A Bivariate Time Series Approach to Anthropogenic Trend Detection in Hemispheric Mean Temperatures

RICHARD L. SMITH

Department of Statistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina

TOM M. L. WIGLEY

National Center for Atmospheric Research, Boulder, Colorado

BENJAMIN D. SANTER

PCMDI, Lawrence Livermore National Laboratory, Livermore, California

(Manuscript received 10 July 2000, in final form 27 June 2002)

ABSTRACT

A bivariate time series regression approach is used to model observed variations in hemispheric mean temperature over the period 1900–96. The regression equations include deterministic predictor variables and lagged values of the two predictands, and two different forms of this basic structure are employed. The deterministic predictors considered are simple linear trends, various climate model–generated time series based on different combinations of greenhouse gas, sulfate aerosol, and solar forcing, and the Southern Oscillation index (SOI). With linear trends as the only predictors, the best model is a fourth-order bivariate autoregressive model including lagged Southern Hemisphere (SH) to Northern Hemisphere (NH) dependence, as in previous work by Kaufmann and Stern. The estimated NH and SH trends are both $+0.67$°C century^{-1}, and both are highly statistically significant. If SOI is included as an additional predictor, however, a first-order time series model, with no SH to NH dependence, is an adequate fit to the data. This shows that SOI may be an important covariate in this kind of analysis. Further analysis uses climate model–generated forcing terms representing greenhouses gases, sulfate aerosols, and solar effects, as well as SOI. The statistical analysis makes extensive use of Bayes factors as a device for discriminating among a wide spectrum of possible models. The best fits to the data are obtained when all three forcing terms are included. Total sulfate aerosol forcing of -1.1 W m^{-2} (with a corresponding climate sensitivity of $\Delta T_{2\times} = 4.2$°C) is preferred to -0.7 W m^{-2} (with sensitivity of 2.3°C), but the Bayes factor discrimination between these cases is weak.

1. Introduction

A major theme of current climatological research is to examine how well observed changes in climate are represented by the output of climate models under various assumptions about the influences of different forcing factors. Until recently, most studies of this nature involved analyses of spatial patterns (e.g., Barnett and Schlesinger 1987; Santer et al. 1995, 1996; Hegerl et al. 1996a,b; Allen and Tett 1999; Tett et al. 1999; Stott et al. 2000). Recently, however, it has been suggested that considerable evidence to discriminate between anthropogenic signals and natural forcing factors is available in just the hemispheric mean temperatures (Kaufmann and Stern 1997).

Wigley et al. (1998) examined lagged autocorrelations and cross correlations in Northern Hemisphere (NH) and Southern Hemisphere (SH) mean temperatures, both in the raw data and in the residuals after removal of an ENSO signal, and compared them with corresponding autocorrelations and cross correlations computed from unforced (control run) simulations of two ocean–atmosphere general circulation models (OAGCMs). The discrepancies between the observed data and the control run data were considerable, with the observed data showing much larger autocorrelations over time lags of up to 20 yr. This effect, however, largely disappeared when the observed data were ‘‘corrected’’ by subtracting trends, not necessarily linear, based on a number of hypotheses about forcing factors. The hypotheses were (a) forcing due to anthropogenic influences [greenhouse gases (GHGs) and sulfate aerosols], (b) forcing due to solar effects, and (c) forcing due to both kinds of effects. In addition, comparisons were made for a number of values of $\Delta T_{2\times}$, the climate

Corresponding author address: Dr. Richard L. Smith, Dept. of Statistics, University of North Carolina at Chapel Hill, CB #3260, New West, Chapel Hill, NC 27599-3260.
E-mail: rls@email.unc.edu

sensitivity under doubled CO_2. The best agreement between the autocorrelations of the detrended observed temperatures and those of the OAGCM control runs was achieved when both natural and anthropogenic forcings were considered. Based on this, Wigley et al. (1998) claimed to have identified strong evidence for a combined anthropogenic and solar effect.

In this paper, we take these conclusions further by analyzing the hemispheric mean data as a bivariate autoregressive (hereafter, BVAR) time series. Specifically, we represent the observed series as a sum of deterministic trend plus random error components, where the random errors are BVAR. There are two reasons for pursuing such an approach as an alternative to that of Wigley et al. (1998). The first is that we can now use analytical methods of statistics, such as maximum likelihood estimation and Bayes factors, to characterize the quality of fit for different assumptions regarding the mixture of climate forcings. A second reason is that our judgement of the quality of fit is freed from comparisons with those of an OAGCM run, which may itself be problematic if the variability in the model run does not correspond well with that of the observational data. We also reconsider some analyses of Kaufmann and Stern (1997), who also examined hemispheric mean temperature data with a view to detecting anthropogenic effects, agreeing with some aspects of the Kaufmann–Stern analysis but differing from them in a number of other respects.

2. Bivariate autoregressive models

Kaufmann and Stern (1997) considered models of the form

$$N_t = \alpha_1 + \sum_{j=1}^{k} \beta_{1j}x_{tj} + \sum_{j=1}^{p_1} \gamma_{1j}N_{t-j} + \sum_{j=1}^{q_1} \delta_{1j}S_{t-j} + \epsilon_{1t},$$

$$S_t = \alpha_2 + \sum_{j=1}^{k} \beta_{2j}y_{tj} + \sum_{j=1}^{p_2} \gamma_{2j}N_{t-j} + \sum_{j=1}^{q_2} \delta_{2j}S_{t-j} + \epsilon_{2t}, \tag{1}$$

in which N_t and S_t denote the observed NH and SH temperature averages in year t, and $\{x_{tj}, y_{tj}, 1 \leq j \leq k\}$ denote k covariates, identifiable as the deterministic components of their model. These covariates may either be time t, in which case the deterministic components would be simple linear trends, or they may reflect more complex temporal changes, due to forcing factors such as greenhouse gases, etc. The simplest form of their analysis is the linear trend case $k = 1$, $x_{t1} = y_{t1} = t$. The γ_{ij} and δ_{ij} coefficients represent various lagged terms in the model that represent serial correlation both within and between the two hemispheres. The ϵ_{1t} and ϵ_{2t} terms represent error terms that are assumed normally distributed with mean 0, independent from one value of t to another and with variances σ_1^2, σ_2^2, say. The analysis of Kaufmann and Stern implicitly assumed that ϵ_{1t} and ϵ_{2t} were also independent of each other, though we shall consider the more general form in which Corr$(\epsilon_{1t}, \epsilon_{2t}) = \rho$, where ρ may be any number between -1 and 1.

An alternative form of the model is

$$N_t = \alpha_1 + \sum_{j=1}^{k} \beta_{1j}x_{tj} + W_t,$$

$$S_t = \alpha_2 + \sum_{j=1}^{k} \beta_{2j}y_{tj} + Z_t,$$

$$W_t = \sum_{j=1}^{p_1} \gamma_{1j}W_{t-j} + \sum_{j=1}^{q_1} \delta_{1j}Z_{t-j} + \epsilon_{1t},$$

$$Z_t = \sum_{j=1}^{p_2} \gamma_{2j}W_{t-j} + \sum_{j=1}^{q_2} \delta_{2j}Z_{t-j} + \epsilon_{2t}. \tag{2}$$

Whereas Eq. (1) contains autoregressive terms directly in the variables of interest, N_t and S_t, Eq. (2) first forms detrended series W_t and Z_t by subtracting the deterministic trend terms from N_t and S_t, respectively, and then models (W_t, Z_t) as a stationary bivariate autoregressive series. Although either form of the model, (1) or (2), is plausible for the kind of data we deal with here, in the discussion to follow we shall argue that (2) is preferable.

A number of earlier authors have used models of the form of (1) or (2), or their obvious analogs for univariate time series, in testing for trends in climatological time series. Bloomfield and Nychka (1992) and Woodward and Gray (1993, 1995) used univariate versions of model (2), with the trend term external to the autoregressive equation. On the other hand, Tol (1994) used an equation similar to (1), with the trend terms internal to the autoregressive equation.

Our principal method of model fitting is maximum likelihood estimation (MLE). For given model order and values of the parameters α_1, α_2, $\{\beta_{1j}, 1 \leq j \leq k\}$, $\{\beta_{2j}, 1 \leq j \leq k\}$, $\{\gamma_{1j}, 1 \leq j \leq p_1\}$, $\{\delta_{1j}, 1 \leq j \leq q_1\}$, $\{\gamma_{2j}, 1 \leq j \leq p_2\}$, $\{\delta_{2j}, 1 \leq j \leq q_2\}$, σ_1, σ_2, and ρ, we extract the values of $\{\epsilon_{1t}\}$ and $\{\epsilon_{2t}\}$ using (1) or (2) and calculate their joint density—this is the likelihood function for a given set of parameters. In the time series literature (see, e.g., Brockwell and Davis 1991), this is known as a conditional likelihood approach. The calculation requires some lagged values (before time 1) of the NH and SH series, but this will not be a problem because, for every analysis we do, we have values available from years prior to the beginning of our analysis period. The likelihood is maximized with respect to all the unknown parameters to obtain the MLEs. In practice, this is usually carried out by numerical minimization applied to the negative of the log likelihood function, which we shall henceforth refer to as NLLH. Ignoring an irrelevant constant of proportionality, the NLLH is derived from the ϵ_{1t}, ϵ_{2t}, values by the formula

$$\mathrm{NLLH} = N\log\sigma_1 + N\log\sigma_2 + \frac{N}{2}\log(1-\rho^2) + \frac{1}{2(1-\rho^2)}\sum_{t=1}^{N}\left(\frac{\epsilon_{1t}^2}{\sigma_1^2} - \frac{2\rho\epsilon_{1t}\epsilon_{2t}}{\sigma_1\sigma_2} + \frac{\epsilon_{2t}^2}{\sigma_2^2}\right),$$

in which N denotes the total number of years used for the calculation, ignoring the initial lagged values. Note that if $\rho = 0$, the central part of the calculation reduces to selecting the regression coefficients to minimize the residual sums of squares $\Sigma\ \epsilon_{1t}^2$ and $\Sigma\ \epsilon_{2t}^2$, which appears to have been what Kaufmann and Stern (1997) did.

The matrix of second-order derivatives of NLLH, evaluated at the MLE, is known as the observed information matrix and its inverse is widely used as an estimator of the covariance matrix of the MLEs (Cox and Hinkley 1974). The square roots of the diagonal entries of the inverse observed information matrix are estimated standard errors.

For comparison among models, two widely used methods are the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), defined by

$$\begin{aligned}\mathrm{AIC} &= 2\mathrm{NLLH} + 2p,\\ \mathrm{BIC} &= 2\mathrm{NLLH} + p\log N,\end{aligned} \qquad (3)$$

where p is the total number of estimated parameters and N the length of the series. In either case we seek the model that minimizes the value of (3). Both criteria use the NLLH but impose a penalty term to prevent p from getting too large—BIC imposes the larger penalty and therefore tends to select a model with fewer parameters. BIC is sometimes preferred on the grounds that this leads to consistent model selection (i.e., for choosing among a finite collection of models, one of which is correct, the probability that the correct model is chosen, using BIC, tends to 1 as $N \rightarrow \infty$, where N is the sample size) but not all statisticians regard this property as the only one that matters, and both AIC and BIC, among several other criteria, are used in practice.

As an alternative to AIC and BIC, when two models are nested (one being derived from the other by fixing some of the parameters of the larger model) it is possible to compare the models via a formal hypothesis test in which we define the null hypothesis H_0 that the smaller model is correct, against the alternative H_1 that the larger model is correct. The likelihood ratio statistic (LRS) is defined as

$$\mathrm{LRS} = 2(\mathrm{NLLH}_0 - \mathrm{NLLH}_1), \qquad (4)$$

where NLLH_0 and NLLH_1 represent the values of NLLH under the smaller and larger models, respectively. Assuming the model satisfies various regularity conditions, when the null hypothesis is true, the distribution of the LRS should be approximately χ_ν^2, where ν is the difference in the number of parameters between the two models (Cox and Hinkley 1974).

Models (1) and (2) are fitted here to an updated version of the temperature dataset employed by the Intergovernmental Panel on Climate Change (IPCC; Nicholls et al. 1996) covering the years 1900–96. For the remainder of this section we discuss a number of specific issues that seem to be important in selecting a suitable form of model.

a. Initial selection of covariates

The suitability of the model obviously depends on the selection of predictor variables or regressors $\{x_{tj}\}$ and $\{y_{tj}\}$. The simplest thing is just to specify $x_{t1} = y_{t1} = t$ to represent a linear trend. In sections 3 and 4, we shall also include terms based on forcing factors due to different influences. In the present section, however, we concentrate on two terms: linear trends and the El Niño–Southern Oscillation (ENSO) influence, as characterized by the Southern Oscillation index (SOI).

The possible influence of ENSO terms has been noted by a number of previous authors, including Kaufmann and Stern (1997) and Wigley et al. (1998). The SOI (Können et al. 1998) is a widely used numerical measure of ENSO activity, and a logical way to incorporate its effect is to include it as a covariate in the model. In the present study, we use 6-month lagged values (i.e., the value used for a particular year is the average of the last six months of the previous year and the first six months of the current year) as this has been found in previous studies (Jones 1989; Wigley 2000) to give the best correlation with observed temperature; we return to the question of optimum lag in section 4.

b. Order of autoregressive components

The first analyses considered assume that the model orders p_1, p_2, q_1, and q_2 are all the same. Figure 1 plots AIC and BIC values for model orders from 1 to 10, both with and without the linear trend and SOI terms. AIC is minimized by either a first-order or a fourth-order model with both linear and SOI terms. The more conservative BIC criterion identifies a first-order model and marginally prefers one without a linear trend, but including SOI, over the model containing both regressors.

In trying to pin down some of these choices more precisely, it is useful also to consider the results of hypothesis tests. Concerning the choice between model orders 1 and 4, for the model with both linear trend and SOI term, we find $\mathrm{NLLH} = -352.3$ (13 parameters) for model order 1, and $\mathrm{NLLH} = -363.3$ (25 parameters) for model order 4. Thus we have $\mathrm{LRS} = 2(363.3 - 352.3) = 22.0$ with 12 degrees of freedom, which corresponds to a p value of 0.038 when assessed according to the χ_{12}^2 distribution. Thus, at significance level 5%, we reject the null hypothesis $p_1 = p_2 = q_1 = q_2 = 1$ in favor of the alternative $p_1 = p_2 = q_1 = q_2 = 4$. Kaufmann and Stern (1997) also preferred a fourth-order model. Similar LRS tests have been conducted for

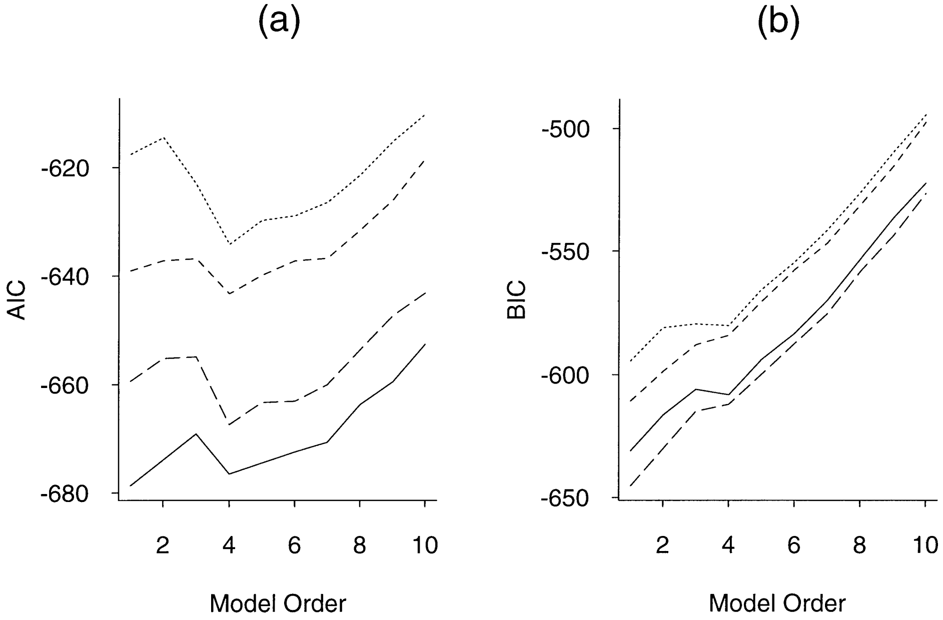

FIG. 1. (a) AIC vs common model order $p_1 = q_1 = p_2 = q_2$ for model (2), including SOI and a linear trend (solid line), SOI with no linear trend (long dashes), linear trend with no SOI (short dashes), and no trend or SOI component (dotted). (b) Same plot with but BIC instead of AIC.

the linear and SOI terms, and also for the correlation coefficient ρ—in our analyses, all of these terms are statistically significant and are therefore included in subsequent analyses.

c. Model (1) or model (2)?

All of the analyses discussed so far could equally well be performed using either (1) or (2) as the basic form of the model [though the actual results reported above have all been for model (2)]. We can use the NLLH as a means of comparing the results under both forms of model, assuming that the other choices (which covariates to include, the orders of the autoregressive terms, and the inclusion of ρ) are all the same. For the model with linear trend only, the NLLH values for models (1) and (2) are virtually identical. For the model with both linear trend and SOI terms, model (2) has a significantly lower value of NLLH (-363.3 against -356.9). This implies that model (2) is preferable if SOI is included.

For the model with linear trend only, $p_1 = p_2 = q_1 = q_2 = 4$, and including ρ, the estimated trend coefficients in (2) are $\hat{\beta}_{11} = 0.0067$ (standard error 0.0013), $\hat{\beta}_{21} = 0.0067$ (standard error 0.0009), in units of °C yr^{-1}. These values differ slightly from those obtained by direct linear regression without any autoregressive terms (0.0058 and 0.0064, respectively), which shows that the omission of autoregressive terms can bias the estimates of trends. In model (1), however, the estimates are completely different: $\hat{\beta}_1 = 0.0011$ (standard error 0.0010) and $\hat{\beta}_2 = 0.0026$ (standard error 0.0008), which do not have any direct interpretation as a global warming trend. It can, in fact, be shown that these two sets of estimates are consistent—if model (2) is rewritten in the form of model (1) with all error terms omitted, and all the parameters replaced by their numerical estimates, we indeed get the values of β_1 and β_2 just stated—but the lack of direct intepretation of these parameters as trends in model (1) is a practical reason for preferring model (2). In models with additional regression terms such as SOI, it appears that model (2) fits better than model (1).

d. Inclusion of cross-correlation terms

A considerable part of the paper by Kaufmann and Stern (1997) is concerned with the significance of the cross-correlation terms $\{\gamma_{2j}\}$ and $\{\delta_{1j}\}$, which represent, respectively, a north to south directional dependence, and a south to north dependence. In their analysis of model (1) with simple linear trend, they claimed that the north to south dependence is not statistically significant, but the south to north dependence is.

Within the models of form (2) with linear trends and model order 4, but without including SOI, we can consider four different model types, with associated negative log likelihood values:

(a) $p_1 = q_2 = p_2 = q_2 = 4$ (dependence in both directions), NLLH $= -344.6$;
(b) $p_1 = p_2 = q_2 = 4, q_1 = 0$ (north to south dependence only), NLLH $= -339.6$;
(c) $p_1 = q_1 = q_2 = 4, p_2 = 0$ (south to north dependence only), NLLH $= -342.1$;

(d) $p_1 = q_2 = 4$, $q_1 = p_2 = 0$ (no interhemispheric dependence), NLLH = −337.3.

From the NLLH values, we can see that the ranking order of the models is (d) → (b) → (c) → (a), with model (a) (the one with both directions of dependence) fitting best. To determine the statistical significance of these differences, we perform a series of tests using the χ^2 test described earlier. This test shows, for example, that model (c) is significantly superior to model (d) at the 5% level (the actual *p* value is 0.045). Similarly, (a) is superior to (b) (*p* value 0.038). In other words, the south to north dependence is significant whether or not the north to south dependence is also included in the model. However, (a) is not significantly superior to (c) (*p* value 0.29) and (b) is not significantly superior to (d) (*p* value 0.33), or in other words, the north to south dependence is not significant. These results agree qualitatively with those of Kaufmann and Stern.

If the same tests are repeated for the models including SOI, however, the results are a little different: the *p* value for (c) versus (d) is now 0.071, and that for (a) versus (b) is 0.085. Although we would not wish to overstate the importance of these slight changes in *p* values, it is noticeable that the changes are in the direction of larger *p* values (i.e., less significant results), and indeed, that the interhemispheric terms are no longer significant at 5% level. It should be noted in passing that Kaufmann and Stern (1997) also discussed whether ENSO could be an explanatory factor but they did not perform a direct test as we have done here.

e. Further remarks and summary

With the interhemispheric dependence terms omitted, we can repeat the order-determination analysis given earlier, including the linear trend and SOI terms. Once again the main choice is between model orders 1 and 4 with AIC preferring model order 4 and BIC preferring model order 1. A hypothesis test between the two models, using a likelihood ratio statistic, results in a *p* value of 0.055, indicating that model order 1 is just accepted at the 5% level of significant (since $0.055 > 0.050$). Since at this stage of the analysis it is not clear which is the better model order, we retain both model orders in subsequent analyses.

In summary, for the models including just linear trend and SOI as covariates, we find that both terms are significant, and the optimal model order is either 1 or 4. We find ρ to be statistically significant and, in comparing models (1) and (2), our preference is for model (2) both because of ease of interpretation of the parameters and because, in the model with SOI, it provides a better fit to the data as determined by NLLH. In considering the influence of cross correlations, we find that cross correlations are not significant in either direction, provided SOI is included in the model. We also found that the underlying linear trends (0.67°C in each hemisphere) are somewhat higher than a naïve analysis, ignoring the autoregressive structure of the data, would imply.

3. Comparison of different forcing hypotheses

We now consider possible alternative models in which the linear trend hypothesis of the previous section is replaced by the hemispheric-mean temperature responses to a range of forcing factors for different values of the climate sensitivity $\Delta T_{2\times}$. Kaufmann and Stern (1997) also included in their statistical models a variety of anthropogenically influenced terms, such as the radiative forcings of carbon dioxide, methane, chlorofluorocarbons (CFCs) and tropospheric sulfate aerosols, and also solar activity. Our approach differs from theirs by using temperature responses based on physical models incorporating different combinations of anthropogenic forcing factors, rather than using anthropogenic forcings directly as covariates. We believe that this is a more reasonable approach because the response to anthropogenic forcing is typically nonlinear and therefore may not be captured accurately by a linear statistical model using forcings as predictor variables.

We refer to each of the forcing cases as a specific ''model,'' although we note that each response time series was generated by a single (physical) model based on Wigley and Raper (1992), as modified in Raper et al. (1996). The central idea of our analysis is to use the goodness of fit of the time series models under various trend terms generated by different forcing hypotheses as a measure of the plausibility of those hypotheses in explaining climatic data. We consider six forcing models A–F and four different values (for each forcing model) of $\Delta T_{2\times}$, as described in Table 1.

All anthropogenic forcing factors thought to be important are considered, but only solar forcing is considered as a possible natural forcing factor. The anthropogenic forcings used are the best-estimate values employed in the IPCC Second Assessment Report (Kattenberg et al. 1996), and an additional case with lower sulfate aerosol forcing. For solar forcing, we initially examined results based on the irradiance reconstructions of Hoyt and Schatten (1993, corrected and updated) and Lean et al. (1995), but used only the former for our more detailed analyses. Table 1 summarizes the different forcing cases.

Table 1 also gives, for each case, the best-fit climate sensitivity and root-mean-square error (rmse) determined by minimizing, over 1899–1998, the rmse between the climate model output for different sensitivities and raw observed global-mean temperatures [as described in Wigley et al. (1997)]. These results by themselves are of some interest: they show that the inclusion of solar forcing improves the fit (cf. Wigley et al. 1997; Stott et al. 2000); and that the higher sulfate aerosol cases give a better fit—issues that will arise again later. They also show that the fits obtained using the Hoyt and Schatten irradiance data are better than those ob-

Table 1. Summary of models used to define deterministic trend terms.

Forcing case[a]	GHGs and biomass aerosols[b]	Sulfate aerosols	Solar[a]	Optimum[c] $\Delta T_{2\times}$ (°C)	Rmse[d] (°C)
A: GHGs	Yes	No	No	1.36	0.143
B: Anthropogenic[g]	Yes	Best[e]	No	11.90	0.129
C: Low SO_4	Yes	Low[f]	No	3.19	0.135
D: Anthropogenic[g] + Hoyt	Yes	Best[e]	Yes	4.16	0.114
Anthropogenic[g] + Lean	Yes	Best[e]	Yes	3.99	0.124
E: Low SO_4 + Hoyt	Yes	Low[f]	Yes	2.31	0.119
Low SO_4 + Lean	Yes	Low[f]	Yes	2.10	0.129
F: Hoyt alone	No	No	Yes	15.41	0.124
Lean alone	No	No	Yes	9.09	0.162

[a] Hoyt refers to the irradiance reconstruction of Hoyt and Schatten (1993); Lean refers to the construction by Lean et al. (1995).

[b] Forcing values as used by IPCC (Kattenberg et al. 1996).

[c] Based on the period 1899–1998. Note that the optimum climate sensitivities quoted in Wigley et al. (1997) were based on a different time period.

[d] Root-mean-square error of the difference between modeled and observed global-mean temperatures for the optimum sensitivity.

[e] The 1990 direct forcing -0.3 W m^{-2}, indirect forcing -0.8 W m^{-2} (Kattenberg et al. 1996).

[f] The 1990 direct forcing -0.3 W m^{-2}, indirect forcing -0.4 W m^{-2}.

[g] Includes GHGs and sulfate aerosols.

tained with the Lean et al. data. This, however, should not be taken as an endorsement of one irradiance reconstruction over the other, an issue that is more properly judged on the basis of the reconstruction methods used. Which reconstruction we employ in our analysis is somewhat arbitrary, since both would serve adequately to illustrate the statistical methods that are the primary focus of this paper. We use the Hoyt and Schatten data, since these provide a clearer separation of the different statistical models that we compare.

Beyond solar forcing, the next most important factor in the natural forcing category is likely to be volcanic forcing (Robock 2000; Stott et al. 2000), which we will investigate further in a later paper. When volcanic forcing is considered in the physical model used to define the deterministic trend terms, the goodness of fit (at the global-mean level) is degraded, possibly reflecting uncertainties in volcanic forcing estimates and/or deficiencies in the physical model. Given this problem, a comprehensive consideration of volcanic forcing effects is not possible in the present paper. The four values of $\Delta T_{2\times}$ used for each model were 1.5°C (labeled 1), 2.5°C (labeled 2), 4.5°C (labeled 3) and the abovementioned optimal values (labeled 4; see Table 1).

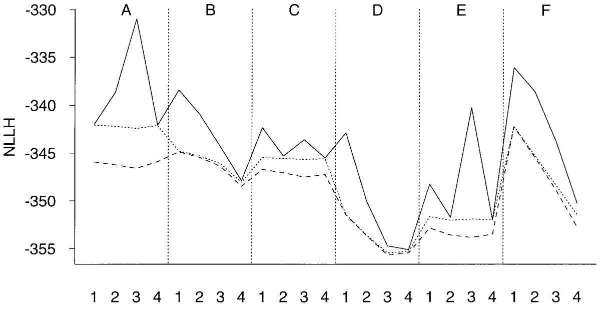

Fig. 2. Comparisons of NLLH under 24 combinations of forcing model (A–F, see Table 1) and climate sensitivity (1–4), and under three assumptions about the trend coefficients β_{11} and β_{21}. Solid lines: $\beta_{11} = \beta_{21} = 1$. Dotted lines: estimated coefficients assuming $\beta_{11} = \beta_{21}$. Dashed lines: estimated coefficients allowing β_{11} and β_{21} to be different. Model fits are for model (2) with SOI signal and $p_1 = q_2 = 1$, $p_2 = q_1 = 0$. The four model sensitivities are 1.5° [model (1)], 2.5° [model (2)], 4.5°C [model (3)], and optimized sensitivity [model (4)]. Lower (more negative) values indicate a better fit.

For each combination of forcing model and $\Delta T_{2\times}$, we use the climate model's hemispheric-mean temperature output to define regressors x_{t1} and y_{t1}, and then fit these to the observed data using model (2). We do this for each of three submodels using different assumptions about the regression coefficients β_{11} and β_{21}: (i) $\beta_{11} = \beta_{21} = 1$, (ii) β_{11} and β_{21} estimated assuming $\beta_{11} = \beta_{21}$, and (iii) β_{11} and β_{21} estimated with no constraints. In many ways, the most interesting comparison is (i), since this corresponds to the hypothesis that the time series model identifies the trend with no need for any scaling adjustment to the climate model output. Assumption (ii) implies that the climate model is off by a constant scaling factor that is the same for the two hemispheres. Assumption (iii) is the worst because it requires separate adjustment for each hemisphere. Our ideal conclusion would be if assumption (i) held without any need for adjustment. In addition, for each of the models, we assume SOI terms are included as x_{t2} and y_{t2}, as in section 2.

In Fig. 2, the NLLH values are plotted for each of the 24 combinations of forcing model and climate sensitivity, for each of the submodels (i)–(iii). Within each of the three submodels, the 24 NLLH values are based on the same number of estimated parameters and are therefore directly comparable. These plots are based on $p_1 = q_2 = 1$, $p_2 = q_1 = 0$, since the earlier discussion showed that a first-order autoregressive model appears adequate to fit the data when SOI is included. Similar results were, however, obtained for a fourth-order model, and in later discussion (Fig. 3) we shall directly compare results based on first-order and fourth-order autoregressive time series models.

From the plots, it appears that models D3 and D4 are the best fitting, and also that for those two models, though not for many of the others, submodel (i) is as good as (ii) or (iii). For D4, under submodel (ii), the estimated common value of β_{11} and β_{21} is 0.95, with a standard error of 0.08. Under submodel (iii), which allows β_{11} and β_{21} to be different, β_{11} is estimated as 0.91 with a standard error 0.10, and β_{21} is estimated as 0.99 with standard error 0.10 (Table 2, row 4). None of these three parameter estimates is significantly different from 1. These results indicate that the hemispheric differential in the D4 climate model–based predictors is consistent with the observations.

In general, the numerical values of the regression (i.e., trend) coefficients are consistent with our a priori expectations for the different forcing models, A–F. This is especially so for sensitivity case 4 [i.e., where the x_{t1} and y_{t1} predictor variables in Eq. (2) are defined using sensitivities optimized using global means; see Table 1]. To demonstrate this consistency, we use submodel (iii), in which the external-forcing regression coefficients (β_{11} and β_{21}) are unconstrained. These results are shown in Table 2.

For the external-forcing regression coefficients, the average of the two hemispheric values is close to 1 in all cases. This is as expected because the predictors, x_{t1} and y_{t1}, have already been optimized in a global-mean sense. The average values are generally less than 1 for two reasons: because some of the overall warming trend is captured by the ENSO terms in the regression, and because the global-mean temperature is not the arithmetic mean of the hemispheric values but the area-weighted mean. Since coverage is greater in the NH, the average regression coefficient must be ‘‘biased’’ toward the NH values, which are lower, relative to what would be obtained from a global-mean analysis.

The individual SH and NH regression coefficients, however, differ markedly from 1 in all cases except model D, with the SH coefficient always greater than the NH coefficient (i.e., $\beta_{21}/\beta_{11} > 1$; see Table 2). The ratio $\beta_{21}/\beta_{11} > 1$ implies that the NH predictor variable (x_{t1}) values used were too large compared with the SH predictor values (y_{t1}).

To try to explain this, note that in each forcing model the NH and SH predictor series are different. The observed hemispheric-mean temperature time series also differ; and the regression coefficients reflect differences between the observed and predictor variable NH to SH differentials. Predictor variable differences arise from both radiative forcing and climate response differences. In general, the forcing differs in each hemisphere (except for model F) because of differences in sulfate aerosol forcing (more negative in the NH) and tropospheric ozone forcing (more positive in the NH); see Kattenberg et al. (1996) and Raper et al. (1996). The corresponding

TABLE 2. Regression coefficients for different forcing models, with optimized climate sensitivity and unconstrained external forcing regression coefficients. Here β_{11}, β_{21} are the coefficients of external forcing for NH, SH, respectively; $\overline{\beta} = (\beta_{11} + \beta_{21})/2$. Also, β_{12}, β_{22} are the SOI components for NH, SH, respectively. Finally, σ_1 and σ_2 are the std dev of the residuals for the NH and SH components, respectively. Solar factors are ‘‘Hoyt’’ as defined in Table 1.

	External forcing				SOI terms		Residual std dev	
Forcing case	β_{11}	β_{21}	$\overline{\beta}$	β_{21}/β_{11}	β_{12}	β_{22}	σ_1	σ_2
A: GHGs	0.654	1.331	0.992	2.034	−0.0651	−0.0654	0.1235	0.0866
B: Anthropogenic	0.838	1.002	0.920	1.196	−0.0638	−0.0653	0.1211	0.0860
C: Low SO_4	0.741	1.112	0.927	1.502	−0.0645	−0.0653	0.1222	0.0863
D: Anthropogenic + Hoyt	0.914	0.993	0.953	1.087	−0.0616	−0.0655	0.1146	0.0842
E: Low SO_4 + Hoyt	0.837	1.104	0.920	1.319	−0.0627	−0.0651	0.1176	0.0837
F: Hoyt alone	0.777	1.034	0.906	1.331	−0.0681	−0.0677	0.1147	0.0864

responses are further affected by hemispheric differences in both the climate sensitivity and in lag effects due to oceanic thermal inertia. In the NH, the climate sensitivity is slightly larger than in the SH (because of a greater land area and greater sensitivity over land than ocean) and the response is more rapid (also a result of greater land area). The net effect on the temperature predictor variables may be characterized by their relative SH/NH temperature changes (e.g., as evidenced by the linear trends) over 1900–96. The ratios are A, 0.62; B, 0.87; C, 0.76; D, 0.89; E, 0.80; and F, 0.76. Note that the NH change exceeds the SH change in all cases, even for the higher sulfate aerosol cases. Models B and D (with highest sulfate forcing) have the highest SH/NH ratios.

It can be seen now that the smallest values of the SH/NH predictor variable change ratios (A, C, E, and F) correspond to the largest values of the SH/NH regression coefficient ratios (Table 2). In these cases, the NH predictor changes are too large relative to the observed temperature changes. The two other cases (B and D) where the SH/NH predictor variable change ratio is larger (around 0.9) are the two cases where the regression coefficient ratio (β_{21}/β_{11}) is closest to unity, the value expected if the hemispheric predictor time series were correct. Comparing B and D, it can be seen that the inclusion of solar forcing (case D) improves the regression coefficient ratio. Additionally, as both of these cases have relatively large sulfate aerosol forcing, these results imply that sulfate aerosol forcing is necessary in order to explain the observed temperature changes at the hemispheric level, and that the larger aerosol forcing case is preferred to the smaller aerosol forcing case. Santer et al. (1996) and Wigley et al. (1997) came to similar conclusions.

For the SOI predictor, the negative coefficient (Table 2) indicates that negative values of the SOI (warm events in the eastern equatorial Pacific) are associated with globally anomalous warmth (cf. Jones 1989). The SOI ‘‘sensitivity’’ is virtually independent of the assumed external forcing, largely because it is determined by higher-frequency variability than is captured by any of the other predictors, and the sensitivity is almost the same in each hemisphere. The values are all highly statistically significant. Slightly lower, but still highly significant, values arise in the fourth-order models, because the higher-order autoregressive terms are able to ‘‘explain’’ part of the ENSO-related variability.

The final results shown in Table 2 are those for the standard deviations of the residuals, σ_1 and σ_2. It is clear from the table that cases with both solar and anthropogenic forcing included have smaller residual standard deviations, implying that solar forcing helps in explaining observed variations in hemispheric-mean temperature. Also, the residual standard deviations are substantially smaller in the SH. These residuals should represent the effects of internally generated variability in the observations if all external forcing factors were correctly included in the model. We know that this is not the case, since volcanic effects have not been considered. Nevertheless, the residuals are similar to the internally generated variability produced by unforced control runs with coupled ocean–atmosphere general circulation models, most of which have NH variability greater than SH variability. For example, the Geophysical Fluid Dynamics Laboratory (GFDL) model (Manabe and Stouffer 1994) has NH and SH standard deviations of 0.13° and 0.10°C, respectively. These are values for complete hemispheric coverage, and slightly different values would be obtained if the observed coverage mask were applied. Since there are quite large differences between models, knowledge of the background level of natural variability is subject to considerable uncertainty. Nevertheless, the general similarity between our residuals and coupled model results provides a reassuring consistency check for the regression model.

The above results support our a priori selection of model D as the ‘‘best’’ model among those considered. In terms of interhemispheric temperature differences, it appears that sulfate aerosol forcing similar to that used as ‘‘best guess’’ by IPCC (Kattenberg et al. 1996) is required (cases B and D). The standard deviations of the residuals are lower for case D than for the other models, but it is difficult to interpret this as a quantitative measure of the extent to which model D fits better than the others. In the next section, we extend this analysis using a Bayes factor approach, to try to quantify to what extent the NLLH results in Fig. 2 can really be considered evidence either for or against the different climate forcing models.

4. Model selection using Bayes factors

In this section, we suggest an alternative interpretation of the model-fitting results in terms of Bayes factors (Kass and Raftery 1995). Suppose we have M different models to choose from. Suppose that the mth model ($1 \le m \le M$) defines a probability density function $f_m(y;\, \theta_m)$ for observed data y in terms of model parameter θ_m. Suppose that the prior probability that model m is correct is $\Pi(m)$, and that conditional on model m being correct, the prior density of θ_m is $\pi_m(\theta_m)$. Then the posterior probability that model m is correct, given the data y, is

$$\Pi(m|y) = \frac{\Pi(m) \displaystyle\int f_m(y;\, \theta_m)\pi_m(\theta_m)\, d\theta_m}{\displaystyle\sum_{m'} \Pi(m') \int f_{m'}(y;\, \theta_{m'})\pi_{m'}(\theta_{m'})\, d\theta_{m'}}. \quad (6)$$

The denominator of (6) is the sum over all alternative models m'; this ensures that (6) defines a proper probability distribution over all M models (proper in the sense that the probabilities sum to 1). The prior prob-

TABLE 3. Jeffreys's table of interpretation of Bayes factors, adapted from Kass and Raftery (1995).

Value of $B(m; m')$	Value of $\log_{10} B(m; m')$	Strength of evidence against model m'
1–3.2	0–0.5	Barely worth a mention
3.2–10	0.5–1	Substantial
10–100	1–2	Strong
>100	>2	Decisive

abilities for the different models, $\Pi(m)$, must be specified, and implicit in the whole formulation is the assumption that one of the M models is correct. The latter assumption is questionable in the present context, since it is obvious that many other models besides those in Table 1 could have been considered.

An alternative formulation, however, is to rewrite (6) in the form

$$\frac{\Pi(m\,|\,y)}{\Pi(m'\,|\,y)} = \frac{\Pi(m)}{\Pi(m')} \times \frac{\displaystyle\int f_m(y;\,\theta_m)\pi_m(\theta_m)\,d\theta_m}{\displaystyle\int f_{m'}(y;\,\theta_{m'})\pi_{m'}(\theta_{m'})\,d\theta_{m'}} \qquad (7)$$

for the comparison of two given models m and m'. The advantage of this approach is that (7) separates out the influence of the prior probabilities of the models (the Π factors on the right-hand side) from the likelihood components (the ratio of integrals). If we ignore the Π components in (7), the ratio of integrals, which we shall denote by $B(m; m')$, is called the *Bayes factor* of model m relative to model m' and represents the relative ''weight of evidence'' of the two models. This therefore represents a direct means of comparing two models without any prior assumption that one of them must be correct.

Bayes factors were first popularized in the classic treatise of Jeffreys (1961), who gave the interpretation in Table 3, as slightly modified by Kass and Raftery (1995).

In practice, we have assumed the prior densities $\pi_m(\theta_m)$ to be constant and have evaluated the integrals in (7) using Laplace's integral approximation (Kass and Raftery 1995),

$$\int f_m(y;\,\theta_m)\pi_m(\theta_m)\,d\theta_m \approx (2\pi)^{p_m/2}|\mathbf{V}_m|^{1/2} f_m(y;\,\hat{\theta}_m), \qquad (8)$$

where $\hat{\theta}_m$ is the MLE under model m, p_m is the dimension of the model, and $\mathbf{V}_m$ is the inverse observed information matrix under model m.

Taking model D4 as a reference model, since this is the best fitting according to NLLH, the Bayes factor for model D4 relative to each of the other models is plotted in Fig. 3. This is for $\beta_{11} = \beta_{21} = 1$, and we have shown both the case with $p_1 = q_2 = 1$ (the version of the model used to compute Fig. 2) and the alternative form of model with $p_1 = q_2 = 4$.

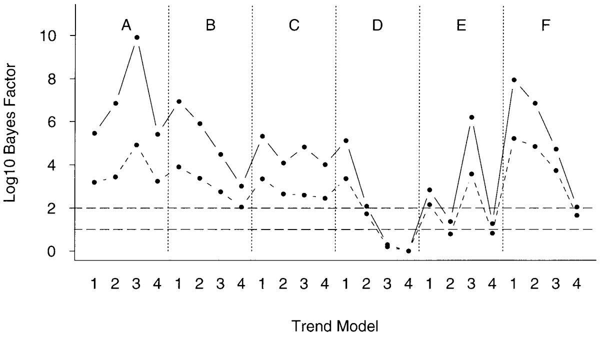

FIG. 3. Approximate $\log_{10}$ Bayes factors computed for 24 combinations of climate model and climate sensitivity (same as in Fig. 2), all computed relative to model D4 (for which the $\log_{10}$ Bayes factor is fixed at 0). All fits are for model (2) including SOI, with $p_2 = q_1 = 0$ and $\beta_{11} = \beta_{21} = 1$. Dashed lines: $p_1 = q_2 = 4$. Solid lines: $p_1 = q_2 = 1$. The horizontal dashed lines represent Bayes factor 10 (above which there is strong evidence against the alternative model, compared with D4, according to the Jeffreys interpretation) and Bayes factor 100 (the lower bound for decisive evidence against).

Consider first the results using the first-order model ($p_1 = q_2 = 1$), represented by solid lines in Fig. 3. According to the Jeffreys interpretation of Bayes factors, if we take D4 as our reference model, there is ''strong'' evidence against any model with a Bayes factor bigger than 10, and ''decisive'' evidence against any model with a Bayes factor bigger than 100. The horizontal lines on Fig. 3 delineate these two boundaries. Under this criterion, only four models, besides D4 itself, fail to be decisively worse than model D4. These are D3, E2, E4, and F4 (and F4 is borderline).

To give this result a climatological interpretation, we see that the best models are those (D and E) that include all three forcing components, that is, greenhouse gases, sulfate aerosols, and solar, with model D (high aerosol forcing) superior to model E (low forcing), though this comparison is not decisive. Moreover, both models perform best at near their optimum climate sensitivity values. In contrast, model F4 is also (just) competitive in terms of fit to the data, but using a highly unrealistic value of $\Delta T_{2\times}$, whereas a solar-only model with a more realistic $\Delta T_{2\times}$ (such as model F3) is very much inferior. Thus, we reject the ''solar forcing only'' model because it is either a much inferior fit to the data (model F3) or uses a physically meaningless value of $\Delta T_{2\times}$ (model F4). However, when combined with the anthropogenic and sulfate aerosol components, solar forcing *is* important, because either of models D and E is a vastly superior fit to the data than A, B, or C.

Thus, our overall conclusion is that a model that incorporates all three forms of forcing factors—greenhouse gases, aerosols, and solar terms—fits the observed data much better than any of the alternatives that do not include all those three factors.

The conclusions based on a fourth-order model ($p_1 = q_2 = 4$) are shown as dashed lines in Fig. 3. For these calculations, the Bayes factor calculation relative to model D4 has been repeated with the fourth-order time series model. Note that the comparisons here are among different combinations of the climate forcing, made separately for each of the two forms of time series model; we are not using Bayes factors to compare first-order time series models with fourth-order time series models, since formula (8) is problematic for comparing models of different dimensions.

In general, Bayes factors computed using the fourth-order models are smaller than those using the first-order models. This is to be expected: if we incorporate higher-order terms into the time series models, then the time series models on their own contain more degrees of freedom to explain the external forcing; consequently, the discrimination among different forms of external forcing, as explained by the Bayes factors, will be weaker if we adopt a fourth-order time series model than if we adopt a first-order model. The *qualitative* conclusions, however, are unchanged. The Bayes factors for models E4, F4, and F3, each computed with respect to D4, are 6.8, 46, and 5492. Thus, model E4 could still be entertained as an alternative to D4, model F4 is clearly inferior to D4 but still a possibly acceptable fit to the data, while the more physically realistic model F3 is decisively worse. All the Bayes factors derived from models A, B, and C are again decisively worse than those derived from models D and E. The climatological conclusions are the same as under the first-order model.

Overall, our conclusions strengthen those of Wigley et al. (1998), reinforcing the message that within a realistic range of $\Delta T_{2\times}$, the models involving a combination of anthropogenic and solar effects clearly outperform those based on either anthropogenic or solar effects on their own.

Fig. 4 shows a number of ''diagnostic'' plots, aimed at visually assessing the fit of the regression equation, where we have used model D4 for the anthropogenic model. Panels (a) and (b) show that the anthropogenic model traces the irregularities of the observed data considerably better than the straight line model, though it is clear that there are still some periods in the data, notably the 1940s and 1950s, where it is capable of improvement. The remaining panels of Fig. 4 show some standard statistical diagnostics based on residuals from the anthropogenic model fit. Panels (c) and (d) show residuals plotted against fitted values; (e) and (f) show residuals plotted against time; (g) and (h) show quantile–quantile (QQ) plots for the residuals against a normal distribution; the fact that these plots stay very close to the straight line of unit slope through the origin indicates a close fit to the normal distribution. The only concern in any of these plots is that in (d) [and to a lesser extent in (f)] the last few values of the residuals seem to show a downward trend, corresponding to several successive values in (b) where the observed value is below the fitted trend value. Given that these residuals lie well within the range of residuals established elsewhere in the plot, we do not believe these indicate anything wrong with the model.

Referees have raised some other questions of a diagnostic nature, concerning (a) heteroscedasticity and (b) multicollinearity. Heteroscedasticity refers to the residuals having nonconstant variance, of which there is absolutely no visual evidence in Fig. 4c–f, but as an additional check, we performed the Godfrey–Koenker test given by Wetherill (1986, p. 203), which produced test statistics of 0.26 for the NH data, 0.01 for the SH data, each nominally χ_1^2 under the null hypothesis of constant variance. Since both the test statistics are much smaller than 1 (mean of the χ_1^2 distribution), we conclude that there is no evidence at all for heteroscedasticity. Multicollinearity concerns the possibility of misleading regression coefficients because of correlations among the regressors. One of the most widely used tests is based on the singular value decomposition of the matrix of regressors (Belsley et al. 1980). In the present setting of a time series model in which some of the parameters enter nonlinearly, the Belsley–Kuh–Welsch procedure is not directly applicable, but we have adopted the following procedure that should be equivalent to it: after the maximum likelihood fit is completed and with $\mathbf{H}$ (the Hessian matrix of the negative log likelihood, which estimates the covariances of the maximum likelihood estimators), compute the eigenvalues of $\mathbf{H}$, say $\lambda_1 \geq \cdots \geq \lambda_p$ where p is the number of parameters, and let the kth condition number be $\nu_k = \sqrt{\lambda_k/\lambda_p}$. In the case of a linear regression, the matrix $\mathbf{H}$ is equivalent to $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ and this would be an equivalent definition to that of Belsley et al. For the anthropogenic model with AR(1) residuals, the largest condition index according to this definition is 16.6. Belsley et al. (1980, p. 105) suggest that a condition index of 5–10 is associated with weak dependencies and that ''moderate to strong relations are associated with condition indices of 30 to 100.'' On the basis of this, we conclude that there is no problem with multicollinearity in this analysis.

Finally, in this section we return to one of the questions considered at the beginning: what is the optimum lag of the SOI signal? This question can also be considered within our overall modeling framework, by fitting different covariates corresponding to different lags and choosing the best model on the basis of NLLH. In Fig. 5, we show the results of this, for our final preferred model (model D4 including SOI, $p_1 = q_2 = 1$, $p_2 = q_1 = 0$). The results show a sharp improvement in the fit of the model over lags 0–4 months, and a sharp decline in lags of greater than 7 months, with lags between 4 and 7 months virtually equivalent. This justifies our earlier decision to base the SOI analysis on a 6-month lag.

5. Conclusions

Section 2 revisited the question of bivariate time series models for hemispheric data, re-examining several of the issues considered by Kaufmann and Stern (1997).

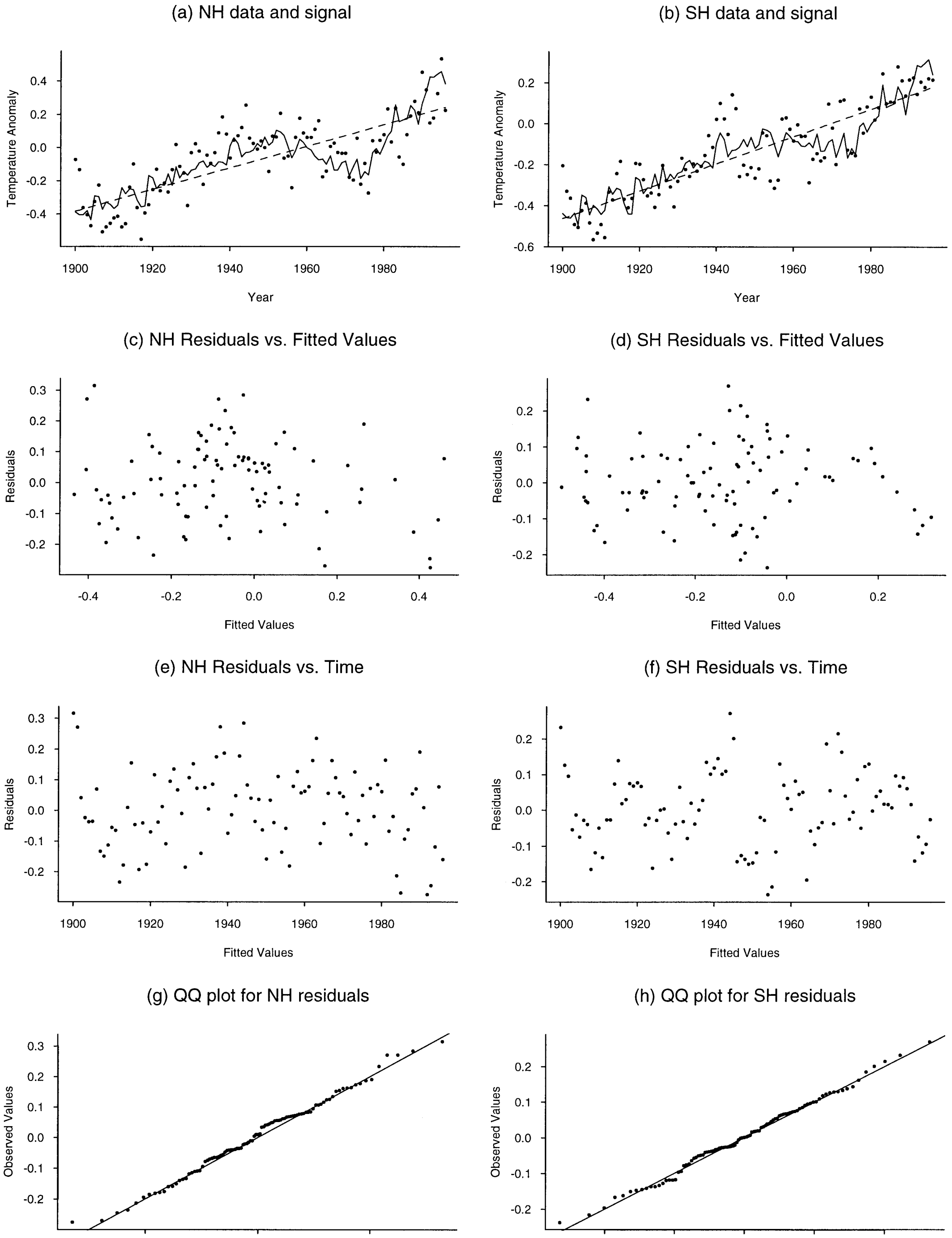

FIG. 4. Diagnostic plots. (a),(b) Raw data with best fitting straight-line and model-based trends. (c),(d) Residuals (from regression on anthropogenic + solar model) vs fitted values. (e),(f) Residuals vs time. (g),(h) QQ plots of residuals.

Like Kaufmann and Stern, we found evidence for a south to north dependency in the hemispheric mean temperatures, but the evidence for this is weaker if SOI terms are included in the model. We also established that in this case, it appears that the time series dependence can be adequately represented by a first-order model, provided SOI is included. This contrasts both with the results of Kaufmann and Stern and with our own models when SOI is omitted, both of which pointed toward a fourth-order model. We also find a preference

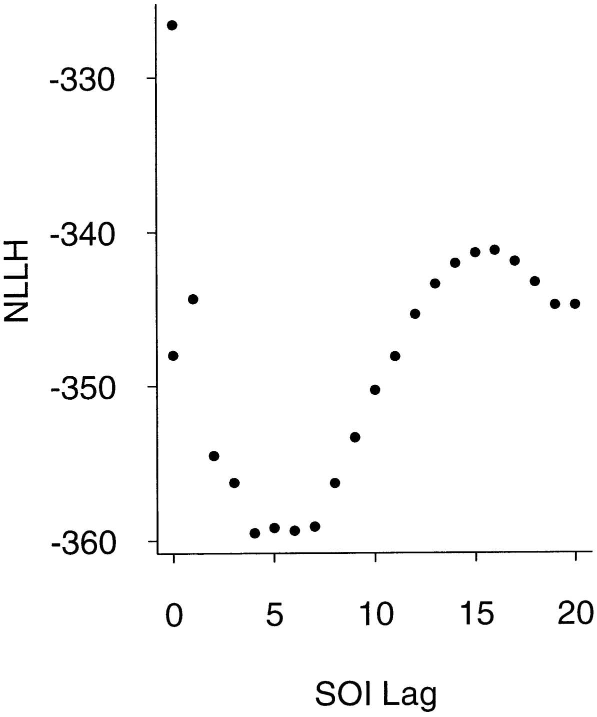

Fig. 5. NLLH values for model with various lags of SOI vs the lag in months. Results are for model (2) with $p_1 = q_2 = 1$, $p_2 = q_1 = 0$, based on model D4 for the forcing component.

for models of form (2) over those of form (1), and for including a cross correlation between ϵ_{1t} and ϵ_{2t}.

The results regarding SOI are seemingly at variance not only with those in Kaufmann and Stern (1997), but also with Wigley et al. (1998). However, neither of the earlier two papers tested for the SOI component directly by including it as an additional term in the statistical model. The contrast with the earlier results arises because of the use of a more sensitive statistical test, which allows us to reassess this component of the model fit. We include SOI in our subsequent comparisons of climate models.

The south to north dependency was shown to be significant by Kaufmann and Stern (1997), and our results agree with theirs in the case of a model without SOI. With SOI, however, the south to north dependence is no longer significant at the 5% level. This result in itself may not be too important—for example, it is quite plausible that with a few more years' data, the south to north dependence would reappear as a significant component—but it is relevant to the physical interpretation of the result that the observed south to north dependence is partly explained by the SOI. In this respect, our conclusions differ from Kaufmann and Stern (1997). While there may be an anthropogenic influence on low-frequency variations in ENSO (Trenberth and Hoar 1997; Timmerman et al. 1999), we judge that any anthropogenic component would be sufficiently small that the primary ENSO variations that we see are due to natural fluctuations. The claim that the south–north dependency is solely a result of anthropogenic climatic influence is not supported by our analysis.

In sections 3 and 4, we have extended the analysis of section 2 to include different forms of forcing terms generated by climate models. Various combinations of greenhouse gas, sulfate aerosol, and solar terms were considered, and also different values of the climate sensitivity $\Delta T_{2\times}$. The main conclusion here is that a model that includes all three forms of forcing term is clearly better than any other model. The solar forcing model alone is apparently competitive, but only with a completely unrealistic climate sensitivity parameter (15°C). Climate sensitivities within the normally accepted range (1.5°–4.5°C) do not produce an adequate fit under this model. All of these conclusions are based on an approximate Bayes factor method of comparison, and essentially the same results are obtained using fourth-order time series models as under first-order time series models.

For further work along these lines, we anticipate that alternative forms of Bayesian model comparison might be considered. Kass and Raftery (1995) reviewed a number of alternative approaches to Bayes factors; Hasselmann (1998), Levine and Berliner (1999), and Berliner et al. (2000) have also considered the application of Bayesian statistical methods to climatological data. Finally, we would consider it of great interest to extend the methodology to spatiotemporal data; as such, it would provide a powerful alternative to the pattern correlation and fingerprint detection methods that have been developed by Hegerl et al. (1996a,b), Santer et al. (1995, 1996), and Allen and Tett (1999), among others.

Acknowledgments. RLS's work was carried out primarily during a series of visits to the Geophysical Statistics Project at the National Center for Atmospheric Research (NSF Grant DMS-9312686), and was also supported by NSF Grant DMS-9705166. TMLW was supported by the NOAA Office of Climate Programs (Climate Change Data and Detection) Grant NA87GP0105 and U.S. Department of Energy (DOE) Grant DE-FG02-98ER62601. BDS's work was performed under the auspices of the Environmental Sciences Division of DOE, Contract W-7405-ENG-48.

REFERENCES

Allen, M. R., and S. F. B. Tett, 1999: Checking for model consistency in optimal fingerprinting. *Climate Dyn.,* **15,** 419–434.

Barnett, T. P., and M. E. Schlesinger, 1987: Detecting changes in global climate induced by greenhouse gases. *J. Geophys. Res.,* **92,** 14 772–14 780.

Belsley, D. A., E. Kuh, and R. E. Welsch, 1980: *Regression Diagnostics.* Wiley, 292 pp.

Berliner, L. M., R. A. Levine, and D. J. Shea, 2000: Bayesian climate change assessment. *J. Climate,* **13,** 3805–3820.

Bloomfield, P., and D. Nychka, 1992: Climate spectra and detecting climate change. *Climatic Change,* **21,** 275–288.

Brockwell, P. J., and R. A. Davis, 1991: *Time Series: Theory and Methods.* 2d ed. Springer-Verlag, 577 pp.

Cox, D. R., and D. V. Hinkley, 1974: *Theoretical Statistics.* Chapman and Hall, 512 pp.

Hasselmann, K., 1998: Conventional and Bayesian approach to climate-change detection and attribution. *Quart. J. Roy. Meteor. Soc.,* **124,** 2541–2565.

Hegerl, G., K. Hasselmann, U. Cubasch, J. F. B. Mitchell, E. Roeckner, R. Voss, and J. Waszkewitz, 1996a: On multi-fingerprint detection and attribution of greenhouse gas and aerosol forced climate change. MPI Rep. 207, 70 pp.

——, H. von Storch, K. Hasselmann, B. D. Santer, U. Cubasch, and P. D. Jones, 1996b: Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. *J. Climate,* **9,** 2281–2306.

Hoyt, D. V., and K. H. Schatten, 1993: A discussion of plausible solar irradiance variations, 1700–1992. *J. Geophys. Res.,* **98** (A11), 18 895–18 906.

Jeffreys, H., 1961: *Theory of Probability.* 3d ed. Oxford University Press, 459 pp.

Jones, P. D., 1989: The inflence of ENSO on global temperatures. *Climate Monit.,* **17,** 80–89.

Kass, R. E., and A. E. Raftery, 1995: Bayes factors. *J. Amer. Stat. Assoc.,* **90,** 773–795.

Kattenberg, A., and Coauthors, 1996: Climate model projections of future climate. *Climate Change 1995: The Science of Climate Change,* J. Houghton et al., Eds., Cambridge University Press, 285–357.

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

Können, G. P., P. D. Jones, M. H. Kaltofen, and R. J. Allan, 1998: Pre-1866 extensions of the Southern Oscillation index using early Indonesian and Tahitian meteorological readings. *J. Climate,* **11,** 2325–2339.

Lean, J., J. Beer, and R. Bradley, 1995: Reconstructions of solar irradiance since 1610: Implications for climate change. *Geophys. Res. Lett.,* **22,** 3195–3198.

Levine, R. A., and M. Berliner, 1999: Statistical principles for climate change studies. *J. Climate,* **12,** 564–574.

Manabe, S., and R. J. Stouffer, 1994: Multiple century response of a coupled ocean–atmosphere model to an increase of atmospheric carbon dioxide. *J. Climate,* **7,** 5–23.

Nicholls, N., G. V. Gruza, J. Jouzel, T. R. Karl, L. A. Ogallo, and D. E. Parker, 1996: Observed climate variability and change. *Climate Change 1995: The Science of Climate Change,* J. Houghton, et al., Eds., Cambridge University Press, 133–192.

Raper, S. C. B., T. M. L. Wigley, and R. A. Warrick, 1996: Global sea level rise: Past and future. *Sea-Level Rise and Coastal Subsidence: Causes, Consequences and Strategies,* J. D. Milliman and B. U. Haq, Eds., Kluwer Academic, 11–45.

Robock, A., 2000: Volcanic eruptions and climate. *Rev. Geophys.,* **38,** 191–219.

Santer, B. D., K. E. Taylor, T. M. L. Wigley, J. E. Penner, P. D. Jones, and U. Cubasch, 1995: Towards the detection and attribution of an anthropogenic effect on climate. *Climate Dyn.,* **12,** 77–100.

——, and Coauthors, 1996: A search for human influences on the thermal structure of the atmosphere. *Nature,* **382,** 39–46.

Stott, P. A., S. F. B. Tett, G. S. Jones, M. R. Allen, W. J. Ingram, and J. F. B. Mitchell, 2000: Attribution of twentieth century temperature change to natural and anthropogenic causes. *Climate Dyn.,* **17,** 1–21.

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

Timmermann, A., J. M. Oberhuber, A. Bacher, M. Esch, M. Latif, and E. Roeckner, 1999: Increased El Niño frequency in a climate model forced by future greenhouse warming. *Nature,* **398,** 694–696.

Tol, R. S. J., 1994: Greenhouse statistics—Time series analysis: Part II. *Theor. Appl. Climatol.,* **49,** 91–102.

Trenberth, K. E., and T. J. Hoar, 1997: El Niño and climate change. *Geophys. Res. Lett.,* **24,** 3057–3060.

Wetherill, G. B., 1986: *Regression Analysis with Applications.* Chapman and Hall, 311 pp.

Wigley, T. M. L., 2000: ENSO, volcanoes and record-breaking temperatures. *Geophys. Res. Lett.,* **27,** 4101–4104.

——, and S. C. B. Raper, 1992: Implications for climate and sea level of revised IPCC emissions scenarios. *Nature,* **357,** 293–300.

——, P. D. Jones, and S. C. B. Raper, 1997: The observed global warming record: What does it tell us? *Proc. Natl. Acad. Sci.,* **94,** 8314–8320.

——, R. L. Smith, and B. D. Santer, 1998: Anthropogenic influence on the autocorrelation function of hemispheric-mean temperatures. *Science,* **282,** 1676–1679.

Woodward, W. A., and H. L. Gray, 1993: Global warming and the problem of testing for a trend in time series analysis. *J. Climate,* **6,** 953–962.

——, and ——, 1995: Selecting a model for detecting the presence of a trend. *J. Climate,* **8,** 1929–1937.