Statistical Procedures for Making Inferences about Climate Variability

RICHARD W. KATZ

Environmental and Societal Impacts Group, National Center for Atmospheric Research, Boulder, Colorado*

(Manuscript received 23 February 1988, in final form 26 April 1988)

ABSTRACT

A statistical procedure is described for making inferences about changes in climate variability. The fundamental question of how to define climate variability is first addressed, and a definition of intrinsic climate variability based on a "prewhitening" of the data is advocated. A test for changes in variability that is not sensitive to departures from the assumption of a Gaussian distribution for the data is outlined. In addition to establishing whether observed differences in variability are statistically significant, the procedure provides confidence intervals for the ratio of variability. The technique is applied to time series of daily mean surface air temperatures generated by the Oregon State University atmospheric general circulation model. The test application provides estimates of the magnitude of change in variability that the procedure should be likely to detect.

1. Introduction

The detection of a climate change in the historical record or in general circulation model (GCM) climate experiments has commonly been treated as a problem of estimating differences between location parameters (e.g., means or medians). To take into account the manner in which an atmospheric variable is distributed about a location parameter, differences between scale parameters (e.g., variances or interquartile ranges) also need to be estimated. In many applications, the potential impacts of changes in climate variability may be as great or greater than the impacts of any changes in climate means. Further, decisions about what strategy to take under a change in a climate mean cannot be made without also making an assumption about how climate variability would change. Although it is ordinarily assumed, by default, that the level of climate variability would remain the same, Mearns et al. (1984) have illustrated how sensitive the likelihood of extreme meteorological events of importance to society is to changes in variability.

The primary purpose of this paper is to describe a statistical procedure for making inferences about differences in scale that would be appropriate when dealing with time series of historical climate data or time series generated by a GCM. This procedure could be used, for example, to test whether a doubling of carbon dioxide levels in the atmosphere would result in a change in the variability of surface air temperature as well as an increase in mean surface air temperature (as is now often claimed, e.g., WMO 1986). It can be viewed as an extension of the parametric time series modeling approach employed by Katz (1982) to make inferences about changes in means. Other studies that are concerned primarily with testing for changes in means simulated by GCM climate experiments include Chervin and Schneider (1976), Laurmann and Gates (1977), and Livezey (1985).

The problem of estimating changes in scale is more difficult than that of estimating changes in location, especially when working with real or simulated atmospheric time series. Because such time series are autocorrelated, the conventional definition of scale as simply the variance of a random variable has some drawbacks. Instead, a definition of intrinsic variability based on a "prewhitening" of the data will be advocated. This approach requires fitting a parametric time series model—an autoregressive process (as in Katz 1982), for example—to the data. The fitted model is then used to obtain residuals that are approximately uncorrelated. The definition of climate variability refers to the variance of the corresponding theoretical errors (or innovations) for which the residuals are estimates.

An additional problem arises because atmospheric variables at best have distributions that are only approximately Gaussian in shape. Although tests for changes in means are not particularly dependent on the Gaussian assumption, standard tests for changes in variance are extremely sensitive to departures from the Gaussian assumption. In fact, Box (1953) has suggested that one particular test for equality of variances could be better used as a test for Gaussian distributions. In another context he has commented that making the

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: Dr. Richard W. Katz, Environmental and Societal Impacts Group, NCAR, P.O. Box 3000, Boulder, CO 80307.

standard test for equality of variances is "like putting to sea in a rowing boat to find out whether conditions are sufficiently calm for an ocean liner to leave port!" Nevertheless, tests for changes in climate variability have relied on such procedures; for example, Chervin (1980) and Hayashi (1982) have applied the F-test for the equality of two variances to GCM simulated data. Zwiers and Thiebaux (1987) have mentioned this issue of possible nonrobustness to the Gaussian assumption of common tests for scale, although they do not explicitly deal with the problem. In this paper, a test procedure will be recommended that corrects for non-Gaussian distributions.

The basic question of how to define climate variability is discussed in section 2. Using this definition, the appropriate statistical methodology for making inferences about changes in variance is presented in section 3. The application of this methodology to climate observations or to GCM climate experiments is described in section 4, and section 5 gives the results of a test application of the statistical procedure to control data generated by the Oregon State University (OSU) atmospheric GCM. Finally, section 6 consists of some concluding remarks and suggestions for future research. We note that Wilson and Mitchell (1987) have applied this procedure to compare GCM simulated climate variability with observed climate variability.

2. Definition of climate variability

In attempting to devise a meaningful definition of climate variability, several questions arise. In viewing a real or simulated time series for an atmospheric variable as a realization of a stochastic process, variability could refer to the variance of the process or to the variance of a time average of the process. If the data were generated by an independent process, then the time-average variance would be directly proportional to the process variance. Because real or simulated atmospheric time series are autocorrelated, there is no simple relationship between the time-average variance and the process variance. The time-average variance depends instead on the extent of the autocorrelation as well as on the process variance (e.g., Jones 1975). The concept of an underlying uncorrelated process is the basis for modeling the dependence over time of an atmospheric variable. Thus, it is natural to consider the variance of this underlying process, called the "innovation variance," as one alternative to the variance of the original process for representing climate variability. Roughly speaking, the innovation variance can be thought of as the variation of the process that still remains after variation attributed to past values of the process is removed.

To define the innovation variance, we need to consider a class of models that allow for autocorrelation. Following the parametric time series modeling approach employed in Katz (1982), we assume that the climate time series can be represented as an autoregressive process of order p, $p \geq 1$, denoted by AR(p). A stationary stochastic process $\{X_t: t = \cdots, -1, 0, 1, \cdots\}$, with mean $\mu = E(X_t)$ and variance $\sigma^2 = \text{var}(X_t)$, is an AR($p$) process if it can be expressed as

$$X_t - \mu = \sum_{k=1}^{p} \phi_k (X_{t-k} - \mu) + a_t. \tag{1}$$

Here the autoregression coefficients ϕ_k, $k = 1, 2, \cdots, p$, are unknown parameters satisfying certain constraints in order for the X_t process to be stationary (Box and Jenkins 1976, p. 53). In (1) the a_t's are called "innovations" or "shocks" and are assumed to constitute a "white noise" process; that is, they are uncorrelated random variables with zero mean and constant variance $\sigma_a^2 = \text{var}(a_t)$. For reasons that will be discussed later, we do not require that the innovation a_t have a Gaussian distribution.

The expression (1) for an AR(p) process can be viewed as a transformation of an autocorrelated process, namely the X_t's, into an uncorrelated process, namely the a_t's. Such an operation is sometimes called "prewhitening," with "white" referring to the fact that a white-noise process is produced and "pre" to the fact that this operation is performed prior to analysis of the data. The common variance σ_a^2 of the a_t's is called the "innovation variance." It is related to the variance of σ^2 of the original process by

$$\sigma_a^2 = \sigma^2 \prod_{k=1}^{p} \{1 - [\phi_k(k)]^2\} \tag{2}$$

(Durbin 1960). Here $\phi_k(k)$ denotes the kth-order partial autocorrelation coefficient (for its definition, see Box and Jenkins 1976, p. 64).

We note that a stationary AR(p) process has an equivalent representation as an infinite-order moving average process (Box and Jenkins 1976, p. 46); that is,

$$X_t - \mu = \sum_{j=0}^{\infty} \psi_j a_{t-j}. \tag{3}$$

Here the ψ-weights are functions of the p autoregression coefficients ϕ_k, $k = 1, 2, \cdots, p$. For example, (3) reduces to

$$X_t - \mu = \sum_{j=0}^{\infty} \phi_1^j a_{t-j} \tag{4}$$

in the special case of an AR(1) process. Provided the autoregression coefficients are known, the representation (3) shows explicitly how the original time series (X_t) can be generated from the innovation time series (a_t) alone. In this sense, working with the prewhitened data is equivalent to dealing with the original time series.

The innovation variance also arises as a fundamental quantity in the context of forecasting and predictability studies for atmospheric variables. By (1), σ_a^2 can be

interpreted as the mean-squared error in predicting one time step into the future using the p-most recent values of the process. Specifically,

$$\text{var}(X_t | X_{t-1}, X_{t-2}, \cdots, X_{t-p}) = \sigma_a^2. \quad (5)$$

Moreover, the forecasting error l time steps ahead when using an autoregressive process can be most conveniently represented in terms of the ψ-weights attached to the innovations (3) and the innovation variance (Chu and Katz 1987). The skill of other more complex schemes for forecasting the next state of an atmospheric variable sometimes is measured relative to this "persistence" standard. In fact, Pierce (1979) has proposed that an R^2 measure for time series (i.e., proportionate reduction in variation) should be defined relative to the innovation variance σ_a^2, rather than relative to the process variance σ^2. Given this interpretation of the innovation variance as the variation of an atmospheric variable that remains to be "explained," testing for changes in climate variability in terms of innovation variances would certainly be meaningful.

In accordance with its interpretation as a measure of intrinsic variability, the innovation variance is closely related to other variances of interest when dealing with atmospheric variables. The relationship between the variance σ^2 of an AR(p) process and its innovation variance σ_a^2 is specified by (2). Because such a process is autocorrelated, σ_a^2 necessarily is strictly less than σ^2. Only for an uncorrelated [or AR(0)] process would these two variances be equal.

The time-average variance of an AR(p) process can be expressed as a function of the innovation variance. For large sample size n,

$$\text{var}(\bar{X}) \approx \frac{1}{n} \frac{\sigma_a^2}{(1 - \sum_{k=1}^{p} \phi_k)^2} \quad (6)$$

(Jones 1975; Katz 1982). Equation (6) involves a somewhat complicated function of the autoregression parameters for an AR(p) process. By comparing (6) to the expression for the variance of a time average of an uncorrelated process, an estimate of the "effective degrees of freedom" or "effective number of independent sample values" (e.g., Madden 1979) for an atmospheric variable could be obtained. Typically, atmospheric variables exhibit "persistence," making the effective number of independent sample values less than the original number n of correlated sample values. As an alternative approach, the concept of underlying innovations provides a direct mechanism for producing what could be regarded as "equivalent independent sample values." The original time series of length n could be viewed as equivalent to n uncorrelated innovations with common variance σ_a^2 smaller than the process variance σ^2.

3. Statistical methodology

a. Estimation of innovation variance

It is assumed that a time series, $X_1, X_2, \ldots, X_n$, of length n is available. The innovation variance σ_a^2 routinely is estimated as part of fitting an AR(p) process to a given time series. We assume, for now, that the order p of the autoregressive process is known; in practice, the choice of a value of p would constitute the first step in the fitting of the process. Let $\hat{\phi}_k$, $k = 1, 2, \cdots, p$, be estimates of the autoregression coefficients. These parameter estimates could be obtained by the Yule–Walker recursion or by several other methods. Substituting the estimates $\hat{\phi}_k$ in place of ϕ_k and the time average $\bar{X}$ in place of μ in (1), the tth residual is given by

$$\hat{a}_t = (X_t - \bar{X}) - \sum_{k=1}^{p} \hat{\phi}_k (X_{t-k} - \bar{X}), \quad t = 1, 2, \cdots, n. \quad (7)$$

To allow (7) to hold for $t = 1, 2, \cdots, p$, set $X_{t-k} = \bar{X}$ if $t - k < 0$. Then an unbiased estimator $\hat{\sigma}_a^2$ of the innovation variance is

$$\hat{\sigma}_a^2 = \frac{1}{n - p - 1} \sum_{t=1}^{n} \hat{a}_t^2. \quad (8)$$

We refer to the residual time series $\hat{a}_t$, $t = 1, 2, \cdots, n$, as the "prewhitened" data. Because the prewhitening operation (7) involves adjusting the current "anomaly" (i.e., departure from the mean) $X_t - \bar{X}$ to take into account the previous p anomalies, $X_{t-1} - \bar{X}, X_{t-2} - \bar{X}, \ldots, X_{t-p} - \bar{X}$, the prewhitened data can be viewed as a sequence of "adjusted anomalies." In this regard, the prewhitened quantity $\hat{a}_t$ represents a departure from the expected anomaly; that is, in some sense it is an "anomalous anomaly."

b. Inferences about innovation variance

We will now present a procedure for making inferences about the innovation variance on the basis of its estimator (8) and its associated standard error. In the case of an independent sample, several techniques have been proposed to correct for departures from the assumption of a Gaussian distribution. Here we will apply one of the simplest of these so-called "robust" procedures to the prewhitened data. Because the prewhitened data (7) are based on the estimated autoregression parameters rather than on the true autoregression parameters, they are only estimators of the true innovations. Even though the true innovations are assumed to be uncorrelated, the innovation estimators are only approximately uncorrelated. A weak correlation is present because of constraints on the prewhitened data; for instance, the prewhitened data must sum to zero. This result is analogous to the constraint in regression analysis that the residuals must sum to zero.

Nevertheless, Davis (1977, 1979) has shown that several robust procedures for making inferences about variances still have the same asymptotic properties when applied to prewhitened data. He has also studied their performance when applied to small samples by means of simulations. The estimator $\hat{\sigma}_a^2$, defined by (8), has a distribution that is asymptotically (i.e., as the sample size n tends to infinity) Gaussian about the true innovation variance σ_a^2. Its asymptotic variance, however, depends on the kurtosis γ_2 of the distribution of the innovation a_t. Formally, the population kurtosis is defined as the standardized fourth moment

$$\gamma_2 = \frac{1}{\sigma_a^4} E(a_t^4) - 3. \tag{9}$$

For a Gaussian distribution, $\gamma_2 = 0$. Roughly speaking, kurtosis is a measure of the peakedness ($\gamma_2 > 0$) or flatness ($\gamma_2 < 0$) of a distribution relative to the Gaussian, although this interpretation is not always correct. An estimator of the population kurtosis γ_2 is the sample kurtosis

$$\hat{\gamma}_2 = \frac{\frac{1}{n}\sum_{t=1}^{n} \hat{a}_t^4}{\left(\frac{1}{n}\sum_{t=1}^{n} \hat{a}_t^2\right)^2} - 3, \tag{10}$$

with the prewhitened $\hat{a}_t$ specified by (7).

Davis (1977) shows that the statistic

$$\frac{\ln\hat{\sigma}_a^2 - \ln\sigma_a^2}{s(\ln\hat{\sigma}_a^2)}, \tag{11}$$

with standard error

$$s(\ln\hat{\sigma}_a^2) = \left[\frac{1}{n}(2 + \hat{\gamma}_2)\right]^{1/2}, \tag{12}$$

has a distribution that is asymptotically standard Gaussian (i.e., zero mean and unit variance). The statistic (11) is in terms of the logarithms of the innovation variance and its estimator to improve the Gaussian approximation for moderate sample size n. In section 4, (11) will be employed to construct tests of significance and confidence intervals for the comparison of two innovation variances. Because this statistic contains an adjustment involving the sample kurtosis $\hat{\gamma}_2$ (10) of the prewhitened data, procedures based on (11) are asymptotically robust against the innovations having a non-Gaussian distribution.

Other more complex procedures, such as the "jackknife" (Miller 1968) or the "bootstrap" (Efron 1982), could be used instead of this simple "standard error" technique. The jackknife involves systematically deleting data from the sample, for example, one value at a time, computing the statistic (e.g., the logarithm of the sample variance) for each subsample, and then combining the subsample statistics into one overall estimate. The bootstrap is based on a resampling scheme that differs slightly from that for the jackknife. Such procedures would be more powerful in detecting changes in variability, especially for moderate sample sizes, but would be somewhat more complex to implement and would require a considerable additional amount of computational effort. Nonparametric tests for changes in scale could also be applied (e.g., Hollander and Wolfe 1973, Chapter 5). Because the primary purpose of this paper is to make climatologists aware of the problems inherent in making inferences about climate variability on the basis of conventional techniques and to provide an alternative procedure that is as simple as possible, we do not consider these other procedures.

4. Application to climate time series

Suppose that we are given two climate time series of length n_1 and n_2, consisting either of observations of a climate variable over two different time periods or of the output of GCM control and climate experiment runs. It is assumed that both of the time series are generated by autoregressive processes of the form (1) with the parameters, as well as the order, of the two processes permitted to differ. It is further assumed that the two time series are mutually independent. The statistical problem of concern is to make inferences about the unknown difference between the innovation variances, $\sigma_a^2(1)$ and $\sigma_a^2(2)$ say. It is convenient to make this comparison in terms of the ratio, τ say, of the two innovation variances defined as

$$\tau = \frac{\sigma_a^2(2)}{\sigma_a^2(1)}. \tag{13}$$

a. *Model fitting*

Model identification is the first step of the testing procedure. The Bayesian information criterion (BIC) (Katz 1982; Schwarz 1978) or some other technique, such as Akaike's information criterion (Akaike 1974), is applied to the two time series, yielding the selection of AR(p_1) and AR(p_2) processes to model the data. Using (7), the two prewhitened time series are obtained. Then the two innovation variances are estimated from the prewhitened data by (8), giving $\hat{\sigma}_a^2(1)$ and $\hat{\sigma}_a^2(2)$.

b. *Tests of significance*

The test for equality of the two innovation variances,

$$\text{null hypothesis: } \tau = 1 \text{ [i.e., } \sigma_a^2(1) = \sigma_a^2(2)]$$

$$\text{alternative hypothesis: } \tau \neq 1 \text{ [i.e., } \sigma_a^2(1) \neq \sigma_a^2(2)],$$

can be constructed on the basis of (11). Under the null hypothesis, the distribution of the statistic

$$Z = \frac{\ln\hat{\sigma}_a^2(2) - \ln\hat{\sigma}_a^2(1)}{\{s^2[\ln\hat{\sigma}_a^2(1)] + s^2[\ln\hat{\sigma}_a^2(2)]\}^{1/2}} \tag{14}$$

is approximately standard Gaussian for large sample sizes n_1 and n_2. We note that this test statistic (14) does not require that the true means of the two time series, say μ_1 and μ_2, be equal. If desired, the test for equality of means given in Katz (1982) could be conducted concurrently with this test for equality of innovation variances.

c. *Confidence intervals*

Along with a formal test of significance, (14) can be employed to provide a confidence interval for the ratio τ of the two innovation variances. For $\min(n_1, n_2)$ sufficiently large, an approximate $[100(1 - \alpha)]\%$ confidence interval for the difference in the logarithms of the innovation variances is

$$\ln\hat{\sigma}_a^2(2) - \ln\hat{\sigma}_a^2(1) \pm z_{\alpha/2}\{s^2[\ln\hat{\sigma}_a^2(1)] + s^2[\ln\hat{\sigma}_a^2(2)]\}^{1/2}. \quad (15)$$

Here $z_{\alpha/2}$ satisfies

$$\Pr\{Z > z_{\alpha/2}\} = \alpha/2, \quad (16)$$

with Z assumed to have a standard Gaussian distribution. The lower and upper limits of the interval (15) can then be converted into the corresponding limits of an approximate $[100(1 - \alpha)]\%$ confidence interval for the variance ratio τ simply by exponentiation.

5. Preliminary test calculations

To test the operational feasibility of the proposed procedure for making statistical inferences about changes in innovation variance, a data sample from a 3-yr control integration of the OSU atmospheric GCM is used. Three consecutive January simulations and three consecutive July simulations were analyzed. Daily mean surface air temperature for nine grid points at scattered locations in the United States at both 34° and 46°N were examined. The 3-yr January and July datasets each have a sample size of 93 (=3 × 31).

a. *Model fitting*

Table 1 summarizes the results of fitting autoregressive processes to the January and July GCM control time series of daily mean surface air temperature at the nine locations. Time periods consisting of a single month, January or July, were chosen to minimize the effects of seasonal cycles, so that the assumption of stationarity on which the testing procedure is based holds at least approximately. The AR(1) or AR(2) processes were always selected by the BIC procedure, with the orders for the same location sometimes differing between January and July. Except for one grid point (34°N, 115°W), the sample variance S^2 is always greater in January than in July. The estimated innovation variance, computed by means of (8), is also greater in January than in July, with the exception of the same single location.

To demonstrate the effects of prewhitening, we consider the January temperature time series at one particular grid point (34°N, 100°W). Figure 1a shows the time series of daily mean temperature at this location for the second of the 3 yr of a GCM control run. As indicated by the relatively long runs of either above average or below average temperatures, a relatively high degree of autocorrelation is present. An AR(2) process,

TABLE 1. Results of fitting autoregressive processes to time series of January and July GCM control daily mean surface air temperature.

Location	Month	Order selected (p)	Time average (°C) ($\bar{X}$)	Sample variance (°C)^2 (S^2)	Estimated innovation variance (°C)^2 ($\hat{\sigma}_a^2$)
34°N, 115°W	Jan	2	5.40	14.17	5.63
	Jul	2	21.24	25.63	5.76
34°N, 100°W	Jan	2	4.19	29.12	16.17
	Jul	1	29.74	11.62	2.23
34°N, 90°W	Jan	2	6.92	28.29	15.51
	Jul	2	35.33	9.01	2.76
34°N, 80°W	Jan	1	6.36	21.58	14.26
	Jul	2	36.03	7.32	3.27
46°N, 120°W	Jan	1	−1.84	31.96	15.72
	Jul	2	16.90	2.36	0.81
46°N, 100°W	Jan	1	−7.74	39.78	26.47
	Jul	2	19.75	22.65	3.74
46°N, 90°W	Jan	1	−5.97	34.40	18.86
	Jul	2	25.32	26.35	2.73
46°N, 80°W	Jan	2	−7.10	29.08	15.24
	Jul	2	24.13	17.38	4.07
46°N, 65°W	Jan	1	−6.41	37.14	19.47
	Jul	2	23.20	22.52	3.84

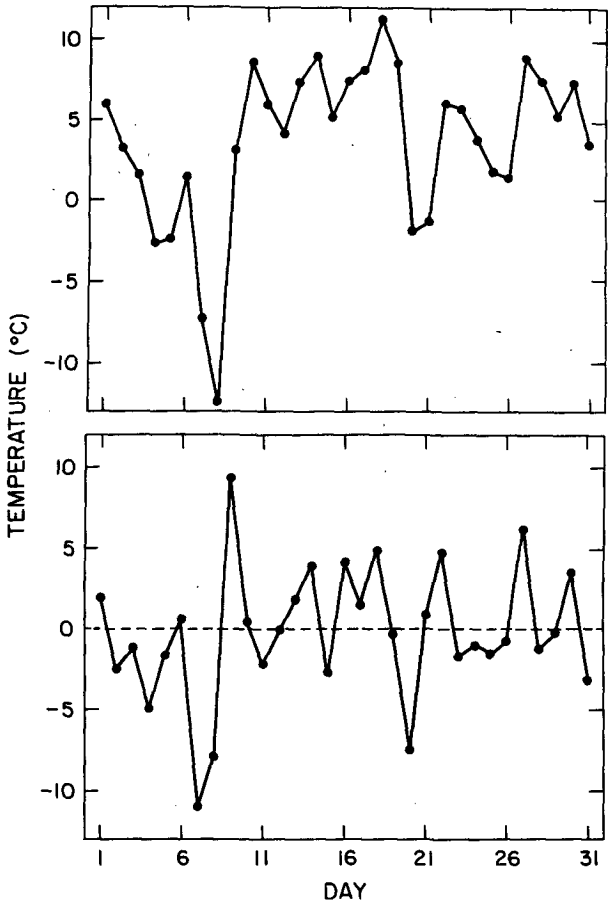

FIG. 1. Time series of January daily mean surface air temperature at one OSU GCM grid point (34°N, 100°W) for the second year of the control run (a) original time series, and (b) prewhitened data with horizontal line indicating mean (necessarily equal to zero).

with estimated autoregression parameters $\hat{\phi}_1 = 0.829$ and $\hat{\phi}_2 = -0.297$, was selected by the BIC procedure to model this time series. The associated prewhitened data are obtained by the transformation (7), which in this case reduces to

$$\hat{a}_t = (X_t - \bar{X}) - 0.829(X_{t-1} - \bar{X}) + 0.297(X_{t-2} - \bar{X}), \quad (17)$$

where $\bar{X} = 4.19$°C. Comparing the prewhitened data shown in Fig. 1b to the original temperature time series, the transformation (17) has evidently removed much, if not all, of the autocorrelation and resulted in a somewhat less variable time series.

b. *Tests of significance*

To illustrate how tests for significant differences between innovation variances may be conducted, the procedure described in section 4b was applied to January and July GCM control runs. Here we are testing whether the January innovation variance equals the July innovation variance. The results of this test of significance are summarized in Table 2. *P*-values [i.e., the probability of obtaining a difference in logarithms of estimated innovation variances larger (in absolute value) than the difference actually observed when, in fact, the true innovation variances are equal] are included in Table 2 as one means of conveying the outcomes of the tests of significance. For eight of the nine locations, the difference in innovation variances is clearly statistically significant (i.e., small *P*-value).

These tests were performed by using the statistic (14), involving a correction for the kurtosis of each of the two samples. In January the sample kurtosis (10) is positive for seven out of nine locations and ranges from about −0.33 to 2.13, whereas in July it is positive for eight out of nine locations and ranges from −0.20 to 16.09. It is interesting to examine the effect of kurtosis on the standard errors given in Table 2. If the correction for kurtosis is not made, then the standard error [i.e., (12) with $\hat{\gamma}_2 = 0$] depends only on the sample size $n = 93$ and is equal to 0.1467. When the January and July standard errors are combined to form the denominator of the test statistic (14), the effect of the correction for kurtosis is to inflate this denominator for each of the nine locations. Not correcting for kurtosis thus would have resulted in *P*-values that were erroneously small. In other words, the effect of unequal innovation variances would be confounded with the effect of non-Gaussian distributions.

c. *Confidence intervals*

Table 3 lists approximate 95% confidence intervals for the ratio τ of January to July innovation variances

TABLE 2. Tests for equality of January and July innovation variances.

Location	$\ln\hat{\sigma}_a^2$ (Standard error) January	July	Test statistic (Z)	*P*-value
34°N, 115°W	1.728 (0.1534)	1.751 (0.4410)	−0.05	0.96
34°N, 100°W	2.783 (0.1604)	0.800 (0.3327)	5.37	$<10^{-4}$
34°N, 90°W	2.741 (0.1373)	1.106 (0.2396)	6.25	$<10^{-4}$
34°N, 80°W	2.658 (0.1673)	1.185 (0.1702)	6.17	$<10^{-4}$
46°N, 120°W	2.755 (0.2107)	−0.207 (0.1391)	11.73	$<10^{-4}$
46°N, 100°W	3.276 (0.1341)	1.319 (0.3450)	5.29	$<10^{-4}$
46°N, 90°W	2.937 (0.1520)	1.005 (0.1966)	7.77	$<10^{-4}$
46°N, 80°W	2.724 (0.1773)	1.403 (0.2663)	4.13	$<10^{-4}$
46°N, 65°W	2.969 (0.1500)	1.345 (0.1513)	7.62	$<10^{-4}$

TABLE 3. Confidence intervals for ratio (January to July) of innovation variances.

Location	Estimated ratio of variances	95% confidence interval for ratio of variances τ	
		Lower limit	Upper limit
34°N, 115°W	0.977	0.391	2.441
34°N, 100°W	7.265	3.522	14.985
34°N, 90°W	5.613	3.267	9.642
34°N, 80°W	4.362	2.732	6.965
46°N, 120°W	19.337	11.788	31.718
46°N, 100°W	7.078	3.427	14.620
46°N, 90°W	6.903	4.242	11.235
46°N, 80°W	3.747	2.001	7.016
46°N, 65°W	5.073	3.342	7.702

for each of the nine locations. Since eight of the tests for equality of innovation variances had P-values less than 0.05 (Table 2), the corresponding confidence intervals necessarily do not include the value $\tau = 1$ (i.e., equal January and July innovation variances). The intervals, nevertheless, are rather wide. If the correction for kurtosis in (15) had not been made, the resultant confidence intervals would have been too short. As mentioned in section 5b, the standard errors would be underestimated, thus yielding intervals whose associated confidence coefficients $1 - \alpha$ would actually be smaller than 0.95.

The standard errors included in Table 2 can be converted into approximate estimates of the magnitude of change in innovation variance that could be detected in an atmospheric GCM climate experiment. The smallest proportionate increase (decrease) in innovation variance (i.e., $\tau - 1$) that has roughly a 50% chance of being identified as statistically significant (i.e., P-value smaller than 0.05) ranges from about 0.45 to 0.79 (-0.44 to -0.31) in January and from 0.47 to 2.39 (-0.71 to -0.32) in July. Hence, quite substantial changes in monthly innovation variances (a 40% increase or a 30% decrease, for example) would not likely be detected in a GCM climate experiment consisting of 3 years of control runs and 3 years of experiment runs. If more GCM runs were available, smaller changes in innovation variances could, of course, be identified as statistically significant. Another way to increase the control and experiment sample sizes would be to consider time periods larger than a month (for example, a winter or summer season 3 months in length). Unfortunately, in this case the assumption that the GCM time series are stationary (in particular, that the innovation variance is constant) would be more questionable.

6. Concluding remarks

A statistical procedure for making inferences about changes in climate variability has been presented. By defining climate variability in terms of the innovation variance, the difficulties that arise because atmospheric time series are autocorrelated have been circumvented. By making an adjustment involving the sample kurtosis of the prewhitened data, the procedure is not sensitive to departures from the assumption of a Gaussian distribution. The application of the technique has been demonstrated through the use of control data generated by the OSU atmospheric GCM. The test application provides standard errors of January and July estimated innovation variances. These standard errors can be employed to obtain estimates of the magnitude of change in climate variability that the procedure should be likely to detect.

The consideration of how to define climate variability suggests that the problem of testing for a change in the autocorrelation structure of climate time series needs to be addressed. Besides being of interest for its own sake, this problem must be resolved in order to make inferences about changes in measures of climate variability, other than the innovation variance, that might also be relevant. These other measures of climate variability include the process variance and time-average variances. We note that Zwiers and Thiebaux (1987) have studied the performance of the F-test for the equality of variances of time averages. Further, it would be of interest to make inferences about changes in other climate statistics such as the relative frequency of extreme events (e.g., Mearns et al. 1984).

The definition of intrinsic variability suggests that a systematic examination of the statistical characteristics of prewhitened climate data might be of interest. Such statistics should be useful, for instance, in climate predictability studies (e.g., Chu and Katz 1987). Further, prewhitening has proved helpful in identifying the nature of the relationships among autocorrelated time series (e.g., Granger and Newbold 1977). Katz (1988), in particular, has advocated the use of prewhitening in teleconnection studies.

Acknowledgments. The author thanks W. Lawrence Gates for suggesting this topic, William McKie and Robert Mobley for providing programming assistance, and Steven Esbensen for supplying comments on this work. Helpful suggestions by two referees are also acknowledged. Some of this research was reported in "Procedures for Determining the Statistical Significance of Changes in Variability Simulated by an Atmospheric General Circulation Model," Climatic Research Institute Report No. 48, Oregon State University (September 1983). Part of this research was conducted while the author was a member of the Department of Atmospheric Sciences at Oregon State University. This research was partially supported by the National Science Foundation under Grants ATM 80-01702 and ATM 82-05992.

REFERENCES

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

Box, G. E. P., 1953: Non-normality and tests on variances. *Biometrika,* **40,** 318–335.

——, and G. M. Jenkins, 1976: *Time Series Analysis: Forecasting and Control.* Rev. ed., Holden-Day, 575 pp.

Chervin, R. M., 1980: On the simulation of climate and climate change with general circulation models. *J. Atmos. Sci.,* **37,** 1903–1913.

——, and S. H. Schneider, 1976: On determining the statistical significance of climate experiments with general circulation models. *J. Atmos. Sci.,* **33,** 405–412.

Chu, P.-S., and R. W. Katz, 1987: Measures of predictability with applications to the Southern Oscillation. *Mon. Wea. Rev.,* **115,** 1542–1549.

Davis, W. W., 1977: Robust interval estimation of the innovation variance of an ARMA model. *Ann. Statist.,* **5,** 700–718.

——, 1979: Robust methods for detection of shifts of the innovation variance of a time series. *Technometrics,* **21,** 313–320.

Durbin, J., 1960: The fitting of time series models. *Rev. Int. Inst. Statist.,* **28,** 233–244.

Efron, B., 1982: *The Jackknife, the Bootstrap and Other Resampling Plans.* Society for Industrial and Applied Mathematics, 92 pp.

Granger, C. W. J., and P. Newbold, 1977: *Forecasting Economic Time Series.* Academic Press, 333 pp.

Hayashi, Y., 1982: Confidence intervals of a climatic signal. *J. Atmos. Sci.,* **39,** 1895–1905.

Hollander, M., and D. A. Wolfe, 1973: *Nonparametric Statistical Methods.* Wiley, 503 pp.

Jones, R. H., 1975: Estimating the variance of time averages. *J. Appl. Meteor.,* **14,** 159–163.

Katz, R. W., 1982: Statistical evaluation of climate experiments with general circulation models: A parametric time series modeling approach. *J. Atmos. Sci.,* **39,** 1446–1455.

——, 1988: Use of cross correlations in the search for teleconnections. *J. Climatol.* (in press).

Laurmann, J. A., and W. L. Gates, 1977: Statistical considerations in the evaluation of climate experiments with atmospheric general circulation models. *J. Atmos. Sci.,* **34,** 1187–1199.

Livezey, R. E., 1985: Statistical analysis of GCM climate simulation, sensitivity, and prediction experiments. *J. Atmos. Sci.,* **42,** 1139–1149.

Madden, R. A., 1979: A simple approximation for the variance of meteorological time averages. *J. Appl. Meteor.,* **18,** 703–706.

Mearns, L. O., R. W. Katz and S. H. Schneider, 1984: Extreme high-temperature events: Changes in their probabilities with changes in mean temperature. *J. Climate Appl. Meteor.,* **23,** 1601–1613.

Miller, R. G., 1968: Jackknifing variances. *Ann. Math. Statist.,* **39,** 568–582.

Pierce, D. A., 1979: R^2 measures for time series. *J. Amer. Statist. Assoc.,* **74,** 901–910.

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Statist.,* **6,** 461–464.

Wilson, C. A., and J. F. B. Mitchell, 1987: Simulated climate and CO_2-induced climate change over Western Europe. *Climatic Change,* **10,** 11–42.

WMO, 1986: Report of the international conference on the assessment of the role of carbon dioxide and of other greenhouse gases in climate variations and associated impacts, Villach, Austria. WMO No. 661, World Meteorological Organization, Geneva, 78 pp.

Zwiers, F. W., and H. J. Thiebaux, 1987: Statistical considerations for climate experiments. Part I: Scalar tests. *J. Climate Appl. Meteor.,* **26,** 464–476.