EVALUATING EXPLANATORY MODELS OF THE SPATIAL PATTERN OF SURFACE CLIMATE TRENDS USING MODEL SELECTION AND BAYESIAN AVERAGING METHODS

Ross McKitrick
Department of Economics
University of Guelph
rmckitri@uoguelph.ca
Tel +1-519-824-4120 x52532
Fax +1-519-763-8497

Lise Tole
Department of Economics
Strathclyde University
Glasgow, UK

Citation: McKitrick, Ross R. and Lise Tole (2012) "Evaluating Explanatory Models of the Spatial Pattern of Surface Climate Trends using Model Selection and Bayesian Averaging Methods" *Climate Dynamics*, 2012, DOI: 10.1007/s00382-012-1418-9

Running Title: *Testing Models of Spatial Climate Trends*

Abstract

We evaluate three categories of variables for explaining the spatial pattern of warming and cooling trends over land: predictions of general circulation models (GCMs) in response to observed forcings; geographical factors like latitude and pressure; and socioeconomic influences on the land surface and data quality. Spatial autocorrelation (SAC) in the observed trend pattern is removed from the residuals by a well-specified explanatory model. Encompassing tests show that none of the three classes of variables account for the contributions of the other two, though 20 of 22 GCMs individually contribute either no significant explanatory power or yield a trend pattern negatively correlated with observations. Non-nested testing rejects the null hypothesis that socioeconomic variables have no explanatory power. We apply a Bayesian Model Averaging (BMA) method to search over all possible linear combinations of explanatory variables and generate posterior coefficient distributions robust to model selection. These results, confirmed by classical encompassing tests, indicate that the geographical variables plus three of the 22 GCMs and three socioeconomic variables provide all the explanatory power in the data set. We conclude that the most valid model of the spatial pattern of trends in land surface temperature records over 1979-2002 requires a combination of the processes represented in some GCMs and certain socioeconomic measures that capture data quality variations and changes to the land surface.

Key Words: GCM testing; spatial trend patterns; climate data contamination, spatial autocorrelation; non-nested tests; encompassing tests; Bayesian Model Averaging.

Evaluating Explanatory Models of the Spatial Pattern of Surface Climate Trends using Model Selection and Bayesian Averaging Methods

1 Introduction

1.1 Model Testing in Spatial and Temporal Dimensions

General Circulation Models (GCMs) are the basis for modern studies of the effects of greenhouse gases and projections of future global warming. Reliable trend projections at the regional level are essential for policy guidance, yet formal statistical testing of the ability of GCMs to simulate the spatial pattern of climatic trends has been very limited. This paper applies classical regression and Bayesian Model Averaging methods to test this aspect of GCM performance against rival explanatory variables that do not contain any GCM-generated information and can therefore serve as a benchmark.

In 20th century reproduction tests, the GCM takes as inputs the observed post-1900 changes in greenhouse gases, aerosols, solar output and so forth, and generates a predicted sequence of temperatures throughout the atmosphere. GCMs have mainly been tested on their ability to explain a univariate global or regional average temperature anomaly series (herein T_t) over the post-1900 interval (e.g. Knutson et al. 2006, CCSP 2008). But this does not appear to be all that informative as to the overall power of the model. Kaufmann and Stern (2004) analysed T_t using an error-correction model containing a GCM-generated prediction of T_t as well as a small number of GCM inputs, namely observations on greenhouse gas concentrations, solar irradiance, volcanic dust and atmospheric aerosols, rescaled to represent forcing units on temperature change processes. They could not reject the hypothesis that the GCM added no information to the forecast of T_t other than that inherited from the observed forcing series.

In addition, conflicting hypotheses can explain the 20th century evolution of T_t. Shukla et al. (2006) showed that models with greater propensity to warm in response to greenhouse gases tended to fit a globally-averaged temperature trend better, but Knutti (2008), CCSP (2008, p. 44), Knutti and Hegerl (2008), Kiehl (2007), Hegerl et al. (2007 p. 678), Schwartz et al. (2007) and others have all noted that the observed global trend can be consistent with stronger or weaker sensitivity to greenhouse gas-induced warming if paired with offsetting assumptions about aerosol-induced cooling, oceanic heat uptake or other mechanisms.

The evaluation of climate models in Chapter 8 of the Fourth Intergovernmental Panel on Climate Change (IPCC) report (Randall et al. 2007) primarily consists of static reproduction tests, that is, the ability to reproduce the distribution of mean temperature and precipitation levels, and diurnal temperature ranges, but not temperature trends, around the world; and *a priori* process checks, that is, whether certain known meteorological processes are coded into the models. The IPCC report notes (p. 594) that relatively few studies have looked at whether empirical fidelity between model simulations of historical periods and observations improves the accuracy of climate trend forecasts. Gleckler et al. (2008) note that the ability of a climate model to replicate a mean climate state has little correlation to measured fidelity on interannual trend measures, a point shown clearly in Jun et al. (2008).

GCMs differ in their handling of physical processes, parameterization of subgrid scale processes and assumptions about poorly-observed forcing effects, such as aerosols. The range of methodologies implies that they cannot all be “correct.” Knutti (2008) points out that testing fidelity on both the spatial and temporal scales is essential, yet Berk et al. (2001) noted that quantitative comparison of model outputs to observed data was rare and “relies very heavily on eyeball assessments” (Berk et al p. 126). Neither the Climate Change Science Program review of GCMs (CCSP 2008) nor the 2007 IPCC report provided quantitative tests of how well climate models reproduce the spatial pattern of temperature trends in recent decades, relying instead on “eyeball assessments.” Chapter 9 of the IPCC report (Hegerl et al. 2007) presents a diagram and accompanying discussion (Figure 9.6, pp. 684-686) of the averaged output from 58 GCM runs and the spatial pattern of temperature trends over land from 1979-2005, comparing model runs under the assumption that greenhouse gases do not warm the climate versus runs that assume they do. It is asserted that the latter assumption fits the data better, but no quantitative evidence is provided. CCSP (2008) presents a visual comparison of the fit between observed trend patterns over 1979-2003 and those generated by the GISS ModelE. Again the discussion is entirely qualitative—readers are not even given a correlation coefficient, much less a suite of significance tests.

CCSP (2008), referring to Covey et al. (2003), reports a 95-98% correlation between modeled and observed temperatures over space and time. These were ahistorical, no-forcing control runs, not a test of predictive capability. Instead Covey et al. compared the 12 monthly means from GCM control runs on a gridcell-by-gridcell basis to late 20th century monthly temperature means. The models did a good job reproducing the spatial pattern of the mean across grid cells, the amplitude of the seasonal cycle within each grid cell and the different seasonal amplitudes across grid cells. However the ability to predict trends in the gridded monthly or annual means over time in response to observed forcing changes was not tested.

Knutson et al. (2006) present a comparison of the spatial trend pattern between an ensemble average of model simulations for 1949—2000 and corresponding observations. In their Figure 5d (p. 1635) the differences are denoted as significant or not based on a *t*-test. They report that in 31%

of the locations, the t-statistic rejects the hypothesis that the modeled and observed trends are the same. However, failure to reject differences elsewhere does not imply much about actual explanatory power. Consider for example, a test for differences among two lists of Gaussian random numbers: failure to reject would not imply that one explains the other. Model evaluation against a specific alternative is more useful for assessing explanatory power, as we will discuss in the next section. Jun et al. (2008, Figure 7) contrast observed and model trends but the significance of the mismatches is not reported.

1.2 Testing Framework

The vector we seek to explain, denoted y, contains i = 1,…,428 trend terms y_i representing the least squares trends through the monthly temperature anomaly data series in the 428 available grid cells i in the land surface data set produced by the Climatic Research Unit (CRU, see Brohan et al. 2006). The data set is commonly denoted CRUTEM3v which we will shorten to CRU3v herein. Trends are computed over 1979-2002, an interval which was dictated by the availability of some key data (see section 2.1). There are three categories of explanatory variables. Fixed geographic observations (such as latitude and mean air pressure) are denoted by the matrix **G**. Measures of regional socioeconomic development (such as Gross Domestic Product (GDP) and population growth) that can be conjectured to induce climatic effects through local changes to the land surface are herein denoted by the matrix **X**. A vector of surface temperature trends projected by GCM j is denoted z^j and the collection of these vectors is the matrix **Z**. We are concerned herein with understanding the extent to which the individual z^j's, and all the possible linear combinations thereof up to the entire matrix **Z**, can explain y, in comparison to the elements of **X**, in tests in which the explanatory power of **G** are also taken into account.

We address three issues that arise in such testing. First, spatial autocorrelation is shown to affect the dependent variable but not the residuals of the fully-specified testing model, allowing use of familiar least-squares estimation. Second, while standard tests of zero restrictions allow rejection of nested alternatives, they do not help choose when rival models are considered mutually exclusive. To conduct more generalized tests among alternatives we use encompassing (Mizon 1984) and non-nested tests (Davidson and MacKinnon 1981, 2004). Third, we recognize that GCMs might have optimal performance only in unknown linear combinations, and selection of rival variables likewise may be arbitrary, so we apply a Bayesian Model Averaging (BMA) method that searches over all possible linear combinations of explanatory variables and generates posterior coefficient distributions robust to model uncertainty.

The matrix of geographic factors **G** includes presence of a major coastline, mean air pressure, latitude, and measures of the local influence of major climatic cycles such as El Nino and the Pacific Decadal Oscillation. The matrix of socioeconomic data **X** includes measures of population growth, income growth, energy consumption and other variables known to affect local air temperature trends and data quality. Note that the effects of these variables are supposed to have been removed during the data set construction process, an issue that we discuss in detail below. The matrix **Z** of GCM terms z_i^j, sometimes called the "guess pattern" (e.g. Hegerl et al. 1997) represents the prediction by climate model j of the i-th gridcell temperature trend over the same period in response to observed changes in the main climate forcing input series, chiefly greenhouse gases, solar irradiance, atmospheric aerosols and (in some cases) ozone depletion. Note that the GCM takes as inputs the observed changes in greenhouse gases, solar flux and other global climatic influences over a target time interval, then generates as outputs the projected pattern of warming and cooling trends in each location, which can be compared to the observed trends. Because we are analysing a spatial pattern of trend coefficients this is a cross-sectional, rather than a time-series, test.

Interest in the role of socioeconomic variables arises because of findings in de Laat and Maurellis (2004, 2006), McKitrick and Michaels (2004, 2007, herein MM07), and McKitrick and Nierenberg (2010), all of whom modeled the surface trend patterns in the data of Brohan et al. (2006) and similar data products, to test the null hypothesis of independence from measures of local socioeconomic development, such as population growth, urbanization, equipment changes, data quality problems in developing countries, variations in local air pollution levels, etc. The belief that these factors have been removed from climatic data is widely assumed and is integral to climatic research (see., e.g., Yun et. al. 2008, p. 935). MM07 regressed the observed 1979-2002 trends in surface grid cells on the same set of fixed geographical variables as are employed herein. They strongly rejected the independence of the surface temperature trends and socioeconomic variables, concluding that the climatic data are contaminated by the effects of industrialization on local temperature records. Schmidt (2009, herein S09) critiqued this interpretation on the grounds that a GCM with no representation of socioeconomic surface processes could generate a spatial pattern of trends that is also apparently correlated with socioeconomic measures. McKitrick and Nierenberg (2010) replied that, on average, GCMs predict a spatial pattern of correlations between temperature trends and socioeconomic variables opposite to the one observed in CRU data, and also that the regression model using GCM output exhibited spatial autocorrelation in the residuals which, if corrected, rendered the apparent correlations insignificant.

McKitrick and Nierenberg (2010) included GCMs in rather simplistic ways, as the mean of individual runs by the GISS-E model (Schmidt 2009) and as the mean of runs from 22 climate models used for IPCC (2007). The limitation of this approach as a test of climate model accuracy is that it requires arbitrary selection of the regressors, hence model uncertainty (in other words uncertainty over which variables to include in the model) is not taken into account. For instance, while one GCM might have weak explanatory power on its own, it may also lack a physical process (such as ozone depletion) covered by another model, so that a linear combination of the two models might perform well. In the same way it might or might not be appropriate to include linear combinations of the other variables. *A priori* theory is not decisive on this point, and arbitrary inclusion or exclusion can lead to contrasting risks of overfitting and omitted variable bias.

After introducing the data and methods in the next section we illustrate the model selection problem using nested and non-nested hypothesis tests of rival models of surface trends. While some GCMs clearly demonstrate significant explanatory power, these results are always conditional on model selection decisions, a limitation we then tackle using BMA to obtain confidence intervals and inclusion probabilities robust to model uncertainty. We show in section 3.4 that the results of the BMA analysis align well with those from classical methods in identifying a subset of climate models and socioeconomic variables necessary to yield a good model of the spatial pattern of temperature trends over land.

2 Data and Methods

2.1 Data Set

All data are taken from McKitrick and Nierenberg (2010) and McKitrick (2010). The data consist of 428 records, one for every 5x5 degree grid cell over land for which adequate observations were available in CRU3v to identify a trend over the 1979-2002 interval. The locations are plotted in Figure 1. The time interval was determined by the start of the satellite record (1979) and the terminus of the most recently available socioeconomic data during the preparation of McKitrick and Michaels (2007). Future work on this topic will involve use of the additional decade of data now available at each location. Each record contains the linear surface trend y_i expressed as degrees C per decade, supplemented with the variables listed in Tables 1 and 2, which were drawn from a variety of sources (see Appendix for source information and other details). Since we are analyzing the pattern of trends over space, temporal autocorrelation within gridcells does not need to be addressed. We do not make use of the variance of the trend terms, just the least squares trends themselves. If we were to estimate a model on the pooled cross-sectional time series panel data set, then we would need to address the temporal autocorrelation problem. Panel estimators and non-parametric variance estimators for climatic data sets are discussed in McKitrick et al. (2010). In future work the analysis of spatial performance of climate models will be examined in a pooled cross-sectional time series framework, but for now the temporal data are collapsed into trend terms so our data set is cross-sectional.

msu_i is the lower troposphere (LT) temperature trend in gridcell *i* taken as the mean of the UAH (Spencer and Christy 1990) and RSS (Mears et al. 2003) data products. slp_i is the mean sea level air pressure. dry_i is a dummy variable denoting when a grid cell is characterized by predominantly dry conditions (which is indicated by the mean dewpoint being below 0 °C). $dslp_i$ is $dry_i \times slp_i$. $water_i$ is a dummy variable indicating the grid cell contains a major coastline. $abslat_i$ denotes the absolute latitude of the grid cell. ao_i, nao_i, pdo_i and soi_i denote, respectively, the correlation over time in grid cell *i* between the grid cell temperature anomaly and a standard index of the Arctic Oscillation, the North Atlantic Oscillation, the Pacific Decadal Oscillation and the El Nino-Southern Oscillation Index. p_i, m, q_i and c_i are, respectively, percent changes over 1979 to 2002 in population, real per capita income, total Gross Domestic Product (GDP) and coal consumption. g_i is GDP density (national Gross Domestic Product per square kilometer) as of 1979, e_i is the average level of educational attainment as late in the interval as possible, and x_i is the number of missing months in the observed temperature series. z_i^j denotes the 1979-2002 surface trends projected by model *j*, where *j* = 1,…,22, the number of available models in the PCMDI archive (see Appendix).

2.2 Model Evaluation Methods

A simple test of the explanatory power of GCM *j* would involve estimating the regression

$$y = a + \gamma_j z^j + e \qquad (1)$$

where e is a vector of independent and identically distributed residuals. Rejection of $\hat{\gamma}_j \leq 0$ would imply significant explanatory power for model *j*. The sign restriction is imposed since a significant estimate $\hat{\gamma}_j < 0$ is not a validation of the GCM, but an extreme form of model failure. If no alternative is specified the test may not tell us much. Suppose that a fixed, exogenous factor like latitude were added to Equation (1). If $\hat{\gamma}_j$ fell to insignificance we would then conclude that the apparent explanatory power attributed to the GCM was due to it serving as a (rather expensive) proxy for latitude. So tests must control for plausible alternatives, namely a list of those variables for which the GCM is expected to provide explanatory power "over and above." If the GCM contains scientifically valid climatic information it ought to provide explanatory power over and above latitude, as well as other geographic and socioeconomic variables that might be considered.

Since there are 22 models in our sample, testing at 5% significance leaves open the possibility that even if all of them were devoid of explanatory power, nonetheless one might appear to achieve significance by chance. However most of our key results are obtained at very low significance levels so we are not overly concerned about Type I errors.

Identification of the contribution of GCMs relative to an alternative can be done using conventional nested regressions, as well as encompassing tests and non-nested tests. A nested test of explanatory power can take the form of the regression

$$y = \mathbf{G}\mu + \mathbf{Z}\gamma + \mathbf{X}\beta + e. \tag{2}$$

where **G** is augmented with a column of 1s. The conventional *F*-statistic on the restrictions $\hat{\gamma} = 0$ will test whether the GCM data contribute significant explanatory power over and above that contributed by **G** and **X**. The concept of *encompassing* (Mizon 1984) provides a deeper interpretation of the results of such a test. Ignoring the role of **G** for a moment, suppose there are two rival models M1 and M2:

$$\text{M1: } y = \mathbf{Z}\gamma + e_1. \tag{3a}$$

$$\text{M2: } y = \mathbf{X}\beta + e_2. \tag{3b}$$

M1 carries the strong implication that the elements of **X** have no explanatory power, and in the current context can be viewed as the maintained hypothesis that non-climatic factors relating to surface disruption have been removed from observed climatic data y; likewise M2 implies that **Z** has no explanatory power. Below we also consider a non-nested approach that offers less extreme contrast by forming a weighted pair of equations (3a) and (3b), but for now we consider the case in which a researcher wants to consider mutually exclusive alternatives.

Under the null hypothesis that (3a) is true, the ordinary least squares (OLS) estimator $\hat{\gamma}$ converges to the true value γ as the sample size grows. The principle of encompassing states that if equation (3a) is the true model, it ought to explain y in terms of **Z**, but should also explain the behavior of $\hat{\beta}$ in M2. Thus when we say that one model encompasses another, it not only provides explanatory power for the dependent variable but also accounts for any apparent explanatory power of the rival model. OLS Estimation of model (3b) yields $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$. If equation (3a) is true then the limiting behavior of $\hat{\beta}$ can be explained using $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\gamma$. Substituting in $\hat{\gamma}$ we obtain

$$\tilde{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'y = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{P_Z}y$$

where $\mathbf{P_Z}$ denotes the OLS projection matrix and $\mathbf{P_Z}$ y is therefore the vector of fitted values from a regression of y on $\mathbf{Z}$. If the differences between the estimators $(\hat{\beta} - \tilde{\beta})$ are small (that is, have a probability limit of zero), then M1 is said to *encompass* M2. The relationship between the two models is quite simple, so in this case, though not generally, the encompassing test turns out to be equivalent to the F test on $\hat{\beta} = 0$ from the regression

$$y = \mathbf{X}\beta + \mathbf{Z}\gamma + e$$

(Davidson and MacKinnon 2004, p. 672). Now suppose each model is augmented with $\mathbf{G}\mu$, yielding models M3 and M4:

$$\text{M3: } y = \mathbf{G}\mu + \mathbf{Z}\gamma + e_1 \quad (4a)$$

$$\text{M4: } y = \mathbf{G}\mu + \mathbf{X}\beta + e_2. \quad (4b)$$

Then the F test on $\hat{\beta} = 0$ from equation (2) tests whether the explanatory variables unique to M4, namely the columns of $\mathbf{X}$, are encompassed by M3. Likewise the F test on $\hat{\gamma} = 0$ from equation (2) would test whether M3 is encompassed by M4. In the implementation in Section 3 we introduce each GCM separately as well as collectively in $\mathbf{Z}$.

The encompassing approach helps clear up an ambiguity that can arise due to the possible relationship between $\mathbf{Z}$ and $\mathbf{G}$. Suppose the GCMs predict that the response to increased greenhouse gases is a spatial warming pattern exactly stratified by some or all of the elements of $\mathbf{G}$, such as latitude and proximity to a coastline, and the variables in **X** have no explanatory power. Then any apparent explanatory power of **G** might equally be interpretable as evidence for the models that generated the predictions in **Z**, and likewise an observation of insignificance of $\hat{\gamma}$ might be attributed to multicollinearity with **G**, rather than lack of explanatory power of the models. We can resolve this ambiguity by testing if M1 encompasses M4, or in other words whether any explanatory power of **G** and **X** together is attributable to **Z**. This can be done using the F test on whether $\hat{\beta}$ and $\hat{\mu}$ from equation (2) are jointly zero.

A related approach that does not involve either-or comparisons is called *non-nested* testing (Davidson and MacKinnon 2004). It involves taking equations (4a-b) and forming the weighted pair

$$y = (1-\alpha)(\mathbf{G}\mu + \mathbf{X}\beta) + \alpha(\mathbf{G}\mu + \mathbf{Z}\gamma) + e. \quad (5)$$

If equation (5) could be estimated directly a test of $\hat{\alpha} = 1$ would indicate if the GCM's have full explanatory power while the socioeconomic variables have none, and likewise a test of $\hat{\alpha} = 0$ would test if the socioeconomic variables have complete explanatory power. However α is not identifiable in equation (5). Davidson and MacKinnon (1981) showed that if $(\mathbf{G}\mu + \mathbf{Z}\gamma)$ in (5) is replaced with the predicted values from a regression of y on **G** and **Z** (denoted $\hat{y}_{GZ}$) then estimation of

$$y = \mathbf{G}\mu + \mathbf{X}\beta + \alpha\hat{y}_{GZ} + e \quad (6)$$

yields an estimate $\hat{\alpha}$ corresponding to α from equation (5), the *t*-statistic for which is asymptotically N(0,1) under the null hypothesis that $\alpha = 0$. This is called the Davidson-MacKinnon *J*-test. It is customary to do the test twice, switching the α positions to double check the result. Our interest is in whether the confidence interval around α includes 0 or 1, so we report tests of each null.

2.3 Spatial Autocorrelation

Estimation of models in the form of (1) yield residuals with strongly autocorrelated disturbances. Spatial autocorrelation in the residual vector can be modeled using

$$\mathbf{u} = \lambda \mathbf{W}\mathbf{u} + e \qquad (7)$$

Where λ is the autocorrelation coefficient, **W** is a symmetric matrix of weights that measure the influence of each location on the other, and e is a vector of homoskedastic Gaussian disturbances, (Pisati 2001). The rows of **W** are standardized to sum to one. A test of $H_0: \lambda = 0$ measures whether the error term is spatially independent. Anselin et al. (1996) point out that if the alternative model allows for possible spatial dependence of y then conventional tests of $\lambda = 0$ will be biased towards over-rejection of the null. Anselin et al. (1996) derive an LM test of $\lambda = 0$ robust to nonzero value of a nuisance parameter measuring the spatial lag, and that is the form of test we use herein. Hypothesis tests, and any subsequent parameter estimations, are conditional on the assumed form of the spatial weights matrix **W** in equation (7). We use elements $u_{ij}^{-\mu}$ denoting the great circle distance between grid cells i and j where μ determines the rate at which the relative influence of one cell on adjacent cells declines. Based on the maximum likelihood results in McKitrick and Nierenberg (2010) we used μ =2.3. Our tests (not reported) on regressions of the form (2) using individual GCM terms as well as the full matrix **Z** show that while the dependent variable is spatially autocorrelated the residuals are not, so in our estimations we do not need to apply the correction in equation (7).

2.4 Bayesian Model Averaging

Specification of the GCM component and the rival variables can be arbitrary due to the lack of theoretical guidance as to the correct model. For instance, there are 22 GCMs, but as mentioned above, the best choice may not be one particular model, but a linear combination. Since there are 2^{22} such combinations, it is computationally infeasible to test each one. Likewise, there are 16 geographical and socioeconomic variables, and cases can be made for or against inclusion of each one. Estimation of any one regression model entails a choice among a large number of potential alternatives, and this uncertainty needs to be accounted for in the variances of the coefficient estimates. Model averaging surmounts this problem by including information from every potential model. Results are a weighted average of estimates from every model, where the weights are proportional to the support each model gets from the data. Since it treats models (and parameters) as random variables, it is easier to implement model averaging in a Bayesian framework.

In this study, we implement Bayesian model averaging as outlined in Fernandez, Ley and Steel (2001) and we use an algorithm referred to as Markov chain Monte Carlo model composition (MC^3). Intuitively, this method involves randomly drawing models in a way that a given model is drawn with frequency proportional to $p(M_r|Data)$. The algorithm focuses on the models with high probability, which receive high weight in the model averaging procedure, while avoiding those with low probability. See also Hoeting et al (1999) for more details. The dependent variable in each case is y and contains N observations. k is the total number of potential explanatory variables and these are stacked in a $N \times k$ variable matrix $\mathbf{V}$. We have r=1,..,R models, denoted by M_r. These are all Normal linear regression models which differ in their explanatory variables,

$$y = \varphi\iota_N + \mathbf{V}_r b_r + e \qquad (8)$$

where ι_N is an N×1 vector of ones, $\mathbf{V}_r$ is an N×k_r matrix containing some (or all) of the potential explanatory variables (i.e. the columns of **G**, **X**, and **Z**) and b_r is a k_r×1 vector of coefficients on these explanatory variables. The N-vector of errors, e, is assumed to be $N(0_N, \sigma^2\mathbf{I}_N)$ where 0_N is an N-vector of zeros and $\mathbf{I}_N$ is the N×N identity matrix. Note that we are assuming all models contain an intercept φ.

We define the models by the choice of explanatory variables (i.e. by the choice of $\mathbf{V}_r$). BMA assumes different models are defined by the inclusion or exclusion of each variable. This leads to 2^k models. If k is at all large, the enormous number of potential models to be estimated imposes huge demands on computation. These computational demands motivate our choice of the Normal linear regression model.

We use a Normal-Gamma natural conjugate prior with hyperparameters chosen in the objective fashion described in Fernandez, Ley and Steel (2001). Specifically, our error variance uses the standard noninformative prior:

$$p(\sigma) \propto 1/\sigma. \tag{9}$$

We centre all the explanatory variables by subtracting off their means. As a consequence, we use a flat prior ($p(\sigma) \propto 1$) for the intercept. For the slope coefficients we assume a g-prior of the form:

$$b_r \sim N(0_k, \sigma^2[g_r \boldsymbol{V}'_r \boldsymbol{V}'_r]^{-1}) \tag{10}$$

where g_r is a scalar. Following Fernandez, Ley and Steel (2001), who relate these choices to common information criteria, we choose

$$g_r = \begin{cases} \frac{1}{k^2} if\ N \leq k^2 \\ \frac{1}{N} if\ N > k^2 \end{cases}. \tag{11}$$

The resulting posterior for β_r follows a multivariate t-distribution with mean:

$$E(b_r|Data, M_r) = [(1+g_r)\mathbf{V}_r'\mathbf{V}_r']^{-1}\mathbf{V}_r{}'y \quad (12)$$

covariance matrix:

$$var(b_r|Data, M_r) = \frac{N\bar{s}^2}{N-2}[(1+g_r)\boldsymbol{V}_r'\boldsymbol{V}_r'] \quad (13)$$

where

$$\bar{s}^2 = (y'\mathbf{M}_{Vr}y) + g_r(y-\bar{y}\iota_N)'(y-\bar{y}\iota_N)/(N(g_r+1)) \quad (14)$$

where $\mathbf{M}_{Vr} = \mathbf{I}_N - \mathbf{V}_r(\mathbf{V}'_r\mathbf{V}_r)^{-1}\mathbf{V}'_r$. The posterior model probabilities are given by:

$$p(M_r|Data) = c\left(\frac{g_r}{g_r+1}\right)^{\frac{k_r}{2}}(N\,\bar{s}^2)^{-(N-1)/2} \quad (15)$$

where c is a constant common to all models. Equation (15) is related to familiar information criteria since the Schwartz Information Criterion is an approximation to its logarithm. If desired, the fact that $\sum_{\gamma=1}^{R} p(M_r|Data) = 1$ can be used to evaluate c.

If the number of models, R, is relatively small, these equations can be evaluated for every possible model and BMA or model selection can be implemented directly. In traditional applications of BMA, $R=2^k$ and where $k>20$ direct implementation of BMA is computationally infeasible. Accordingly, we adopt the MC^3 algorithm described in Madigan and York (1995). This is a Metropolis algorithm which is very simple to implement. In particular, if the current model in the chain is M_s then a candidate model, M_j, which is randomly (with equal probability) selected from the set of models including M_s and all models containing one more or one less explanatory variable (i.e.

the algorithm randomly either adds or subtract one column from $\mathbf{V}_s$), is drawn. M_j is accepted with probability:

$$\min\left\{1, \frac{p(M_j|Data)}{p(M_s|Data)}\right\} \quad (16)$$

We monitor convergence of the chain by calculating the probability of the ten most probable models drawn in two different ways. First, we calculate them analytically using the equation above. Then we approximate this probability using output from the MC^3 algorithm. When these probabilities (computed using equation 15) are the same to two decimal places, we deem convergence to have taken place. The number of draws required for the various models considered varied from 1,000,000 to 2,000,000.

3 3. Results

3.1 GCM Selection

All computations herein were done using Stata version 12.0 except for the BMA analysis which was done using Matlab. Data and code are available in the online supplement. All linear regressions controlled for heteroskedasticity and clustered errors.

The first step was to estimate equations (1) and (2), examining one GCM at a time, to assess if the model-predicted spatial trend pattern z_i^j is anticorrelated with the data or not. Table 3 shows the results. The first column lists the model number. The second column shows the estimate of γ_j from equation (1) and the third column shows the same term when the regression is augmented with **G** and **X**. The fourth column indicates whether the model will be kept (1) or not (0) in subsequent analysis, based on the requirement that γ_j is non-negative in both columns. The final column indicates the p value on γ_j from the multivariate regression used for the third column. On this criterion 12 of 22 models are retained, while the other 10 are found not to reliably yield a positive correlation with the observed data and so were discarded. Conditional on having a positive coefficient, five of 11 models are significant at 10%, and two (models 11 and 19) are significant at 5%, indicating that they contributed significant explanatory power in the presence of the rival explanatory variables.

3.2 Nested and Encompassing Tests

The results in Table 4 are all based on regressions in the form of equation (2). The first column either indicates the model number, for a case in which it is included individually, or "all" denoting that all models retained as per Table 3 were included jointly. The next six columns present three pairs of F statistics and associated p values. The first pair refers to the joint test of the coefficients in γ, the second pair refers to the joint tests of the coefficients in β and the third pair refers to the joint tests of the coefficients in μ. The last two columns present F statistics and associated p values for a test that model M1 encompasses M4.

The first row reports on result of an estimation in which only the retained columns of $\mathbf{Z}$ are included and $\hat{\gamma} = 0$ is (jointly) tested. The second row reports on a result in which $\mathbf{Z}$ and $\mathbf{G}$ are included and the tests are $\hat{\gamma} = 0$ and $\hat{\mu} = 0$. The third row reports on a result in which $\mathbf{Z}$ and $\mathbf{X}$ are included and the tests are $\hat{\gamma} = 0$ and $\hat{\beta} = 0$. The fourth row presents the results when all the retained columns of $\mathbf{Z}$ are included along with $\mathbf{X}$ and $\mathbf{G}$. In each case the tests reject, indicating that none of the variable groupings can be dropped, and these inferences are robust to basic variations in the model. The implication is that no two of the variable groups encompasses the third.

The next 22 rows present results of estimating equation (2) in which the only retained element of $\mathbf{Z}$ is, sequentially, $z^1 - z^{22}$. In all 22 cases we strongly reject dropping $\mathbf{X}$ or $\mathbf{G}$, but the hypothesis that the individual GCM does not contribute significant explanatory power is rejected at 5% only three times, and in one of those cases (model 22) the coefficient is negative. So we only find two cases (models 11 and 19) in which the individual model contributes significant explanatory power. In other words, a model consisting only of the geographic and socioeconomic variables encompass a model containing any one of the retained GCMs except models 11 and 19 , accounting not only for some of the variability in temperatures but also the apparent explanatory power of the GCMs.

It is notable that, in every case, the last two columns show that we can massively reject the hypothesis that M1 encompasses M4. In other words the GCM data (either from an individual model or from the set of retained models) cannot account for the explanatory power of a model that includes **G** and **X**. Hence the concern about the potential ambiguity between **Z** and **G** confounding the test, as noted in Section 2.2, can be set aside.

3.3 Non-nested Tests

Tables 5 and 6 show the results from the non-nested tests based on estimation of versions of equation (6) constructed two ways so that each variable group is tested both for inclusion and exclusion:

$$y = \mathbf{G}\mu + \mathbf{X}\beta + \alpha\hat{y}_{GZ} + e \qquad (17)$$

and

$$y = \mathbf{G}\mu + \mathbf{Z}\mu + \alpha\hat{y}_{GX} + e. \qquad (18)$$

The first column shows the model number when a single GCM is tested. The last row reports results when all 12 retained models are included in **Z**. Each remaining column shows the p values of the indicated tests. The first two show the results from testing $\hat{\alpha} = 0$ and $\hat{\alpha} = 1$ from equation (17), which tests that the variable group containing GCMs has, respectively, zero explanatory power or 100% of the explanatory power. In five of 12 cases (models 10, 11, 13, 18, and 19) the hypothesis of zero explanatory power is rejected at 10% and in two cases at 5% (models 11 and 19). In no case is the hypothesis of full explanatory power rejected, but this is likely due to the wide variances on $\hat{\alpha}$ rather than an indication that the **X** variables have no explanatory power. Estimation of equation (18) confirms this. The test of $\hat{\alpha} = 0$ (socioeconomic variables have no explanatory power) strongly rejects in every case. Across all 12 estimations in Table 5 the mean *p* value is 4.61E-07 (Table 6). Also the *p*-values on the test of $\hat{\alpha} = 1$ in equation (18) are higher than in equation (17) in almost all cases, with a mean value of 0.775 versus 0.455 (Table 6).

Overall the nested and non-nested tests identify models 11 and 19 (respectively from the Institute of Atmospheric Physics, China, and the National Center for Atmospheric Research, USA) as the only ones that have the correct sign, a non-zero level of explanatory power against a rival that only includes geographic and socioeconomic variables, and whose apparent explanatory power is incapable of being accounted for by a rival that makes no use of GCM processes. This indicates, somewhat surprisingly, that twenty of the twenty two GCMs demonstrated no significant explanatory power for the spatial pattern of 1979-2002 warming trends over land. However, the climate models were tested either one-at-a-time or jointly. We noted above that models might perform best in some linear combination, or when tested against a partial linear combination of rival variables. We now turn to the BMA analysis results to see whether model uncertainty is masking some aspects of potential explanatory power of our explanatory variables.

3.4 BAYESIAN MODEL AVERAGING

Our main interest is in the comparative explanatory power of **Z** and **X** when **G** is included in the model, such as models M3 and M4 in equation (4). To implement this in BMA we forced the columns of **G** to be included by assumption. Hence the BMA analysis searched only over the 19 columns of **X** and **Z** (including only the retained GCMs 2, 3, 10, 11, 13, 14, 16, 17, 18, 19, 20 and 21), requiring evaluation of 2^{19} models. Table 7 shows the results, with the variables ranked (high to low) by Posterior Inclusion Probabilities (PIPs). The asterisks denote one or two standard deviations from zero. These are not the same as t-statistics, but are included for convenience.

The first column lists the variable name and the second lists the PIP. The third and fourth columns list the BMA point estimates and standard deviations. The fifth and sixth columns show, respectively, the point estimates and standard deviations using the model that obtained the highest support in the data, which coincides with the model that includes only those variables in **X** and **Z** that obtain PIPs greater than 0.5, namely *e*, *g*, ms11 and ms13.

Table 8 provides some additional estimates for comparative purposes. These are classical regression estimates under three specifications: the full model (**G**, **X** and **Z**), the model with only **G** and the socioeconomic variables **X**, and the model with only those variables that obtain PIP scores above 0.2. As is the case in the classical tests, a combination of GCMs and socioeconomic variables are necessary to explain the surface temperature trends. The BMA analysis indicates that only two columns from each of **X** and **Z** obtain PIPs over 0.5, and thus only a subset of each category is more likely than not to be in the true model.

The regression models receiving the most support in the data yield point estimates for three atmospheric oscillation terms (NAO, AO and SOI) that are near zero; while the PDO term has a point estimate just under 0.1. GCMs 2, 3, 16, and 27 all receive essentially no support in the data, obtaining PIP scores below 0.05.

One noticeable discrepancy between the classical tests and the BMA results is that model 19 is no longer the top-ranked GCM, instead its PIP is only 0.215, substantially smaller than that for models 11 and 13. In other words, whereas model 19 was unencompassed by the non-GCM data when tested alone, and therefore appeared to have unique explanatory power, in the BMA analysis GCMs 11 and 13 emerge with the highest probability of jointly belonging in the correct model.

Another, and even more unexpected, discrepancy is the disappearance of an effect due to population growth, since it is typically considered a good indicator of socioeconomic influences on temperature trends, and in a model with only **G** and **X** it obtains a coefficient of about 0.29 and a p value of 0.034 (see Table 8, second and third columns). Table 7 shows that this result is not robust to model uncertainty: models receiving the strongest support in the data yield a low probability that the population growth variable belongs in the model, and the BMA point estimate is only 0.053. A possible explanation is that the space spanned by the entire group of socioeconomic variables is nearly spanned by the subset consisting only of GDP density and education, so that these serve as an effective proxy for the full effects of variations in socioeconomic conditions and rates of change, including population change. If so, the penalty for adding other variables imposed in the model support criterion (marginal likelihood, equation 15) would override the improvement in explanatory power, resulting in greater posterior support for the simpler specification.

The classical and Bayesian results can be reconciled using encompassing tests. The BMA analysis isolates variables **G**, e, g, ms11 and ms13 as constituting the model with the highest support score in the data, net of penalties for overspecification (log of the marginal likelihood). The question then arises whether such a model can encompass the remaining variables. An F test on the excluded variables rejects this hypothesis (*P*=0.0005). However, by adding in the next two variables on the list, namely *c* (change in coal consumption) and ms19, the log marginal likelihood does not change much and we no longer reject the hypothesis that the resulting model encompasses all the remaining variables (*P*=0.1837). Moving sequentially down the list no further additions yield significant rejections of the encompassing test, indicating that the model consisting of **G**, ms11, ms13, ms19, *e*, *g* and *c* provides all the explanatory power found in the complete data set. The estimates are shown in the last column of Table 8. Hence by using a rule that variables are retained if they have a PIP score of 0.2 or higher, we find the Bayesian and classical results concur.

4 CONCLUSIONS

In view of the generally weak performance of climate models at predicting the spatial pattern of surface temperature trends it might be supposed that we are testing their accuracy at something they were never designed to do. Our tests asks, if we are given the observed pattern of LT trends over a 24 year period plus some fixed geographical measures, can climate models, when provided with observed forcings over the same interval as inputs, provide any information that significantly reduces the forecast errors of the observed pattern of climate trends at the Earth's surface over the same interval. If users believe the models were never designed to meet this standard, then our results will not be surprising, and could be interpreted merely as confirmation that this is not something that should be expected of climate models. But since the socioeconomic variables do significantly improve such forecasts, it is arguable that the climate models ought to be able to do so as well, and their extensive use for the prediction and interpretation of the spatial patterns of temperature change at the Earth's surface, and the use of such projections in reports for policymakers (e.g. Parry et al. 2007), leads us to the view that it is appropriate to assess their usefulness in this regard.

When testing GCM's one at a time, 10 of 22 do not reliably yield positive correlations with the observed spatial pattern of trends at the Earth's surface and are removed from the joint analysis. 10 of the 12 remaining models are encompassed by simpler models that do not have GCM information in them. When treating variables jointly within their groupings, nested and non-nested test results indicate that none of the three data categories encompass the other two, and therefore a good model of the spatial pattern of trends at the land surface likely requires some combination of GCM output, measures of socioeconomic contamination of climate data, and some fixed geographic detail. Such a model would be statistically superior to one that depends on GCM outputs alone.

This is confirmed in a BMA exercise, which searches over 2^{19} possible regression specifications and finds that some of both the GCM and socioeconomic variable groups likely belong in the model. Of the GCMs, classical tests identify models 11 (IAP FGOALS 1.0g) and 19 (NCAR CCSM 3.0) as having the most robust explanatory power, while the Bayesian analysis identifies models 11 and 13 (INM 3.0) as having the most, with model 19 coming third on the list. Models 2, 3, 16 and 17 (CCCMA CGCM 3.1 T47& T63, MIROC 3.2 MEDRES and MPI ECHAM5) have very little probability (<0.05) of being included in a valid explanatory model of the surface pattern of climate change, and their outputs are encompassed by a simple empirical model that does not include any GCM processes, meaning that they do not contribute any significant explanatory power.

The Bayesian analysis yields robust posterior inclusion probabilities (PIPs) that indicate the likelihood that each variable belongs in the true model. The geographic variables are forced in by assumption (though we tested that they are not encompassed by the GCMs). Of the GCM and socioeconomic data, BMA analysis assigns PIPs greater than 0.2 to educational attainment, GDP density and the change in local coal consumption, and GCMs 11, 13 and 19, which are from, respectively, China, Russia and the USA. A model consisting only of these measures encompasses all remaining GCM output and socioeconomic measures. Thus we conclude that a good model of the spatial pattern of surface climatic trends likely requires a combination of the processes represented in some GCMs and measures of data contamination induced by regional socioeconomic variations.

5 REFERENCES

Anselin, L, Anil K. Bera, Raymond Florax and Mann J. Yoon (1996). "Simple diagnostic tests for spatial dependence." *Regional Science and Urban Economics* 26: 77—104

Berk, Richard A., Robert G. Fovell, Frederic Schoenberg and Robert E. Weiss (2001) "The use of statistical tools for evaluating computer simulations." *Climatic Change* 51: 119-130.

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

CCSP (2008) *Climate Models: An Assessment of Strengths and Limitations*. A Report by the U.S. Climate Change Science Program and the Subcommittee on Global Change Research [Bader D.C., C. Covey, W.J. Gutowski Jr., I.M. Held, K.E. Kunkel, R.L. Miller, R.T. Tokmakian and M.H. Zhang (Authors)]. Department of Energy, Office of Biological and Environmental Research, Washington, D.C., USA, 124 pp.

Covey, C., K.M. AchutaRao, U. Cubasch, P. Jones, S.J. Lambert, M.E. Mann, T. J. Phillips, K.E. Taylor (2003) "An overview of results from the Coupled Model Intercomparison Project." *Global and Planetary Change* 37 103—133.

Davidson, R. and J.G. MacKinnon (2004), *Econometric Theory and Methods* Toronto: Oxford.

Davidson, R. and J.G. MacKinnon (1981) "Several Tests for Model Specification in the Presence of Alternative Hypotheses" *Econometrica*, Vol. 49, No. 3 781-793

De Laat, A.T.J., and A.N. Maurellis (2004), "Industrial CO_2 emissions as a proxy for anthropogenic influence on lower tropospheric temperature trends," *Geophys. Res. Lett.* Vol. 31, L05204, doi:10.1029/2003GL019024.

De Laat, A.T.J., and A.N. Maurellis (2006), "Evidence for influence of anthropogenic surface processes on lower tropospheric and surface temperature trends," *Int. J. Climatol.* 26:897—913.

Easterly W. and M. Sewadeh (2003) World Bank Global Development Network Growth Data Base http://www.worldbank.org/research/growth/GDNdata.htm. Accessed fall 2003.

Fernandez, C., Ley, E. and Steel, M. (2001), "Benchmark priors for Bayesian model averaging." *Journal of Econometrics* 100 381-427.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux (2008), "Performance metrics for climate models," *J. Geophys. Res.*, *113*, D06104, doi:10.1029/2007JD008972.

Hegerl G, Hasselmann K, Cubash U, Mitchell J, Roeckner E, Voss R, Waszkewitz J (1997) "Multi-fingerprint detection and attribution analysis of greenhouse gas, greenhouse gas-plus-aerosol and solar forced climate change." *Climate Dynamics* 13(9):613–634

Hegerl, G.C., F. W. Zwiers, P. Braconnot, N.P. Gillett, Y. Luo, J.A. Marengo Orsini, N. Nicholls, J.E. Penner and P.A. Stott, (2007). Understanding and Attributing Climate Change. In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis,K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Hoeting, J., Madigan, D., Raftery, A. and C Volinsky. (1999) "Bayesian model averaging: A tutorial" *Statistical Science* 14 382-417

Jenne, R.L. (1974), Jenne's northern hemisphere climatology, monthly, 1950-64. National Center for Atmospheric Research Dataset DS205.0, National Center for Atmospheric Research, Boulder CO.

Jun, Mikyoung, Reto Knutti and Douglas W. Nychka (2008) "Spatial analysis to quantify numerical model bias and dependence: How many climate models are there?" *Journal of the American Statistical Association* 108 No. 483 934—947 DOI 10.1198/016214507000001265.

Kaufmann, Robert K. and David I. Stern (2004) "A Statistical Evaluation of Atmosphere -Ocean General Circulation Models: Complexity vs. Simplicity." Rensselaer Polytechnic Institute Department of Economics Working Paper 0411, May 2004.
Kiehl, J.T. (2007) "Twentieth century climate model response and climate sensitivity." *Geophysical Research Letters* 34 L22710, doi:10.1029/2007GL031383, 2007.
Knutson, T. R., T. L. Delworth, K. W. Dixon, I. M. Held, J. Lu, V. Ramaswamy, M.D. Schwartzkopf, G. Stenchikov and R.J. Stouffer (2006) Assessment of twentieth-century regional surface temperature trends using the GFDL CM2 coupled models. *J. Clim.* 19, 1624–1651 (2006).
Knutti, R. (2008) "Why are climate models reproducing the observed global surface warming so well?" *Geophysical Research Letters* 35, L18704, doi:10.1029/2008GL034932, 2008.
Knutti, R. and G. Hegerl (2008) "The equilibrium sensitivity of the Earth's temperature to radiation changes." *Nature Geoscience* doi: 10.1038/ngeo337 735—743.
Madigan, D. and J. York. (1995). "Bayesian graphical models for discrete data." *International Statistical Review* 63 215-232.
McKitrick, R.R. and P.J. Michaels (2007), "Quantifying the influence of anthropogenic surface processes and inhomogeneities on gridded global climate data," *J. Geophys. Res.*, 112, D24S09, doi:10.1029/2007JD008465.
McKitrick, Ross and Patrick J. Michaels (2004). "A Test of Corrections for Extraneous Signals in Gridded Surface Temperature Data" *Climate Research* 26 pp. 159-173.
McKitrick, Ross R. (2010) "Atmospheric Oscillations do not Explain the Temperature-Industrialization Correlation." *Statistics, Politics and Policy*, Vol 1 No. 1, July 2010.
McKitrick, Ross R., Stephen McIntyre and Chad Herman (2010) "Panel and Multivariate Methods for Tests of Trend Equivalence in Climate Data Sets." *Atmospheric Science Letters* DOI: 10.1002/asl.290.
McKitrick, Ross R. and Nicolas Nierenberg (2010) "Socioeconomic Patterns in Climate Data." *Journal of Economic and Social Measurement*, 35(3,4) pp. 149-175. DOI 10.3233/JEM-2010-0336
Mears, Carl A., Matthias C. Schabel, and Frank J. Wentz. (2003). A Reanalysis of the MSU Channel 2 Tropospheric Temperature Record. *Journal of Climate* 16, no. 22 (November 1): 3650-3664 .
Michaels, P.J, P.C. Knappenberger, R.C. Balling Jr, and R.E. Davis (2000), "Observed warming in cold anticyclones," *Climate Research*. 14:1-6.
Mizon, G. E. (1984). 'The encompassing approach in econometrics', in Hendry D. F. and Wallis K. F. (eds), *Econometrics and Quantitative Economics*, Basil Blackwell, Oxford, pp. 135–172.
Parry, M.L., O.F. Canziani, J.P. Palutikof, P.J. van der Linden and C.E. Hanson (eds) *Contribution of Working Group II to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 2007*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
Pisati, Maurizio (2001) "Tools for spatial data analysis." *Stata Technical Bulletin* STB-60, March 2001, 21—37
Randall, D.A., R.A. Wood, S. Bony, R. Colman, T. Fichefet, J. Fyfe, V. Kattsov, A. Pitman, J. Shukla, J. Srinivasan, R.J. Stouffer, A. Sumi and K.E. Taylor (2007). "Climate Models and Their Evaluation." In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.
Santer, B. D., P. W. Thorne, L. Haimberger, K. E. Taylor, T. M. L. Wigley, J. R. Lanzante, S. Solomon, M. Free, P. J. Gleckler, and P. D. Jones (2008) "Consistency of Modelled and Observed Temperature

Trends in the Tropical Troposphere." *International Journal of Climatology* DOI: 10.1002/joc.1756.
Schmidt, Gavin (2009) "Spurious correlation between recent warming and indices of local economic activity." *International Journal of Climatology* 10.1002/joc.1831
Schwartz, Stephen E., R.J. Charlson and H. Rodhe (2007) "Quantifying climate change—too rosy a picture?" *Nature reports climate change* 2 23—24.
Shukla, J. T. DelSole, M. Fennessy, J. Kinter and D. Paolino (2006) Climate model fidelity and projections of climate change. *Geophysical Research Letters* 33, L07702, doi:10.1029/2005GL025579, 2006.
Spencer, R.W. and J.C. Christy (1990), "Precise monitoring of global temperature trends from satellites," *Science* 247:1558—1562.

6 TABLES

Variable Name	Definition	Mean	Std. Dev.	Min	Max
y	Surface temperature trend (°C/decade) by gridcell	0.3030	0.2527	-0.717	1.042
msu	Lower tropospheric temperature trend (°C/decade) by gridcell	0.2429	0.15165	-0.0921	0.6588
dry	Air predominantly dry (mean dewpoint < 0°C)	0.4673	0.4995	0	1
water	Grid cell includes major coastline	0.6028	0.4899	0	1
abslat	Absolute latitude	40.72	17.81	2.5	82.5
aorv	Correlation between Atlantic Oscillation and gridcell temperature	-0.0406	0.5615	-2.5281	1.621
naorv	Likewise for North Atlantic Oscillation	-0.2584	0.7374	-3.5433	1.6746
pdorv	Likewise for Pacific Decadal Oscillation	-0.0835	0.2722	-1.1770	0.9650
soirv	Likewise for Southern Oscillation Index	0.0416	0.1728	-0.4977	0.7192
g	1979 GDP density ($million/km^2.)	0.3006	0.6056	0.0014	3.0023
e	Sum of literacy rate and post-secondary attainment	106.78	26.19	11.6	144.2
x	Number of missing months in gridcell record	0.7313	2.4419	0	24
p	% growth in population	0.2774	0.2092	-0.0691	1.2353
m	% growth in real per capita income	0.3781	0.6156	-0.7901	2.1472
q	% growth in GDP	0.7663	0.8410	-0.6686	3.0025
c	% growth in coal consumption	1.031	4.1106	-1	39.333

TABLE 1. SUMMARY STATISTICS FOR VARIABLES IN **G** AND **X**. SAMPLE SIZE = 428

Model Number	Name	Modeling Team	Mean	Std. Dev.	Min	Max
1	BCCR BCM 2.0	Bjerknes Centre for Climate Research, Norway	0.127	0.209	-0.774	0.720
2*	CCCMA CGCM 3.1 T47	Canadian Centre for Climate Modeling and	0.308	0.123	-0.120	0.972
3*	CCCMA CGCM 3.1 T63	Analysis, University of Victoria	0.284	0.221	-0.537	1.259
4	CSIRO MK 3.0	Commonwealth Scientific and Industrial	0.149	0.275	-1.610	1.058
5	CSIRO MK 3.5	Research Organization, Australia	0.176	0.312	-1.323	0.958
6	GFDL CM 2.0	Geophysical Fluid Dynamics Laboratory,	0.451	0.459	-0.141	2.947
7	GFDL CM 2.1	National Oceanic and Atmos. Admin., USA	0.337	0.298	-0.364	1.746
8	GISS AOM	Goddard Institute of Space Studies, National	0.179	0.157	-0.246	1.194
9	GISS EH	Aeronautics and Space Administration, USA	0.229	0.152	-0.139	0.735
10*	GISS ER		0.220	0.136	-0.110	0.610
11*	IAP FGOALS 1.0g	Institute of Atmospheric Physics, China	0.094	0.164	-0.312	0.808
12	INGV ECHAM 4	Max Planck Institute for Meteorology, Germany	0.198	0.238	-0.512	1.083
13*	INM CM 3.0	Institute for Numerical Mathematics, Russia	0.230	0.276	-0.430	1.127
14*	IPSL CM 4	Institut Pierre Simon Laplace, France	0.308	0.296	-0.272	3.190
15	MIROC 3.2 HIRES	Center for Climate System Research	0.248	0.249	-0.795	1.077
16*	MIROC 3.2 MEDRES	(The University of Tokyo), Japan	0.208	0.153	-0.181	1.042
17*	MPI ECHAM 5	Max Planck Institute for Meteorology, Germany	0.206	0.088	0.002	0.671
18*	MRI CGCM 2.3.2a	Center for Climate System Research, Japan	0.185	0.138	-0.124	0.767
19*	NCAR CCSM 3.0	National Center for Atmospheric Research,	0.291	0.174	-0.060	1.119
20*	NCAR PCM 1	Colorado USA	0.188	0.123	-0.121	0.609
21*	UKMO HAD CM 3	Hadley Centre for Climate Prediction and	0.134	0.274	-2.495	1.040
22	UKMO HADGEM 1	Research / Met Office, UK	0.451	0.442	-0.484	1.944

TABLE 2. GCM NAMES AND SUMMARY STATISTICS FOR VARIABLES z^j. NUMBER OF OBSERVATIONS = 428. *-INCLUDED IN BMA ANALYSIS. Column 4 shows the average temperature trend per grid cell in oC/decade. For further details see http://www-pcmdi.llnl.gov/ipcc/model_documentation/ipcc_model_documentation.php.

Model	$\hat{\gamma}$ from regression: univariate	multivariate	Keep GCM	Prob ($\hat{\gamma} = 0$)
1.	-0.1980	0.0370	0	-
2.	0.4934	0.0914	1	0.2985
3.	0.0093	0.0282	1	0.5552
4.	0.0440	-0.0439	0	-
5.	-0.1991	0.0244	0	-
6.	0.0736	-0.0024	0	-
7.	-0.0538	-0.0051	0	-
8.	0.2327	-0.0086	0	-
9.	0.5826	-0.0255	0	-
10.	0.5868	0.1157	1	0.0778
11.	**0.0577**	**0.1677**	**1**	**0.0437**
12.	-0.1033	-0.1012	0	-
13.	0.3893	0.1222	1	0.0538
14.	0.2021	0.0469	1	0.1745
15.	-0.0766	-0.0626	0	-
16.	0.2720	0.0019	1	0.9816
17.	0.1045	0.0735	1	0.7445
18.	0.4523	0.1848	1	.0916
19.	**0.6344**	**0.1526**	**1**	**0.0195**
20.	0.5378	0.1369	1	0.3504
21.	0.0639	0.0566	1	0.3453
22.	0.1703	-0.0569	0	-

TABLE 3. TESTS OF BASIC CORRELATION BETWEEN GCM GUESS PATTERNS AND OBSERVED TRENDS OVER LAND.

Columns show coefficient values on z^j in univariate and multivariate cases, corresponding to equations (1) and (2). GCM being tested indicated by row number. "Keep GCM" indicates 1 if both coefficients are positive, 0 otherwise. Prob value in 5th column, "-" indicates model not retained in joint analyses. Bold face denotes significant at 5%.

Model	Coefficient being tested:						Test of whether M1	
included	$H0\colon \hat{\gamma} = 0$		$H0\colon \hat{\beta} = 0$		$H0\colon \hat{\mu} = 0$		Encompasses M4	
in **Z**	*F*	*p*	*F*	p^{+}	*F*	p^{++}	*F*	p^{+++}
all	**18.6372**	**2.42e-18**						
all	**3.9719**	**0.0001**			**25.670**	**0**		
all	**22.9645**	**5.92e-21**	**3.0036**	**0.0075**				
all	**2.7208**	**0.0040**	**4.7540**	**0.0002**	**21.812**	**0**	**23.579**	**0**
1	0.7290	0.3958	**9.1119**	**0**	**66.110**	**0**	**75.428**	**0**
2*	1.0952	0.2985	**8.8202**	**0**	**43.800**	**0**	**65.799**	**0**
3*	0.3510	0.5552	**7.6954**	**0**	**61.491**	**0**	**75.307**	**0**
4	1.2965	0.2583	**9.3598**	**0**	**60.905**	**0**	**81.779**	**0**
5	0.1531	0.6966	**10.988**	**0**	**69.424**	**0**	**65.924**	**0**
6	0.0069	0.9338	**9.5369**	**0**	**63.387**	**0**	**76.650**	**0**
7	0.0523	0.8197	**9.0582**	**0**	**53.824**	**0**	**63.650**	**0**
8	0.0420	0.8381	**9.1646**	**0**	**58.156**	**0**	**73.512**	**0**
9	0.0807	0.7771	**8.3841**	**0**	**38.232**	**0**	**60.512**	**0**
10*	3.1928	0.0778	**7.8993**	**0**	**47.687**	**0**	**64.577**	**0**
11*	**4.2026**	**0.0437**	**7.8604**	**0**	**56.803**	**0**	**80.900**	**0**
12	3.4947	0.0653	**8.9041**	**0**	**83.508**	**0**	**76.669**	**0**
13*	3.8327	0.0538	**7.4947**	**0**	**66.181**	**0**	**75.910**	**0**
14*	1.8771	0.1745	**9.1613**	**0**	**58.462**	**0**	**71.825**	**0**
15	1.3561	0.2477	**9.6312**	**0**	**56.445**	**0**	**64.299**	**0**
16*	0.0005	0.9816	**9.3193**	**0**	**64.077**	**0**	**80.757**	**0**
17*	0.1070	0.7445	**8.8075**	**0**	**62.830**	**0**	**76.124**	**0**
18*	2.9165	0.0916	**8.8827**	**0**	**54.778**	**0**	**69.442**	**0**
19*	**5.6880**	**0.0195**	**8.1350**	**0**	**51.187**	**0**	**68.539**	**0**
20*	0.8824	0.3504	**6.4801**	**0**	**72.635**	**0**	**71.748**	**0**
21*	0.9016	0.3453	**7.5206**	**0**	**60.921**	**0**	**74.979**	**0**
22	**6.4741**	**0.0129**	**9.0865**	**0**	**52.571**	**0**	**76.678**	**0**

TABLE 4. TESTS OF ENCOMPASSING RELATIONS AMONG EXPLANATORY VARIABLE CATEGORIES.
Column 1: GCM being tested. Columns 2 and 3: *F* and *p* values, respectively, for test that indicated column or columns of **Z** have no explanatory value, or are encompassed by other variables in the model. Columns 4 and 5, same for test that columns of **X** are

encompassed by other explanatory variables. Columns 6 and 7, same for test that columns of **G** are encompassed. Bold face denotes significant at 5%. Columns 8 and 9: F and p values for test that model M1 ($y = \mathbf{Z}\gamma + e_1$) encompasses model M4 ($y = \mathbf{G}\mu + \mathbf{X}\beta + e_2$), in other words GCM's account for apparent explanatory power of non-GCM variables. *-GCM explanatory pattern has positive correlation with observations. +: values below 5e-6 reported as 0. ++ values below 7.4e-19 reported as zero. +++Values below 4.4e-24 reported as 0.

	p-values of non-nested test of % explanatory power of			
	GCMs		Socioeconomic Vars	
GCM Number	0%	100%	0%	100%
2	0.2985	0.3930	**0**	0.9027
3	0.5552	0.2646	**0**	0.9231
10	0.0778	0.4050	**0**	0.9042
11	**0.0437**	0.3846	**0**	0.5853
13	0.0538	0.1574	**0**	0.5473
14	0.1745	0.1567	**0**	0.8151
16	0.9816	0.9744	**0**	0.9999
17	0.7445	0.8925	**0**	0.9822
18	0.0916	0.9802	**0**	0.9475
19	**0.0195**	0.3820	**0**	0.8332
20	0.3504	0.2603	**0**	0.7025
21	0.3453	0.4776	**0**	0.7900
all	**0.0063**	0.1843	**0**	0.1421

TABLE 5. NON-NESTED TESTS OF GCM AND SOCIOECONOMIC VARIABLE EXPLANATORY POWERS. Rows included only if model exhibits positive correlation with observed surface trend pattern. 2nd and 3rd columns: p-value on tests of α=0 and α =1, respectively, in equation (17). 4th and 5th columns: test of α =0 and α =1, respectively, in equation (18). Bold faced denotes significant at 5%. Values below 3.1e-8 reported as zero.

Test	Mean	Std. Dev	Min	Max
GCM: 0%	0.2879	0.3068	0.0063	0.9816
GCM: 100%	0.4548	0.2999	0.1567	0.9802
Socioecon: 0%	4.61e-07	1.64e-06	9.37e-10	5.92e-06
Socioecon: 100%	0.7750	0.2375	0.1421	0.9999

TABLE 6. SUMMARY STATISTICS FOR SCORES IN TABLE 5.

Variable	Posterior Inclusion Probability	Point estimate (BMA)	Standard Deviation (BMA)	Point Estimate (Model Selection)	Standard Deviation (Model Selection)
msu	1.0000	0.9757	0.0820**	0.9525	0.0716**
aorv	1.0000	0.0424	0.0750	0.0388	0.0708
naorv	1.0000	-0.0296	0.0541	-0.0238	0.0497
pdorv	1.0000	0.0966	0.0685*	0.0735	0.0644*
soirv	1.0000	0.0429	0.0973	0.0111	0.0921
slp	1.0000	0.0078	0.0033**	0.0078	0.0031**
dry	1.0000	4.5443	4.3336*	3.0458	4.0987
dslp	1.0000	-0.0044	0.0043*	-0.0029	0.0040
water	1.0000	-0.0190	0.0196	-0.0162	0.0191
abslat	1.0000	0.0011	0.0009*	0.0009	0.0008*
e	0.9999	-0.0022	0.0004**	-0.0021	0.0004**
g	0.8576	0.0435	0.0229*	0.0504	0.0150**
ms11	0.7610	0.1367	0.0923*	0.1851	0.0571**
ms13	0.5124	0.0650	0.0723	0.1277	0.0477**
c	0.2460	0.0011	0.0022	0	0
ms19	0.2150	0.0315	0.0690	0	0
ms18	0.1070	0.0134	0.0470	0	0
ms21	0.0873	0.0041	0.0165	0	0
ms10	0.0778	0.0083	0.0375	0	0
ms14	0.0715	0.0032	0.0149	0	0
ms20	0.0695	0.0073	0.0355	0	0
p	0.0528	0.0034	0.0218	0	0
q	0.0520	0.0005	0.0067	0	0
m	0.0500	0.0007	0.0085	0	0
ms2	0.0446	0.0026	0.0234	0	0
ms16	0.0436	-0.0021	0.0198	0	0
ms3	0.0398	0.0005	0.0095	0	0
ms17	0.0364	0.0003	0.0236	0	0
x	0.0345	0.0000	0.0007	0	0

TABLE 7. BAYESIAN MODEL AVERAGING RESULTS.
** = two posterior standard deviations from zero. * = one posterior standard deviation from zero. Column 1: Variable name. Column 2: Posterior Inclusion Probability (PIP). Column 3: Point Estimate of slope coefficient (equation 12). Column 4: Standard Deviation of point estimate (equation 13). Columns 5 and 6: Point Estimates and Standard Errors of model obtaining the maximum support in the data.

	Dependent variable: *y*		
	Model based on:		
Variable	Use of all variables	Inclusion only of **G** and **X**	Variables identified in BMA
msu	**0.8968**	**0.9792**	**0.913**
	(0.000)	(0.000)	(0.000)
slp	**0.0088**	**0.0079**	**0.0092**
	(0.028)	(0.033)	(0.012)
dry	4.2086	5.2144	4.4492
	(0.304)	(0.226)	(0.300)
dslp	-0.0041	-0.0051	-0.0043
	(0.309)	(0.232)	(0.306)
water	-0.0251	-0.0244	-0.0206
	(0.243)	(0.287)	(0.333)
abslat	0.0014	0.0023	0.001
	(0.260)	(0.085)	(0.208)
aorv	0.0056	0.071	0.0153
	(0.961)	(0.571)	(0.904)
naorv	0.0032	-0.0636	0.006
	(0.971)	(0.470)	(0.949)
pdorv	0.101	**0.1386**	0.0839
	(0.142)	(0.032)	(0.225)
soirv	0.0352	0.0891	0.0374
	(0.740)	(0.383)	(0.738)
g	**0.0412**	**0.0425**	**0.0531**
	(0.006)	(0.001)	(0.000)
e	**-0.002**	**-0.0025**	**-0.0021**
	(0.000)	(0.000)	(0.000)
x	-0.0011	0.002	
	(0.786)	(0.451)	
p	0.1866	**0.2897**	
	(0.161)	(0.034)	
m	0.2335	0.2895	
	-(0.125)	(0.072)	
q	-0.1735	-0.212	
	(0.150)	(0.097)	
c	**0.0055**	**0.0059**	**0.0044**
	(0.001)	(0.001)	(0.000)
ms2	0.0455		
	(0.651)		
ms3	-0.0388		

	(0.344)		
ms10	0.0178		
	(0.814)		
ms11	0.1387		**0.1801**
	(0.104)		(0.003)
ms13	0.0891		**0.1128**
	(0.093)		(0.044)
ms14	0.0293		
	(0.365)		
ms16	-0.0847		
	(0.303)		
ms17	-0.0137		
	(0.952)		
ms18	0.0654		
	(0.671)		
ms19	0.0908		**0.1362**
	(0.504)		(0.025)
ms20	0.0971		
	(0.430)		
ms21	0.0421		
	(0.365)		
constant	**-8.7895**	**-7.8746**	**-9.2155**
	(0.030)	(0.036)	(0.013)
N	428	428	428
R^2	0.594	0.573	0.587

TABLE 8 : REGRESSION RESULTS FOR 3 MODELS.

Regression of temperature trend pattern y on various groups of dependent variables. Shown are coefficients and underneath in parentheses the p-value of the associated t-statistic. Bold face denotes significant at 5%. Error terms corrected for heteroskedasticity and clustering.

7 FIGURE

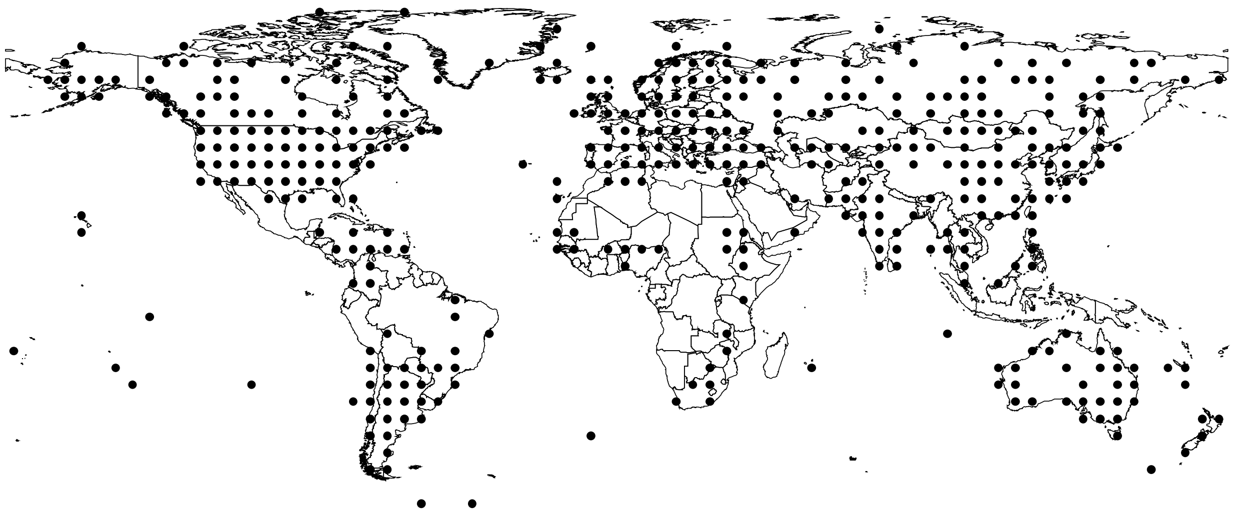

Figure 1: Locations of 428 grid cells in data base.

8 APPENDIX: FURTHER DETAILS ON DATA SET

Temperature Trends: The observed surface temperature trends y_i are linear (Ordinary Least Squares) trends through monthly temperature anomalies (not subject to annual averaging) within 5x5 degree grid cells over 1979:1 to 2002:12 in the land-based grid cells in the CRU data, versions 2 and 3, as well as in the GCM-generated data and the tropospheric data. Because of the need for a trend across 23 years we required each cell to have data for at least ninety percent of the years, where a year is considered intact if at least 8 months are available. In the CRU version 2 data this left 451 usable locations. Antarctic cells were removed, leaving 440 observations in the final version 2 data set but only 428 in CRU version 3.

GCM Data: We used all available (55) runs from 22 GCMs used in the IPCC report (Hegerl et al. 2007). The archive is at http://www-pcmdi.llnl.gov. Multiple runs from a single model were averaged. HadCM3 wasn't used because it did not represent its data in the required IPCC pressure levels. MUIB ECHO G wasn't processed because no atmospheric temperature data was available, thus synthetic MSU brightness temperatures couldn't be calculated. A single run is a deterministic computation of a climate model sing as inputs the observed climatic forcings over the historical interval, and solving for predicted climate fields including temperature, pressure and precipitation.

The calculation of tropospheric temperature from the models was done using the same algorithm and weighting functions implemented in Santer et al. (2008), which are designed to yield layer averages corresponding to observational ones measured by weather satellites. Model trend fields in degrees C/decade for the surface and lower tropospheric temperature were calculated as follows.

1. Extract all monthly GCM-generated data on temperature by grid cell from the surface through to the mid-troposphere from Jan 1979 - Dec 2002.

2. Compute the climatology (gridcell averages by month) for the same period.
3. Subtract the climatology from the original data, yielding deviations or “anomalies.”
4. Calculate the least squares trend field for each grid point only if all the data points are valid.
5. Collect only the trends that correspond to the McKitrick and Michaels (2007) set of lat/lon coordinates.
6. Multiply the resulting annual trends by 10 to obtain decadal trends

There was no missing data for the surface temperature variable in models, but there was some missing data in some runs for the lower tropospheric (LT) temperatures. This is because the models originally didn't represent the atmospheric temperature on the same set of pressure levels that the IPCC mandated. Interpolation was required and this resulted in some missing data points in the lower atmosphere. To calculate the LT temperature, the atmospheric temperature profile was multiplied by a set of weights specific to a given atmospheric layer (TLT, TMT, TLS). The weighted temperatures were then added up and divided by the sum of weights that correspond to non-missing temperature values. If this total weight did not equal or exceed 0.5 or 50%, then the temperature at that grid point was flagged as missing.

Geographic Data: $press_i$ is the mean sea level air pressure in grid cell *i.* The source of the pressure data is the climatology of Jenne (1974). DRY_i is a dummy variable denoting when a grid cell is characterized by predominantly dry conditions (which is indicated by the mean dewpoint being below 0 °C). $DSLP_i = DRY_i \times PRESS_i$. Surface warming due to greenhouse gases is hypothesized to occur faster in regions with relatively dry air and high atmospheric pressure (Michaels et al. 2000) so pressure enters the regression model as a linear spline function with a different intercept and slope in dry regions versus moist regions. $WATER_i$ is a dummy variable indicating the grid cell contains a major coastline. $ABSLAT_i$ denotes the absolute latitude of the grid cell.

Oscillation data: The measures for the Arctic Oscillation, North Atlantic Oscillation, Pacific Decadal Oscillation and Southern Oscillation are taken from McKitrick (2010), who obtained them in turn from http://www.cdc.noaa.gov/Correlation, the website of the National Oceanic and Atmospheric Administration). There is a single value of each oscillation index for the whole planet each period. What is reported at the grid cell level is the correlation between the temperatures in that grid cell and the index value over the 1979-2001 interval, thus representing a measure of the influence of the oscillation over space. The correlation can be computed in two ways, as simple Pearson correlation term, or as a regression coefficient. McKitrick (2010) reports that the latter formula yielded stronger results for the oscillation terms in the regression models so that is the form used herein.

Socioeconomic data: Each grid cell was assigned to a country. Annual real (inflation adjusted) GDP for 1979, 1989 and 1999 for each country was obtained primarily from Easterly and Sewadeh (2003) or the Central Intelligence Agency (CIA) World Fact Book. Conversions from local currency to US dollars was done using the purchasing power parity method. There were small adjustments made to the economic data for some countries to provide consistency in quantities where direct measures were unavailable. In most cases the adjustment took the form of using an available observation for one or two years after the desired year, and adjusting it backwards. Population data are obtained from Easterly and Sewadeh (2003) and the percent change p_i is measured from 1979 to 1999. Income growth m_i is the percentage change in real GDP per capita from 1979 to 1999. GDP growth y_i is defined as the percentage change in real GDP from 1979 to 1999. National coal consumption data were obtained from the US Energy Information Administration and the coal growth measure is the percentage growth of short tons of coal consumed between 1980 and 2000. The 1999 (or closest year) national literacy rate and the percentage completing post-secondary education was obtained from UNESCO. The two measures are summed together to yield e_i. Land area estimates (excluding water) for each country were obtained from the CIA World Fact Book. GDP density g_i is measured as \$million/km^2. The 1979 value was used to help ensure the right-hand side variables are predetermined with respect to the dependent variable. x_i is the number of months over the period 1979-2002 in which an observation was missing for a grid cell.