 1935, being 58 years and 41.5% of the total, and the second, from 1936 to 2016, being 82 years and 58.5% of the total. Of the total number of instances where the MSLP at Tahiti was unchanged from one month to the next 64.7% occurred in the first period and just 35.3% in the second, whereas for Darwin the split is very close to 40% to 60%, which is very close to evenly distributed given the duration of the two periods. The metric is indicative rather than absolute but it suggests critical

differences between the recording of MSLP at Tahiti prior to 1936 and the recording after that year.

Thirdly, and perhaps related, the negative correlation between MSLP at Darwin and at Tahiti is considerably weaker in the early period. The monthly MSLP anomalies shown earlier in Figure 10.2 are derived from the 1961-1990 long-term average MSLP at each site but while the correlation between Darwin and Tahiti MSLP anomalies is -0.341 from 1936 to 2016 the correlation for 1877-1935, as shown in Figure 10.5, is weaker at just -0.206.

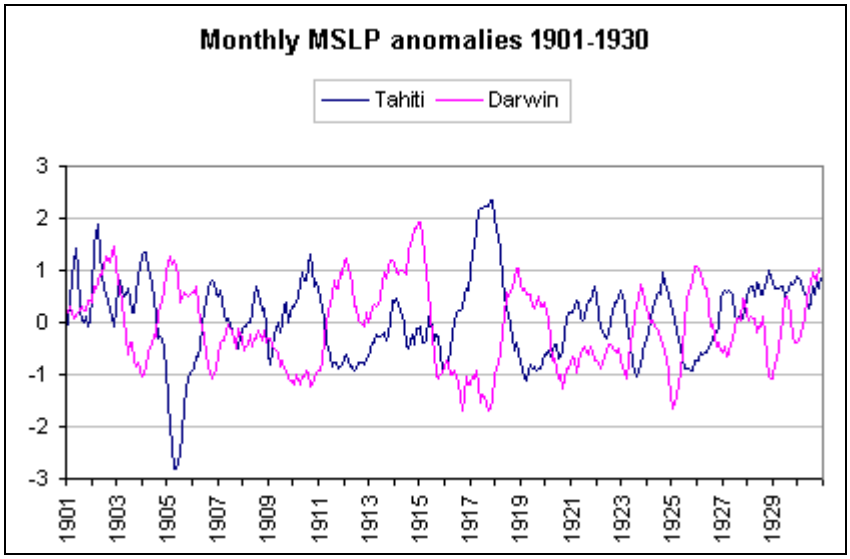


Figure 10.5 Monthly anomalies in MSLP, derived from the 1961-1990 averages, at Tahiti and Darwin.

On the basis of the above it seems unwise to make any use whatsoever of the Troup SOI data prior to 1935.

10.3.3 Background to the derivation of an improved MSLP-based index

It was shown earlier that as the MSLP at Darwin rises the MSLP at Tahiti falls. The nature of this oscillation means that the MSLP at either Darwin or Tahiti can be used individually to generate a reasonable approximation of the SOI because a similar but inverted MSLP state will very likely be occurring at the at the other location.

A 'local SOI' can therefore be created, based on the anomaly in mean local monthly MSLP. In order to take into account the variable range of MSLP in each month, this local SOI can be calculated by standardisation, in a similar manner to the Troup SOI, namely:

$$\text{Normalised MSLP} = \frac{[\ \text{MSLP} - \text{avMSLP}\ ]}{\text{SD(MSLP)}}$$

where:
MSLP = mean MSLP for the month,
avMSLP = long term average of MSLP for the month in question, and
SD(MSLP) = standard deviation associated with avMSLP for the month in question

Cane (2011) shows that the SLP for Darwin has a strong negative correlation with the Niño 3.2 ENSO Index, which is based on sea surface temperatures in the central Pacific. The paper simply says "SLP" but the figure in question, its Figure 1, shows the SLP in the range +4 to -2, which suggests it is the standardised anomaly as discussed here.

As we might expect, the normalised MSLP (NMSLP) for Darwin is close to a mirror-image of that for Tahiti (Figure 10.5). (The Figure, like the analysis that follows for other locations, uses MSLP data from the KNMI Climate Explorer available at

http://climexp.knmi.nl.) The correlation coefficient from comparing Troup SOI to Darwin's NMSLP is -0.933 and to Tahiti's NMSLP is 0.876, the slight asymmetry probably attributable to the different annual ranges of MSLP and different local weather at the two locations.

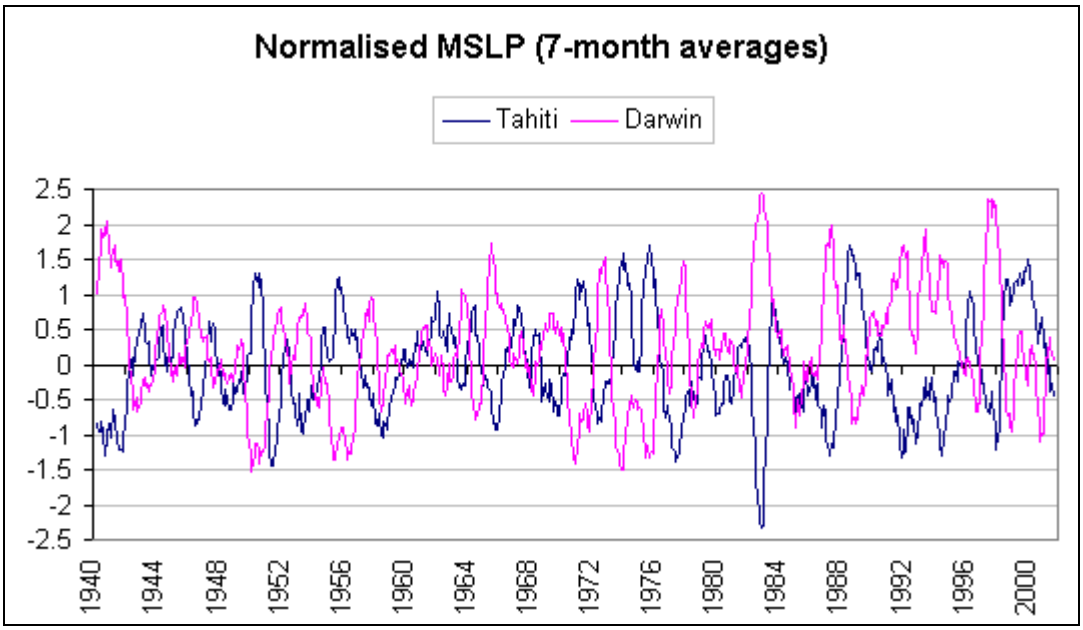


Figure 10.6 Normalised MSLP anomalies at Tahiti and Darwin, based on average MSLPs and standard deviations calculated from 1951-1980 data.

Comparing the normalised MSLP at various locations to the Troup SOI has the problem that the exercise is undertaken on the assumption that the Troup SOI is distorted by local weather conditions. This is not a major problem because the aim is to identify locations where the MSLP is similar but not the same as the Troup SOI, in fact any locations with perfect correlation to it would seem to be suffering from identical weather when the aim is to reduce that weather influence. The complete removal of weather influences is impossible because the MSLP at any location will be impacted by local weather.

10.3.4 Analysis of MSLP data around the Pacific

MSLP data from more than 80 locations in and around the Pacific Ocean was analysed to determine the correlations to the Troup SOI, the results indicating both the extent of direct ENSO influence on MSLP and the locations with the strongest positive or negative correlations.

Apart from drawing on the KNMI Climate Explorer as mentioned above, the SLP data for three locations at sea, all grid cells of 2° latitude x 2° longitude, was extracted from the International Comprehensive Ocean Atmospheres Data Set (ICOADS) database and converted to monthly mean values. Details of the three locations are given in Table 10-3.

No.	Latitude	Longitude	From	To	Location	Comments
1	6°S to 8°S	76°E to 78°E	1954	2011	Indian Ocean due South of India	Excludes 1.5% of the data when obs per month <10
2	6°N to 8°N	84°W to 86°W	1950	2014	Off the south coast of Costa Rica, near the Panama Canal	All data processed; all months <=15 obs
3	26°N to 28°N	116°W to 118°W	1940	2014	Between Baja Sur (Mexico) and Hawaii	Excludes data when obs per month <5. Poor correlation with Troup SOI.

Table 10-3 Details of locations at sea for which SLP data was obtained and processed

The MSLP data for all other locations was obtained via the KNMI Climate Explorer and were of variable duration, often ending in 1991 and often with various months where the data was missing. In some cases the amount or sequence of missing data was such that for practical reasons several years of data were trimmed from the record. An arbitrary constraint placed on this data was for a minimum period of 25 years so that the long-term averages and standard deviations would be meaningful

even when the periods varied slightly depending on the available data. The first of the following criteria that the data satisfied was used to define the period:

(a) when all calendar months have at least 20 values for 30-year period 1951-1980
(b) ditto for 16 values for 25-year period 1956-1980
(c) ditto for 20 values for 30 years from first year of data
(d) ditto for 20 values for 30 years back from last year of data
(e) ditto for 16 values for 25 years from first year of data
(f) ditto for 16 values for 25 years back from end of data
(g) ditto for 16 values over the entire period of the data

In practice most locations satisfied the first criteria and therefore shared the same period, but there were exceptions.

Extreme outliers, probably caused by data noise due to local weather events such as storms, were trimmed from the record by excluding any standardised values equalling or exceeding ±4, which according to the probabilities associated with Gaussian distribution means a probability of exclusion of ~6 in 100,000. This very rarely excluded more than 5 months of the minimum of 25 years of data from any of the more than 80 locations that were examined and no exclusions were required for most locations.

Details for many observation stations, including the correlations of their NMSLP to the Troup SOI (both as 7-month averages) are shown in Table 10-4. NMSLPs were calculated for several other stations but their correlations to the Troup SOI fell outside the range ±0.5 and showed no sign of significant improvement in even a recent subset of the data (see Appendix 4 to this thesis).

The full list of sites is split between Table 10-4, which shows locations with strong correlation and a list of locations with weak correlation is given in Appendix 2.

Name	Country	Lat	Long	Beg yr	End yr	Correl Coeff	Comments
Darwin	Australia	12.42S	130.88E	1940	2001	-0.933	
Broome	Australia	17.95S	122.22E	1951	2001	-0.907	
Singapore	Singapore	1.37N	103.92E	1948	1998	-0.812	
Goa	India	15.48N	73.82E	1965	2001	-0.776	correlation 1986-2001
Townsville	Australia	19.25S	146.75E	1951	2001	-0.763	
Geraldton	Australia	28.78S	114.70E	1961	2001	-0.754	
Rabaul	Papua-New Guinea	4.22S	152.18E	1951	1990	-0.735	
Penang	Malaysia	5.30N	100.27E	1951	2001	-0.711	
Palau	Caroline Islands	7.33N	134.48E	1961	2001	-0.710	
Adelaide	Australia	34.93S	138.52E	1961	1997	-0.705	
Perth	Australia	31.90S	116.00E	1940	2001	-0.699	
Noumea	New Caledonia	22.27S	166.45E	1941	2001	-0.693	correlation 1993-2001
Mannar	Sri Lanka	8.98N	79.92E	1951	1989	-0.691	
Mildura	Australia	34.22S	142.08E	1961	2001	-0.676	
Sea grid cell 1	Indian Ocean	6-8S	76-78E	1954	2011	-0.637	
Brisbane	Australia	27.4S	153.1E	1940	1998	-0.636	
Ceduna	Australia	32.12S	133.70E	1951	2001	-0.599	
Chumphon	Thailand	10.48N	99.18E	1951	2001	-0.570	
San Francisco	USA	33.93N	118.40E	1971	2001	0.500	
Apia	American Samoa	13.8S	171.8W	1940	1992	0.504	correlation 1965-1992
Christobal	Panama	9.40N	79.90W	1940	1960	0.537	
San Diego	USA	32.73N	117.17W	1940	2001	0.585	correlation 1973-2001
Trujillo	Peru	8.10M	79.03E	1949	1980	0.589	downward trend (error?)
Penrhyn	Cook Islands	9.03S	158.05E	1940	1990	0.637	correlation 1955-1990
Sea grid cell 2	near Panama Canal	6-8N	84-86W	1950	2014	0.682	
La Serena	Chile	29.9S	71.2W	1960	2001	0.761	correlation 1973-2001
Tahiti	French Polynesia	17.55S	149.64W	1940	2001	0.876	
Atuona	French Polynesia	9.8S	139.03W	1959	2000	0.903	correlation 1967-2000

Table 10-4 Locations with normalised MSLP correlating to Troup SOI (both 7-month averages) less than or equal to -0.5 or greater than or equal to +0.5.

Figure 10.7 shows some examples of reasonable positive correlation - Penrhyn (Cook Islands), La Serena (Chile) and Trujillo (Peru), with coefficients of 7-month centred averages (applied to reduce data noise) of 0.637 since 1955, 0.761 since 1973 and 0.589 respectively, but a persistent downward trend in Trujillo MSLP suggesting instrument problems. Figure 10.8 illustrates some corresponding examples of negative correlations with the Troup SOI - Perth (Australia), Mannar (Sri Lanka) and Singapore, with correlation coefficients of -0.699, -0.691 and -0.812 respectively.

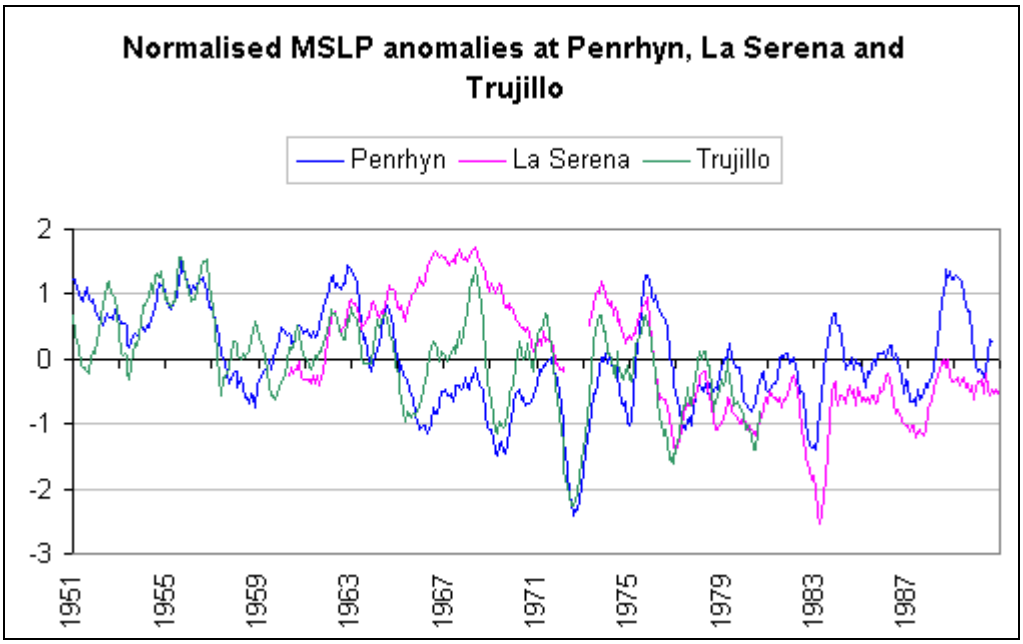


Figure 10.7 Example of normalised MSLP anomalies showing positive correlations between local NMSLP and Troup SOI for at least 20 years - Penrhyn (Cook Is.), La Serena (Chile) and Trujillo (Peru). (7-month centred averages)

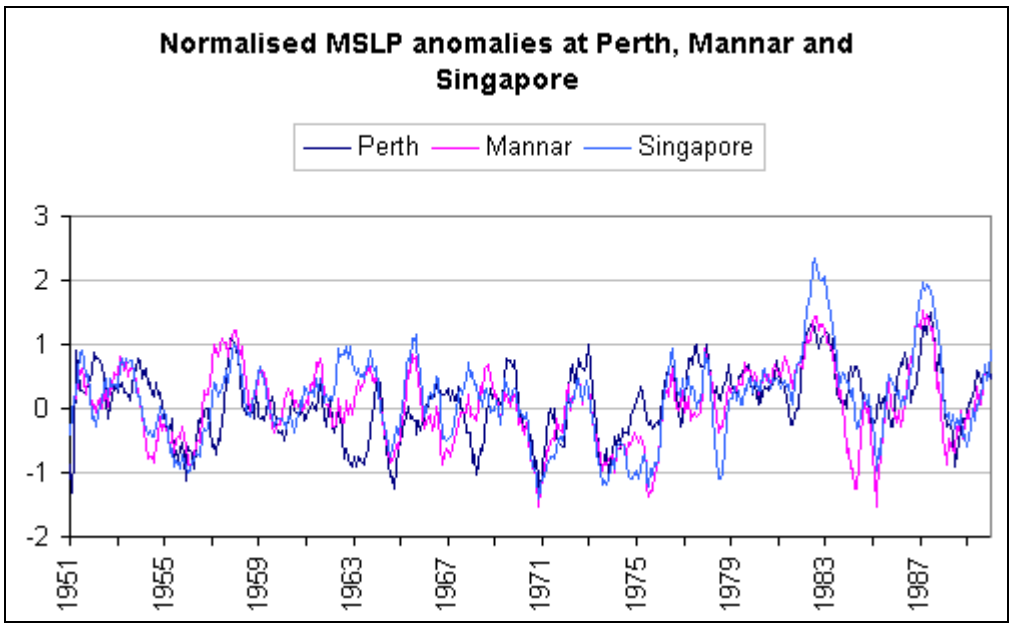


Figure 10.8 Examples of normalised MSLP anomalies showing negative correlations between local NMSLP and Troup SOI for at least 20 years - Perth (Australia), Mannar (Sri Lanka) and Singapore. (7-month centred averages)

Figure 10.9 maps the locations noted in Table 10-3 along with the approximate regions of positive correlation of ≥0.5 and negative correlation of ≤-0.5. Applying the

HadCRUT4 method of determining coverage from grid cells, these regions together amount to ~20% of the Earth's surface. Berlage and de Boer (1959) shows a similar figure based on a correlation with MSLP at Easter Island rather than the Troup SOI as shown here. The Figure in Berlage and de Boer (1959) is derived from data covering just nine years (1949-1957) compared to the 25 (or more) years analysed here and the limited availability of MSLP data, especially at sea and away from usual shipping routes, makes the accuracy of the figure uncertain.

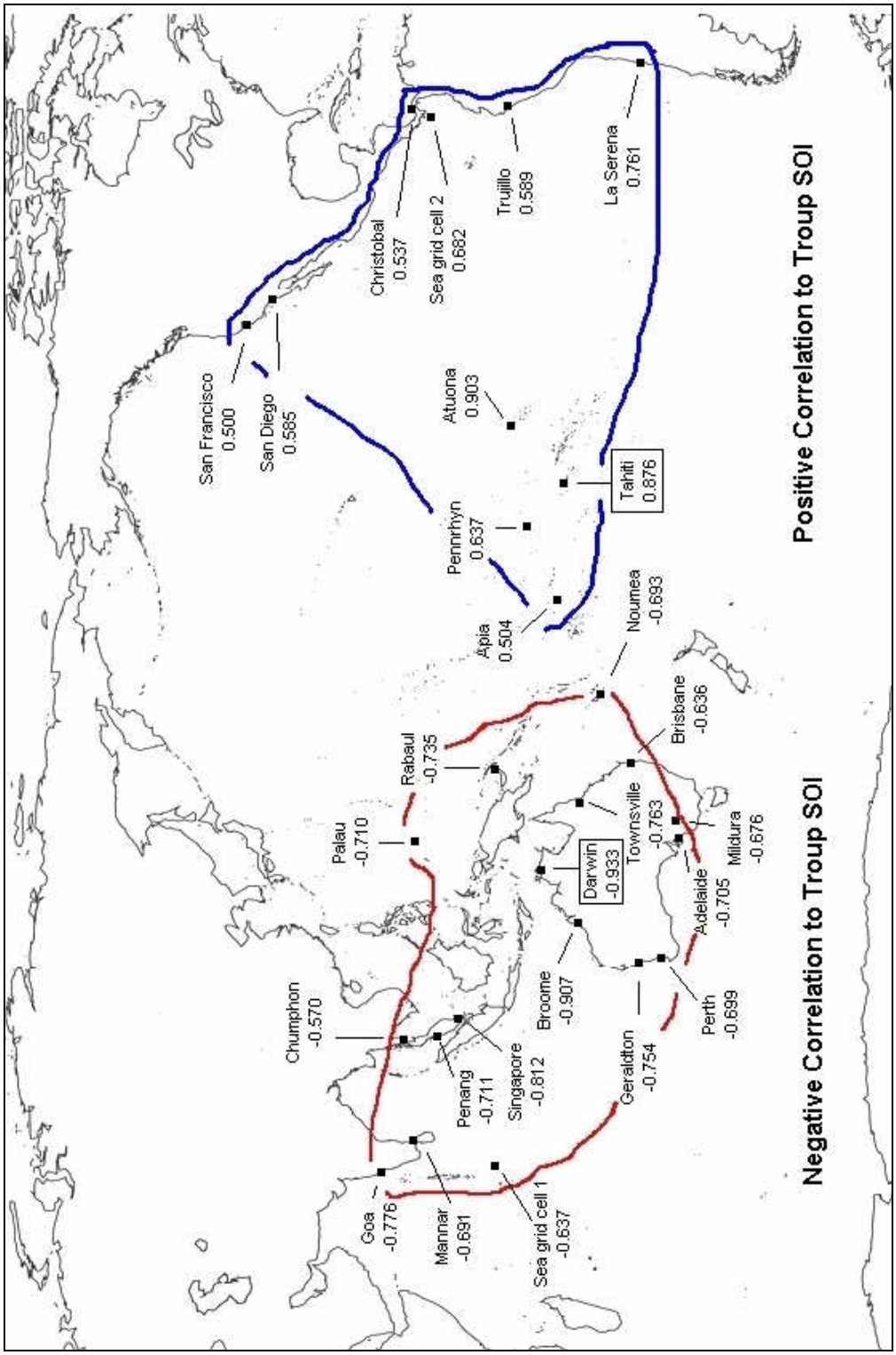


Figure 10.9 The two regions corresponding to the data shown in Table 10-3, i.e. locations where the correlation coefficient for the comparison of the Troup SOI and the local NMSLP is ≤ -0.5 or ≥ 0.5. (The boundaries of the two regions can only be estimated from the given locations.)

10.4 Proposal for an improved MSLP-based index

Figure 10.9 shows that the area directly influenced by ENSO conditions is substantial, stretching from Panama across the dateline to India and from southern Chile to California. The indication from data beyond the regions illustrated here are that the Andes, Rocky Mountains and the Central American mountain range spine largely confine the relationship in the east of the Pacific. Using the same approach as HadCRUT4 to determine global surface coverage, it seems that 21% of the Earth's surface is directly impacted, although that figure is uncertain because the exact limits of the region cannot be identified.

Figure 10.9 also shows that the "oscillation" of MSLPs at Darwin and Tahiti is likely to be mirrored elsewhere, with one of the pair of locations from the region with negative correlation to the Troup SOI and the other from the region of positive correlation. Ropelewski and Jones (1987) discusses using indices based on MSLP at Darwin (Australia) and either Apia (Samoa), Suva (Fiji) or Santiago (Chile) but for reasons that are not made clear finds no reasons to change the Darwin-Tahiti pairing.

Simply changing the pair of locations for which the MSLP can be used to derive an index of ENSO conditions seems unlikely to greatly alter the occurrence of "noise" in the data.

It is suggested here that a better approach would be to create an "ensemble" SOI based on several location pairs, not only Darwin and Tahiti. Three such paired locations are shown in Table 10-5, along with the correlation coefficients from the comparison of the base Troup SOI and indices calculated as for the Troup SOI but using MSLP at the given location pairs. The correlation of both monthly values and the 7-month centred averages (as used earlier in this document) are shown.

The three additional pairings greatly extend the range of the SOI from just the Darwin and Tahiti pairing where the stations are at 12.4°S 130.9°E and 17.55°S 149.6°W. The latitude of the Geraldton-La Serena pair is at approximately 29°S, Atuona at 9°S,

Singapore at 0°S, the centre of the Panama grid cell at 7°N and Mannar at 9°N. Longitudinally the range stretches from Mannar at almost 80°E to La Serena at 71°W. The three pairings, when added to the Darwin-Tahiti pairing, therefore provide better coverage of the regions where the MSLP correlates with ENSO conditions.

	Data Period		Correlation Coefficient	
Location pair	**from**	**to**	**Monthly data**	**7-mon avg**
Singapore - Panama	1950	1998	0.723	0.875
Mannar - Atuona	1970	1989	0.783	0.936
Geraldton - La Serena	1973	2001	0.720	0.884

Table 10-5 - Correlation coefficients for the comparison of the Troup SOI (based on Darwin-Tahiti location pair) and indices calculated using the same method but for other location pairs

The Panama location mentioned here is for the grid cell at sea off the west coast of Panama and south of Costa Rica (and labelled "Sea grid cell 2" in Figure 10.9). While being sourced from ships' observations has an advantage of using multiple instruments to measure air pressure, which means that errors in one or few instruments are of little consequence, for practical reasons the MSLP data from Christobal, at the northern end of the Panama Canal, might be preferable.

Figure 10.10 shows the Troup SOI and the ensemble index based on the average of several indices. The KNMI Climate Explorer contains data for all four pairs of sites only for the period 1973 to 1989 but data for a minimum of three pairs is available from 1961 to 1998 and that data is used here. (The publishing of MSLP data is less common than the publishing of temperature and rainfall data although it is certainly recorded for weather predictions.) At times there is considerable difference between the ensemble index and Troup SOI, such as November 1973, when the Troup SOI is 31.6 but the ensemble index is just 18.6, and May 1983 when the Troup SOI is 6.0 but the ensemble index is -12.2. Figure 10.11 shows the monthly differences between the two indices for the entire period of Figure 10.10.

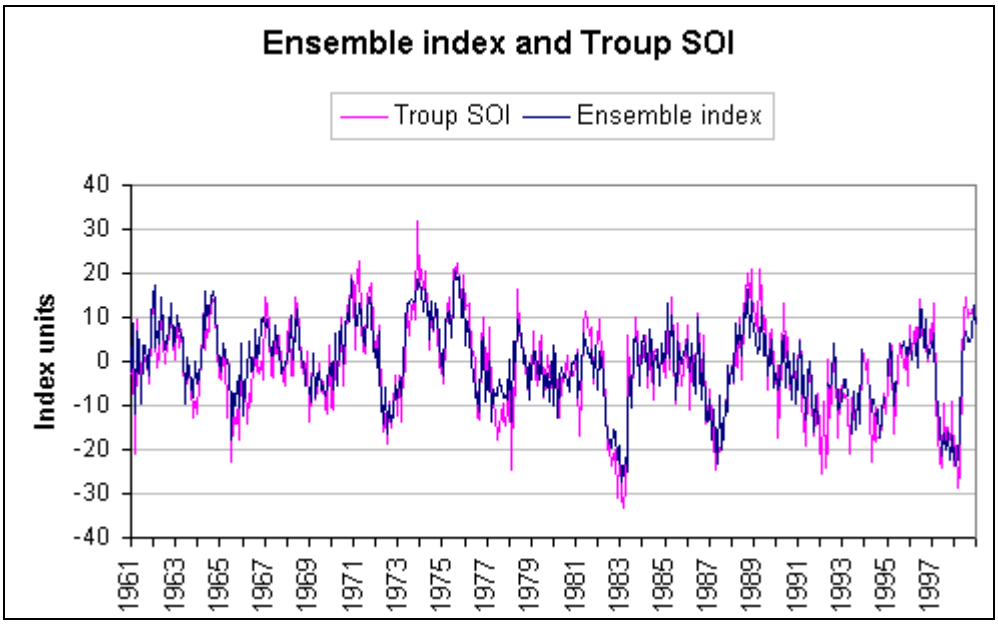


Figure 10.10 Monthly values of Troup SOI and the ensemble SOI from 1961 to 1998

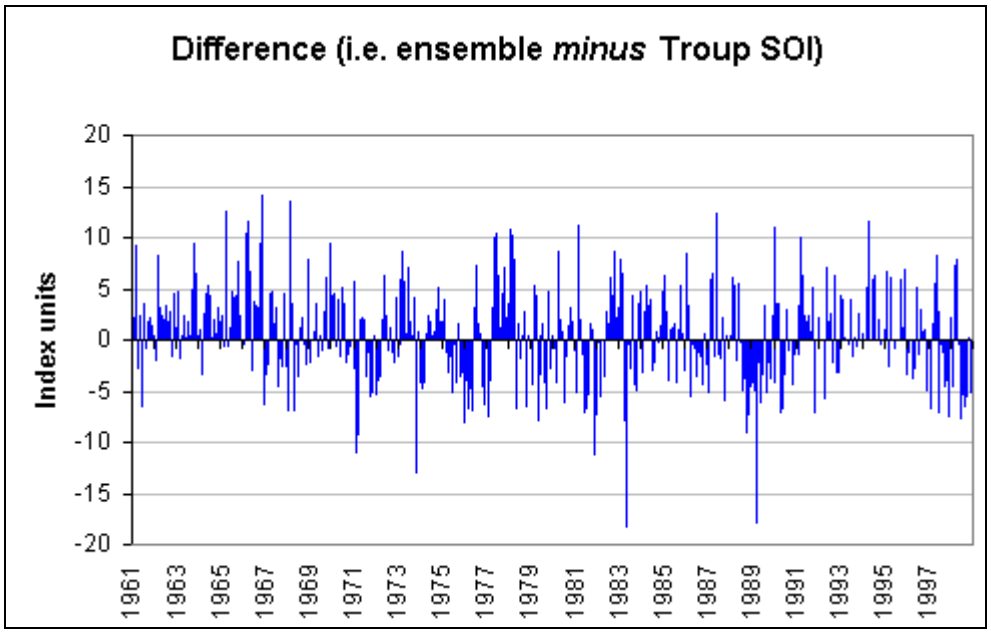


Figure 10.11 Difference between the ensemble index and the Troup SOI for the period shown in the previous Figure

10.5 Comparison of the Troup SOI and the Extended SOI

A quantitative comparison of the original Troup SOI and the Extended Troup SOI described here requires other ENSO indices as a reference against which the two can be compared.

One frequently used index is the Nino 3.4 index derived from the average SST anomalies in the region covered by latitudes 5°N to 5°S over longitudes 170°W to 120°W, using a 5-month running mean of the average anomalies.

For the period 1961 to 1998, the period for which a pair SOI can be calculated for minimum of three of the four pairs of locations in the Extended Troup SOI, the correlation in monthly values between the Nino3.4 index and the Troup SOI is -0.71 whereas it is -0.74 for the Extended Troup SOI.

Another ENSO index is the Oceanic Nino Index (ONI) uses the same data as the Nino 3.4 index but uses a 7-month (cf. 5-month) running mean. The correlations over the same period above are -0.74 and -0.76 respectively, again showing a small improvement in correlation for the Extended Troup SOI.

Two issues arise with these comparisons. Firstly SST changes more slowly than does air pressure, the latter partly a product of the transfer of heat from the ocean to the atmosphere but partly also a product of the Earth's rotation bringing atmospheric pressure patterns to the region. Correlating a variable that varies only slightly from one datum to the next in the sequence with one that varies much more in its sequence is always problematic and there is an element of chance in corresponding shifts.

Second is the problem that both the Nino 3.4 and ONI indices are based on average sea surface temperature from the Nino 3.4 region of the Pacific Ocean (as defined above) and that there are issues with SST data from this region. For the period prior to 1950 the two indices use interpolated sea surface temperature data, presumably estimated from the available measurements and according to derived relationships in

data since that time. SST observations are used to derive the values since 1950 but to do so ignores the sometimes-small number of observations that were made. The region covers 20 of the 5° latitude by 5° longitude used in the HadSST3 gridded dataset, 10 in the region from latitude 5°N to the equator and 10 parallel below them. Figure 10.12 shows, for the period 1950 to 2015, the annual average number of grid cells reporting SST data (maximum is 20 as per above), split into the number of grid cells with fewer than five observations in the month and those with five or more.

Further, in all but three years of the period from 1950 to 1972 the annual averages indicate that less than five observations were made in 25% or more of the reporting grid cells (Figure 10.13). These high percentages, coupled in the 1950s with a shortfall in the number of grid cells for which data was reported, cast doubt on the accuracy of both the Nino 3.4 and ONI indices from 1950 to 1972.

The extended Troup SOI therefore is about as accurate as the Troup SOI but there are indices in which we have high confidence against which it can be compared.

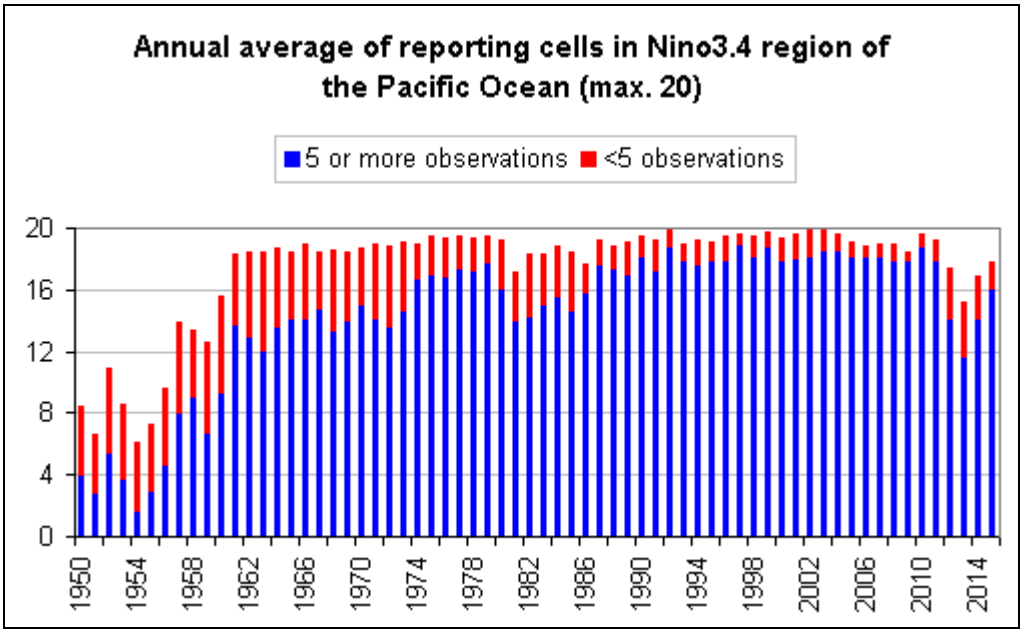


Figure 10.12 Annual average number of reporting grid cells in the Nino 3.4 region broken into cells with less than five observations and those with five or more

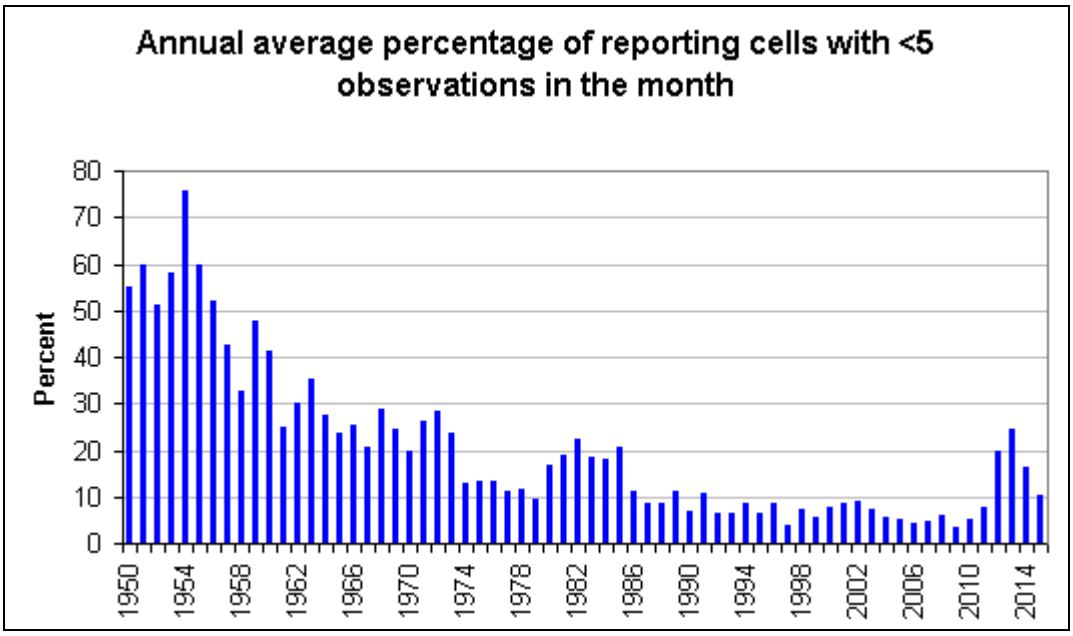


Figure 10.13 Annual average numbers of reporting cells with less than five observations per month expressed as a percentage of all reporting cells in the same month

10.6 Summary

This chapter has discussed the indices used to describe the state of the El Nino-Southern Oscillation (ENSO) and some of the shortcomings of each, before addressing the Troup SOI in more detail. It has shown that the reliance on mean sea level pressure data at just two sites is unwise because that data might be impacted by local weather conditions and has shown other reasons for scepticism about the Troup SOI prior to 1935.

It has been shown that the fundamental "oscillation" in mean sea level pressure at Darwin and Tahiti indicates that data from just one of those stations would be indicative of ENSO conditions. With this in mind MSLP data from more than 80 locations was examined and it was found that the MSLP at many locations in the eastern Pacific Ocean show similar patterns in normalised MSLP to Tahiti and that many in the western Pacific Ocean and into the Indian Ocean show similar patterns to Darwin.

It was proposed that Troup-like SOI values could be derived from MSLP data from pairs of locations with strong correlations, one of the pair in the eastern region and one in the western region. It is suggested that the average of the three calculated indices plus the Troup SOI, based on Darwin and Tahiti MSL data, will provide a useful ensemble SOI.

Deriving the ensemble SOI from individual indices from station pairs rather than average MSLP of the four locations in the eastern region and the four in the western region takes into account that the natural range of SLP might differ between locations, as was shown for Darwin and Tahiti in Figure 10.2. Moreover we might find that the individual indices are also of value, especially with the Singapore-Panama pairing being equatorial and 9°N and the Geraldton-La Serena pairing being at almost 30°S.

The ensemble Troup SOI has three significant advantages. Firstly it is not a radical shift to a new set of ENSO-related factors with new threshold values but merely a refinement of an existing index. Secondly, it provides better coverage of the entire region where mean sea level pressure is directly related to ENSO conditions.

Thirdly, and somewhat related to the second, is that data noise such as that caused by small-scale weather conditions at one or two locations is less likely to have a major impact on the index value than in the Troup-SOI, which is calculated from MSLP data from just two locations. The risk of false positives and false negatives regards ENSO thresholds is therefore reduced and this provides greater confidence about sequences of index values, not only that they persisted across periods when the Troup SOI indicates a disruption but also confidence is increased regarding the starting and end months of such sequences.

Chapter 11: On the likelihood of historical coral bleaching on the Great Barrier Reef

11.1 Introduction

The Great Barrier Reef (GBR) stretches for approximately 2300 km along the north-east coastline of Australia, from latitude 11°S to 24°S, with most of the reef about 100km offshore. The coral that makes up the GBR is prone to bleaching when the water is unusually warm, the bleaching being due to the coral expelling the symbiotic zooxanthallae algae that live in coral tissues and supply it with nutrients. The expelling appears to be a normal mechanism for coping with warmer water but unless those algae, or alternative algae, are acquired again the coral is likely to die.

Coral bleaching can be caused by unusually cold water (Hoegh-Guldberg & Fine, 2004; Rodriguez-Troncoso et al, 2010; Lirman et al, 2011) however the emphasis on GBR bleaching has focused on warm water.

Anthony & Kerswell (2007) report that tide height can be a significant influence on bleaching. Summer tidal patterns along the GBR typically see the lowest tide during the night and the second lowest during mid afternoon, which is often the warmest time of the day. Low tides in mid-afternoon might contribute to bleaching but the underlying factor is still the elevated water temperature that might be a consequence of low tides and little cloud cover.

Severe bleaching occurred in parts of the GBR four times in the last 20 years (1998, 2002, 2016 and 2017) and has been declared or implied by some researchers to be very recent phenomena caused primarily by manmade climate change[32] (Lough, 2000; Lough, 2008; De'ath et al, 2012; Hughes et al 2017). One such example is that of Professor Terry Hughes, a coral ecologist at James Cook University in Townville,

[32] The frequent use of the term "climate change" to mean "manmade climate change" causes substantial confusion. Natural climate change, in the form of the 1977 Pacific Climate shift that resulted in an increase in El Niño events, certainly changed the climate in and around the Pacific (and further).

Australia. This professor, who has been responsible for much of the publicity about the bleaching event of early 2016, stated on Australian national radio:
"*a critical issue here is that these bleaching events are novel. When I was a PhD student 30 years ago regional scale bleaching events were completely unheard of, they're a human invention due to global warm*ing."[33].

Bleaching was in fact first recorded early last century by Sir Maurice Yonge in the first major scientific study of the Great Barrier Reef (Yonge and Nicholls 1931). Oliver et al (2009) reports 26 records of coral bleaching before 1982. Many early reports are anecdotal and refer to only small areas of the reef but that is not surprising given that monitoring of the reef prior to about 1985 was largely random and by opportunity, rather than a methodical and regular event.

Understanding whether severe bleaching, or at least conditions associated with severe bleaching, occurred prior to 1998 would put recent bleaching into context and perhaps prompt a re-interpretation of these episodes, so this is the aim of this study.

A complicating factor in this work is the acclimatisation of corals to earlier periods of bleaching (Maynard et al, 2008; Middlebrook et al, 2008; Schoepf et al, 2015). Acclimatisation cannot be taken into account in this study because it would depend on whether extensive bleaching occurred prior to 1998, which is what this study attempts to ascertain. It does however follow that if earlier extensive bleaching did occur and acclimatisation is a significant issue then any earlier extensive bleaching likely occurred at a lower temperature than the bleaching episodes since 1998.

[33] ABC Radio National 2 March 2016, audio available at http://www.abc.net.au/radionational/programs/breakfast/widespread-coral-bleaching-detected-on-the/7212760
The quoted passage above starts 3min 13secs into the audio and follows about 80 seconds (starting at 1min 33secs) of Hughes saying that the El Nino event at the time was causing temperatures "hotter than average, not much wind and lots of sunshine. These conditions have pushed temperatures up." He went on to describe how the weather over the next few weeks would determine if the mild bleaching at the time would worsen.

11.2 Methods, data and limitations

With the absence of observational evidence of bleaching prior to 1998 this study will rely on four different analyses of past weather conditions to establish the likelihood of bleaching in those years. These analyses are (i) a probabilistic approach based on Willis Island mean summer temperature, (ii) the identification of summers in which El Niño episodes occurred in some or all of those months, (iii) the identification of large numbers of hot summer days and successive hot summer days recorded at Willis Island and (iv) a similar identification of hot days according to sea surface temperature observations recorded in the International Comprehensive Ocean Atmosphere Data Set (ICOADS) database.

Two of these analyses rely on familiar data for this thesis, the El Nino episodes that can be identified by the Troup SOI data published by the Bureau of Meteorology and the ICOADS database of observations made at sea. By its nature the ICOADS data is dependent on the movement of ships, which for the GBR means at irregular time intervals and located in designated shipping routes on the east and west sides of the reef, plus various passages between the two, these routes by necessity being of deeper water when faster warming occurs in shallow water, suggesting that corals in shallow water are more susceptible to bleaching.

Data from the Bureau of Meteorology's Willis Island observation station (Lat: 16.2878°S, Long: 149.9652°E) (Figure 11.1) is used in the other two analyses. Like the GBR this location is surrounded by ocean, with particular exposure to the east, rather than on mainland Australia where it might be subject to land influences or urbanisation. Although there is reason to think that the Willis Island observation station has been relocated short distances on the small island (approx 500m x 100m) the microenvironment around the station has essentially been unchanged since observations began in 1921. According to data from the Bureau of Meteorology an automatic weather station was installed in June 1991, and was replaced with a new type in June 2000.

Daily data is available from the Willis Island station since 1921 except for certain periods (May 1933-Dec 1934, May 1936-Dec 1936, May 1937-Dec 1938, Nov 2004-Sep 2006) and occasional short sequences of days or even months. While this is a significant advantage over the intermittent data from ICOADS the disadvantage of the Willis Island data is that it is for air temperature rather than sea temperature, which means that it is only generally indicative of the environment around the coral.

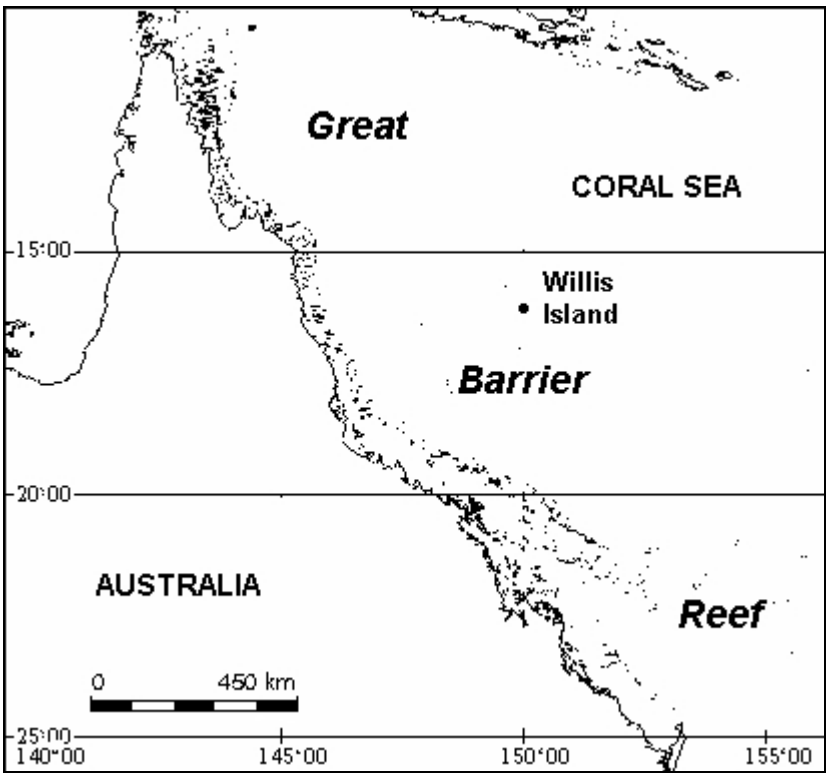


Figure 11.1 Map of the Great Barrier Reef, Willis Island and the three latitude bands discussed in this chapter.

This study focuses on temperature data across the period from December to March inclusive and for convenience refers to this period as "summer", although strictly speaking it extends beyond the normal summer period. Figure 11.2 shows the summer[34] mean maximum temperatures from Willis Island from 1922 to 2017. The temperature trend across the entire period is 0.059°C/decade but that includes a shift in temperature from 2002 onwards because the trend from 1940 (after two years of missing data in the 1930s) to 2001 is just 0.006°C/decade with an average of 27.96°C, whereas the average for 2002 to 2017 is 0.5°C higher at 28.46°C.

[34] As with other references to "summer" in this chapter, the year displayed is that for the January of the summer.

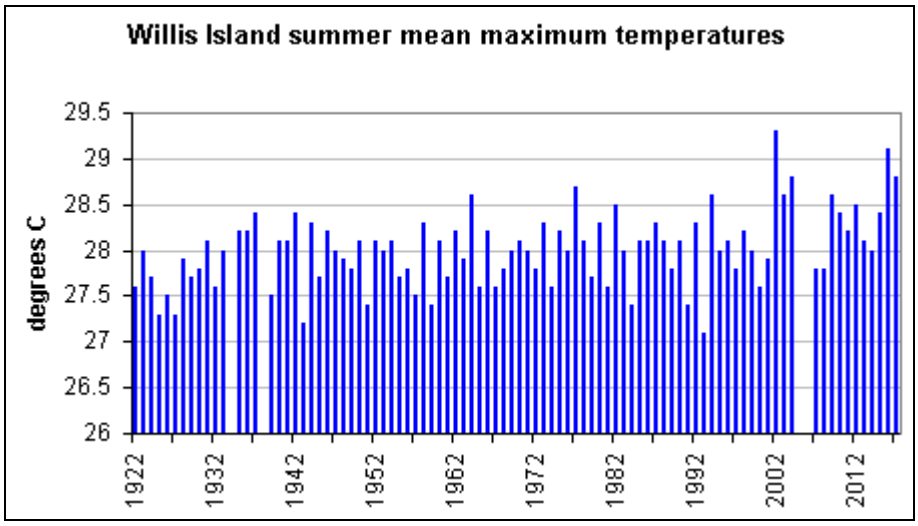


Figure 11.2 Willis Island mean maximum summer (Dec to Mar) temperatures 1922-2017

The correlations of average NOAA OI SSTs for three latitudinal bands of the GBR (see Figure 11.1) and Willis Island summer mean temperatures in individual months and the average for each summer are shown in Table 11-1. Probably as a consequence of local conditions (eg. ocean currents) the correlations are stronger for the central and southern thirds of the reef.

	Northern GBR 10°S-15°S	Central GBR 15°S-20°S	Southern GBR 20°S-24°S
December	0.638	0.773	0.740
January	0.426	0.644	0.637
February	0.430	0.733	0.759
March	0.560	0.677	0.508
Summer mean	0.561	0.794	0.747

Table 11-1 Correlation coefficients between Willis Island monthly mean temperatures and average NOAA OI SSTs for the GBR in three latitudinal bands.

Sea surface temperature (SST) data reported by ships is problematic for this study because with rare exception the ships were travelling designated channels of deeper water where corals would not be damaged, whereas coral beaching is common in shallow water where over time and in the absence currents, such as in pools, the heat at the sea surface is progressively forced lower and eventually contacts the coral. SST data obtained by satellites offers better coverage of both deep and shallow water but that data is available only since 1982.

11.3 Analyses

11.3.1 Probabilistic approach

Severe bleaching has been reported in 4 of the last 20 years (1998-2017), meaning that the present probability of severe bleaching in any given year is 4/20 or 20%. If we assume normal (i.e. Gaussian) distribution of temperatures at Willis Island the probability corresponds to the temperature being at least 0.842 standard deviations (i.e. 0.842σ) greater than the mean (Figure 11.3). After taking into account the mean December-March temperature over the last 20 years and the standard deviation associated with that mean, 28.71°C is the Willis Island average December-March temperature at which bleaching occurs, i.e. the bleaching threshold.

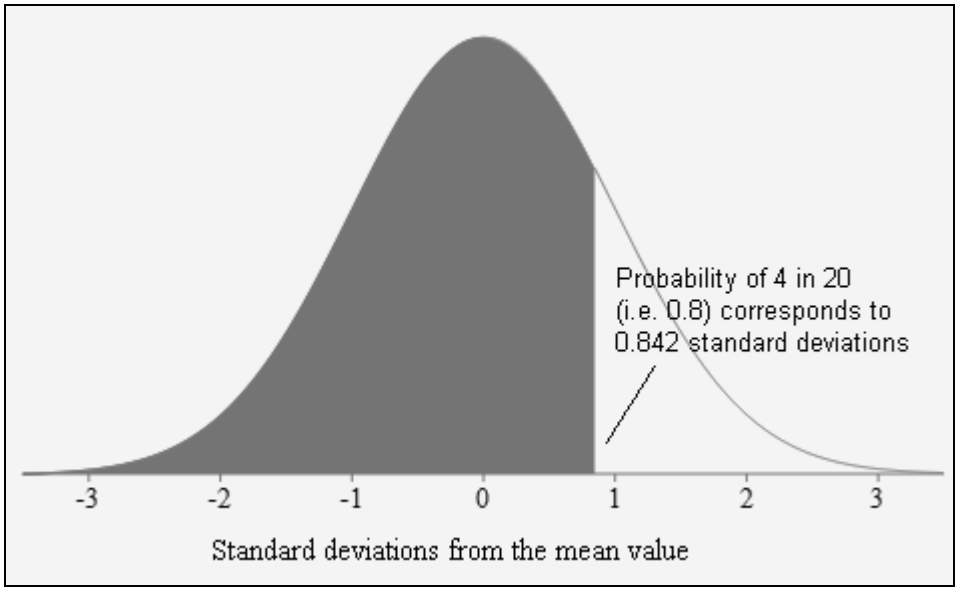


Figure 11.3 Deriving the number of standard deviations from the probability

The data for Willis Island from 1922 to 1997 can be split into an initial period of 16 years followed by three periods each of twenty years. The probability of bleaching threshold average summer temperature occurring in each year of these periods is determined from the mean temperature for that period and the number of standard deviations from the mean that corresponds to the bleaching threshold of 28.714°C. For example, the mean temperature for 1978-97 was 27.961°C with standard deviation 0.377°C, meaning that the bleaching threshold was (28.714 - 27.961)/0.377 (i.e. 1.995) standard deviations from the mean, which has a corresponding probability of 0.023.

From the probability of the threshold being exceeded in *any* given summer (p) the probability of it not being exceeded in *any* given summer (i.e. 1 - p) can be calculated and from that the probability of it not being exceeded in *all* summers of the period (i.e. $(1 - p)^n$ where n is the number of summers). Subtracting that probability from 1.0 gives the probability that in at least one summer of the period the threshold was exceeded. Table 11-2 summarises the calculations for each period and the 1922-1997 period as a whole.

Factor					Full per.
Period	1922-37	1938-57	1958-77	1978-97	1922-97
Years of data	15	19	20	20	74
Mean Summer Temperature (℃)	27.802	27.886	28.006	27.961	27.921
Standard deviation of MST (℃)	0.334	0.311	0.337	0.377	0.347
Bleaching point above mean Temp (℃)	0.912	0.829	0.709	0.753	0.793
No of standard deviations of bleaching threshold above the mean temperature	2.735	2.662	2.103	1.995	2.282
Probability of no bleaching (per summer)	0.997	0.996	0.982	0.977	0.989
Probability of bleaching (per summer)	0.003	0.004	0.018	0.023	0.011
Probability of no bleaching in the period	0.953	0.928	0.698	0.628	0.435
Probability of bleaching in at least one year of the period	0.047	0.072	0.302	0.372	0.565

Table 11-2 Summary of analysis of probability of Willis Island mean summer temperature reaching the notional bleaching temperature according to the period 1998-2017

The rightmost column of Table 11-2 shows data for the single composite period from1922 to 1997, concluding a probability of 0.565 of bleaching occurring in one or more years of the entire period. Calculating this probability from the probability of no bleaching in each of the periods, i.e. 1.0 - (0.953 x 0.928 x 0.698 x 0.628), produces a similar but different probability of 0.6121 because the periods contain different sample sizes, but the division into periods is useful for indicating in which periods bleaching was more likely.

Therefore, assuming Gaussian distribution, the probability of the Willis Island mean summer temperature bleaching threshold being reached in at least one year of the period 1922 to 1997, excluding the three years for which data is unavailable, is ~60% with the strongest possibility that it occurred during 1978-1997.

11.3.2 El Niño episodes

The second approach considers the occurrence of El Niño events. Hendy et al (2003), Arthur et al (2005), Guzman & Cortes (2007), del Mónaco et al (2012) and many

others have established the link between such events and coral bleaching. Strong El Nino events persisted through the 1997-1998 summer, the 2002-2003 summer and the 2015-2016 summer, all being summers with severe bleaching on the GBR.

The Troup Southern Oscillation Index (SOI) is based on the standardised difference in mean monthly sea level pressure at Tahiti and Darwin, calculated on a calendar month basis. The Australian Bureau of Meteorology advises that Troup SOI ranges between about -35 and +35 with values of less than -7 for three consecutive months indicating El Niño conditions. It advises that local weather events (e.g. summer storms) can make it a "noisy" index.

Table 11-3 shows all instances since 1935 of summer months that were the last of three consecutive months when the Troup SOI was below -7. Some summers feature multiple times (1940-1, 1982-3, 1991-2 and 1997-8) because the El Nino event persisted for longer than three months. By this approach the most likely times for severe widespread bleaching appear to be in 1941, at some point during the 1983 summer, when the January-March Troup SOI averaged a very low -30.6, and again in the 1992 summer.

Year	Month	SOI	**Year**	Month	SOI
1939	12	-10.43	1992	1	-16.47
1941	2	-18.17	1992	2	-17.13
1941	3	-11.90	1992	3	-19.63
1941	12	-12.70	1993	3	-8.20
1942	1	-10.30	1994	12	-11.00
1963	12	-11.27	1997	12	-14.03
1977	12	-12.70	1998	1	-15.93
1982	12	-24.20	1998	2	-17.27
1983	1	-27.67	1998	3	-23.73
1983	2	-28.40	2010	2	-10.53
1983	3	-30.63	2010	3	-11.73
1991	12	-12.30	2016	2	-16.17

Table 11-3 Last month of three consecutive months, ending in December to March inclusive, when the Troup SOI was below -7 in each month, which according to the Bureau of Meteorology indicate El Niño conditions.

As table 11-3 indicates, strong El Nino events during summer have become more common since 1977. The Australian Institute of Marine Science (AIMS) reports severe bleaching at Myrmidon Reef (18.274S 147.381E) in 1983, the same year in which widespread coral bleaching was reported in the eastern Pacific (Glynn, 1984) and Indonesia (Brown & Suharsonso, 1990), with both reports attributing bleaching to the El Niño event at the time. Whether severe bleaching occurred in other parts of the reef in that year or across any parts of the reef during earlier years with low Troup SOI (eg. 1941) is unknown because extensive surveying of the GBR was not the practice until well into the 1990s.

11.3.3 Willis Island daily maximum temperatures

The two approaches above are both broad-brush generalisations about the occurrence of conditions under which bleaching is likely, using very coarse temporal granularity of one or more months and treating the reef as a single unit. We turn now to more finely detailed analyses, the first considering daily temperatures and the second dealing with temperature recordings that might be just hours apart.

The first approach is to consider daily maximum and mean temperatures at Willis Island temperatures, identifying summers with high numbers of hot days and the duration of any hot spells. Data for 97 summers (December-March inclusive) were analysed, with six years having incomplete summer data (unknown reasons 1934-38 and cyclone damage 2004-05).

Tables 11-4(a) and 11-4(b) show summary information for maximum and mean temperatures respectively, with the total number of days above the specified temperature to the left and the longest sequence of consecutive days with temperatures equal to or above the given value to the right, the latter not precluding the possibility of other sequences of shorter duration meeting the same criteria.

Weightings of 1, 2, 4, 8 and 16 were applied to the five columns of "Total days" above each temperature, i.e. ≥31°C to ≥35°C in 11-4(a) and ≥28°C to ≥32°C in

11-4(b), before the data was sorted to rank the years. This ranking system is an arbitrary means of elevating summers with higher numbers of warm days to the top of the list. Only the first 12 entries are shown in each instance, this being sufficient to illustrate likely periods of bleaching prior to 1998.

End	Days in	Total days					Maximum sequence of days				
Year	period	≥31°C	≥32°C	≥33°C	≥34°C	≥35°C	≥31°C	≥32°C	≥33°C	≥34°C	≥35°C
2002	120	91	64	27	5	2	23	16	13	2	1
2004	122	79	40	12	2	0	32	13	4	1	0
2016	122	96	41	9	1	0	38	10	4	1	0
1944	121	63	35	13	2	0	10	7	4	1	0
1964	122	74	48	7	0	0	35	15	2	0	0
1977	121	72	39	8	0	0	26	17	5	0	0
1942	121	88	34	2	0	0	36	10	1	0	0
1960	122	54	19	7	4	0	12	6	6	4	0
2017	121	84	27	3	0	0	27	8	3	0	0
1982	121	57	31	7	0	0	48	16	3	0	0
1940	122	55	29	5	1	0	18	7	3	1	0
1958	121	83	22	1	0	0	16	3	1	0	0

Table 11-4(a) Total number of days and maximum sequence of days when the daily maximum temperature was greater than or equal do the listed temperature

End	Days in	Total days					Maximum sequence of days				
Year	period	≥28°C	≥29°C	≥30°C	≥31°C	≥32°C	≥28°C	≥29°C	≥30°C	≥31°C	≥32°C
2002	120	103	75	33	6	0	43	16	12	2	0
2016	122	107	66	19	3	0	22	13	11	2	0
2004	122	96	56	16	3	0	37	10	6	3	0
1977	121	88	56	15	0	0	48	11	5	0	0
2017	121	104	53	9	0	0	31	7	2	0	0
1994	121	80	48	14	0	0	25	15	5	0	0
2009	120	78	48	10	0	0	20	11	4	0	0
1964	122	90	42	7	0	0	35	10	2	0	0
2012	122	88	47	5	0	0	16	15	4	0	0
1944	118	80	39	7	0	0	16	11	3	0	0
1982	121	82	34	9	0	0	48	12	4	0	0
2015	121	88	36	4	0	0	20	7	2	0	0

Table 11-4(b) Total number of days and maximum sequence of days when the daily mean temperature was greater than or equal do the listed temperature

In Table 11-4(a) year 2002 stands out with daily maximum temperatures ≥31°C on just over 75% of days from December 1st to March 31st, more than 66% of those having maximum temperatures ≥ 32°C. Year 2016 had even more daily maximums above 31°C than 2002 but almost half of those did not reach 32°C. The summer of 1944 is second only to 2002 for the number of days in which temperatures were ≥33°C.

The conclusion from Tables 11-4(a) and 11-4(b) is that while high daily maximum and daily mean temperatures have occurred in recent years they also occurred in earlier years, such as over the summers that ended in 1942, 1944 and 1964. The first of these, 1942, came at the end of almost 18 months of El Niño conditions whereas ENSO conditions in summer 1944 were more neutral. In relation to the summer of 1964 the last of three months with Troup SOI below -7.0 occurred in December 1963 but the collapse of an El Niño can bring warm water from the central Pacific Ocean to the reef thus reducing the input required to warm the water to bleaching temperature.

While 2002 and 2016 both rate highly in Tables 11-4(a) and (b), the year 1998, in which bleaching was severe, rates 67th of 97 years for Table 11-4(a) (maximum daily temperatures) and 33rd for Table 11-4(b) (mean daily temperatures). This suggests that either other factors contribute to coral bleaching, perhaps quite localised conditions like clear sky and the absence of cooling winds, or the less likely prospect that local conditions at Willis Island somehow suppressed temperatures for much the summer in that year despite the El Niño conditions.

11.3.4 ICOADS sea surface temperature observations

Data for individual observations of sea and meteorological conditions along the GBR since 1901 was accessed from the ICOADS database in order to try to identify periods of elevated sea surface temperatures. The data is somewhat problematic because the observations are irregular, varying in number each month, and that the measurements of SSTs were usually made when the ships were either in port or travelling defined shipping routes, which are deeper passages of water. Despite these issues the data can

be taken as broadly indicative of conditions in a given month, although confidence in the data diminishes when few observations were made that month.

In this study the data was split into almost equal thirds along the length of the reef, the northern region from 10S (northern limit of GBR marine park) to 15S (near Cape Flattery), the middle region from 15S to 20S (near the township of Bowen) and the southern region from 20S to the southern end of the marine park. Again only data for the months December to March were used.

Table 11-5 shows instances when, for certain months prior to 1998, the percentages of SST observations recording ≥29°C and ≥30°C were relatively high (>35% for the northern region and >30% for the middle region), indicating periods of warm SST. No such instances meeting those parameters were identified in the southern region during that time and only three instances after 1998.

Table 11-5 shows in 1970 and 1983 multiple months with high percentages of warm SSTs were reported in the northern third of the reef and for both years a single month of warm SSTs in the middle region of the reef. Bleaching in 1983, during a strong El Nino event, was mentioned earlier. The Troup SOI in the first two months of 1970 were -10.1 and -10.7 respectively, coming after two months of very neutral -0.1 and +3.7, suggesting a tenuous link if the Troup SOI is assumed to be accurate rather than as being data noise.

Year	Month	Count SSTObs	Count ≥29℃	Count ≥30℃	% SST ≥29℃	% SST ≥30℃
Northern region						
1963	1	35	28	19	80.00	54.29
1964	1	39	26	14	66.67	35.90
1966	2	29	21	14	72.41	48.28
1970	1	37	32	15	86.49	40.54
1970	2	49	35	28	71.43	57.14
1970	3	37	33	25	89.19	67.57
1971	1	36	25	23	69.44	63.89
1973	1	39	28	17	71.79	43.59
1977	1	56	30	23	53.57	41.07
1980	3	30	19	18	63.33	60.00
1981	12	38	22	17	57.89	44.74
1983	1	57	48	25	84.21	43.86
1983	3	32	29	20	90.63	62.5
1983	12	46	32	19	69.57	41.30
1984	2	44	29	21	65.91	47.73
1985	12	46	33	23	71.74	50.00
1987	1	47	42	28	89.36	59.57
Middle region (latitudes 15℃S to 20℃S)						
1961	2	30	16	9	53.33	30.00
1970	3	63	35	21	55.56	33.33
1971	1	48	29	22	60.42	45.83
1973	2	37	26	12	70.27	32.43
1987	1	60	47	23	78.33	38.33
1991	1	39	20	14	51.28	35.90
1995	2	27	17	11	62.96	40.74

Table 11-5 Instances of a high proportion of recorded SSTs in the given month being ≥29°C and ≥30°C. (Minimum observations per month = 25).

Consecutive days of high temperatures can pose a threat to corals. Table 11-6 shows some instances of extended periods of temperatures ≥29°C and ≥30°C based on at least one SST observation in the given region reporting that SST on each day.

	≥29℃			≥30℃	
from	**to**	**days**	**from**	**to**	**days**
Northern					
31/01/1970	5/02/1970	6	31/01/1970	5/02/1970	5
5/01/1972	13/01/1972	9	5/01/1972	8/01/1972	4
16/03/1978	26/03/1978	11	17/03/1978	21/03/1978	5
22/03/1980	27/03/1980	6	22/03/1980	27/03/1980	6
23/03/1983	31/03/1983	7	25/03/1983	31/03/1983	7
11/03/1990	21/03/1990	11	12/03/1990	17/03/1990	6
Middle					
25/12/1923	3/01/1924	10	25/12/1923	2/01/1923	9
7/03/1970	13/03/1970	7	7/03/1970	13/03/1970	7
8/02/1978	17/02/1978	10	25/03/1978	28/03/1978	4
4/01/1986	16/01/1986	13			<4
17/01/1989	28/01/1989	12			<4
Southern					
16/01/1971	21/01/1971	6			<4
2/02/1971	6/02/1971	5			<4
14/02/1973	19/02/1973	6			<4
11/01/1986	20/01/1986	10			<4
20/02/1986	23/02/1986	4			<4
23/01/1987	29/01/1987	7			<4

Table 11-6 Number of consecutive days (minimum of 4) when SSTs were reported of ≥29°C and ≥30°C

As with the other analyses presented here, year 1983 features in this table with consecutive periods of seven days in which SSTs greater than or equal to 30°C were recorded in the northern third of the reef. Surpassing the length of that period though is 1923 when nine consecutive days of maximum temperatures exceeding 30C were reported for the central third of the reef.

11.4 Conclusions

While it has been sometimes claimed or implied that widespread severe coral bleaching has only occurred since 1998 the four methods of analysis shown here indicate a substantial likelihood that severe bleaching occurred but was unobserved earlier in the twentieth century.

The first analysis technique established a probability of 0.6 that severe bleaching occurred prior to 1998 and the subsequent three showed that land and sea summer temperatures reported after 1998 also occurred prior to that year, suggesting a strong possibility of earlier bleaching. Three of the four methods deal directly with temperatures either on land or at sea and the fourth with El Nino events, which are recognised as causing elevated temperatures. All four methods indicate likely severe bleaching episodes prior to 1998.

As noted earlier, this analysis does not take into account the impact of winds, especially cooling sea breezes, tides in any sense other than generally and that acclimatisation to any bleaching prior to 1998 would suggest that the earlier bleaching occurred at a lower sea temperature.

No approach shown here is favoured over any of the other approaches because all four have various strengths and weaknesses. It is left to readers to form their own judgement as to which approach is most appropriate to their needs.

References

Aguilar E., I. Auer, M. Brunet, T.C. Peterson and J. Wieringa, (2003) "Guidance on Metadata and Homogenization" *WMO Tech. Doc*. 1186 53pp

Alexandersson, H., 1986. A homogeneity test applied to precipitation data. *J. Climatol.*, 6, pp661-675.

Alexandersson, H. and A. Moberg, 1997: Homogenization of Swedish temperature data. Part I: Homogeneity test for linear trends. *Int. J Climatol*., 17, 25–34.

Anthony, K.R.N. and A.P. Kerswell (2007) Coral mortality following extreme low tides and high solar radiation *Mar Biol* 151 pp1623–1631 doi: 10.1007/s00227-006-0573-0

Barnston, AG., M. Cheliah and S.B. Goldenberg (1997) Documentation of a highly ENSO-related SST region in the equatorial Pacific. *Atmosphere-Ocean* 35, pp367-383.

Belkin, I, M and A.L. Gordon (1996) *Southern Ocean Fronts from the Greenwich Meridian to Tasmania,* J. Geophys. Res., *doi: 10.1029/95JC02750*

Berlage, H.F. and H.J. de Boer (1959) On the extension of the Southern Oscillation throughout the world during the period July 1, 1949 up to July 1, 1957, *Geofis. Pura e appl*. 44, pp287-295

Brohan, P., R. Allan, J. E. Freeman, A. M. Waple, D. Wheeler, C. Wilkinson and S. Woodruff, (2009), Marine Observations of Old Weather. *Bull. Amer. Meteor. Soc.*, 90, 219–230. doi: http://dx.doi.org/10.1175/2008BAMS2522.1

Brohan, P., J.J.Kennedy, I.Harris, S.F.B.Tett and P.D.Jones (2006) Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850, *J. Geophys. Res*. 111: D12106. doi:10.1029/2005JD006548

Brooks, C.F. (1926) Observing Water-Surface Temperatures at Sea, *Mon. Wea. Rev*., 54, 241–253, doi:10.1175/1520-0493(1926)54<241:OWTAS>2.0.CO;2

Brown, B.E., Suharsono (1990) Damage and recovery of coral reefs affected by El Nino related seawater warming in the Thousand Islands, Indonesia *Coral Reefs* 8(4) pp163-170 doi:10.1007/BF00265007

Cane, M.A. (2011) The Evolution of El Nino, past and future, *Earth and Planetary Science Letters* 164 (2004) pp1-10

Christy, J.R., R.W. Spencer and W.D. Braswell (2000), MSU tropospheric temperatures: Dataset construction and radiosonde comparison, *J. Atmos. Oceanic Technol*., 17, pp1153–1170, doi:10.1175/1520-0426(2000)017<1153:MTTDCA>2.0.CO;2.

Clement A., P. N. Dinezio and C. Deser (2011) Rethinking the Ocean's Role in the Southern Oscillation, *J. Clim* 24(15) pp4056-4072

De'ath, G., Fabricius, K. E., Sweatman, H., & Puotinen, M. (2012) The 27–year decline of coral cover on the Great Barrier Reef and its causes. *Proceedings of the National Academy of Sciences of the United States of America* (PNAS), 109(44), pp17995–17999

de Freitas, C.R., M.O.Dedekind and B.E.Brill (2014) A reanalysis of long-term surface air temperature trends in New Zealand, *Environmental Modelling and Assessment*, 20(4), 399-410 DOI:10.1007/s10666-014-9429-z

de Freitas, C.R. and J.D.McLean (2013) Update of the Chronology of Natural Signals in the Near-Surface Mean Global Temperature Record and the Southern Oscillation Index, *IJG*, 4, 234-239 doi:10.4236/ijg.2013.41A020 Published Online January 2013 (http://www.scirp.org/journal/ijg)

Dijkstra, H.A. (2006) The ENSO Phenomenon: theory and mechanisms, *Advances in Geosciences*, 6, pp3-15.

Donlon, C. J. and the GHRSST-PP Science Team (2005) The Recommended GHRSST-PP Data Processing Specification GDS (Version 1 revision 1.6). *GHRSST-PP Report Number 17*. The GHRSST-PP International Project Office, Exeter, U.K., 245 pp. (available at: ftp://podaac.jpl.nasa.gov/allData/ghrsst/docs/GDS-v1.0-rev1.6.pdf)

Emery, W.J., D.J.Baldwin, P.Schlüssel and R.W. Reynolds, (2001) Accuracy of in situ sea surface temperatures used to calibrate infrared satellite measurements *J. Geophys. Res.* v106 no. C2, pp2387-2405

Ewing, G. and E.D.MacAlister (1960) On the thermal boundary layer of the ocean *Science* 131 pp1374-1376

Fairall, C.W., E.F.Bradley, D.P.Rogers, J.B.Edson and G.S.Young (1996) Bulk parameterization of air-sea fluxes for Tropical Ocean-Global Atmosphere Coupled Ocean-Atmosphere Response Experiment. *J. Geophys. Res.*, 101, pp3747–3764.

Fall, S., A.Watts, J.Nielsen-Gammon, E.Jones, D.Niyogi, J.Christy and R.A. Pielke Sr., (2011) Analysis of the impacts of station exposure on the U.S. Historical Climatology Network temperatures and temperature trends, *J. Geophys. Res*, 116, D14120, doi:10.1029/2010JD015146

Farrar, J. T., C. J. Zappa, R. A. Weller, and A. T. Jessup (2007), Sea surface temperature signatures of oceanic internal waves in low winds, *J. Geophys. Res.*, 112, C06014, doi:10.1029/2006JC003947.

Fedorov, K.N., and A.I.Ginsburg (1992), The near-surface layer of the ocean, Koninldije Wohrmann, Utrecht, Netherlands, pp63-108

Folland, C.K. and F.E.Kates, (1984) Changes in decadally averaged sea-surface temperature over the world 1861-1980: pp. 721-727 in "Milankovitch & climate", pr 2, Eds. A. L. Berger et al., D. Reidel

Folland, C.K. and D.E.Parker (1995) Correction of instrumental biases in Historical sea surface temperature data *Q.J.R. Meteorolol. Soc.,* 121, pp. 319-367

Folland, C. K., N.A. Rayner, S. J. Brown, T. M. Smith, S. S.P. Shen, D. E. Parker, I. Macadam, P. D. Jones, R. N. Jones, N. Nichols and D.M. H. Sexton (2001) Global temperature change and its uncertainties since 1861, *Geophys. Res. Lett.*, 28(13), 2621–2624

Frank, P. (2010) Uncertainty in the Global Average Surface Air Temperature Index: A representative lower limit, *Energy & Environment,* vol 21, No. 8, pp 969-989

Fujibe, F. (2009) Detection of urban warming in recent temperature trends in Japan, *Int. J. Climatol.*, 29, pp1811–1822, doi:10.1002/joc.1822.

Giese, B.S., G.A.Chepurin, J.A.Carton, T.P.Boyer and H.F.Seidel (2011) Impact of Bathythermograph Temperature Bias Models on an Ocean Reanalysis *J. Clim.* v24 pp84-93 doi:10.1175/2010JCLI3534.1

Glynn, P.W. Widespread Coral Mortality and the 1982-83 El Nino Warming Event *Environmental Conservation* 11(2) pp133-146 doi:10.1017/S0376892900013825.

J. Hansen, J., R. Ruedy, M. Sato, M. Imhoff, W. Lawrence, D. Easterling, T. Peterson and T. Karl (2001) A closer look at United States and global surface temperature change, *J.. Geophys. Res.* v106, pp 23947-23963

Hasse, L. (1963) On the cooling of the sea surface by evaporation and heat exchange *Tellus* 15 pp363-366

Hausfather, Z., M.J.Menne, C.N.Williams, T. Masters, R. Broberg, and D. Jones (2013) Quantifying the effect of urbanization on U.S. Historical Climatology Network temperature records, *J. Geophys. Res*:*Atmos*, v118, pp 481–494, doi:10.1029/2012JD018509

Hessell, J.W.D (1980) Apparent trends in mean temperatures in New Zealand since 1930, *New Zealand Journal of Science*, v23, p1-9

Hinkel, K.M., F.E. Nelson, A.E. Klene and J.H. Bell (2003) The Urban Heat Island in winter at Barrow, Alaska, *Int. J. Climatol*. 23 pp 1889-1905

Hoegh-Guldberg, O. and M. Fine (2004) Low temperatures cause coral bleaching, *Coral Reefs*, v23 3 pp444

Hughes, T.P., J.T. Kerry, M. Álvarez-Noriega (plus 40 others) (2017) Global warming and recurrent mass bleaching of corals *Nature*, 543(7645), pp373-377, doi: 10.1038/nature21707

IPCC (2013) Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, eds. Stocker, T.F., D. Qin, G-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp.

Jones, P. D. (1994) Hemispheric surface air temperature variations: a reanalysis and an update to 1993. *J. Climate*, 7, 1794–1802.

Jones, P. (2016) The reliability of global and hemispheric surface temperature records *Advances in Atmospheric Sciences* v33, pp269-282

Jones, P.D., D.H. Lister, and Q. Li (2008) Urbanization effects in large-scale temperature records, with an emphasis on China, *J. Geophys. Res.*, 113, D16122, doi:10.1029/2008JD009916.

Jones, P.D., D.H. Lister, T.J. Osborn, C. Harpham, M. Salmon and C.P. Morice (2012) Hemispheric and large-scale land surface air temperature variations: an extensive revision and an update to 2010. *J. Geophys. Res.* 117, D05127, doi:10.1029/2011JD017139.

Kalnay, E., M. Cai, H. Li, and J. Tobin (2006) Estimation of the impact of land-surface forcings on temperature trends in eastern United States, *J. Geophys. Res.*, 111, D06106, doi:10.1029/2005JD006555.

Kawai, Y and A Wada (2007) Diurnal Sea Surface Temperature Variation and its Impact on the Atmosphere and Ocean: A Review, *Journal of Oceanography*, v63, pp721-744

Kennedy J.J., N.A. Rayner, R.O. Smith, D.E. Parker and M. Saunby (2011a) Reassessing biases and other uncertainties in sea-surface temperature observations measured in situ since 1850 part 1: measurement and sampling uncertainties *J. Geophys. Res.* 116, D14104, doi:10.1029/2010JD015220

Kennedy J.J., N.A. Rayner, R.O. Smith, M. Saunby and D.E. Parker (2011b) Reassessing biases and other uncertainties in sea-surface temperature observations measured in situ since 1850 part 2: biases and homogenisation. *J. Geophys. Res.* 116, D14104, doi:10.1029/2010JD015220

Kent, E.C and P.K.Taylor (2006) Toward Estimating Climatic Trends in SST Part I: Methods of Measurement *Journal of Atmospheric and Oceanic Technology* v23 pp 464-475

Kizu, S., C.Sukigara1 and K.Hanawa1 (2011) Comparison of the fall rate and structure of recent T-7 XBT manufactured by Sippican and TSK *Ocean Sci.*, 7, pp231–244

Kousky, V.E., M.T. Kagano and I.F.A. Cavalcanti (1984) A review of the Southern Oscillation: oceanic-atmospheric circulation changes and related rainfall anomalies, *Tellus*, 36A, pp490-504.

Kug, J-S. and In-sik Kang (2006) Interactive Feedback between ENSO and Indian Ocean *Jour. Clim.* 19(9).

Kumar, K.K., B. Rajagopalan, M. Hoerling, G. Bates and M. Cane (2006), Unraveling the mystery of Indian monsoon failure during El Nino, *Science*, v 3134, pp115-119

Lirman D, S. Schopmeyer, D. Manzello, L.J. Gramer, W.F. Precht et al. (2011) Severe 2010 Cold-Water Event Caused Unprecedented Mortality to Corals of the Florida Reef Tract and Reversed Previous Survivorship Patterns. *PLoS ONE* 6(8): e23047. doi:10.1371/journal.pone.0023047

Lough, J. (2008) 10th Anniversary Review: a changing climate for coral reefs. *J. Environ Monit.* 10(1):21-9. doi: 10.1039/b714627m.

Matthews, J.B.R (2013) Comparing historical and modern methods of sea surface temperature measurement - Part 1: Review of methods, field comparisons and dataset adjustments, *Ocean Sci*., 9, pp683-694

Matthews, J.B.R and J.B. Matthews (2013) Comparing historical and modern methods of sea surface temperature measurement - Part 2: Field comparison in the central tropical Pacific, *Ocean Sci*., 9, pp695-711

Maynard, J.A., Anthony, K.R.N., Marshall, P.A. and Masiri, I. (2008) Major bleaching events can lead to increased thermal tolerance in corals *Marine Biology (Berlin)* 155 pp173-182

McKitrick, R. (2010) A critical review of global surface temperature data products, *SSRN Working paper 1653928*, (75 pages), abstract online at http://ssrn.com/abstract=1653928

McKitrick, R. (2013) Encompassing Tests of Socioeconomic Signals in Surface Climate Data *Climatic Change* doi 10.1007/s10584-013-0793-5 vol 120, No. 1-2

McLean, J (2014) Late Twentieth-Century Warming and Variations in Cloud Cover, *Atmospheric and Climate Sciences*, open access at http://www.scirp.org/journal/PaerInformation.aspx?PaperID=50837 pp 727-742

McLean, J.D., C. R. de Freitas and R. M. Carter (2009) Influence of the Southern Oscillation on Tropospheric Temperature, *J. Geophys. Res.* Vol. 114, No. D14. doi:10.1029/2008JD011637

Menne, M. J. and C.N. Williams Jr (2009) Homogenization of temperature series via pairwise comparisons *J. Clim*., 22 pp1700–1717, doi:10.1175/2008JCLI2263.1.

Middlebrook, R., O. Hoegh-Guldberg and W. Leggat (2008) The effect of thermal history on the susceptibility of reef-building corals to thermal stress *The Journal of Experimental Biology* 211 pp1050-1056

Mobasheri, M.R. (1995) Heat Transfer in the upper layer of the ocean with application to the correction of satellite sea surface temperature, PhD Thesis (unpublished), James Cook University

Morice, C.P., J.J.Kennedy, N.A.Rayner and P.D.Jones (2012) Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set *JGR:Atmospheres* Volume 117, Issue D8 27 April 2012, doi:10.1029/2011JD017187

Mullan, A.B., S.J. Stuart, M.G. Hadfield and M.J. Smith (2010) "Report on the Review of NIWA's 'Seven-Station' Temperature Series" *NIWA Information Series No. 78*. p175, online as https://www.niwa.co.nz/sites/default/files/import/attachments/Report-on-the-Review-of-NIWAas-Seven-Station-Temperature-Series_v3.pdf

Newhall, C. G., and S. Self (1982), The volcanic explosivity index (VEI): An estimate of explosive magnitude for historical volcanism, *JGR:Oceans and Atmospheres*, 87(C2), 1231– 1238, doi:10.1029/JC087iC02p01231.

Osborn T.J. and P. Jones (2014) The CRUTEM4 land-surface air temperature data set: construction, previous versions and dissemination via Google Earth *Earth Syst. Sci. Data*, 6, 61–68, 2014 www.earth-syst-sci-data.net/6/61/2014/ doi:10.5194/essd-6-61-2014

Parker, D.E. (2010) Urban heat island effects on estimates of observed climate change *WIREs Clim. Change*, 1(1), pp123–133, doi:10.1002/wcc.21

Parker,D.E., M.Jackson, and E. B. Horton (1995b) The GISST2.2 sea surface temperature and sea ice climatology. *Climate Research Tech. Note CRTN 63*, Hadley Centre, Met Office, Exeter, United Kingdom, 35pgs.

Peterson, T.C. and D.R. Easterling (1994) Creation of homogeneous composite climatological reference series. *Int. J. Climtol*., 14, pp671-679.

Peterson, T.C., D.R.Easterling, T.R.Karl, P. Groisman, N.Nicholls, N.Plummer, S.Torok, I. Auer, R. Böhm, D.Gullett, L.Vincent, R.Heino, H. Tuomenvirta, O. Mestre, T. Szentimrey, J.Salinger, E.J. Førland, I. Hanssen-Bauer, H. Alexandersson, P.E. Jones, and D.Parker (1998). Homogeneity adjustments of in situ atmospheric climate data: a review, *Int. J. Climatol*., 18, pp1493-1517.

Rayner,N.A., P.Brohan, D.E.Parker, C.K.Foland,J.J.Kennedy,M.Vanicek,T.J.Ansell and S.F.B.Tett (2006) Improved Analyses of Changes and Uncertainties in Sea Surface Temperature Measured In Situ since the Mid-Nineteenth Century: The HadSST2 Dataset *J. Clim.* v19(3) pp446-469

Reason, C.J.C, R.J. Allan, J.A. Lindesay and T.J. Ansell (2000) ENSO and climatic signals across the Indian Ocean Basin in the global context: Part 1, Interannual composite patterns, *Int. J. Climatol*. 20, pp1285-1327.

Rhoades, D. A., & Salinger,M. J. (1993). Adjustment of temperature and rainfall records for site changes. *Int J Climatol*, 13, pp899–913

Rodríguez-Troncoso, A.P., E. Carpizo-Ituarte and A.L. Cupul-Magaña (2010) Differential response to cold and warm water conditions in *Pocillopora* colonies from the Central Mexican Pacific, *Journal of Experimental Marine Biology and Ecology*, v391 –2, pp57–64.

Ropelewski, C.F. and P.D. Jones (1987) An Extension of the Tahiti-Darwin Southern Oscillation Index, *Monthly Weather Review* v115 pp2161-2165

Rumsey D.J (2011), Statistics for Dummies, *Wiley Publishing Inc*, Indianapolis

Salinger,M. J. (1980) The New Zealand temperature series. *Climate Monitor*, 9(4), pp112–118.

Saur, J.F.T. (1963) A Study of the Quality of Sea Water Temperatures Reported in Logs of Ships' Weather Observations, *J. Appl. Meteor.*, 2, 417–425, doi:10.1175/1520-0450(1963)002<0417:ASOTQO>2.0.CO;2

Schoepf, V., M. Stat, L. Falter and M.T. McCulloch (2015) Limits to the thermal tolerance of corals adapted to a highly fluctuating, naturally extreme temperature environment *Scientific Reports (Nature Publisher Group)* v5 pp17639-17652

Shapiro, S. S. and M.B. Wilk (1965). An analysis of variance test for normality (complete samples) *Biometrika* 52 (3–4): 591–611. doi:10.1093/biomet/52.3-4.591 available via https://www.jstor.org/stable/2333709

Simkin, T. and L. Siebert (1994), "Volcanoes of the World", 2nd ed., 349 pp., Geoscience Press, Tucson, Ariz., 1994.

Simkin, T., L.Siebert, L.McClelland, D.Bridge, C.G.Newhall, and J.H.Latter (1981), Volcanoes of the World, 232 pp., Van Nostrand Reinhold, New York.

Smith, T.M. and R.W.Reynolds (2004) Improved Extended Reconstruction of SST (1854-1997), *J. Clim.* v 17, pp 2466-2477

Smith, T.M. and R.W.Reynolds (2005) A Global Merged Ladn-Air-Surface Temperature Reconstruction Based on Historical Observations (1880-1997), *J. Clim.* v18 pp2021-2036

Soloviev, A. and R. Lukas (1997) Observation of large diurnal warming events in the near-surface layer of the western Equatorial Pacific warm pool, *Deep Sea Research I*, v 44 pp 1055-1076

Spencer, R.W., and J.R. Christy, (1992) Precision and radiosonde validation of satellite gridpoint temperature anomalies, Part II: A tropospheric retrieval and trends during 1979-90. *J. Climate*, 5, pp858-866.

Steirou, E. and D. Koutsoyiannis (2012) Investigation of methods for hydroclimatic data homogenization *Geophysical Research Abstracts* Vol. 14, EGU2012-956-1, 2012, presented at European Geosciences Union General Assembly 2012 Vienna, Austria, 22-27 April 2012 (see also http://itia.ntua.gr/1212)

Stothers, R.B. (2001) Major optical depth perturbations to the stratosphere from volcanic eruptions: Stellar extinction period, 1961-1978. *J. Geophys. Res.*, 106, 2993-3003, doi:10.1029/2000JD900652.

Thompson, W.D.J., J.J.Kennedy, J.M.Wallace and P.D.Jones (2008) A large discontinuity in the mid-twentieth century in observed global-mean surface temperature *Nature* Vol 453 pp 646-649 doi:10.1038/nature06982

Toole, J.M. and B.A. Warren (1993) A hydrographic section across the subtropical South Indian Ocean, *Deep Sea Research Part I: Oceanographic Research Papers*, v40, No. 10, pp1973-2019.

Toreti,A., F.G.Kuglitsch, E. Xoplaki, J. Luterbacher and H. Wanner (2010) A Novel Method for the Homogenization of Daily Temperature Series and Its Relevance for Climate Change Analysis *J. Clim* v23, pp 5325-5331

Toreti,A., F.G.Kuglitsch, E.Xoplaki, P.M. Della-Marta, E. Aguilar, M. Prohom and J.Luterbacher (2011) A Note on the use of the standard normal homogeneity test to detect inhomogeneities in climatic time series, *Int. J. Climatol.* 31 pp630-632

Trewin, B (2012) Techniques involved in developing the Australian Climate Observations Reference Network - Surface Air temperature (ACORN-SAT) dataset, CAWCR Technical Report No. 049. The Centre for Australian Weather and Climate Research.

Trewin, B. (2013) A daily homogenized temperature data set for Australia *Int. J. Climatol.* 33: pp 1510–1529 (2013), DOI: 10.1002/joc.3530

Troup, A. J. (1965), The 'Southern Oscillation,' *Q. J. R. Meteorol. Soc.*, 91, 490– 506, doi:10.1002/qj.49709139009.

Tuomenvirta, H (2002) Homogeneity Testing and Adjustment of Climatic Time Series in Finland *Geophysica* (2002), 38(1-2), pp15-41

Webster, P.J., C.A. Clayson, and J.A. Curry (1996) Clouds, Radiation, and the Diurnal Cycle of Sea Surface Temperature in the Tropical Western Pacific, J. Clim., v9 pp1712-1730

Wells, A.J., C.Cenedese, J.T.Farrar and C.J.Zappa (2009) Variations in Ocean Surface Temperature due to Near-Surface Flow:Straining the Cool Skin Layer, *Journal of Physical Oceanography*, v39, pp2685-2710

Whetton P and I.D.Rutherford (1994) Historical ENSO teleconnections in the Eastern Hemisphere, *Climatic Change*, 28, pp221-253

Witte R.S. and J.S. Witte (2010) Statistics, *John Wiley & Sons Inc*, Hoboken, 9th ed.

WMO (1989) WMO-TD/No. 341: Calculation of Monthly and Annual 30-year standard models, *World Meteorological Organization*, Geneva

WMO (2010) WMO-No. 8 Guide to Meteorological Instruments and Methods of Observation *World Meteorological Organization*, Geneva

WMO (2011) WMO-No. 100: Guide to Climatological Practices *World Meteorological Organization*, Geneva

Wolter, K., and M.S. Timlin (1993) Monitoring ENSO in COADS with a seasonally adjusted principal component index in *Proc. of the 17th Climate Diagnostics Workshop, Norman, OK*, NOAA/NMC/CAC, NSSL, Oklahoma Clim. Survey, CIMMS and the School of Meteor., Univ. of Oklahoma, pp52-57.

Wolter, K. and M.S. Timlin (2011) El Nino/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext) *Int. J. Climatol.*, 31, pp 1074-1087.

Woodcock, A.H. (1941) Surface cooling and streaming in shallow and salt water *J.Mar.Res* 4 p153

Wyrtki, K (1989) Some thoughts about the west Pacific Warm Pool, in Picault, Lukas and Decroix (eds) *Proc. West Pacific Int. Meeting and Workshop on TOGA, COARE, ORSTOM, Noumea, New Caledonia*, pp99-109

Yonge, C.M. and Nicholls (1931), G.Barrier Reef Expedition, Sci. Repts., Brit. Mus., 1, 135

Appendix 1 - Observation station site classes according to WMO standards

(refers chapter 6)

Chapter 1 of WMO-No. 8 (WMO, 2010) Part 1 describes five site classes of temperature observation stations, noting that class 3 has "additional estimated uncertainty added by siting up to 1°C", class 4 up to 2°C and class 5 up to 5°C. An appendix to that document defines the various classes as follows:

CLASS 1

(a) Flat, horizontal land, surrounded by an open space, slope less than ⅓ (19°);

(b) Ground covered with natural and low vegetation (< 10 cm) representative of the region;

(c) Measurement point situated:

  (i) At more than 100 m from heat sources or reflective surfaces (buildings, concrete surfaces, car parks, etc.);

  (ii) At more than 100 m from an expanse of water (unless significant of the region);

  (iii) Away from all projected shade when the sun is higher than 5°.

A source of heat (or expanse of water) is considered to have an impact if it occupies more than 10 per cent of the surface within a circular area of 100 m surrounding the screen, makes up 5 per cent of an annulus of 10–30 m, or covers 1 per cent of a 10 m circle.

CLASS 2

(a) Flat, horizontal land, surrounded by an open space, slope inclination less than ⅓ (19°);

(b) Ground covered with natural and low vegetation (<10 cm) representative of the region;

(c) Measurement point situated:

(i) At more than 30 m from artificial heat sources or reflective surfaces (buildings, concrete surfaces, car parks, etc.);

(ii) At more than 30 m from an expanse of water (unless significant of the region);

(iii) Away from all projected shade when the sun is higher than 7°.

A source of heat (or expanse of water) is considered to have an impact if it occupies more than 10 per cent of the surface within a circular area of 30 m surrounding the screen, makes up 5 per cent of an annulus of 5–10 m, or covers 1 per cent of a 5 m circle.

CLASS 3

(a) Ground covered with natural and low vegetation (< 25 cm) representative of the region;

(b) Measurement point situated:

(i) At more than 10 m from artificial heat sources and reflective surfaces (buildings, concrete surfaces, car parks, etc.);

(ii) At more than 10 m from an expanse of water (unless significant of the region);

(iii) Away from all projected shade when the sun is higher than 7°.

A source of heat (or expanse of water) is considered to have an impact if it occupies more than 10 per cent of the surface within a circular area of 10 m surrounding the screen or makes up 5 per cent of an annulus of 5 m.

CLASS 4

(a) Close, artificial heat sources and reflective surfaces (buildings, concrete surfaces, car parks, etc.) or expanse of water (unless significant of the region, occupying:

(i) Less than 50 per cent of the surface within a circular area of 10 m around the screen;

(ii) Less than 30 per cent of the surface within a circular area of 3 m around the screen;

(b) Away from all projected shade when the sun is higher than 20°.

CLASS 5

Site not meeting the requirements of class 4.

Appendix 2 - Alternative methods for identifying inhomogeneities in temperature data

(refers Chapter 6)

The following methods of identifying inhomogeneities in temperature data records are given in Aguilar et al (2012):

(i) Caussinus-Mestre

– Assumes that a dataset for a given site is homogenous between one break and the next (or to its end) and compares other data from that same site to it.

(ii) Bushland range test

– Sums the temperature anomalies calculated against the mean of the temperature across the data being considered and looks for peaks.

(iii) Craddock test

– Uses a homogenous reference series and accumulates the normalised difference between the test series and reference series.

(iv) Expert Judgement

– Relies on expert interpretation of usually a graphical display of data – target series, neighbouring series and perhaps the difference between the two – to identify discontinuities.

(v) Instrument comparison

– Side-by-side comparison of instruments in their enclosures; useful particularly when changing instruments

(vi) Multiple Analysis of Series for Homogenisation (MASH)

– Compares target series to multiple other series from the same climatic area, essentially using the composite of the others as a reference series.

(vii) Multiple Linear Regression

- Applies four regression models to test for steps and trends in the data using surrounding stations as references.

(viii) Petit Test

– A non-parametric rank test that then draws on significance tables.

(ix) Potter's Method

– tests the hypothesis that the combined population before and after a target year has the same bivariate normal distribution as the before and after periods considered separately.

(x) Radiosonde data

– Uses radiosonde data obtained by balloons to estimate surface temperatures and compares these estimates to the measured temperatures.

(xi) Rank-Order Change point test

– Uses a non-parametric test related to Wilcoxon-Mann-Whitney test and based on the sum of the ranks of values from the start of data to the target point

(xii) Standard Normal Homogeneity test

– Takes the difference at each point between a target series and a reference series, normalises those differences (i.e. subtract from mean and divide by standard deviation) then tests for symmetry before and after each target value.

(xiii) Stop-Trend method

– Splits data into intervals and compares differences in consecutive values in each interval.

(xiv) Two phase regression

– Looks at the difference between lines of regression before and after a target value, which might be a temperature measurement or might be the difference between data from a target site and corresponding data at a reference site.

Appendix 3 - Diagrams describing recognised ENSO states

(refers Chapter 9)

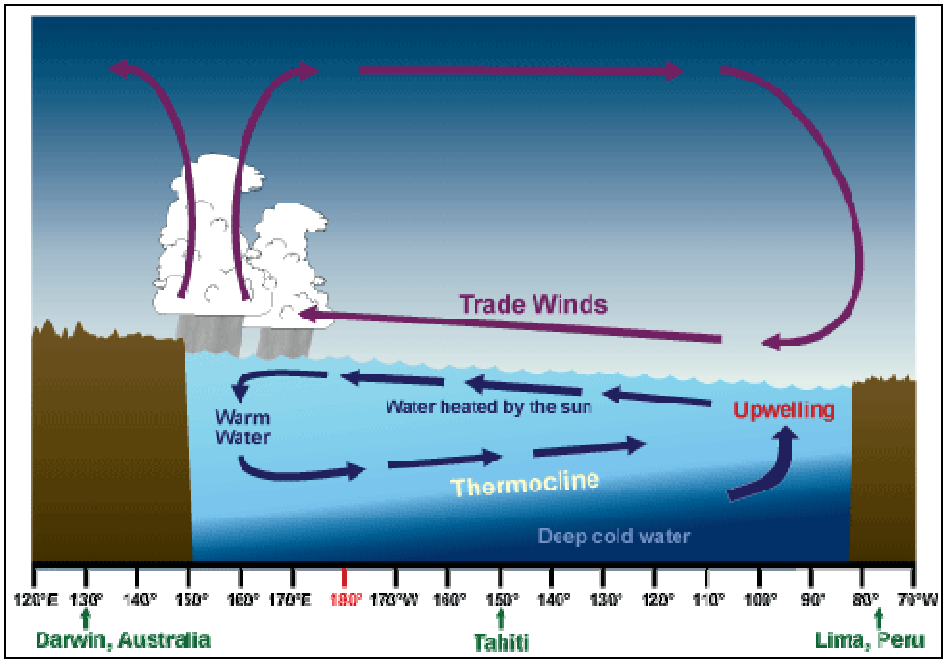


Figure A2.1 Neutral ENSO state, with easterly winds across the Pacific and circulating water. Image credit: NOAA

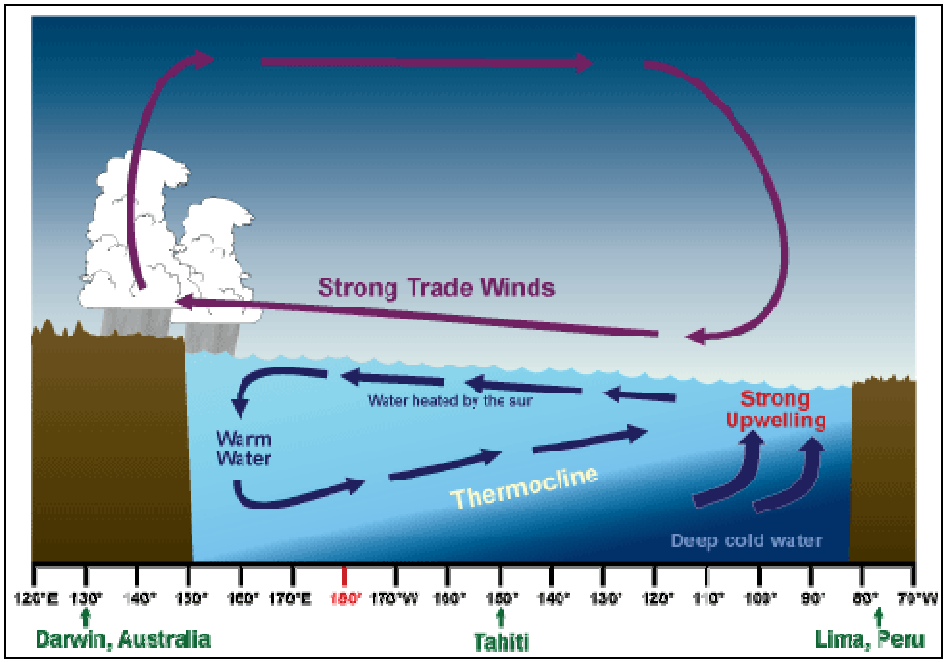


Figure A2.2 La Niña state, with an increase in cold water upwelling and strengthened winds across the Pacific. Image credit: NOAA

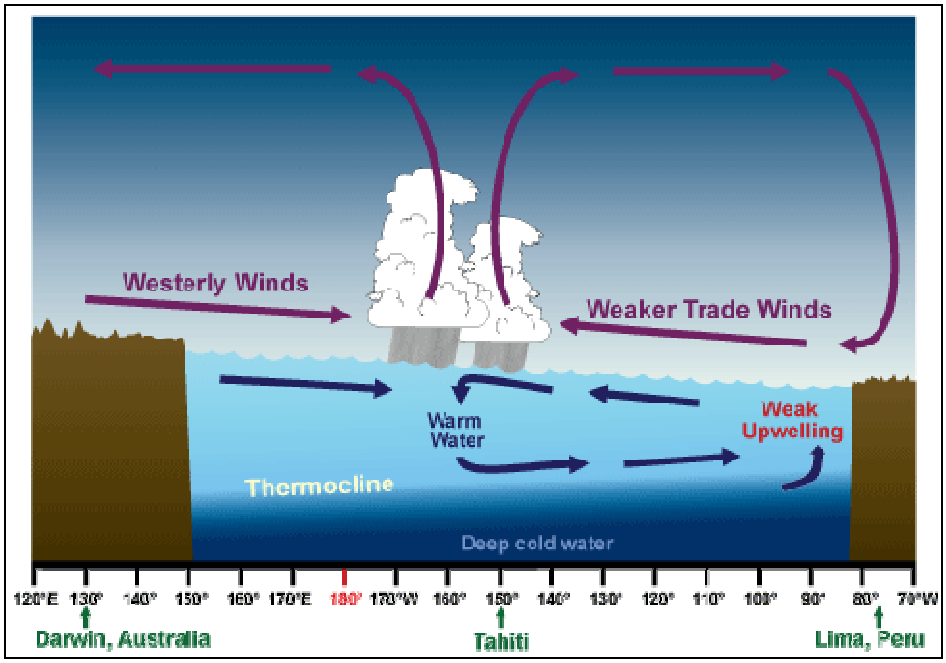


Fig A2.3 El Niño state, with reduced upwelling and reduced winds. Image credit: NOAA

Appendix 5 - Correlation coefficients for comparison of Troup SOI and MSLP

The table below lists the correlation coefficient at locations additional to those shown in Chapter 9, table 9-3.

Name	Country	Lat	Long	Yr beg	Yr end	Correl Coeff	Comments
Auckland	New Zealand	37.02S	174.08E	1940	1992	0.109	
Bangkok	Thailand	13.73N	100.57E	1940	2001	-0.485	
Brownsville	USA	25.92N	97.42W	1951	2001	0.053	Texas-Mexico border
Calcutta	India	22.53N	88.33E	1940	2001	-0.213	1940-1981 coeff = -0.479
Cebu	Philippines	10.30N	123.90E	1951	1982	-0.400	downward MSLP trend
Chittagong	Bangladesh	22.35N	91.82E	1948	1987	-0.401	15% of data missing
Christchurch	New Zealand	43.48S	172.52E	1940	1993	0.274	
East Sale	Australia	38.10S	147.13E	1961	2001	-0.396	
Funafuti	Tuvalu	8.52S	179.22E	1940	1993	0.273	
Guam	US Territory	13.55N	144.83E	1965	1993	-0.433	with SOI 5 months previously, -0.753
Guangzhou	China	23.13N	113.23E	1961	2001	-0.246	
Hihifo	Wallis and Fortuna Is.	13.23S	176.7W	1951	1994	0.097	
Hilo	USA	19.72N	155.07W	1961	2011	0.387	with SOI 6 months ahead, 0.590
Honiara	Solomon Islands	9.42S	159.97E	1951	1976	0.116	
Honolulu Obs.	USA	21.30N	158.10W	1940	1985	0.320	with SOI 5 months ahead 0.483
Kwajalein	Marshall Islands	8.73N	167.73E	1971	2001	-0.202	with SOI 5 months previous, -0.529
Legazpi	Philippines	13.13N	123.73E	1947	2000	-0.379	
Martin de Vivies	French territories	37.78S	77.52E	1951	1990	-0.029	
Melbourne	Australia	37.82S	144.97E	1940	1982	-0.382	
Nadi	Fiji	17.75S	177.45E	1947	2001	0.136	
Oita	Japan	33.23N	131.62E	1951	2001	-0.168	
Pamban	India	9.27N	79.30E	1940	2001	-0.268	1981 onwards, -0.766
Ponape	Micronesia	6.97N	158.22E	1961	2001	-0.265	with SOI 4 months previously, -0.401
Port Vila	Vanuatu	17.8S	168.30E	1954	1981	-0.426	
Raizet	Guadaloupe	16.27N	61.52W	1955	1989	0.063	
Raoul Island	New Zealand	29.25S	177.92E	1940	1988	-0.070	
Rodriguez	Mauritius	19.68S	63.42E	1954	2001	-0.310	
Sea grid cell 3	btwn Baja Sur and Hawaii	26-28N	116-118W	1966	2014	0.398	
Shantou	China	23.40N	116.68E	1961	2001	-0.396	

							with SOI 5 months
Wake Island	US territory	19.28N	166.65E	1949	1996	-0.226	previously, -0.378
Williamtown	Australia	32.78S	151.82E	1957	2001	-0.393	

Table A3-1 Alphabetic listing of other locations where the correlation between Troup SOI and NMSLP was determined.