A Statistical Evaluation of GCMs:

Modeling the Temporal Relation between Radiative Forcing and Global Surface Temperature

Robert K. Kaufmann
Center for Energy and Environmental Studies
675 Commonwealth Avenue
Boston University
Boston, MA
02215

Telephone (617) 353-3940
Fax (617) 353-5986
Kaufmann@bu.edu

David I. Stern

Department of Economics
Sage 3208
Rensselaer Polytechnic Institute
Troy, NY 12180-3590

Fax (518) 276-2235
sternd@rpi.edu

Abstract

We evaluate the statistical adequacy of three general circulation models (GCMs) by testing three aspects of a GCM's ability to reconstruct the historical record for global surface temperature: (1) how well the GCMs track observed temperature; (2) are the residuals from GCM simulations random (white noise) or are they systematic (red noise or a stochastic trend); (3) what is the explanatory power of the GCMs compared to a simple alternative model, which assumes that temperature is a linear function of radiative forcing. The results indicate that three of the eight experiments considered fail to reconstruct temperature accurately; the GCM errors are either red noise processes or contain a systematic error, and the radiative forcing variable used to simulate the GCM's have considerable explanatory power relative to GCM simulations of global temperature. The GFDL model is superior to the other models considered. Three out of four Hadley Centre experiments also pass all the tests but show a poorer goodness of fit. The Max Planck model appears to perform poorly relative to the other two models.

1. Introduction

Using general circulation models (GCMs) to attribute changes in surface temperature to human activity and to quantify the uncertainty associated with that effect assumes that GCMs simulate the relation between radiative forcing and surface temperature accurately. Consistent with this notion, GCM temperature reconstructions "look like" the historical temperature record (Figure 1). The spatial and temporal aspects of the similarity between GCM predictions and observations have been investigated using fingerprinting techniques (Mitchell et al, 2001).[1]

Here we use statistical techniques to test three aspects of a GCM's ability to reconstruct the historical record for global surface temperature. First, we test how well the GCMs track observed temperature. This test may be affected by measurement error, which is addressed in detail. Next, we test whether the residual from fitting the GCM simulations to observed temperature are random (white noise) or are systematic (red noise or a stochastic trend). Finally, we test the explanatory power of the GCMs against a simple alternative, which assumes that temperature is a linear function of radiative forcing. These three tests are based on the notion that a model is statistically adequate if its deviations from the data are non-systematic and if the model provides a superior fit than that provided by a simpler alternative (Harvey, 1989). The results indicate that three of the eight experiments that we consider fail to reconstruct temperature accurately. All of the GCM errors are red noise processes and three may contain a random walk. For all experiments, the radiative forcing aggregate used to simulate the GCM has considerable explanatory power relative to the temperature simulation generated by the GCM.

[1] An exception is the study by *Shackley et al*. (1998) who follow a similar approach to our own. Though they criticize GCMs their statistical analysis is applied to a carbon cycle model. An example of an alternative approach that goes beyond the usual fingerprinting technique of computing goodness of fit measures for alternative models

These results and the methods used to generate them are described in five sections. Section 2 describes the statistical theory that is used to evaluate the GCM temperature reconstructions. Section 3 describes the methodology that is used to perform the tests described above. Section 4 describes the results. These results are discussed in section 5. The implications of this research for on-going efforts to attribute temperature change to human activity and quantify the uncertainty of this effect are described in section 6.

2. Statistical Theory

In this section, we briefly review two areas of statistical theory that underlie our methodology: (1) The properties of stochastic trends and the notion of cointegration; and (2) The effects of errors in variables and methods to remedy the difficulties they pose to statistical analyses. GCM simulations of surface temperature, observations of surface temperature, and the radiative forcing of greenhouse gases, solar activity, and anthropogenic sulfur emissions are nonstationary due to the presence of a stochastic trend (Stern and Kaufmann, 2000; Kaufmann and Stern, 2002). Statistical analyses that use these data must take these trends into account and can use their properties to construct statistical tests. Additionally, estimates of regression parameters involving these variables may be biased downward in absolute value due to mismeasurement of the observed variables and inaccuracies in the GCM simulations. Readers familiar with these issues can proceed to section 3.

2.1. Stochastic Trends, Spurious Regressions, and Cointegration

---

and scenarios is the study by *Govindan et al*. (2002) who found that GCMs do not reproduce the universal scaling behavior of the observed records.

Time series can be classified as either white noise, red noise, or stochastically trending. Assume that the series, y, can be approximated by an autoregressive model:

$$y_t = y_0 + \sum_{i=1}^{p} \rho_i y_{t-i} + \varepsilon_t \qquad (1)$$

in which ε is a normally distributed random error, with mean zero, y_0 is a constant term, and the ρ_i are autoregressive coefficients. The classification of the series y depends on the eigenvalues of the matrix $\begin{bmatrix} \rho' \\ I, 0 \end{bmatrix}$ where ρ is the vector of autoregressive coefficients, I is the identity matrix of dimension p-1 and 0 is a vector of zeros. If all the eigenvalues are zero, the process is white noise. If some of the eigenvalues are non-zero but none are unity or greater, then the process is red noise (Gilman, *et al.,* 1963). Finally, if at least one of the eigenvalues equals 1.0, the series y contains a random walk component (i.e. it contains a stochastic trend). For example, in the first order autoregressive case a series is a random walk if the single autoregressive coefficient is unity. Stochastically trending variables do not have a constant mean and are, therefore, nonstationary. The random walk component implies that the effects of a shock do not fade over time—their effect is permanent.

The possible presence of a stochastic trend is critical to four aspects of our methodology. First, classical regression inference using the standard t and F distributions may indicate a statistically significant relationship among observed temperature, radiative forcing, and/or GCM temperature simulations even if no relationship exists (Granger and Newbold, 1974). This happens because a linear combination of time series that each contain a stochastic trend - the regression residual - usually contains a stochastic trend, which affects the distribution of the test statistics (Phillips, 1986). Such a regression is known as a spurious regression (Granger and Newbold, 1974). This problem cannot be eliminated by removing a linear time trend from the

data, but it can be eliminated by differencing the data. But if the regression represents a true linear relation among the variables – a cointegrating relation as described below - differencing destroys information about the long-run relationship.

Second, the omission of a relevant stochastically trending variable interferes with a statistically meaningful interpretation of the regression results. Even if there is a relation among the variables that are included in the regression equation, the omission of a relevant stochastically trending variable introduces a stochastic trend to the regression residual. As explained above, regression inference is, therefore, non-standard. Additionally, regression coefficients are biased due to the well-known problem of omitted variables if the omitted variables are correlated with the variables that are included in the regression. Differencing can be used to remove the stochastic trend from the regression variables and hence the residual. The omitted variable bias will remain, though it may be reduced because the correlations among the differenced variables and the omitted variables usually are smaller.

Third, the presence of a stochastic trend in the regression residual indicates that a relevant nonstationary variable has been omitted, an irrelevant variable has been included, the model includes an inappropriate nonlinearity that distorts the time path of the included stochastically trending variables, and/or does not include such a nonlinearity where it is required. The latter has important implications for tests of the ability of GCMs to reconstruct the observational temperature record.

Finally, the presence of stochastic trends in two or more series can be used to test whether the variables have a functionally dependent relation. If two or more variables are related and no relevant stochastically trending variable(s) is omitted from the regression equation, the series will share the same stochastic trends. In this case, there will be at least one linear combination of

the series that is stationary so that there is no stochastic trend in the regression residual. This phenomenon is known as cointegration (Engle and Granger, 1987). Cointegration allows us to use the stochastic trends as "fingerprints" to detect relations among the GCM simulations for surface temperature, observations of surface temperature, and radiative forcing.

To determine whether a time series or a series of regression residuals contains a stochastic trend, equation (1) can be rearranged as follows:

$$\Delta y_t = y_0 + \gamma_1 y_{t-1} + \sum_{i=1}^{s} \gamma_{i+1} \Delta y_{t-i} + \varepsilon_t \qquad (2)$$

where $\gamma_1 = (\rho_1 - 1)$, Δ is the first difference operator and the other regression coefficients are also based on the original parameters in (1), and s = p-1 is the number of “augmenting lags” of the first differences of y. The number of augmenting lags (s) is chosen using goodness of fit criteria.

The Augmented Dickey Fuller (ADF) statistic (Dickey and Fuller, 1979) is a t-test on the parameter γ_1. The null hypothesis for the ADF test is that the series contains a stochastic trend. The ADF test evaluates the null, $\gamma_1 = 0$ by comparing the t-statistic for γ_1 against a non-standard distribution that was developed for a variety of cases where y is either an observed series or the residual from a regression. If the null hypothesis is rejected, the series or regression residual does not contain a stochastic trend (i.e. it is stationary). If we cannot reject the null the variable is indistinguishable from a stochastically trending variable at the specified significance level for the length of sample and sampling frequency used. Other test statistics are available, but the ADF performs well relative to other tests (Stock, 1994)

The lack of a stochastic trend in the regression residual (i.e. rejecting the null hypothesis of the ADF test) implies that the variables cointegrate. The presence of cointegration indicates that the variables in the regression are related to each other in a statistically significant fashion and

that no important nonlinearity or relevant stochastically trending variable has been omitted. On the other hand, the lack of cointegration (i.e. a stochastic trend in the residual) indicates that: (1) the stochastically trending variables are not related to each other; (2) the regression equation is missing a relevant stochastically trending variable; (3) an irrelevant stochastically trending variable has been included. Regardless of the cause, the lack of cointegration indicates that the regression is spurious and its diagnostic statistics cannot be interpreted reliably.

2.2. Errors in Variables

Estimates of regression coefficients may be biased or estimated less precisely when there are errors in the variables included in the regression. If the errors occur only in the time series that is specified as the dependent variable, the regression coefficients will be unbiased but will be estimated imprecisely —the standard errors will be larger than when no errors are present (Hausman, 2001). If errors are present in the independent variables, the regression coefficients will be biased downward in absolute terms (Hausman, 2001). If errors are present in both the dependent and independent variables, the regression coefficients will be biased downward, but no more so than the case in which errors are present only in the independent variables (Hausman, 2001).

Errors in the historical data for temperature and the atmospheric concentrations and emissions that are used to calculate radiative forcing are well documented (Ramaswamy *et al.*, 2001; Jones 1994; Trenberth et al, 1992). Two errors in variables arguments apply to the temperature data simulated by GCMs. First, there are measurement errors in specifying the initial conditions [2] and the forcings that are used to produce the transient simulation. Second, the

[2] An example of this kind of issue is discussed in Hansen *et al*. (1997). They found that the climate system was in initial disequilibrium equivalent to a net forcing of 0.65 Wm^{-2}. If this initial disequilibrium was not modeled the

first stage of our methodology (described below) uses a regression equation that compares the GCM predictions and the true (but unobserved) temperature record by regressing the former on observed temperature. In this framework, differences between the GCM reconstruction and true temperature due to inaccuracies in the model can be interpreted as measurement errors even though the temperature data generated by the GCM are not a physical measurement of temperature.

To test whether errors in variables bias the ordinary least squares estimates of the regression parameters, Hausman (1978) develops the following statistic (H_M):

$$H_M = (\beta_{OLS} - \beta_{IV})(V_{IV} - V_{OLS})^{-1}(\beta_{OLS} - \beta_{IV})' \qquad (3)$$

which is based on the vectors of regression coefficients, β, and the regression coefficient variance matrices, V, from an ordinary least square regression (OLS) and an instrumental variables (IV) regression.[3] The OLS regression is inconsistent if measurement error is present while the IV regression is not. Therefore, the null hypothesis is that the OLS and IV estimates are equal and this null will be rejected if errors in variables bias the OLS estimates and lead to a significant H_M test statistic.. This test statistic is distributed asymptotically as $\chi^2(k)$ where k is the number of regression parameters. As described by Hausman (1978), the covariance matrix $V_{IV} - V_{OLS}$ is asymptotically positive definite. To guarantee positive definiteness V_{OLS} is computed as: $\sigma^2_{OLS}(X'X)^{-1}$ (Greene, 1993), where X is the matrix of regressors.[4]

If the H_m statistic indicates that the OLS estimate is biased, we can obtain an unbiased estimate of the relation between observed temperature and the temperature simulated by the

---

simulation would not have generated any warming over the period 1979-1995. The simulations we are analyzing here assume initial equilibrium, which may not be the case and this can affect the predicted climate change.

[3] See the next paragraph for an explanation of instrumental variables regression.

[4] In practice we compute heteroskedasticity and autocorrelation consistent estimates of the covariance matrices using the Newey and West (1987) procedure.

GCM from the instrumental variables regression. In this procedure the variable subject to "measurement error" is regressed on variables that are uncorrelated with the measurement error but highly correlated with the variable itself. The fitted values from this regression are then used instead of the original variable in a second stage regression of the two variables.

3. Methodology

3.1 Data

We evaluate eight reconstructions of annual global temperature (both land and sea surface) that are generated by three GCMs. The experiments include the EOGG, EOGS, and EOGI experiments simulated by the ECHAM model at the Max Plank Institute (Roeckner *et al,* 1998), the GS1, GS2, GS3, and GS4 experiments simulated by the Hadley Centre model (Tett *et al*., 1999), and the GS experiment simulated by the GFDL model (Delworth *et al.,* 2002). These are the same eight experiments analyzed by Allen et al. (2000) to quantify the uncertainty associated with the effect of human activity on temperature.

Observed temperature is an average of land and sea surface temperature (Parker et al., 1994; Nicholls et al., 1996). Radiative forcing is calculated from data for the atmospheric concentration of CO_2, CH_4, CFC's, N_2O, anthropogenic sulfur emissions, and solar activity. The sources of these data and the formulae used to calculate their radiative forcing are described in Stern and Kaufmann (2000).

3.2. Regression of GCM Temperature Reconstructions on Observed Temperature

To test how accurately the GCM simulates observed temperature, we start with the assumption that over the long run, the GCM temperature reconstruction, $\hat{T}_t$, is related linearly to the true forcing response, W_t with the addition of a mean zero noise process, η_t:

$$\hat{T}_t = a + bW_t + \eta_t \quad (4)$$

in which a and b are assumed to be fixed parameters. If the GCM reconstruction tracks the true long run forcing response accurately, the parameter b should equal one and the noise component should be white noise (we examine the short run dynamics of this relation in equation 12). We test these two as follows.

Observed temperature (T) is assumed to follow a similar model, except that we assume *a priori* that the coefficient associated with the true forcing response is unity:

$$T_t = a + W_t + \upsilon_t \quad (5)$$

where a is a fixed parameter and υ_t is a random error. There is no *a priori* reason to assume that observed temperature does not track the true forcing response in this way. We use observed temperature as a proxy for the unobserved forcing response (W) in equation (4) and estimate the following regression using ordinary least squares (OLS):

$$\hat{T}_t = \alpha + \beta T_t + e_t \quad (6)$$

We test whether the GCM simulation of global surface temperature tracks observed temperature accurately by testing the null hypothesis $\beta = 1$ with a t-statistic that is calculated as follows:

$$t = \frac{\beta - 1}{se_b} \quad (7)$$

in which se_b is the standard error for of β in equation (6). The standard error is computed using a method developed by Newey and West (1987) that is consistent in the face of heteroskedasticity

and stationary autocorrelation of an unknown form. However, the t-statistic still cannot be relied on if the residual from equation (6) contains a stochastic trend. Rejecting the null hypothesis $\beta = 1$ would indicate that the changes in temperature reconstructed by the GCM either overstate the rate of temperature change ($\beta > 1$) or understate the rate of temperature change ($\beta < 1$). A possible cause for such errors would be an incorrect climate sensitivity.

3.3 Errors in Variables Analysis

The effect of errors in measurement can be demonstrated by solving (5) for the true forcing response and substituting into (4) yielding:

$$\hat{T}_t = a - b\alpha + bT_t - b\upsilon_t + \eta_t \qquad (8)$$

Equation (8) indicates that observed temperature is correlated with the error term in equation (6). This correlation will bias downward the estimate of the regression parameter (β) in equation (6). The (presumably) larger tracking error in the GCM reconstruction, η_t, reduces the efficiency of estimation, but does not add to the bias in the estimate of β.To test whether errors in variables bias the OLS estimate of equation (6), we compute and evaluate the H_M statistic described in section 2.

Next, we use an instrumental variable approach to generate an unbiased estimate for b. As instruments, we use the radiative forcing time series that we use in our previous studies (Kaufmann and Stern, 1997, 2002; Stern and Kaufmann, 1999, 2000). These analyses indicate that the radiative forcing data are strongly correlated with temperature. Though the radiative forcing data also contain measurement error, these errors should not be correlated with the measurement errors in observed temperature because of the differences in the way that the series are compiled. Specifically the instrumental variables used are: radiative forcing due to

greenhouse gases (carbon dioxide, methane, CFCs, and N2O), solar irradiance, anthropogenic sulfur emissions, and stratospheric sulfate levels. The estimate of b is obtained by first regressing observed temperature on a group of instrumental variables and then regressing the temperature reconstructed by the GCM on the fitted values from this first stage regression.

$$\hat{T}_t = \alpha + \beta_{IV}\tilde{T}_t + e_t \qquad (9)$$

Equation (9) is the instrumental variables regression and is identical to (6) except that in place of observed temperature we use $\tilde{T}_t$ which is the fitted series from the first stage regression. The results from equation (9) are used to compute the t-statistic described in section 3.2.[5] Assuming that the instruments are valid [6] this statistic should be unbiased unless the error, e_t, from equation (9) contains a stochastic trend - a question we will address in the next section. 2.

3.4. The Nature of GCM Errors

To test for serial correlation in the GCM errors, i.e. white noise, we use the Q statistic (Box and Pierce, 1970), which is calculated as:

[5] The coefficient variance matrix for an instrumental variable regression would computed as $VAR(\beta_{IV}) = \sigma^2\left(\tilde{T}'\tilde{T}\right)^{-1}$ where σ^2 is computed as $\sigma^2 = \frac{1}{T}\left(\hat{T} - T\beta_{IV}\right)'\left(\hat{T} - T\beta_{IV}\right)$. In practice we use the non-parametric method of Newey and West (1987) as implemented by the RATS econometrics software package.

[6] ; The instruments would be invalid if they are either not orthogonal to the error term (which would occur if radiative forcing and observed temperature share the same measurement error) or the instrumental variables are not strongly correlated with observed temperature in which case they are termed “weak instruments”. Neither case seems likely here for the reasons stated above. We also did some informal tests based on the ideas of *Hahn and Hausman* (2002), which did not suggest any problems.

$$Q_K = N\sum_{k=1}^{K}\left[\frac{\sum_{t=1}^{N-k}(e_t-\bar{e})(e_{t+k}-\bar{e})}{\sum_{t=1}^{N}(e_t-\bar{e})^2}\right]^2 \quad (10)$$

in which e are the residuals from equation (9), and $\bar{e}$ is their mean. The null hypothesis is that there is no serial correlation (the residuals are white noise). The Q statistic is can be evaluated against the $\chi^2(k)$ distribution where k is the order of serial correlation being considered. We only test residuals from the instrumental variables estimates because the OLS regression is biased for each of the eight experiments (see section 4.2).

To test whether the GCM error contains a stochastic trend, we use the ADF statistic (equation 2). Rejecting the null hypothesis of a unit root would indicate that the GCM error term is red noise (assuming that we reject the null hypothesis that the error term is white noise). Based on the statistical rule of thumb that the maximum number of lags should equal the cubic root of the number of time series observations, a maximum of five augmenting lags is used in the ADF test. The optimal lag length is chosen using the adjusted R squared.

As explained above, if the error term is white noise and if the value for β estimated by the instrumental variables regression is not significantly different from one, we would consider the GCM to be a statistically adequate representation of the data. Conversely, the worst case scenario is one in which the error term contains a stochastic trend and the estimate of β from the instrumental variables regression is significantly different from one. The presence of red noise in the residual is to be expected given uncertainty about initial conditions and internal variability in GCM simulations. Nonetheless, this red noise can reduce the usefulness of GCM's for attribution and quantifying uncertainty.

3.5. Comparative Explanatory Power

Our final set of tests investigates whether the GCM temperature reconstructions explain more of the variance in observed temperature than the radiative forcing time series that are used to simulate the GCM experiments. To do so, we estimate a single dynamic regression equation for observed temperature that includes radiative forcing and the GCM temperature reconstruction as explanatory variables and tests restrictions that eliminate individual variables. We specify a dynamic regression model for two reasons. A static linear regression (one in which all impacts are instantaneous – no lagged effects are allowed) is too restrictive to evaluate GCMs because the GCM tracking error is likely to be red noise. A static linear regression is too restrictive to evaluate the explanatory power of the radiative forcing aggregate because surface temperature does not adjust immediately to changes in radiative forcing and so lagged effects must be represented.

We estimate a dynamic regression equation known as an error correction model (ECM). The ECM allows temperature to adjust to shocks nonlinearly in the short-run with a finite long-run response determined by a long-run linear relationship[7]. The regression equation is given by:

$$\Delta T_t = \phi T_{t-1} - \alpha - \zeta \mathrm{Time} - \theta \hat{T}_{t-1} - \Gamma X_{t-1} - \Psi Z_{t-1}$$

$$+ \sum_{i=1}^{s} \lambda_i \Delta T_{t-i} + \sum_{i=0}^{s} \pi_i \Delta \hat{T}_{t-i} + \sum_{i=0}^{s} \gamma_i \Delta X_{t-i} + \sum_{i=0}^{s} \delta_i \Delta Z_{t-i} + \varepsilon_t \qquad (11)$$

in which X is an aggregate for the radiative forcings used to simulate the GCM (which contain a stochastic trend) and Z includes the radiative forcings (which contain a stochastic trend) not used in the GCM simulation. The long-run linear relation among the variables is given by:

$$T_t = -(\alpha/\phi) - (\zeta/\phi) Time - (\theta/\phi)\hat{T}_t - (\Gamma/\phi) X_t - (\Psi/\phi) Z_t \quad (12)$$

X and Z are defined to be consistent with the individual GCM simulations. The X vector can include the radiative forcing of greenhouse gases (CO_2, CH_4, CFC11, CFC12, and N_2O) and/or the direct and indirect radiative forcing associated with anthropogenic sulfur emissions. The Z vector includes components of anthropogenic sulfur emissions not included in the GCM simulation under consideration (all GCM simulations include the radiative forcing of greenhouse gases) and solar irradiance. For example, the EOGG experiment simulates the effect of greenhouse gases only (Table 1), therefore X in (11) includes the radiative forcing of greenhouse gases only when the EOGG experiment is analyzed. For the EOGG experiment, Z includes a time series for the sum of the direct and indirect forcing associated with anthropogenic sulfur emissions and a time series for solar irradiance. Separating the forcing variables used to simulate the GCM (X) from those not included (Z) ensures that the GCM experiments are evaluated fairly. That is, if the radiative forcing associated with anthropogenic sulfur emissions and solar activity affect surface temperature, this "missing" effect from the EOGG simulation will be included in the Z vector and will remain in equation (11) when we eliminate the radiative forcing of greenhouse gases, which makes up the X vector when we test the EOGG experiment's ability to simulate the temporal effect of greenhouse gases on surface temperature.

The number of lags (s) used to estimate equation (11) is chosen using the Akaike (1973) information criterion and the Sims (1980) likelihood ratio, which are given by equations (13) and (14) respectively:

$$AIC = N \log (RSS) + 2k \qquad (13)$$

$$SIMS = (N-c) (\log |\Sigma_r| - \log|\Sigma_u| \qquad (14)$$

---

[7] See other applications of this methodology to climate change analysis by *Kaufmann and Stern* (*1997, 2002*) and *Kaufmann et al*, (*in review*).

where N is the number of observations, RSS is the residual sum of squares for the regression estimate of equation (11), k is the number of estimated parameters, c is the number of parameters estimated for the unrestricted (longer lag length) version of equation (11) and $|\Sigma_r|$ and $|\Sigma_u|$ are the determinant of the restricted (shorter lag length) and unrestricted variance/covariance matrices for equation (11), respectively. These criteria chose lag length by optimizing the trade-off between the parsimony and explanatory power of equation (11). To ensure a "fair" comparison among versions of equation (11) with different lag lengths, equation (11) is estimated over the same sample period, as determined by the longest period for which five lags are possible. When these tests indicate different lag lengths, we choose the shorter length. Choosing the longer lag length does not change any of the conclusions described in the next section.

For each of the eight GCM simulations, we test whether the GCM temperature reconstruction ($\hat{T}$) or the aggregate of GCM inputs (X) can be excluded from equation (11). We also test whether the lagged first differences of these variables have explanatory power about the short-run deviations from equilibrium temperature that is not contained in the lagged first differences of the other explanatory variables. Finally, we test these hypotheses jointly. Because the number of parameters associated with the GCM inputs (X) is the same as the number of parameters associated with the GCM temperature reconstruction ($\hat{T}$), the exclusion tests should not be affected by "overfitting."

Variables are excluded from equation (11) by restricting their regression coefficient(s) to zero. The significance of the restriction is evaluated with the test statistic ϕ:

$$\phi = \frac{(RSS_R - RSS_U)/s}{RSS_U/(N-k)} \qquad (15)$$

in which N is the number of observations, k is the number of regressors in the unrestricted version of equation (11), s is the number of coefficients restricted to zero, RSS_r is the residual sum of squares from the restricted version of equation (11), and RSS_u is the residual sum of squares from the unrestricted version of equation (11). The test statistic φ can be evaluated against the F distribution with s and (N-k) degrees of freedom in the numerator and denominator respectively under the assumption that the residuals in (12) are stationary (Toda and Phillips, 1993) which we test as described below. Values of φ that exceed the critical value indicate that the residual sum of squares for the restricted model increases in a manner that is statistically significant at the relevant level of significance relative to the residual sum of squares for the unrestricted model. In other words, restricting the regression coefficients associated with X or $\hat{T}$ increases the residual sum of squares (i.e. reduces the R-squared) in a statistically significant manner.

The restriction that excludes the temperature reconstruction ($\hat{T}$) from equation (11) tests the null hypothesis that the GCM temperature reconstruction has no explanatory power about observed temperature beyond that provided by the radiative forcing variables that are used to simulate the GCM (X), and the variables not included in the simulation (Z). Rejecting this null hypothesis would indicate that the GCM temperature reconstruction has explanatory power about global surface temperature beyond that of the GCM inputs and the forcings excluded from the simulation (Z).

The restriction that excludes the radiative forcing aggregate (X) tests the null hypothesis that GCM inputs have no explanatory power about observed temperature beyond that provided by the GCM temperature reconstruction ($\hat{T}$) and the variables excluded from the simulation (Z).

Rejecting this null hypothesis would indicate that the GCM inputs have explanatory power about historical temperature that the GCM fails to include.

Under certain circumstances, omitted variables can affect the conclusions that we draw from these tests by causing bias in the parameter estimates. Because the data for radiative forcing, observed temperature, and the GCM temperature reconstructions contain stochastic trends, only the omission of other variables with stochastic trends can seriously bias the estimates for equation (11). The possibility that we have omitted a relevant stochastically trending variable is evaluated by using the ADF statistic to test the residual from equation (12) (Engle and Granger, 1987). If this residual is stationary, it is unlikely that the omission of a relevant variable (including natural variability or systematic errors in the data) causes the parameter estimates to be biased.

Observed temperature also contains fluctuations caused by exogenous variables such as stratospheric sulfates associated with volcanic activity or associated with endogenous variability such as the El Nino-Southern Oscillation. These variables are stationary (Kaufmann and Stern, 2002; Kaufmann et al., in review) and, therefore, their omission should not significantly affect our resultsA referee suggested that there may be long-term cycles in volcanicity. However, Kaufmann and Stern's (2002) statistical tests indicate that such cyclicity is insufficient to lead to the acceptance of the null of nonstationarity.

4. Results

4.1. How well do the GCMs Track the Overall Temperature Increase?

The OLS point estimates of β vary between 0.13 and 0.71 and all are significantly less than 1.0 under the null that the error is stationary. The tendency for the GCM experiments to

underpredict the observed change in temperature implies that the OLS estimates for β may be biased by errors in measurement.

Confirming this suspicion, the H_M statistic indicates that there is a statistically significant difference between the OLS and IV estimate for the relation between GCM and observed temperature for six of the eight experiments (Table 2). Using instrumental variables raises the estimate of β relative to the OLS estimate. This increase is such that in five of the eight experiments, β is not statistically different from unity at the 5% level in a one sided test and not significant at the 5% level in a two sided test in seven out of eight cases. The EOGS, GS1, and particularly the EOGI experiments have the lowest estimates of β.

The goodness of fit as measured by the adjusted R-square varies across the experiments. Of those experiments where β is not significantly different from unity, the GFDL shows superior goodness of fit, while EOGG, GS2, GS3, and GS4 show weaker explanatory power for observed temperature.

4.2. The Nature of GCM Errors

Consistent with a visual appraisal (Figure 2a-h), the statistical results indicate that the residuals are red noise or contain a stochastic trend (Table 3). The Q statistics for serial correlation at one lag and 22 lags are significantly different from zero (Q statistics for intermediate levels of serial correlation also exceed the five percent level). ADF tests on the regression residuals reject the null hypothesis for experiments simulated by the Hadley and GFDL models. These results indicate that the residuals from these models are red noise and do not contain a stochastic trend. The presence of red noise is expected due to the presence of internal variability (e.g. ENSO events) and uncertainty about initial conditions for transient

simulations. For the EOGG, EOGS, and EOGI experiments simulated by the Max Planck model we cannot reject the hypothesis that the residuals contain a stochastic trend. The possible stochastic trend in the EOGI experiment is especially apparent (Figure 2(g)).

4.3. Explanatory Power of GCM's

Each of the eight fitted dynamic regression models has acceptable statistical properties with insignificant serial correlation in the model residuals (Table 4). The ADF statistic tests indicate the variables in the long-run relation cointegrate (12). In all but the GS1 experiment, we cannot reject restrictions that eliminate $\hat{T}$ and/or $\Delta\hat{T}$ from equation (12). These results indicate that the GCM temperature reconstruction does not add significantly to the explanatory power provided by the radiative forcing aggregate that is used to simulate the GCM (Table 4). Conversely, we strongly reject ($p<.01$) restrictions that eliminate X and/or ΔX for all eight experiments. This indicates that the radiative forcing variables used to simulate the GCM have explanatory information about observed surface temperature that is not present in the GCM simulation for global surface temperature.

We test the possibility that GCMs perform poorly relative to GCM inputs due to internal variability by replacing annual values in equation (11) with five or ten year averages, and repeating the analysis (To conserve degrees of freedom, we use only the current lag when using five year averages. When using 10 year averages, no lags are included). Despite the reduction in the degrees of freedom, the results are unchanged. The five or ten year averages of the GCM temperature reconstruction have little explanatory power about the five or ten year averages of observed temperature beyond that in the five or ten year averages for the radiative forcing of GCM inputs. Conversely, the five and ten year averages for the GCM inputs have considerable explanatory power about the five and ten year averages for observed temperature beyond that in the five or ten year averages of the temperature reconstructions generated by the GCM's. These results demonstrate that internal variability is not responsible for the results in table 4. We do not report these results in detail in order to conserve space.

5. Discussion

Our tests of how well the GCM temperature reconstructions track the global average temperature vary across the models and experiments investigated here. Together, these tests indicate which models and simulations perform better and which worse. Our analysis indicates that there is a tendency for some of the GCMs to underestimate the increase in global surface temperature in the period 1860-1990. The IV estimate for β is statistically different from unity at the 5% level in a two sided test for only one simulation, the EOGI experiment, though significant at lower levels for the EOGS and GS1 experiments. The residual process is red noise in the five experiments simulated by the GFDL and Hadley Centre models. The residuals from the three experiments simulated by the Max Planck model, the EOGG, EOGS, and EOGI experiments contain a random walk. Of the remaining experiments, the GFDL experiment has a better goodness of fit than the three remaining Hadley Centre experiments GS2, GS3, and GS4.

Based on these results, we conclude that the GFDL experiment clearly is superior. This experiment simulates the observed increase in temperature in an unbiased fashion, has a high goodness of fit, and a tracking error that is red noise but does not contain a stochastic trend. Conversely the Max Planck Institute ECHAM model has the poorest performance. The simulations generally are biased and there is no evidence that the errors are stationary. This poor performance may be caused by temperature sensitivities that are too small. As indicated in Table 9.1 of Cubasch et al, (2001), the temperature sensitivities of the MPI model generally are smaller than those reported for the HADCM and GFDL models. These low temperature sensitivities may cause the model to understate the rate of temperature change, which is consistent with values for β in equation (10) that are significantly less than unity. In addition, the difficulty

associated with the MPI model may be related to its (in)ability to simulate the effect of anthropogenic sulfur emissions, especially the indirect effect. The results are poorer when the effects of sulfur aerosols are simulated (the EOGS experiment). The EOGI experiment, which includes the indirect effects of the sulfur aerosol has the poorest performance of the three. Finally, the Hadley Centre model has an intermediate performance. Three out of the four of the experiments are satisfactory in terms of tracking performance and none of regression residuals contain a stochastic trend.

As described in section 4.3, none of the GCM's have explanatory power for observed temperature additional to that provided by the radiative forcing variables that are used to simulate the GCM. This result is to be expected if nonlinearities are not important during the sample period even if the GCM simulates the climate system accurately. Under these conditions, a GCM simulation is largely a dynamic linear function of the forcing with the addition of noise. This interpretation is consistent with results that indicate the correlation between GCM simulations and radiative forcing are higher than between the GCM simulations and observed temperature. For the GFDL model these are: 0.86 and 0.68 respectively where radiative forcing includes the variables used in the instrumental regression above. On the other hand, radiative forcing has a 0.80 correlation with observed temperature. In other words, radiative forcing is highly correlated with both the GCM and observed temperature but the GCM simulation is less correlated with observed temperature presumably due to the presence of different noise realizations in the latter two series. This is not a shortcoming of the GCM itself as a model of the climate system. But it does mean that simple linear models relating temperature to the forcing may be more accurate and efficient means for attributing climate change and predicting short-run

climate change when non-linearities are not important. Indeed, the literature suggests that nonlinearities are not important in the historical period (Allen et al, 2000).

Conversely, the exclusion tests in section 4.3 indicate that the radiative forcings used to simulate the GCMs have explanatory power about surface temperature relative to the temperature data generated by the GCMs. The explanatory power of the model inputs relative to model outputs could be due to the presence of noise in the GCM simulations or the GCM being a poor model of the climate system. This could be the case for those simulations that failed to pass our other tests – showing overall underestimation of observed temperature change or stochastic trends in the residual. Nonetheless, we cannot make any precise determinations as to the cause for the explanatory power of the model inputs relative to the GCM output.

6. Conclusion

The results of this statistical analysis highlight the need to match the modeling approach to the information needed. One of the most important policy questions for climate scientists is the issue of attribution: demonstrating that human activity is responsible for the increase in temperature over the last 140 years. Conclusions about the effect of human activity on surface temperature are based in large part on comparisons of observed temperature and GCM simulations (Mitchell et al., 2001). But this may not be most effective means for attribution: the noise in even the best simulation (in this case the GFDL simulation) increases the uncertainty involved in attributing and predicting climate change. This uncertainty could be reduced by using appropriately specified and estimated statistical models to simulate the relation between human activity and observed temperature (Kaufmann and Stern, 1997, 2002; Stern and Kaufmann, 2000; Stern, 2004). Initial results with a new generation of time series models that

are constrained by recently available observations on ocean heat content and energy balance considerations are very promising, yielding realistic rates of adjustment of atmospheric temperature to long-run equilibrium (Stern, 2004). On the other hand, GCMs provide a rich array of spatial variables that can help predict local changes in climate, which is a task for which time series models are not suited. GCMs also may be most effective in long-range forecasts and predictions, when significant nonlinearities may be more important.

Acknowledgment We thank Myles Allen, Reiner Schnur, and Tom Delworth for sending us the data. We thank Mark Friedl, J. Goldfarb, L. Danny Harvey, Stephen Hall, Gabi Hegerl, Guido Salvucci, Richard Smith, John Taylor, and Peter Young, and three anonymous reviewers for their comments on preliminary versions of this manuscript. We are responsible for any errors that remain.

References

Akaike, H, 1973. Information theory and an extension of the maximum likelihood principle, in B. N. Petrov and F. Csaki (eds.) *2nd International Symposium on Information Theory*, Akademini Kiado, Budapest, 267-281.

Allen, M.R., P.A. Stott, J.F.B. Mitchell, R. Schnur, and T.L. Delworth, 2000. Quantifying the uncertainty in forecasts of anthropogenic climate change, *Nature*, 407, 617-620.

Box G. and D. Pierce, 1970. Distribution of residual autocorrelations in autoregressive integrated moving average time series models, *Journal of the American Statistical Association*, 65, 1509-1526.

Craig, S. K. Holmen, and A. Bjorkstrom, 1997. Net terrestrial carbon exchange from mass balance calculations: an uncertainty estimate. *Tellus,* 49B136-148.

Cubasch, U. et al, , 2001. Projections of future climate change, In *Climate Change 2001: The Scientific Basis*. Houghton *et al., (eds.) Cambridge University Press. New York*.

Delworth, T. L., R. J. Stouffer, K. W. Dixon, M. J. Spelman, T. R. Knutson, A. J. Broccoli, P. J. Kushner, and R. T. Wetherald, 2002. Review of simulations of climate variability and change with the GFDL R30 coupled climate model. *Climate Dynamics*, *19*(7), 555-574.

Dickey, D. A. and Fuller, W. A., 1979. Distribution of the estimators for auto regressive time series with a unit root. *Journal of American Statistics Association*, 7, 427-431.

Engle, R. F. and C. W. J. Granger, 1987. Co-integration and error correction: representation, estimation, and testing, *Econometrica*, 55, 251-276.

Gilman, D.L, F.J. Fuglister, and J.M. Mitchell, Jr. 1963. On the power spectrum of "red noise," *Journal of the Atmopsheric Sciences* 20:182-184.

Govindan, R. B., D. Vyushin, A. Bunde, S. Brenner, S. Havlin,. H.-J. Schellnhuber, 2002. Global climate models violate scaling of the observed atmospheric variability, *Physical Review Letters* 89, 28501.

Granger, C. W. J. and P. Newbold, 1974. Spurious regressions in econometrics, *Journal of Econometrics*, 2, 111-120.

Greene, W.H. 1993. *Econometric Analysis* 2nd Edition Macmillan, New York.

Hahn, J. and J. Hausman, 2002. A new specification test for the validity of instrumental variables, *Econometrica*, 70, 163-189.

Hansen, J. et al., 1997. Forcings and chaos in interannual to decadal climate change. *Journal of Geophysical Research*, 102(D22) 25679-25720.

Harvey, A. C., 1989. *Forecasting, Structural Time Series Models, and the Kalman Filter*, Cambridge University Press, Cambridge.

Hausman, J. 1978. Specification tests in econometrics, *Econometrica* 76, 1251-1271,.

Hausman, J. 2001. Mismeasured variables in econometric analysis: problems from the right and problems from the left, *Journal of Economic Perspectives* 54(4), 57-67.

Jones, P. D., 1994. Estimating sampling errors in large-scale temperature averages, *Journal of Climate*, 7, 1794-1802.

Kaufmann, R.K. H. Kauppi and J.H. Stock. In review. Emissions, concentrations, and temperature: a time series analysis. *Journal of Geophysical Research.*

Kaufmann, R. K. and D. I. Stern, 1997. Evidence for human influence on climate from hemispheric temperature relations, *Nature*, 388, 39-44.

Kaufmann, R.K. and D. I. Stern, 2002. Cointegration analysis of hemispheric temperature relations, *Journal of Geophysical Research*, 107(D2), Article No. 4012.

Mitchell, J.F.B et al, 2001. Detection of climate change and attribution of causes, In *Climate Change 2001, The Scientific Basis: The Scientific Basis.* Houghton *et al., (eds.)* Cambridge University Press. *New York.*

Newey, W. K and K. D, West, 1987. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, *Econometrica*, 55(3), 703-708.

Nicholls, N. *et al.* 1996. Observed climate variability and change. in *Climate Change 1995: The Science of Climate Change* (eds. Houghton *et al*.) 138-192, Cambridge Univ. Press.

Parker, D.E., P.D. Jones, C.K. Folland, and A. Bevan, 1994. Interdecadal changes of surface temperatures since the late 19th century, *Journal of Geophysical Research*, 99, 14373-14399,.

Phillips, P. C. B., 1986. Understanding spurious regressions in econometrics, *Journal of Econometrics*, 33, 311-340.

Ramaswamy, V *et al.*, 2001: Radiative Forcing of Climate Change. In: Climate Change 2001 The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change [Houghton, J.T., Y. Ding, D.J. Griggs, M. Noguer, P.J. van der Linden, X. Dai, K. Maskell, and C.A. Johnson (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 881pp.

Roeckner, E., L. Bengtsson, J. Feichter, J. Lelieveld, H. Rodhe, 1998. Transient climate change simulations with a coupled atmosphere-ocean GCM including the tropospheric sulfur cycle. *Journal of Climate*, *12*, 3004-3032.

Shackley S., P. Young, S. Parkinson, and B. Wynne, 1998. Uncertainty, complexity and concepts of good science in climate change modelling: are GCMs the best tools, *Climatic Change* 38, 159-205.

Sims, C., 1980. Macroeconomics and reality, *Econometrica*, 48, 1-49.

Stern D. I., 2004. A multicointegration model of global climate change, Rensselaer Working Papers in Economics 0406, in review *Computational Statistics and Data Analysis*.

Stern D. I. and R. K. Kaufmann, 1999. Econometric analysis of global climate change, *Environmental Modelling and Software* 14, 597-605.

Stern, D. I. and R. K. Kaufmann, 2000. Detecting a global warming signal in hemispheric temperature series: a structural time series analysis, *Climatic Change*, 47, 411-438.

Stock, J.H. 1994. Unit roots, structural breaks, and trend, in *Handbook of Econometrics* R.F Engle and D.L. McFadden (eds.) Elsevier.

Tett, S.F. B., P. A. Stott, M.R. Allen, W.J. Ingram, J.F.B. Mitchell, 1999. Causes of twentieth century temperature change near the Earth's surface, *Nature*, *399*, 569-572.

Toda, H.Y. and P.C.B. Phillips. 1993. The spurious effects of unit roots on exogeneity tests in vector autoregressions: an analytic study *Journal of Econometrics* 59,229-255.

Trenberth, K. E., J. R. Christy, and J. W. Hurrel, 1992. Monitoring global monthly mean surface temperatures, *Journal of Climate*, 5, 1405-1423.

Figure Captions

Figure 1 Observed values for global surface temperature (heavy solid line) and global surface temperature as simulated by the EOGI experiment (dotted line), the GS3 experiment (dashed line), and the GFDL model (dashed-dotted line).

Figure 2 (a) The error from the GS1 experiment (given by *e* in equation (6) **(b)** As in (a) but for the GS2 experiment **(c)** As in (a) but for the GS3 experiment. **(d)** As in (a) but for the GS4 experiment **(e)** As in (a) but for the EOGG experiment **(f)** As in (a) but for the EOGS experiment **(g)** As in (a) but for the EOGI experiment **(h)** As in (a) but for the GFDL experiment

Figure 1

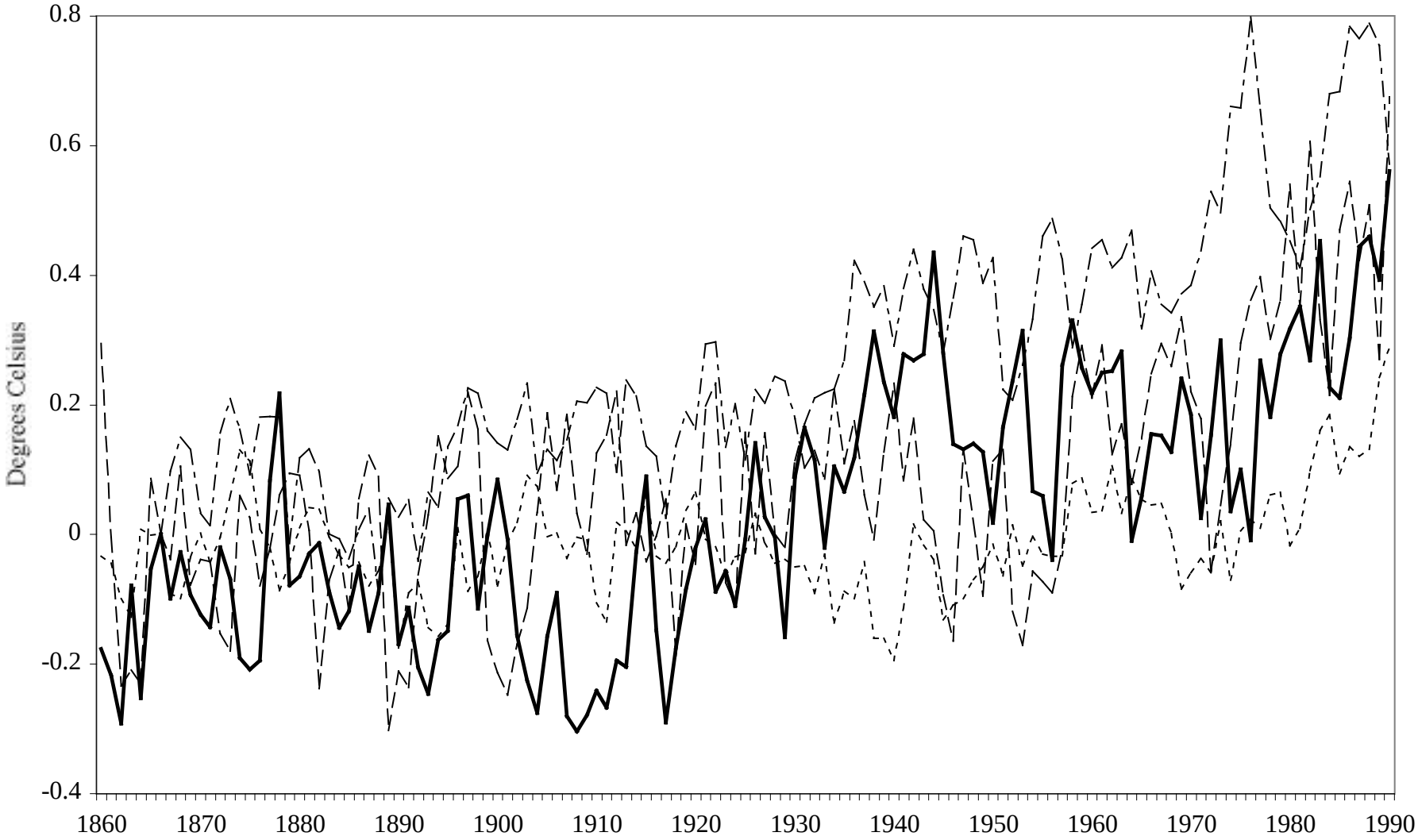
0.8
0.6
0.4
0.2
0
-0.2
-0.4
Degrees Celsius
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990

Figure 2(a)

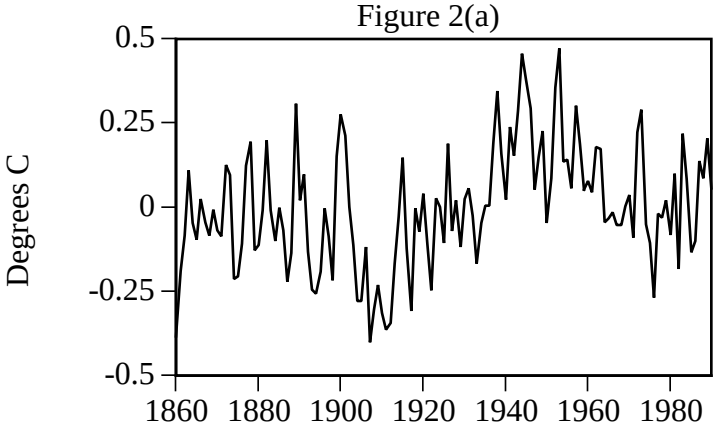

Figure 2(b)

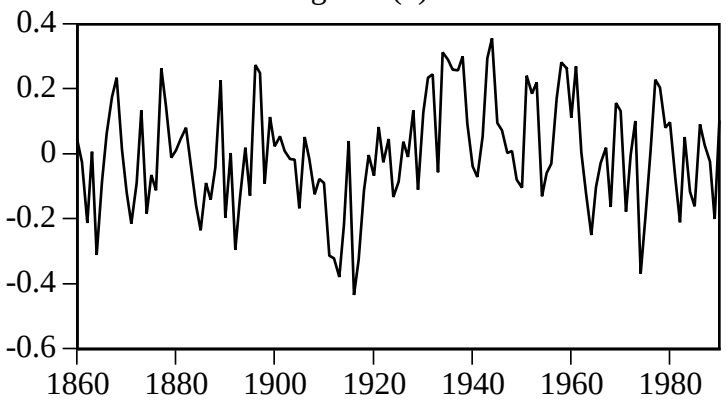

Figure 2(c)

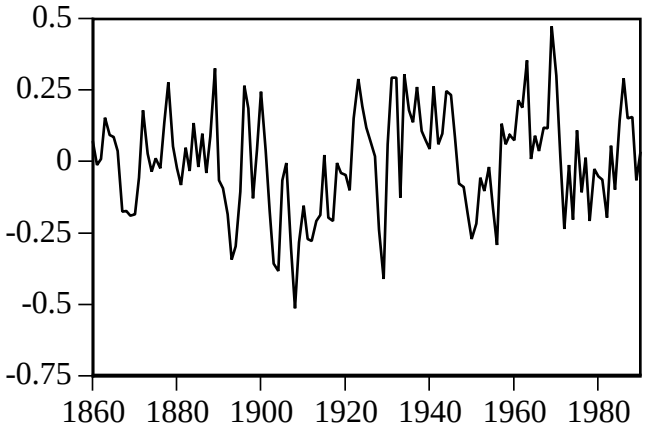

Figure 2(d)

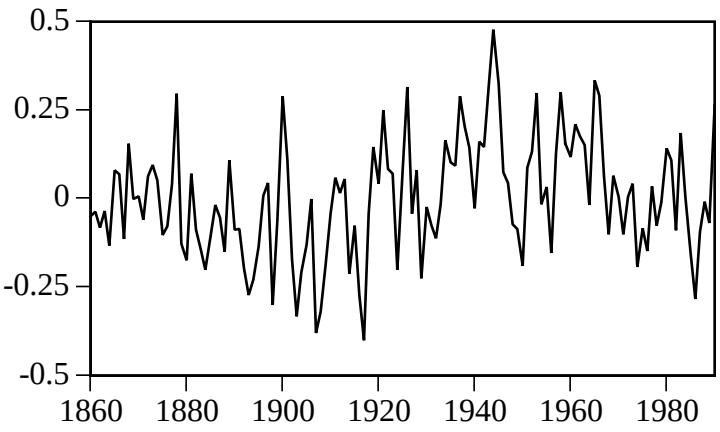

Figure 2(e)

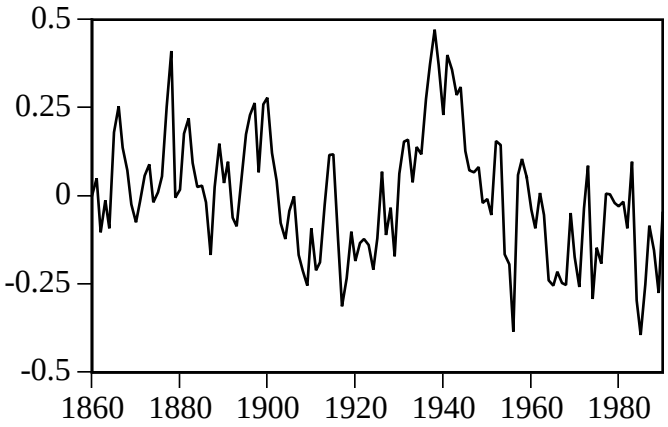

Figure 2(f)

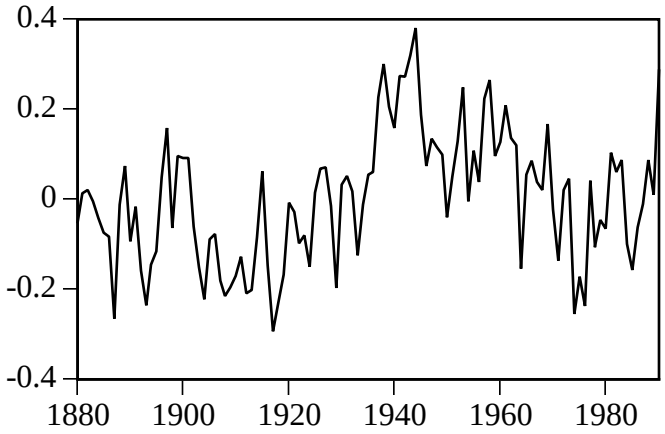

Figure 2(g)

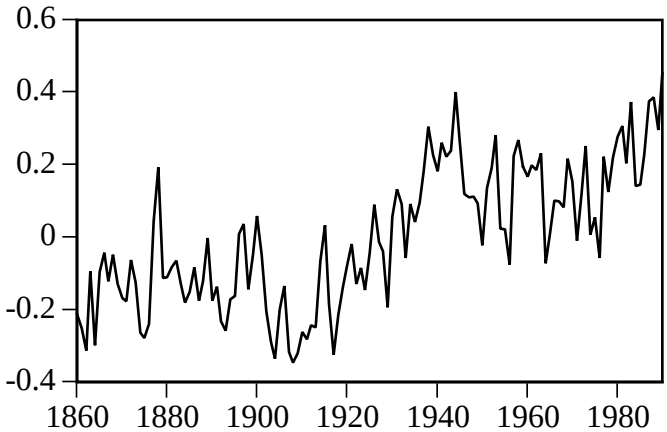

Figure 2(h)

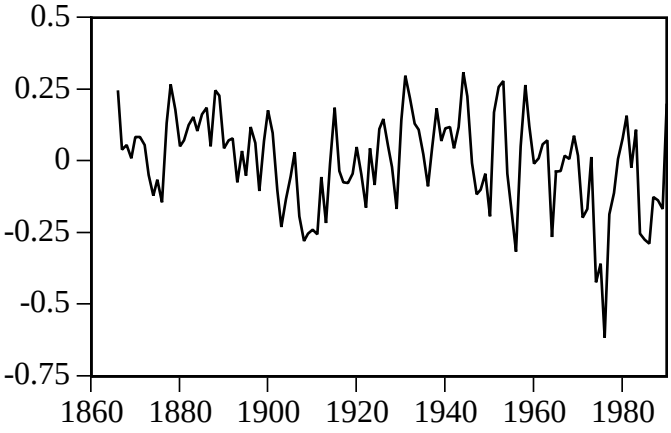

Table 1. Variables used to simulate GCM experiments

Experiment	Greenhouse Gases	Direct Sulfate Forcing	Indirect Sulfate Forcing
EOGG	Yes	No	No
EOGS	Yes	Yes	No
EOGI	Yes	Yes	Yes
GS1	Yes	Yes	No
GS2	Yes	Yes	No
GS3	Yes	Yes	No
GS4	Yes	Yes	No
GFDL:GS	Yes	Yes	Yes

Table 2. Analysis of Overall Tracking Performance

	OLS Results		**Instrumental Variables Results**		**Test for Measurement Error**	**Goodness of Fit**
Experiment	β	t $\beta = 1$	β	t $\beta = 1$	H_M	Adjusted R-square
EOGG	0.70	-3.39**	1.11	1.06	6.82**	0.2869
EOGS	0.56	-6.11**	0.85	-1.75+	8.48*	0.3490
EOGI	0.13	-15.29**	0.22	-11.63**	0.07	0.0357
GS1	0.48	-5.28**	0.78	-1.94+	1.67	0.1406
GS2	0.62	-4.18**	0.91	-0.87	5.45+	0.2901
GS3	0.62	-3.86**	0.90	-0.90	6.24*	0.2477
GS4	0.59	-4.22**	0.87	-1.26	7.89*	0.2729
GFDL	0.71	-3.75**	1.05	0.41	21.96**	0.3530

For the OLS and IV results the first column in each case is the estimate of β. All standard errors were computed with the Newey-West procedure. The final column in each block is the t test of $\beta = 1$. Adjusted R-square is from the instrumental variables regression. Test statistics reject the null hypothesis at the: **1%, *5%, +10% level.

Table 3. Analysis of Residuals

	Statistic		
Experiment	Q(1)	Q(22)	ADF
EOGG	56.70**	179.85**	-2.95†
EOGS	45.39**	167.89**	-2.89†
EOGI	57.56**	218.54**	-2.14†
GS1	32.82**	93.66**	-6.77**
GS2	22.22**	49.14**	-7.36**
GS3	37.67**	81.22**	-4.14*
GS4	25.22**	45.71**	-7.07**
GFDL	36.31**	76.43**	-6.07**

Test statistics reject the null hypothesis at the: **1%, *5%, +10% level. Signifiance levels for the ADF test statistic are calculated from *Mackinnon* [*1994*]. †The ADF test on the first difference of the residual strongly ($p < .001$) rejects the null hypothesis that it contains a stochastic trend in every case.

Table 4. Dynamic Regression Model and Exclusion Tests

	Hadley Centre Model			
	GS1 $r^2 = 0.523$ Q(1) = 0.33 ADF[2] = **-8.06**	GS2 $r^2 = 0.489$ Q(1) = 0.10 ADF[1] = **-8.77**	GS3 $r^2 = 0.515$ Q(1) = 0.09 ADF[1] = **-8.82**	GS4 $r^2 = 0.518$ Q(1) = 0.16 ADF[2] = **-7.84**
Variables (parameters) excluded:				
$\hat{T}$ (θ)	φ(1, 99) 5.17[#] [0.501]	φ(1,99) 0.11 [0.498]	φ(1, 99) 3.25 [0.500]	φ(1, 99) 1.42 [0.511]
X (Γ)	φ(1, 99) 8.14[*] [0.487]	φ(1, 99) 18.81[*] [0.392]	φ(1, 99) 23.41[*] [0.401]	φ(1, 99) 18.90[*] [0.426]
$\Delta\hat{T}$ (π)	φ(4, 99) 0.97 [0.506]	φ(4, 99) 0.13 [0.486]	φ(4, 99) 1.51 [0.486]	φ(4, 99) 1.64 [0.486]
ΔX (γ)	φ(4, 99) 4.07 [0.448][*]	φ(4, 99) 4.41[*] [0.413]	φ(4, 99) 5.24[*] [0.413]	φ(4, 99) 5.00[*] [0.421]
$\hat{T}$, $\Delta\hat{T}$ (θ, π)	φ(5, 99) 1.66 [0.486]	φ(5, 99) 0.11 [0. 486]	φ(5, 99) 1.21 [0.486]	φ(5, 99) 1.34 [0.486]
X ΔX (Γ, γ)	φ(5, 99) 4.02[*] [0.429]	φ(5, 99) 6.16[*] [0. 340]	φ(5, 99) 7.16[*] [0.340]	φ(5, 99) 6.33[*] [0.364]

	ECHAM (Max Planck)			GFDL
	EOGG $r^2 = 0.517$ Q(1) = 0.01 ADF[1] = **-8.62**	EOGS $r^2 = 0.517$ Q(1) = 0.03 ADF[2] = **-7.96**	EOGI $r^2 = 0.502$ Q(1) = 0.14 ADF[1] = **-8.77**	GS $r^2 = 0.512$ Q(1) = 0.001 ADF[1] = **-8.53**
Variables (parameters) excluded:				
$\hat{T}$ (θ)	φ(1,109) 1.97 [0.507]	φ(1,99) 0.26 [0.515]	φ(1,104)2.10 [0494]	φ(1,98) 0.05 [0.502]
X (Γ)	φ(1,109) 19.6[*] [0.421]	φ(1, 99) 11.70[*] [0.459]	φ(1,104) 16.82[*] [0.397]	φ(1,98) 11.0[*] [0.433]
$\Delta\hat{T}$ (π)	φ(4,109) 0.47 [0.507]	φ(4, 99) 1.45 [0.488]	φ(4,104) 0.13 [0.494	φ(4,98) 0.30 [0.495]
ΔX (γ)	φ(4,109) 3.83[*] [0.442]	φ(4, 99) 3.97[*] [0.437]	φ(4,104) 6.65[*] [0.404]	φ(4,98) 5.07[*] [0.428]
$\hat{T}$, $\Delta\hat{T}$ (θ, π)	φ(5,109) 1.27 [0486]	φ(5, 99) 1.25 [0.486]	φ(5,104) 0.58 [0.486]	φ(5,98) 0.27 [0.490]
X ΔX (Γ, γ)	φ(5,109) 5.85[*] [0.374]	φ(5, 99) 4.87[*] [0.397]	φ(5,104) 7.02[*] [0.334]	φ(5,98) 5.3[*] [0.366]

Each column reports the results for (13) for a different GCM model. For all models, three lag were used to estimate the model (equations 14 & 15). The sample period for all models is 1866-1900, except for the GFDL, which is 1872-1990. Values for r^2 and Q are calculated from unrestricted version of equation 13. Q is a first order serial correlation statistic distributed as $\chi 2$ with one degree of freedom. Its critical value ($p = 0.05$) is 3.86. Dates refer to the sample period used to compare versions of equation (13) with 1-5 lags and used to estimate the restricted and unrestricted versions of equation (13). φ values for tests of these restrictions are given in the lower rows. Coefficients are statistically significantly different from zero at the: *1%, #5% level. Consistent with a standard F test, the degrees of freedom in the numerator (the first number in the parenthesis) represents the number of parameters restricted to zero. The degrees of freedom in the denominator (the second number in the parenthesis) represents the degrees of freedom in the unrestricted version of equation (14). The unadjusted r^2 for the restricted versions of equation (13) are given in brackets. ADF refers to the cointegration ADF statistic for equation (14) without a time trend with the number of augmenting lags used to compute the test statistic in brackets. ADF statistics that exceed the $p = .05$ threshold are in bold,.