Global Warming and the Problem of Testing for Trend in Time Series Data

WAYNE A. WOODWARD AND H. L. GRAY

Department of Statistical Science, Southern Methodist University, Dallas, Texas

(Manuscript received 14 October 1991, in final form 27 July 1992)

ABSTRACT

In recent years a number of statistical tests have been proposed for testing the hypothesis that global warming is occurring. The standard approach is to examine one or two of the more prominent global temperature datasets by letting $Y_t = a + bt + E_t$, where Y_t represents the temperature at time t, and E_t represents error from the trend line, and to test the hypothesis that $b = 0$. Several authors have applied these tests for trend to determine whether or not a significant long-term or deterministic trend exists, and have generally concluded that there is a significant deterministic trend in the data. However, we show that certain autoregressive–moving average (ARMA) models may also be very reasonable models for these data due to the random trends present in their realizations. In this paper, we provide simulation evidence to show that the tests for trend detect a deterministic trend in a relatively high percentage of realizations from a wide range of ARMA models, including those obtained for the temperature series, for which it is improper to forecast a trend to continue over more than a very short time period. Thus, we demonstrate that trend tests based on models such as $Y_t = a + bt + E_t$ for the purpose of prediction or inference concerning future behavior should be used with caution.

Of course, the projections that the warming trend will extend into the future are largely based on such factors as the buildup of atmospheric greenhouse gases. We have shown here, however, that based solely on the available temperature anomaly series, it is difficult to conclude that the trend will continue over any extended length of time.

1. Introduction

A common problem, which often has physical significance, is that of testing for the presence of a linear trend in data. The standard approach is to assume the model

$$Y_t = a + bt + E_t, \tag{1.1}$$

where Y_t represents the data at time t and E_t is the deviation of the data from a straight line. In this paper we make the assumption that E_t is a stationary, zero-mean process. A test of hypothesis is then often proposed for testing that $b = 0$. If this hypothesis is rejected at an appropriate significance level, then it is generally accepted that a linear trend is present. By the nature of the model, such a trend is said to be deterministic or long term.

There has been much recent interest in determining whether or not "global warming" is occurring. The issues involved in this area are very complex and involve the role of the so-called "greenhouse gases," clouds, the ocean, relationship to solar activity, etc. In order to determine the effects of these various factors and in order to provide evidence concerning a warming or lack of warming, climatologists have compiled "global" temperature data for approximately the past century from world weather records. Two of the more prominent datasets of this type are the series of Hansen and Lebedeff (1987, 1988) and the series given by Folland et al. (1990). These two datasets are shown in Fig. 1 where the temperatures are calculated by Hansen and Lebedeff as degrees Celsius deviation from the 1951–1980 average while Folland et al. use the 1950–1979 average. Throughout this paper we will refer to these two series as the Hansen and Lebedeff series and the IPCC series, respectively. The Hansen and Lebedeff series is predicated on land-based stations. The IPCC series incorporates land and marine data based on results obtained by Jones et al. (1991) along with a combination of land-based results from Jones (1988) and marine data from Bottomley et al. (1990). The Hansen and Lebedeff series covers the 108 years from 1880–1987 while the IPCC series extends from 1867–1990. Although there are significant differences between these two temperature series, the primary feature of each temperature series is the general appearance of an increasing trend over the time periods for which the data are compiled.

Of course the concept of "average global temperature" is a difficult one to define and is particularly difficult to measure. Problems such as varying station coverage and changing instrumentation cause difficul-

Corresponding author address: Dr. Wayne A. Woodward, Dept of Statistical Sci., Southern Methodist University, Dallas, TX 75275-0332.

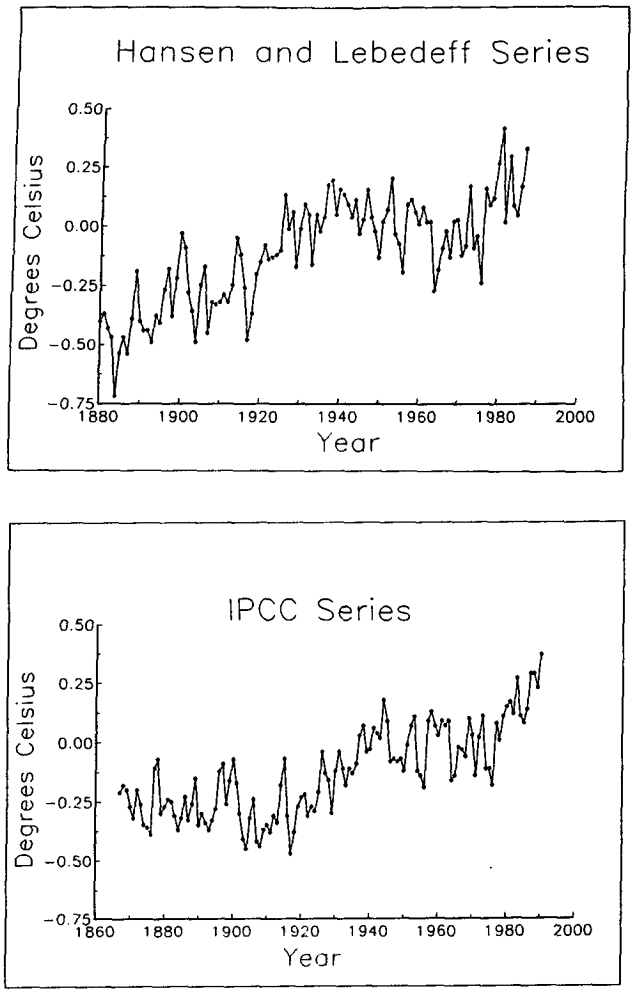


FIG. 1. Global temperature deviation datasets.

ties in the data handling. Additionally, the problem of optimally combining data from spatially and temporally related stations is a very difficult one. These issues have been recently investigated by Gunst et al. (1993). Nevertheless, much interest has been focused on these particular series, and statisticians have been working to determine whether or not the series, assumed to be valid measures of global average temperature deviation, provide evidence of an increasing temperature trend.

Researchers have used a variety of tests of the nature of the linear trend test we described in the first paragraph. In this paper we review many of those tests and apply them to the Hansen and Lebedeff and IPCC temperature datasets. In each case the hypothesis $b = 0$ (i.e., H_0: $b = 0$) is rejected. Before proceeding it is important to discuss the implications of such a finding. One of the goals of time series analysis is to provide forecasts beyond the time frame of the data, that is, to predict Y_{t_0+l} given data $Y_1, \ldots, Y_{t_0}$; consequently, one of the uses for time series models such as (1.1) is to provide equations for optimal forecasts. In the current setting, it is obvious that climate predictions are normally more complex than those obtained from the simple model of (1.1) and certainly involve factors other than the past history of the temperature data. However, the implication of adopting such a model as an appropriate model for slope detection and finding a significant slope is that if conditions do not change, then the fitted model suggests that temperature will continue to climb, roughly along the fitted line. In fact, this is the general form of the optimal forecasts from this model (see Cryer 1986). This is the sense in which the detected trend is called a long-term trend.

In this paper we investigate the performance of statistical tests for testing the hypothesis H_0: $b = 0$ in the model $Y_t = a + bt + E_t$. In particular it is shown that when there is a large correlation between successive values (but no deterministic trend), temporary trends will occur in the data and tests based on the model $Y_t = a + bt + E_t$ will very frequently conclude that a trend exists, whereas the best forecasts based on autoregressive–moving average (ARMA) processes do not support this conclusion.

Further, we fit ARMA models to the warming trend temperature datasets. In each case we see that the correlation structure of the data is not nearly strong enough for the ARMA-based forecasts to predict any continued increase, that is, to support the position of long-term trend. Moreover when data are generated from these ARMA models, we demonstrate that the aforementioned tests quite often incorrectly predict that a trend will continue, that is, the short-term trends in these realizations cause the tests to erroneously conclude $b \neq 0$ (and hence, treat them as if they were long-term trends). We hasten to add that we are not suggesting there is or is not a warming trend. We are simply pointing out that these statistical tests for testing H_0: $b = 0$ have little or no ability to distinguish between realizations from ARMA models with a high correlation between successive values and those from models of the form $Y_t = a + bt + E_t$. Consequently, the conclusion of a warming trend using the above-mentioned tests is clearly heavily dependent on the assumed model.

2. Testing for trend

a. *Regression-based models*

A basic approach to testing for trend in the data is a regression approach in which time, t, (in years) is taken as the independent variable and temperature deviation, Y_t, is taken as the dependent variable. Specifically, the model

$$Y_t = a + bt + E_t, \tag{2.1}$$

is assumed where the E_t are the residuals. Without loss of generality we will assume throughout that temper-

ature readings are taken at $t = 1, 2, \cdots, n$. The least-squares estimator

$$\hat{b} = \frac{\sum_{t=1}^{n} (t - \bar{t})Y_t}{\sum_{t=1}^{n} (t - \bar{t})^2}, \qquad (2.2)$$

was obtained for the series in question resulting in $\hat{b} = 0.00543$ for the Hansen and Lebedeff data and $\hat{b} = 0.00396$ for the IPCC data. (It should be noted that the two temperature series span somewhat different time periods.) Under the usual regression assumptions, that is, when the residuals are independent and normally distributed with mean zero and variance σ^2, the estimated standard error of $\hat{b}$ is given by

$$\widehat{SE}^{(1)}(\hat{b}) = \left[\frac{\sum_{t=1}^{n} (Y_t - \hat{a} - \hat{b}t)^2}{(n-2)\sum_{t=1}^{n} (t - \bar{t})^2} \right]^{1/2}$$

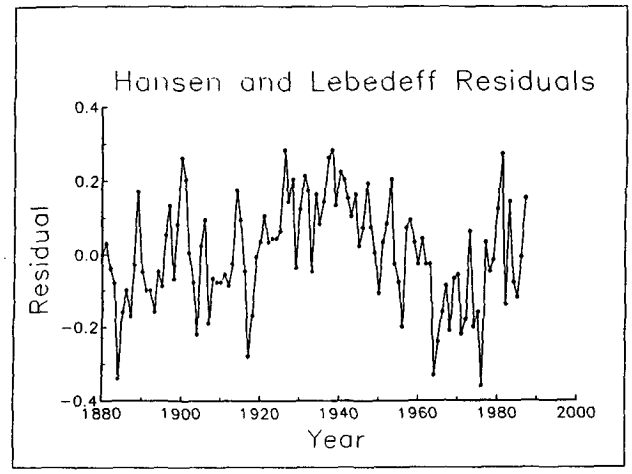
$$= \left[\frac{12 \sum_{t=1}^{n} (Y_t - \hat{a} - \hat{b}t)^2}{(n-2)n(n^2-1)} \right]^{1/2}, \qquad (2.3)$$

where $\hat{a} = \bar{Y} - \hat{b}\bar{t}$. The test of H_0: $b = 0$ is based on the fact that $\hat{b}/\widehat{SE}^{(1)}(\hat{b})$ is distributed as Student's t with $n - 2$ degrees of freedom when the null hypothesis is true. Hanson et al. (1989) use a two-phase regression analysis on Northern Hemisphere temperature data. They also used a nonparametric approach that does not assume normal residuals to test whether or not there was a significant temperature trend in data for the Northern Hemisphere. Their approach is equivalent to ranking independent and dependent variables separately and performing a regression analysis on the ranks. In Table 1 we give the ratios $\hat{b}/\widehat{SE}^{(1)}(\hat{b})$ for the original data and rank-based estimators for the two series. In each case we see that the results strongly indicate an increasing linear trend.

In Fig. 2 we show the residuals associated with the least-squares fits to the two temperature series. It is clear that in both cases there is noticeable autocorrelation among the residuals. Thus, the usual regression analysis results in Table 1 (either on the original data or on ranks) are not appropriate for the data under consideration.

TABLE 1. Ratios $\hat{b}/\widehat{SE}^{(1)}(\hat{b})$ based on regression approaches for global temperature series.

Series	$\hat{b}/\widehat{SE}^{(1)}(\hat{b})$	$\hat{b}_{Rank}/\widehat{SE}^{(1)}(\hat{b}_{Rank})$
IPCC	13.15	13.01
Hansen and Lebedeff	12.441	11.675



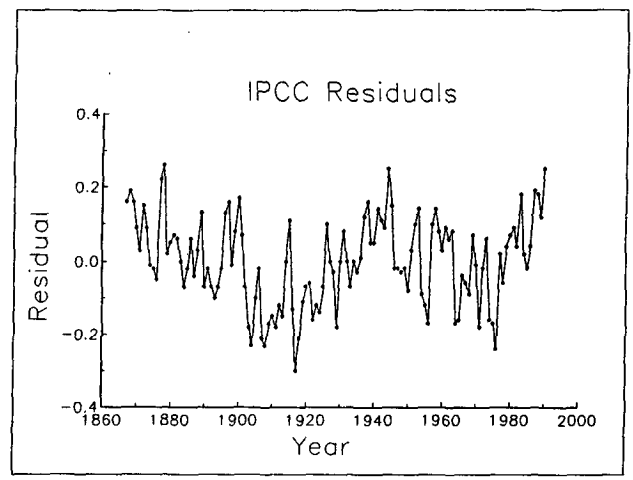


FIG. 2. Residuals of temperature deviation datasets from least-squares lines.

b. *Time series models with constant mean*

Closer examination of the temperature datasets in Fig. 1 suggests the need for time series approaches that take into consideration the correlation structure in the data. Time series analysis of such data involves determining models that describe the manner in which the series evolve in time, and then based on the models, calculating the forecasts of temperature deviations for future years. The key question of interest is whether these models assess the increasing tendency of the temperatures in the observed realizations of Fig. 1 to be of sufficient strength to predict that future temperatures will increase and if not, would tests of H_0: $b = 0$ still often conclude $b \neq 0$?

Box and Jenkins (1976) have popularized the ARIMA (autoregressive integrated moving average) models for describing a wide variety of time series behavior. These models are either stationary or are nonstationary

on the boundary of the stationary region. Letting $B^k Y_t = Y_{t-k}$, the ARMA(p, q) model is given by

$$\phi(B)(Y_t - \mu) = \theta(B)a_t, \tag{2.4}$$

where

$$\phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$$

$$\theta(B) = 1 - \theta_1 B - \cdots - \theta_q B^q,$$

with $\phi_p \neq 0$ and $\theta_q \neq 0$, where the polynomials $\phi(B)$ and $\theta(B)$ share no common factors, and where a_t is zero mean white noise with variance σ_a^2. An ARMA (p, q) process is stationary if and only if all the roots of the characteristic equation $\phi(r) = 0$ lie outside the unit circle. Best forecasts based on a stationary ARMA model eventually tend to $\bar{Y}$.

An interesting class of nonstationary ARMA models is that for which some of the roots of $\phi(r) = 0$ lie on the unit circle. Box and Jenkins (1976) refer to such a model as an ARIMA (p, d, q) model if $\phi(B) = \phi_s(B)(1 - B)^d$ where $\phi_s(B)$ is of order p and all of the roots of $\phi_s(r) = 0$ lie outside the unit circle. Forecasts of Y_{n+l} for $l \geq 1$ based on a realization $Y_1, \ldots, Y_n$ from the process $(1 - B)(Y_t - \mu) = a_t$ are simply the last observed value, that is, Y_n, for each l. Forecasts from an ARIMA$(p, 1, q)$ model $\phi_s(B)(1 - B)(Y_t - \mu) = \theta(B)a_t$ tend to a constant that is not usually equal to $\bar{Y}$. In Fig. 3 we show realizations from a model of this form. The nature of these realizations is discussed in section 3. A model with one root near +1 will have short-term forecasts which are relatively constant.

Forecasts beyond the end of a realization from the model $(1 - B)^2(Y_t - \mu) = a_t$ follow a line determined by the last two data values. Forecasts from the ARIMA model $\phi_s(B)(1 - B)^2(Y_t - \mu) = \theta(B)a_t$ would eventually tend to a line, and thus, these types of models would forecast a linear trend to continue. A nearly nonstationary ARMA(p, q) model that has two roots near +1 and no other roots near the unit circle will have forecasts that approximate a linear trend for the near future.

Gray and Woodward (1981) discussed the ARUMA (autoregressive unit root moving average) models, which are nonstationary models for which the roots on the unit circle are not constrained to be

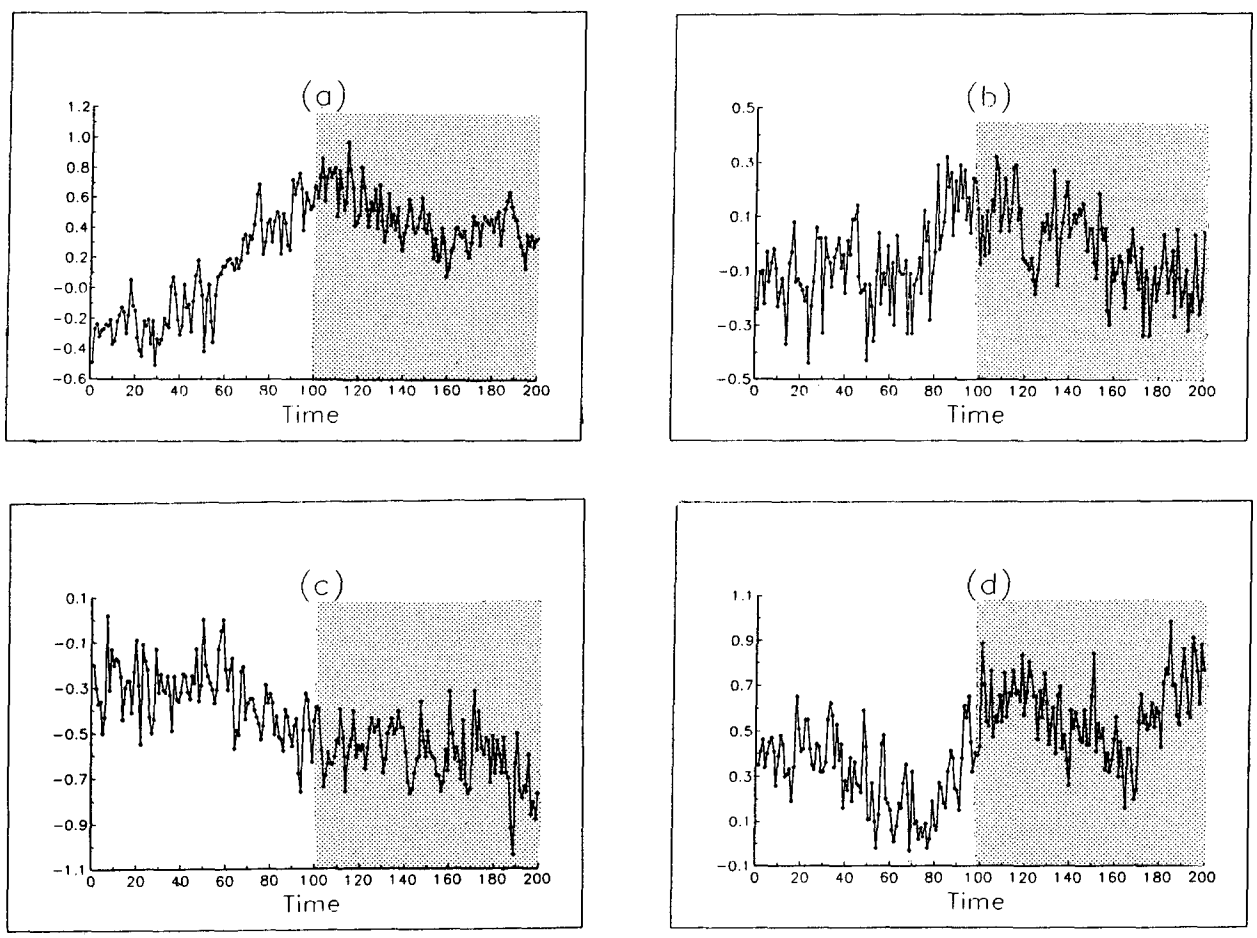


FIG. 3. Realizations of length $n = 200$ from the ARIMA(9, 1, 0) model for the Hansen and Lebedeff series.

TABLE 2. ARIMA models for global temperature series.

IPCC

$$(1 - B)(1 + .5618B + 1.3398B^2 + .6837B^3 + .6966B^4 + .5736B^5 + .3208B^6 + .4653B^7 + .1421B^8 + .2917B^9)(Y_t + .127) = (1 + .2868B + .7678B^2)a_t$$

$$\hat{\sigma}_a^2 = .008035$$

Hansen and Lebedeff data

$$(1 - B)(1 + .4774B + .5473B^2 + .4594B^3 + .3276B^4 + .3680B^5 + .0545B^6 + .2968B^7 + .0303B^8 + .1559B^9)(Y_t + .110) = a_t$$

$$\hat{\sigma}_a^2 = .013188$$

+1. For example, forecasts from the model $\phi_s(B)(1 - (3)^{1/2}B + B^2)(Y_t - \mu) = \theta(B)a_t$ will have a sinusoidal behavior with period 12.

In Table 2 we show the ARIMA models obtained for the two temperature series. Realizations from the ARIMA (9, 1, 0) model fit to the Hansen and Lebedeff data are shown in Fig. 3 and discussed in section 3. In each case the procedure given by Gray and Woodward (1986) was utilized for prefiltering the series to determine nonstationary or near-nonstationary components in the data. In each case this procedure identified one near-unit root, and the series were differenced. The orders of the transformed data were identified using Akaike's information criterion (AIC; Akaike 1974) and generalized partial autocorrelations (GPAC; Woodward and Gray 1981). Both methods independently produced the same models. Maximum likelihood estimates of the parameters in these transformed series were found using International Mathematical and Statistical Laboratories (IMSL) subroutine FTML. For each model, the residuals passed the standard tests for white noise. In Table 3 we show the irreducible first- and second-order factors of the polynomial $\phi(B)$ for the two models. This presentation is similar to that used by Woodward and Gray (1983) and Gray and Woodward (1986). For an irreducible second-order factor, $1 - \alpha_1 B - \alpha_2 B^2$, the associated roots are complex conjugates whose absolute reciprocal is $|\alpha_2|$. The system frequency associated with this factor is

$$f = \frac{1}{2\pi}\cos^{-1}\left(\frac{\alpha_1}{2\sqrt{-\alpha_2}}\right).$$

For a first-order factor, $1 - \alpha_1 B$, the absolute value of the reciprocal of the associated root is $|\alpha_1|$ while the associated frequency is $f = 0$ if $\alpha_1 > 0$ and $f = .5$ if $\alpha_1 < 0$. Frequency is given in cycles per sampling unit which in this case is cycles per year. Nonstationary and nearly nonstationary factors dominate the behavior of an ARIMA or ARMA model in the sense of determining the correlation structure. [See, e.g., Box and Jenkins (1976) and Gray and Woodward (1981).] In Table 3 it is seen that in each case, the most dominant factor is $1 - B$, the factor associated with the unit root. Thus, the ARIMA models for the temperature datasets are accounting for the "trendiness" in the data by incorporating a first-order factor with positive unit root which, as we have discussed, produces random temporary trends. In neither model is there an indication of two roots of +1 or even of two roots near +1. Thus, the ARIMA models fit to the temperature series do not predict that the trend will continue.

It should be noted that Tsonis and Elsner (1989) fit a stationary AR(4) model to the 108-year Jones et al. (1991) data for 1881–1988. The key difference between the model we obtained and that used by Tsonis and Elsner is that their model is stationary with a root near +1 while our model is nonstationary due to a root of +1.

c. *Time series models with nonconstant mean*

Several authors have approached the topic of testing for trend in a time series by assuming that the observed series Y_t can be expressed as $Y_t = \mu_t + E_t$ where E_t is a stationary process with zero mean. See for example, Grenander (1954), Brillinger (1989), Kuo et al. (1990), Cryer (1986), Bloomfield (1992), and Bloomfield and Nychka (1992). Moreover, the work by Bloomfield and Nychka, Bloomfield, and Kuo et al. considered the specific application of testing for trend in global temperature data. If μ_t is considered to be a linear function of t, that is, $\mu_t = a + bt$, then the residuals E_t are the same as those in (2.1). A procedure

TABLE 3. Autoregressive factor tables associated with ARIMA models for temperature data.

ARIMA(9,1,2) fit to IPCC data

AR factors	Absolute reciprocal of root	Frequency
$1 - B$	1.000	.00
$1 + .346B + .922B^2$	.960	.28
$1 - .431B + .764B^2$	.874	.21
$1 + .865B$	.865	.50
$1 + 1.054B + .727B^2$	.853	.36
$1 - 1.271B + .659B^2$	.812	.11

MA factors	Absolute reciprocal of root	Frequency
$1 + .287B + .768B^2$	.876	.28

ARIMA(9,1,0) fit to Hansen and Lebedeff data

AR factors	Absolute reciprocal of root	Frequency
$1 - B$	1.000	.00
$1 + .959B$	.959	.50
$1 + 1.100B + .706B^2$	.840	.36
$1 - .611B + .672B^2$	.820	.19
$1 - 1.279B + .622B^2$	.789	.10
$1 + .308B + .551B^2$	.742	.28

that has been widely employed has been that of estimating b using the least-squares estimator given in (2.2) and testing its significance using a standard error that does not make the independent errors assumption of regression analysis. Grenander (1954), Cryer (1986), and Bloomfield and Nychka (1992), address this problem. Specifically, it can be seen that the standard error of $\hat{b}$ is given by

$$\mathrm{SE}^{(2)}(\hat{b}) = \left[\frac{12}{n(n^2-1)} \times \left(\gamma_0 + \frac{24}{n(n^2-1)} \sum_{s=2}^{n} \sum_{t=1}^{s-1} (t-\bar{t})(s-\bar{t})\gamma_{s-t}\right)\right]^{1/2},$$

where γ_k denotes the kth autocovariance of E_t. Another form for this standard error considered by Bloomfield and Nychka (1992) is

$$\mathrm{SE}^{(3)}(\hat{b}) = \left[2\int_0^{.5} W(f)S(f)df\right]^{1/2}, \quad (2.5)$$

where

$$W(f) = |\sum_{t=1}^{n} b_t e^{-2\pi i f t}|^2,$$

with

$$b_t = \frac{t-\bar{t}}{\sum_{t=1}^{n}(t-\bar{t})^2},$$

that is, $\hat{b} = \sum_{t=1}^{n} b_t Y_t$, and where $S(f)$ denotes the spectrum of E_t. It should be noted that $\mathrm{SE}^{(2)}(\hat{b}) = \mathrm{SE}^{(3)}(\hat{b})$, but we introduce each here because they lead to different standard error estimates. For a given realization, $\hat{a}$ and $\hat{b}$ can be obtained as the usual least-squares estimates, and E_t can be estimated by

$$\hat{E}_t = Y_t - \hat{a} - \hat{b}t. \quad (2.6)$$

We consider the estimator

$$\widehat{\mathrm{SE}}^{(2)}(\hat{b}) = \left[(\widehat{\mathrm{SE}}^{(1)}(\hat{b}))^2 + \frac{288}{n^2(n^2-1)^2} \times \sum_{s=2}^{n} \sum_{t=1}^{s-1} (t-\bar{t})(s-\bar{t})\hat{\gamma}_{s-t}\right]^{1/2} \quad (2.7)$$

where

$$\hat{\gamma}_k = \frac{1}{n}\sum_{t=1}^{n-k} \hat{E}_{t+k}\hat{E}_t.$$

Thus $\widehat{\mathrm{SE}}^{(2)}(\hat{b})$ is the naive estimator obtained by replacing γ_k by $\hat{\gamma}_k$ in $\mathrm{SE}^{(2)}(\hat{b})$ except at $k = 0$ where we use $n\hat{\gamma}_0/n - 2$ to estimate γ_0 as is done in the regression case. It will be seen in the simulations that follow that $\widehat{\mathrm{SE}}^{(2)}(\hat{b})$ in (2.7) is a poor estimator of $\mathrm{SE}^{(2)}(\hat{b})$. The inferior performance of this estimator seems to be based on its use of estimates, $\hat{\gamma}_k$, which for lags near n are quite variable. Alternative estimators that eliminate or downweight the contribution of $\hat{\gamma}_k$ for k near n might be used, but these have not been considered here. We follow the suggestion of Bloomfield and Nychka and estimate $\mathrm{SE}^{(3)}(\hat{b})$ by replacing $S(f)$ in (2.5) with an appropriate estimate. In our implementation we fit an autoregressive model to $\hat{E}_t$ and use the corresponding autoregressive spectral estimator to estimate $S(f)$.

Grenander (1954) has shown that the asymptotic standard error of $\hat{b}$ in the current setting is given by

$$\mathrm{SE}^{(4)}(\hat{b}) = \left[\frac{S(0)}{\sum_{t=1}^{n}(t-\bar{t})^2}\right]^{1/2}, \quad (2.8)$$

where $S(0)$ is the spectrum of E_t at $f = 0$. In our implementation we estimated $S(0)$ using the autoregressive spectral estimator used for $\mathrm{SE}^{(3)}(\hat{b})$.

Ratios $\hat{b}/\widehat{\mathrm{SE}}^{(j)}(\hat{b}), j = 2, 3, 4$ are shown in Table 4 for each of the temperature series in Fig. 1. Since these ratios are smaller than those in Table 1, we see that the standard errors that take into consideration the autocorrelation in the data are larger than those based on usual regression assumptions. However, from Table 4 we see that these approaches all yield ratios ranging from 4.0 to 6.6 and thus indicate a significant linear trend when $\hat{b}/\widehat{\mathrm{SE}}^{(j)}(\hat{b})$ is approximately standard normal under the null hypothesis of no trend. We will discuss these issues further in section 3. The residuals from the estimated trend line, that is, E_t as given in (2.6), were modeled with an autoregressive model of order $p = 10$ for the IPCC data and the Hansen and Lebedeff data. Thus, $\widehat{\mathrm{SE}}^{(3)}(\hat{b})$ and $\widehat{\mathrm{SE}}^{(4)}(\hat{b})$ were obtained using the corresponding autoregressive spectral estimator in each case. Since the slope estimates are significantly different from zero, forecasts based on these models predict an increasing linear trend in future temperature. It should be noted that although we have not implemented their procedure in our analysis, Kuo et al. (1990) also found a significant linear trend in the Hansen and Lebedeff data.

Another approach to estimating the parameters in (2.1) is to assume E_t to be an AR(p) model and obtain estimates of a, b, $\phi_1, \ldots, \phi_p$, and σ_a^2 using maximum likelihood procedures. Bloomfield (1992) uses this technique on the Hansen and Lebedeff and the IPCC data and concludes that there is a significant trend

TABLE 4. Ratios $\hat{b}/\widehat{\mathrm{SE}}^{(j)}(\hat{b}), j = 2, 3, 4$ for time series approaches applied to global temperature series.

Series	$\hat{b}/\widehat{\mathrm{SE}}^{(2)}(\hat{b})$	$\hat{b}/\widehat{\mathrm{SE}}^{(3)}(\hat{b})$	$\hat{b}/\widehat{\mathrm{SE}}^{(4)}(\hat{b})$
IPCC	6.575	4.975	4.780
Hansen and Lebedeff	5.874	4.697	3.976

TABLE 5. Maximum likelihood results for the temperature datasets.

	$\hat{b}_{ML}$	Test statistic ratio
IPCC	.00415	3.779
Hansen and Lebedeff	.00583	4.983

component. Our implementation of this procedure using SAS AUTOREG (1984) again used $p = 10$ in both cases. The results in Table 5 were obtained, and our results agree with those of Bloomfield in finding a significant trend. This approach has the intuitive appeal that the estimation of b is made "simultaneously" with that of the autoregressive components of the model, allowing for the possibility that the estimation scheme can intelligently distinguish between trend-type behavior induced by $a + bt$ and that induced by the autoregressive component of the model. We address the ability of the MLE to make this distinction in the next section.

Brillinger (1989) considers the time series model $Y_t = \mu_t + E_t$ where μ_t is monotonic and where E_t is a stationary process with zero mean. The monotonicity implies that either $\mu_t \leqslant \mu_{t+1}$ for all t or else $\mu_t \geqslant \mu_{t+1}$ for all t. In either case we require that strict inequality holds for some t in order to declare μ_t monotonic. Brillinger (1989) considers the estimator $\sum_{t=1}^{n} c_t Y_t$ where the coefficients c_t are due to Abelson and Tukey (1963) and are given by

$$c_t = \left\{(t-1)\left(1 - \frac{t-1}{n}\right)\right\}^{1/2} - \left\{t\left(1 - \frac{t}{n}\right)\right\}^{1/2}. \quad (2.9)$$

The residuals $\hat{E}_t$ are calculated as $\hat{E}_t = Y_t - \hat{\mu}_t$ where

$$\hat{\mu}_t = \frac{1}{2v+1} \sum_{s=-v}^{v} Y_{t+s},$$

for $t = v + 1, \cdots, n - v$. A test statistic which is approximately standard normal when H_0: $\mu_t = \mu$ is true, is given by Brillinger (1989) as

$$\frac{\sum_{t=1}^{n} c_t Y_t}{[\hat{S}_p(0) \sum_{t=1}^{n} c_t^2]^{1/2}} \quad (2.10)$$

where $\hat{S}_p(0)$ is the smoothed periodogram spectral estimator given by

$$\hat{S}_p(0) = \frac{\sum_{j=1}^{L} \frac{1}{n} |\hat{\xi}_j|^2}{\sum_{j=1}^{L} (1 - a_j)^2},$$

where $\hat{\xi}_j$ denotes the discrete Fourier transform of the residuals, given by

$$\hat{\xi}_j = \sum_{t=v+1}^{n-v} \hat{E}(t) \exp(-2\pi i(t-1)j/n),$$

and the a_j values are given by

$$a_j = \frac{\sin\{2\pi j(2v+1)/2n\}}{(2v+1)\sin(2\pi j/2n)}.$$

The window length, L, should be small with respect to n. The linear combination $\sum_{t=1}^{n} c_t Y_t$ strongly contrasts the beginning and ending of the series and should be positive or negative depending upon whether the monotonic trend is increasing or decreasing, respectively.

In our implementation we have taken $L = n/20$ and $v = n/10$. The resulting test statistics for the IPCC and Hansen and Lebedeff series are 6.650 and 8.468, respectively, again strongly indicating a monotonic trend.

3. Simulations

In the preceding section we considered several techniques for analyzing time series data and determining whether or not an increasing temperature trend is forecast for an extended time into the future. It is important to understand that models (1.1) and (2.4) are philosophically different. Model (1.1) assumes that the observation is a line plus noise while the ARIMA model assumes that no such line exists. That is, the first model accounts for the appearance of a trend by the assumption of the deterministic curve $a + bt$ while the ARIMA model assumes that this behavior is due to correlation in the data. Specifically, for ARIMA models with at least one unit root, it is very common for realizations to trend up for a while and then down for a while as a part of their typical behavior. Since none of these ARMA models will forecast a trend to continue unless at least two unit roots are present, the question of trend when posed with regard to ARMA models is, for the reason discussed in section 2, a question of testing for unit roots. A single unit root would imply that the time series has random temporary trends, that is, changes level. However, as mentioned earlier, forecasts from such a model are relatively constant and do not predict that a trend will continue. Two such roots would produce a linear trend as the best forecast for the future, while two roots near unity or complex roots close to the unit circle with associated frequency near zero would give similar behavior. In the case of one unit root and another positive root, say in the range $\leqslant .8$, the trend will be forecast to continue only for a short time, and long term forecasts soon level off. It is quite clear from Table 3 that neither of the fitted models for these data has two roots sufficiently close to unity to forecast even a short trend.

The preceding discussion does not imply that realizations from ARIMA models with a single unit root, such as those fit to the temperature series, will not exhibit some trend. Quite the contrary, since almost all

realizations of such a process will exhibit some rather lengthy trends, but none of these trends persist. Figure 3 shows four realizations of length $n = 200$ from the ARIMA(9, 1, 0) model obtained for the Hansen and Lebedeff data. In all four of the realizations there are periods of increasing or decreasing trends of varying lengths. The first 100 observations in Figs. 3a and 3b have an appearance similar to that of the temperature series. However, in each case, the trend did not continue for the entire realization length. In Fig. 3a there is somewhat of a leveling off after the initial climb, while in Fig. 3b a more modest upward trend is followed by a noticeable decrease. In Fig. 3c a realization is displayed for which the trend is downward and fairly persistent for the entire 200 time periods of the realization. In Fig. 3d we see a picture of the classic "wandering" behavior (more precisely, random changes in level) associated with models having one unit root. In all cases, had the realizations been sufficiently long, this same type of wandering behavior would have been observed. Paleoclimate obtained from such sources as ice cores, ocean sediments, and data from tree rings suggest that plots of past global temperatures over a much longer record would have very much the same appearance as Fig. 3d.

Based on our analysis of the actual temperature data, we believe that ARIMA models such as those given in Table 3 are plausible models for these datasets. The question then arises concerning the appropriateness of our previously described tests for linearity if in fact the data are generated from an appropriate ARIMA model. In this section we report on simulations designed to examine this question.

The simulations involve generating 100 realizations of length $n = 100$ from selected models. For each realization we test for a significant trend using the tests suggested by the results in section 2. Specifically, the decision criteria used are as follows:

(i) Linear trend, that is, $\mu_t = a + bt$: null hypothesis of no trend is H_0: $b = 0$. We will appeal to asymptotic normality and for each of the associated tests, we reject H_0 at the nominal $\alpha = 0.05$ level whenever $z^{(i)} = \hat{b}/\widehat{\mathrm{SE}}^{(i)}(\hat{b})$ is greater than 1.96 in absolute value. For the "regression setting," that is, $i = 1$, we consider tests based on original data and ranks.

(ii) Monotonic trend: null hypothesis of no trend is H_0: $\mu_t = \mu$. Again we appeal to asymptotic results and reject H_0 at the nominal 0.05 significance level whenever the test statistic in (2.10) is greater than 1.96 in absolute value.

The tests considered above are all approximate tests, so we first consider the actual significance level associated with the nominal significance level of 0.05. To this end we first simulated 100 realizations of length $n = 100$ from the model in which Y_t is normal white noise with zero mean. Thus, in this setting even the regression assumptions apply. In the first row of Table 6 we show the proportion of the realizations for which the null hypothesis of no trend is rejected when the model is white noise. There it can be seen that all of the tests except that using $\widehat{\mathrm{SE}}^{(2)}(\hat{b})$ have observed significance level (false-alarm rate) close to 0.05. The observed significance levels associated with the use of $\widehat{\mathrm{SE}}^{(2)}(\hat{b})$ are unacceptably high which is consistent with our earlier observation that $\widehat{\mathrm{SE}}^{(2)}(\hat{b})$ would be a poorly behaved estimator.

Of course, the situation in which Y_t is white noise is not the only model for which the null hypothesis of

TABLE 6. Proportion of realizations from AR models for which significant trend is incorrectly detected at the nominal 5% level.

Number of realizations = 100
Realization length = 100

	Tests for linear trend based on $\widehat{\mathrm{SE}}^{(j)}(\hat{b})$, j					
	1	1 (ranks)	2	3	4	Test for monotonic trend
(a) $Y_t = a_t$	.04	.04	.26	.06	.07	.07
(b) $(1 - .5B)Y_t = a_t$	.13	.13	.14	.05	.05	.03
(c) $(1 + .5B)Y_t = a_t$	.00	.00	.23	.07	.07	.10
(d) $(1 - .9B)Y_t = a_t$	.64	.63	.43	.20	.17	.25
(e) $(1 + .9B)Y_t = a_t$	.00	.00	.11	.04	.06	.30
(f) $(1 - .95B)Y_t = a_t$	.79	.76	.53	.35	.31	.50
(g) $(1 + .95B)Y_t = a_t$	.00	.00	.11	.05	.12	.27
(h) $(1 - B)Y_t = a_t$	.85	.83	.65	.59	.52	.75
(i) $(1 + B)Y_t = a_t$	.00	.00	.03	.02	.54	.63
(j) $(1 + .5B^2)Y_t = a_t$	.00	.00	.27	.11	.13	.15
(k) $(1 + .9B^2)Y_t = a_t$	.00	.00	.14	.07	.17	.25
(l) $(1 + .95B^2)Y_t = a_t$	.00	.00	.12	.11	.26	.36
(m) $(1 - B)(1 - .7B)Y_t = a_t$	.87	.86	.73	.63	.58	.79

no trend is true. Specifically, if Y_t is any stationary process, then the null hypothesis is technically true. In Table 6 we show the proportion of rejections for several other simple autoregressive models that are stationary or have one unit root. We include models that should show some trending behavior in their realizations and models that should not. Although it is not surprising that the tests would perform poorly on model (m) since it should show some temporary trend behavior, it is amazing that several perform poorly on model (i), which would rarely show trend behavior. Indeed for several of the models considered, the false-alarm rate is surprisingly large. In order to understand this phenomenon, we examine these models more closely. One of the models that shows a large false-alarm rate for the tests for trend is the AR(1) model, $(1 - .95B)X_t = a_t$. For this model, $\rho_k = .95^k$ and thus there is a high positive autocorrelation among values in the series relatively close to each other. Such an autocorrelation structure causes wandering or the rather random temporary trend behavior in realizations, and according to Table 6, resulted in several realizations for which the tests identified a trend in the data. However, the behavior of these realizations is such that trends are equally likely in either direction and very long realizations will typically show both upward and downward trends. The realizations for the nonstationary models $(1 - B)Y_t = a_t$ and $(1 - B)(1 - .7B)Y_t = a_t$ showed an even stronger tendency to have significant trends in realizations. Forecasts from the models $(1 - B)(1 - .7B)Y_t = a_t$ typically suggest a weak trend to continue for fewer than five time periods beyond the end of the observed series. The models of the form $(1 - \phi_2 B^2)Y_t = a_t$, $\phi_2 = -0.5$, -0.9, and -0.95, have peaks in the spectrum at $f = 0.25$ and consequently have realizations that display periodic behavior. Similarly, the models $(1 - \phi_1 B)Y_t = a_t$ for negative ϕ_1 have realizations with periodic behavior associated with $f = 0.5$. These periodic models did not have trends introduced by the wandering behavior associated with models that have a high positive correlation between successive values produced by positive real roots of $\phi(r) = 0$. However, the observed proportions of realizations for which several of the tests detected a significant trend were still substantially higher than the nominal 0.05 level for these models, a surprising result. It should be noted that for all models in Table 5, $\widehat{SE}^{(3)}(\hat{b})$ and $\widehat{SE}^{(4)}(\hat{b})$ were obtained using an AR(2) model for $\hat{E}_t$. Use of an AR(5) model was also examined, and this produced only minor differences from the tabled values.

Again, for all of the models considered in Table 6 it is inappropriate to incorporate a deterministic trend in the forecasts of the form introduced by the models of section 2. The important point is that if a trend is detected in a dataset using the tests discussed in that section, such a finding does not eliminate from consideration ARMA models for which the characteristic equation has a positive real root close to or on the unit circle.

Finally, we examine simulated realizations from the ARIMA models in Table 3 that were fit to the temperature series in order to determine whether or not the realizations have the appearance of the temperature series and whether the tests detect a significant trend. In each case we have simulated 100 realizations of the length in the corresponding temperature series. In Table 7 we show the results of the simulations and it is clear that in both cases, the tests often detect a trend, with the results being similar to those for the case $(1 - B)(1 - .7B)Y_t = a_t$ in Table 6. In Table 7 we also see that the MLE does not distinguish between the two types of trend behavior, since for these two models, which do not have a deterministic trend, at least 50% of the realizations were found to have a significant trend. However, it is important to recall that neither of the fitted models of Table 3 would predict a future trend under a continuation of the conditions that produced the data.

4. Concluding remarks

In this paper we have shown that although there is a major distinction between ARIMA models whose characteristics produce random trends and models of the form (1.1), this difference may be very difficult to ascertain from realizations of length approximately $n = 100$. Several authors (e.g., Bloomfield and Nychka 1992; Bloomfield 1992; Kuo et al. 1990) have applied tests for trend to determine whether or not a significant trend component exists in the widely referenced global temperature datasets, and have generally concluded that there is a significant deterministic, that is, long-term trend in the data. In this paper we apply several tests for trend to the Hansen and Lebedeff and the IPCC temperature series. These tests do indeed suggest the existence of a significant deterministic trend. At first glance this appears to answer the question concerning whether or not a deterministic trend component should be included in the model. However, as we have demonstrated, the ARIMA model is also a plau-

TABLE 7. Proportion of realizations from models fit to temperature deviations series for which a significant trend is incorrectly detected.

	Number of realizations = 100						
	Tests for linear trend based on $\widehat{SE}^{(j)}(\hat{b})$ j						
	1	1 (ranks)	2	3	4	MLE	Test for monotonic trend
IPCC ARIMA(9,1,2)	.81	.81	.66	.57	.49	.50	.69
Hansen and Lebedeff ARIMA(9,1,0)	.85	.84	.75	.67	.57	.60	.74

sible model for these data. Moreover, our simulation results indicate that application of these trend tests to realizations from ARIMA models fit to the temperature series results in the detection of a significant trend in a high percentage of the realizations even though it would be inappropriate to forecast such a future trend for even a short time interval. Let it be clear that we do not claim that our best predictions, using the temperature data alone, are that there will be no increase in the temperature; or that the fixed-mean ARIMA models fit to the temperature series are preferable to fits using model (1.1). We do claim, however, that these ARIMA models are at least plausible models for the temperature series. Realizations from these ARIMA models have random trends that cause the trend tests to detect a long-term trend in a high percentage of the cases. Thus, *if* temperature were behaving according to a "correlation-based" model such as the ARIMA, then there is a high chance that a significant deterministic trend would be incorrectly inferred.

The authors have previously observed a similar phenomenon concerning the use of Priestley's $P(\lambda)$ test for determining whether or not a harmonic component should be included in a model for the Canadian lynx data (see Woodward and Gray 1983). In that paper it was shown that this test (designed to determine whether the peak in the estimated spectra is sufficiently sharp to assume that the model contains a harmonic component) is unable to satisfactorily distinguish between a model with a harmonic component and an ARMA model associated with a pair of complex roots near the unit circle. The results of the current paper are entirely analogous, except here the peak in the spectrum is at the zero frequency.

As stated previously, the projections that the warming trend will extend into the future are largely based on such factors as the buildup of atmospheric greenhouse gases. We have shown here, however, that based solely on the available temperature data, there is no conclusive evidence that the trend should be predicted to continue. This is primarily due to the difficulty in distinguishing between data with long-term trends and those with random trends for series of the lengths of the temperature series. Consequently, tests based on the model $Y_t = \mu_t + E_t$ for the purpose of prediction or inference concerning future behavior should be used with caution.

Acknowledgments. This research was partially supported by DOE Environmental Sciences Division Grant DE-FG05-90ER61015.

REFERENCES

Abelson, R. P., and J. N. Tukey, 1963: Efficient utilization of nonnumerical information in quantitative analysis: General theory and the case of simple order. *Ann. Math. Stat.*, **34,** 1347–1369.

Akaike, H., 1974: A new look at the statistical model identification, *IEEE Trans. Autom. Control,* **19,** 716–723.

Bloomfield, P., 1992: Trends in global temperature. *Clim. Change,* **21,** 1–16.

——, and D. W. Nychka, 1992: Climate spectra and detecting climate change. *Clim. Change,* **21,** 275–287.

Bottomley, M. C., C. K. Folland, J. Hsiung, R. E. Newell, and D. E. Parker, 1990: *Global Ocean Surface Temperature Atlas.* U.K. Meteorological Office, 20 pp., 313 plates.

Box, G. E. P., and G. M. Jenkins, 1976: *Time Series Analysis: Forecasting and Control.* Holden-Day, 575 pp.

Brillinger, D. R., 1989: Consistent detection of a monotonic trend superposed on a stationary time series. *Biometrika,* **76,** 23–30.

Cryer, J. D., 1986: *Time Series Analysis.* Duxbury Press, 286 pp.

Folland, C. K., T. R. Karl, and K. Y. Vinnikov, 1990: Observed climatic variations and change. *Climate Change: The IPPC Scientific Assessment.* J. T. Houghton, G. J. Jenkins, and J. J. Ephraums, Eds., Cambridge University Press, 192–238.

Gray, H. L., and W. A. Woodward, 1981: Applications of s-arrays to seasonal data. *Applied Time Series Analysis II,* D. Findley, Ed., 379–413.

——, and ——, 1986: A new ARMA spectral estimator. *J. Amer. Stat. Assoc.,* **81,** 1100–1108.

Grenander, U., 1954: On the estimation of regression coefficients in the case of an autocorrelated disturbance. *Ann. Math. Stat.,* **29,** 252–272.

Gunst, R. F., S. Basu, and R. Brunell, 1993: Defining and estimating mean global temperature change. *J. Climate,* in press.

Hansen, J., and S. Lebedeff, 1987: Global trends of measured surface air temperature. *J. of Geophys. Res.,* **92,** 13 345–13 372.

——, and ——, 1988: Global surface air temperatures: Update through 1987. *Geophys. Res. Lett.,* **15,** 323–326.

Hanson, K., G. A. Maul, and T. R. Karl, 1989: Are the atmospheric 'greenhouse' effects apparent in the climatic records of the contiguous U.S. (1895–1987)? *Geophys. Res. Lett.,* **16,** 49–52.

Jones, P. D., 1988: Hemispheric surface air temperature variations: recent trends and an update to 1987. *J. Climate,* **1,** 654–660.

——, T. M. L. Wigley, and G. Farmer, 1991: Marine and land temperature data sets: A comparison and a look at recent trends. *Greenhouse Induced Climatic Change: A Critical Appraisal of Simulations and Observations.* M. E. Schlesinger, Ed., 153–172.

Kuo, C., C. Lindberg, and D. J. Thomson, 1990: Coherence established between atmospheric carbon dioxide and global temperature. *Nature,* **343,** 709–714.

SAS Institute, 1988: *SAS/ETS User's Guide, Version 6,* first ed. SAS Institute, Inc., 560 pp.

Tsonis, A. A., and J. B. Elsner, 1989: Testing the global warming hypothesis. *Geophys. Res. Lett.,* **16,** 795–797.

Woodward, W. A., and H. L. Gray, 1981: On the relationship between the s-array and the Box-Jenkins method of ARMA model identification. *J. Amer. Stat. Assoc.,* **76,** 579–587.

——, and ——, 1983: A Comparison of autoregressive and harmonic component models for the lynx data. *J. Roy. Stat. Soc.,* **A146,** 71–73.