Statistical Evaluation of Climate Experiments with General Circulation Models: A Parametric Time Series Modeling Approach

RICHARD W. KATZ

Climatic Research Institute and Department of Atmospheric Sciences, Oregon State University, Corvallis 97331

(Manuscript received 13 October 1981, in final form 19 February 1982)

ABSTRACT

A procedure for making statistical inferences about differences between population means from the output of general circulation model (GCM) climate experiments is presented. A parametric time series modeling approach is taken, yielding a potentially more powerful technique for detecting climatic change than the simpler schemes used heretofore. The application of this procedure is demonstrated through the use of GCM control data to estimate the variance of winter and summer time averages of daily mean surface air temperature. The test application provides estimates of the magnitude of climatic change that the procedure should be able to detect. A related result of the analysis is that autoregressive processes of higher than first order are needed to adequately model the majority of the GCM time series considered.

1. Introduction

Making statistical inferences from the outcomes of general circulation model (GCM) climate experiments is an important, but difficult, task. It is an important task because only if statistical significance is established can the results of an experiment be considered conclusive; otherwise, any "climatic changes" reportedly discovered could just as well be attributed to the chance variation of essentially unpredictable natural fluctuations. It is a difficult task because of the statistical characteristics of time series generated by a GCM. Like real atmospheric measurements, these simulated time series are highly correlated both in time and space. Such correlations make the standard methodological techniques on which statistical inferences are based inappropriate, since they depend on the assumption of independence.

Specific statistical problems that arise with GCM experiments and possible techniques for handling these difficulties have been discussed, for example, by Chervin and Schneider (1976) and Laurmann and Gates (1977). Most of the approaches taken so far have been univariate in nature, treating each grid point and atmospheric variable separately, with the technique proposed by Chervin and Schneider (1976) being most widely employed. To obtain an estimate of the variance of time averages, the Chervin-Schneider technique involves simply comparing the time averages of several GCM runs. This procedure, however, does not provide a very precise estimate of the variance of time averages because relatively few runs are ever available. The procedure, moreover, being based solely on time averages, does not use all of the information provided by GCM simulations. Consequently, it should not be as powerful a technique for detecting climate changes as other, possibly more complex, procedures to be considered in this paper.

The estimation of the variance of time averages of an atmospheric variable through the use of time series models has been proposed by several authors (Jones, 1975, 1976; Leith, 1973; Madden, 1979, 1981). If an adequate time series model can be identified to take into account the autocorrelation (i.e., the dependence through time) of an atmospheric variable, then an estimate of the variance of time averages of the variable can be obtained, even when only a single realization of the time series is available. Such a parametric time series modeling approach will be taken in this paper. One major obstacle to this type of approach is the identification of the appropriate time series model. Some authors (Leith, 1973; Madden, 1979, 1981) have assumed that atmospheric time series can be adequately represented by Markov (i.e., first-order autoregressive) processes. However, recent evidence (Chervin, 1980; Straus and Halem, 1981) suggests that the Markovian assumption may be violated by both real atmospheric measurements and simulated time series. To allow for this possibility, we will fit autoregressive processes of higher than first order, as well as first-order and zero-order (i.e., uncorrelated) processes, to the GCM time series and rely on an automatic model selection criterion to choose the appropriate order.

The parametric time series modeling approach to estimating the variance of time averages is introduced in Section 2. The application of this methodology to GCM climate experiments is discussed

0022-4928/82/071446-10$06.50

in Section 3, and Section 4 presents the results of a test application of the statistical procedure to control data generated by the Oregon State University (OSU) atmospheric GCM. A comparison of the parametric time series modeling approach with other procedures is made in Section 5. Finally, Section 6 consists of some concluding remarks, including suggestions for future research.

2. Parametric time series modeling

In this section the parametric time series modeling methodology necessary to estimate the variance of time averages for an atmospheric variable is described. For clarity, only a single time series is considered. The extension of this methodology to the problem of comparing two or more time series, a problem of specific interest in GCM climate experiments, will be postponed until Section 3a. It should be noted that the procedure to be presented is quite similar to that outlined by Jones (1975).

We are given a single time series that can be viewed as a realization of a stochastic process $\{X_t: t = 1, 2, \cdots\}$. This stochastic process is taken to be stationary with unknown population mean (or ensemble mean) $\mu = E(X_t)$, $t = 1, 2, \ldots$, and unknown population variance (or ensemble variance) $\sigma^2 = \text{var}(X_t)$, $t = 1, 2, \ldots$, with $0 < \sigma^2 < \infty$. Here E denotes expected value and var denotes variance. Further, we assume that X_t has a Gaussian distribution. Additional operations that may need to be applied to GCM simulated time series in order to satisfy these assumptions will be discussed in Section 3b.

The parametric time series modeling approach involves two basic tasks. First, the appropriate model (i.e., order of autoregressive process) to be fitted to the data is identified. To perform this task, one particular automatic model selection criterion is proposed in Section 2a. Second, based on the parametric time series model chosen, an expression for the variance of time averages can be derived. An estimator of this variance and related results regarding the distribution of an associated test statistic are given in Section 2b.

a. *Model identification*

We assume that the time series of concern can be represented as an autoregressive process for some unknown order p, $0 \leqslant p \leqslant p_u$. Here p_u is an upper bound, specified by the experimenter, for this unknown order. More complex stochastic processes, such as autoregressive-moving average (ARMA) processes (e.g., Box and Jenkins, 1976) could also be considered. Such processes might be expected to result in a more parsimonious representation of GCM simulated time series. Nevertheless, we have found that low-order autoregressive processes are adequate to model these time series (see Section 4).

A pth-order autoregressive process (Box and Jenkins, 1976), denoted by AR(p), can be expressed as

$$\sum_{k=0}^{p} \phi_k(X_{t-k} - \mu) = a_t, \quad t = p + 1, p + 2, \ldots, \quad (1)$$

with $\phi_0 = 1$ and the other autoregression coefficients ϕ_k, $k = 1, 2, \ldots, p$, being unknown parameters. In (1) it is required that the a_t's (sometimes called "innovations" or "shocks") constitute a "white noise" process; that is, they are uncorrelated random variables with zero mean $[E(a_t) = 0, t = 1, 2, \cdots]$ and constant variance $[\text{var}(a_t) = \sigma_a^2, t = 1, 2, \cdots]$. To make statistical inferences about the X_t process, the additional assumption that the a_t's have a Gaussian distribution is also necessary. We note that an AR(1) process is commonly referred to as a Markov or "red noise" process, whereas an AR(0) process is an uncorrelated process.

Constraints must be placed on the possible values of the parameters specified in (1) for the X_t process to be stationary as assumed. Specifically, the roots (as a function of x) of

$$\sum_{k=0}^{p} \phi_k x^k = 0 \quad (2)$$

must lie outside the unit circle (Box and Jenkins, 1976, p. 53). For example, when dealing with an AR(1) process, stationarity is ensured by requiring $|\phi_1| < 1$.

Estimates, $\hat{\phi}_k(p)$ say, $k = 1, 2, \ldots, p$, are needed to fit an AR(p) process to the data. To determine the appropriate order, an estimate, $\hat{\sigma}^2(p)$ say, of the white noise variance σ_a^2 is also required for each potential order p, $p = 0, 1, \ldots, p_u$. Given a single time series $\{X_t: t = 1, 2, \ldots, n\}$ of n consecutive values realized by the stochastic process, the following statistics are calculated directly from the data:

(i) The time average

$$\bar{X} = n^{-1} \sum_{t=1}^{n} X_t . \quad (3)$$

(ii) The sample autocovariances

$$c_k = n^{-1} \sum_{t=k+1}^{n} (X_{t-k} - \bar{X})(X_t - \bar{X}), \quad k = 0, 1, \ldots, p_u . \quad (4)$$

We note that c_0 is the sample variance; that is,

$$c_0 = n^{-1} \sum_{t=1}^{n} (X_t - \bar{X})^2. \quad (5)$$

From these statistics [Eqs. (3) and (4)] the Yule-Walker recursive method of calculation (Box and

Jenkins, 1976, p. 82) is used to obtain estimates of the autoregression coefficients and of the white noise variance. This parameter estimation technique is spelled out in the Appendix. One alternative method of estimation would be to employ multiple regression analysis, taking X_t as the dependent variable and $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$, as the independent variables. Another method of parameter estimation, based on the concept of maximum entropy, has been proposed by Burg (see Ulrych and Bishop, 1975).

One procedure for selecting the appropriate order p of the autoregressive process is called the Bayesian information criterion (BIC) introduced by Schwarz (1978). This procedure requires that all possible orders p, $p = 0, 1, \ldots, p_u$, be fitted to the data. Then the value of p is selected that minimizes the quantity

$$\mathrm{BIC}(p) = n \ln\tilde{\sigma}^2(p) + (p+1)\ln n, \quad p = 0, 1, \ldots, p_u, \tag{6}$$

where

$$\tilde{\sigma}^2(p) = \frac{n}{n-p-1}\hat{\sigma}^2(p). \tag{7}$$

Here $\tilde{\sigma}^2(p)$ is an unbiased estimator of the white noise variance for an AR(p) process.

Although $\hat{\sigma}^2(p)$ is most conveniently calculated by the method specified in the Appendix [(A1)-(A5)], we note that

$$\tilde{\sigma}^2(p) \approx \frac{1}{n-p-1}\sum_{t=1}^{n}[X_t - \hat{X}_t(p)]^2, \tag{8}$$

where $\hat{X}_t(p)$ denotes the fitted value of X_t obtained, at least formally, by substituting the estimated autoregression coefficients $\hat{\phi}_k(p)$, $k = 1, 2, \ldots, p$, into (1), along with the time average $\bar{X}$ in place of the population mean μ. Thus, the first term on the right-hand side of (6) can be thought of as a measure of how well an AR(p) process fits the data. The second term on the right-hand side of (6) is a penalty function for the $p+1$ parameters $(\mu, \phi_1, \phi_2, \ldots, \phi_p)$ that need to be estimated. The use of this penalty function results in a parsimonious procedure, balancing the goodness-of-fit of the model against the complexity of the model. Under certain assumptions, (6) is directly related to the approximate posterior probability of an AR(p) process. So, in some sense, the BIC procedure corresponds to choosing the most likely model.

The BIC procedure is a consistent estimator of the order of an autoregressive process (Hannan, 1980); that is, the probability of selecting the correct order converges to one as the sample size n tends to infinity. In addition, this procedure has been applied to model real atmospheric measurements and related time series (Katz and Skaggs, 1981; Katz, 1981). Other model selection criteria, such as Akaike's information criterion (AIC) (e.g., Akaike, 1974) could be employed instead of the BIC. However, the BIC procedure is more parsimonious than the AIC procedure, necessarily resulting in as low or lower selected orders of autoregressive processes.

b. Distribution of time averages

In this section we assume that the order p of the autoregressive process has already been chosen on the basis of the BIC procedure (6), with the selected model having estimated autoregression coefficients $\hat{\phi}_k(p)$, $k = 1, 2, \ldots, p$. Even under dependence, the time average $\bar{X}$ is an unbiased estimator of the population mean μ; that is,

$$E(\bar{X}) = \mu. \tag{9}$$

When the sample size n is large, a convenient expression for the approximate variance of time averages of an AR(p) process is

$$\mathrm{var}(\bar{X}) \approx n^{-1}V_p^2, \tag{10}$$

where

$$V_p^2 = \frac{\sigma_a^2}{[\sum_{k=0}^{p}\phi_k]^2} \tag{11}$$

(Anderson, 1971, Section 5.2).

If the order p and the autoregression coefficients ϕ_k, $k = 1, 2, \ldots, p$, are known, the distribution of the statistic

$$\frac{\bar{X}-\mu}{n^{-1/2}V_p} \tag{12}$$

converges to a standard Gaussian distribution (i.e., zero mean and unit variance) as the sample size n tends to infinity (Anderson, 1971, Section 5.2). This result can be viewed as a generalization of the central limit theorem for independent processes. When the order is known but the autoregression coefficients are unknown and have to be estimated, the variance of time averages can be estimated by substituting $\hat{\phi}_k(p)$ in place of ϕ_k and $\hat{\sigma}^2(p)$ in place of σ_a^2 in (11), yielding

$$\widehat{\mathrm{var}}(\bar{X}) = n^{-1}\hat{V}_p^2, \tag{13}$$

where

$$\hat{V}_p^2 = \frac{\hat{\sigma}^2(p)}{[\sum_{k=0}^{p}\hat{\phi}_k(p)]^2}. \tag{14}$$

The distribution of the statistic

$$\frac{\bar{X}-\mu}{n^{-1/2}\hat{V}_p}, \tag{15}$$

a modification of (12), still converges to a standard Gaussian distribution as the sample size n tends to infinity (Albers, 1978). Alternatively, the variance of time averages could be estimated by using $\tilde{\sigma}^2(p)$, rather than $\hat{\sigma}^2(p)$, in (14).

If, in addition to estimating the autoregression coefficients, the order p must also be selected, the distribution of the statistic (15) may still converge to a standard Gaussian distribution, depending on the properties of the model selection criterion employed. Since the BIC procedure is consistent (see Section 2a), its use results in (13) being a consistent estimator of the variance of time averages (10). In this case, the central limit theorem for the statistic (15) is still in force.

The expression (13) for the estimated variance of time averages for an AR(p) process has been compared, in the special case of an AR(1) process, to the expression for the estimated variance of time averages for an independent process (e.g., Madden, 1979), yielding a quantity commonly referred to as the "effective degrees of freedom" (or "effective number of independent samples"). We emphasize, however, that no theoretical results are available to justify the use of a t-distribution, based on the effective degrees of freedom, for the statistic (15) when the sample size n is relatively small. Jones (1976) mentions this difficulty. Of course, the standard Gaussian distribution can be used to approximate the distribution of (15) when n is large. How large n needs to be for this method of approximation to work well has not been established. Some partial guidance on this unanswered question is the result that the t-distribution is quite closely approximated by a Gaussian distribution for n greater than some lower bound, 50 say, in the case of independent observations.

3. Application to GCM climate experiments

In GCM climate experiments, we are given a control time series $\{X_t(c)\colon t = 1, 2, \ldots, n_c\}$ with unknown population mean $\mu_c = E[X_t(c)]$ and unknown population variance $\sigma_c^2 = \text{var}[X_t(c)]$, and an experiment time series $\{X_t(e)\colon t = 1, 2, \ldots, n_e\}$ with unknown population mean $\mu_e = E[X_t(e)]$ and unknown population variance $\sigma_e^2 = \text{var}[X_t(e)]$. It is assumed that the experiment time series $X_t(e)$ is independent of the control time series $X_t(c)$. The statistical problem of concern is to make inferences about the unknown change in population means $\mu_e - \mu_c$. To apply the results of Section 2, the control and experiment time series must satisfy the requirements stated in that section. For the moment, we assume that these requirements are indeed satisfied.

a. Testing procedure

Model identification is the first step of the testing procedure. The BIC procedure (see Section 2a) is applied to the control time series, yielding the selection of an AR(p_c) process to model the data. Using (13), the variance of the control time average $\bar{X}_c$ can be estimated as

$$\widehat{\text{var}}(\bar{X}_c) = n_c^{-1}\hat{V}_{p_c}^2(c). \tag{16}$$

In a similar fashion, the BIC procedure is applied to the experiment time series, yielding the choice of an AR(p_e) process with an estimate of the variance of the experiment time average $\bar{X}_e$ being

$$\widehat{\text{var}}(\bar{X}_e) = n_e^{-1}\hat{V}_{p_e}^2(e). \tag{17}$$

Next a test statistic can be constructed on the basis of (15). The distribution of the statistic

$$Z = \frac{\bar{X}_e - \bar{X}_c}{[n_c^{-1}\hat{V}_{p_c}^2(c) + n_e^{-1}\hat{V}_{p_e}^2(e)]^{1/2}} \tag{18}$$

under the null hypothesis of no experimental effect (i.e., $\mu_e - \mu_c = 0$), converges to a standard Gaussian distribution as the sample sizes n_c and n_e both tend to infinity. For $\min(n_c, n_e)$ sufficiently large, the test statistic (18) can be used to assess whether the difference between the control and experimental time averages is statistically significant, by means of this Gaussian approximation.

We note that separate autoregressive processes, differing both in estimated autoregression coefficients and possibly order, are fit to the control and experiment time series. This approach is taken because it does not require the assumption that the population variances of the two time series are equal. For this reason, the denominator of the test statistic (18) is not a pooled estimator of the standard deviation of the difference between the two time averages, as is sometimes employed in two-sample t-tests.

Along with a formal test of significance, (18) can be used to derive additional information of interest in assessing GCM climate experiments. The observed significance level or probability value (P-value) associated with the test statistic Z is

$$P = \Pr\{|Z| > |z|\}, \tag{19}$$

where z is the observed sample value of the test statistic (18). It can be computed by employing the Gaussian approximation mentioned earlier. Low P-values, for example, values below 0.05, indicate strong evidence for rejection of the hypothesis of equal population means. These P-values are helpful in making comparisons among the outcomes of several hypothesis tests (e.g., for several grid points).

Further, a confidence interval for the change in expected values $\mu_e - \mu_c$ can be determined. For $\min(n_c, n_e)$ sufficiently large, an approximate $[(100)(1 - \alpha)]\%$ confidence interval for $\mu_e - \mu_c$ is

$$\bar{X}_e - \bar{X}_c \pm z_{\alpha/2}[n_c^{-1}\hat{V}_{p_c}^2(c) + n_e^{-1}\hat{V}_{p_e}^2(e)]^{1/2}, \tag{20}$$

where $z_{\alpha/2}$ satisfies

$$\Pr\{Z > z_{\alpha/2}\} = \alpha/2. \quad (21)$$

For instance, if a 95% confidence interval is desired, using the Gaussian approximation with $\alpha = 0.05$ gives $z_{\alpha/2} = 1.96$. While hypothesis tests simply indicate whether or not a climatic change has occurred, confidence intervals provide information concerning the magnitude of any climatic change.

b. GCM data

GCM simulated time series may need to be modified in several respects before applying the testing procedure outlined in Section 3.1. Here we mention several such considerations.

1) Model order

The upper bound p_u for the order of the autoregressive process fitted to the data needs to be specified. As small as possible a value for p_u should be chosen to minimize the computations required. Based on some preliminary test calculations (see Section 4), setting $p_u = 5$ appears to be satisfactory.

2) Pooled data

The statistical calculations are slightly more complicated if the data consist of GCM runs for more than one year, say l runs each of length n' for a total sample size of $n = n'l$. In this case, the only change required is a revision of (4) for calculating the sample autocovariances; now

$$c_k = n^{-1} \sum_{j=1}^{l} \sum_{t=(j-1)n'+k+1}^{jn'} (X_{t-k} - \bar{X})(X_t - \bar{X}). \quad (22)$$

3) Stationarity

To remove the diurnal cycle, the GCM output should be converted to daily mean data through averaging. The time period considered should also be short enough to minimize the effects of seasonal cycles (i.e., a winter season or a summer season of at most three months). If seasonal time series are modeled, it may be necessary to take into account gradual changes in the mean throughout the season that might be present. It is recommended that monthly sample autocovariances be computed by (22), allowing the time average to differ for each month within the season. Then a single seasonal autocovariance value is obtained by taking a weighted mean (with the weights proportional to the number of days in each month) of these monthly autocovariances. If necessary, more complex methods could be employed to remove the effects of possible nonstationarity on the estimated autocorrelation function.

4) Gaussian distribution

The present statistical testing procedure should only be applied to data which have at least an approximately Gaussian distribution. Some atmospheric variables, and presumably GCM simulated variables, have distributions that are highly non-Gaussian, although sometimes transformations (e.g., logarithm or square root) can be applied to obtain data having an approximately Gaussian distribution (Hinkley, 1977). The statistical testing procedure can then be validly applied to the transformed data.

4. Preliminary test calculations

To test the operational feasibility of the proposed statistical procedure, a data sample from a three-year control integration of the OSU atmospheric GCM is used. Three consecutive winter season simulations (1 December–28 February) and three consecutive summer season simulations (1 June–31 August) were analyzed. The atmospheric variable examined was the daily mean surface air temperature at 3312 locations, consisting of a grid of points covering the entire globe with a spacing of 4° latitude and 5° longitude. The three-year winter season data sets each have 270 ($=3 \times 90$) observations, while the three-year summer season data sets each have 276 ($=3 \times 92$) observations. To simplify the presentation of the results, for the most part only the region consisting of all land from 10 to 80°N latitude and from 150 to 35°W longitude will be considered. These 160 grid points essentially constitute the North American continent.

a. Model identification

The BIC procedure (described in Section 2a) was employed to select the appropriate order autoregressive process for both the three-year winter and summer GCM control data sets. The unknown order p of the autoregressive process was assumed to be no greater than 5 (i.e., $p_u = 5$). Monthly sample autocovariances were calculated by (22) and then weighted (as described in Section 3b) to obtain seasonal autocovariance values. The parameters of the autoregressive processes were estimated using the Yule-Walker recursive method of calculation (see Appendix).

The orders chosen by this procedure for the daily mean surface air temperature time series over the North American continent are summarized in Table 1 and shown in Figs. 1 and 2. The value of $p = 2$ was most frequently selected, with the next most frequent choice being $p = 1$ in winter and $p = 3$ in summer. The value of $p = 0$ (corresponding to an uncorrelated process) was never chosen, and the selected order in the vast majority of cases was greater

TABLE 1. Selection of order of autoregressive process for winter and summer GCM control daily mean surface air temperature time series.

Order	Frequency of selection	
	Winter	Summer
0	0 (0%)	0 (0%)
1	45 (28%)	23 (14%)
2	108 (68%)	86 (54%)
3	7 (4%)	48 (30%)
4	0 (0%)	3 (2%)
5	0 (0%)	0 (0%)
Total	160 (100%)	160 (100%)

than 1. On the other hand, the selected order was rarely higher than 2 in winter or 3 in summer, making the upper bound on the autoregressive order ($p_u = 5$) a reasonable value. When data for the entire globe were considered, similar results were obtained.

b. *Standard deviation of time averages*

Given the selected orders of the autoregressive processes and the estimated autoregression coefficients, the estimated variance (or, equivalently, standard deviation) of the winter and summer time averages for each data set was computed using (16). The estimated standard deviations of the three-year winter and summer time averages of daily mean surface air temperature over the North American continent are shown in Figs. 3 and 4. In the winter these standard deviations range from a minimum of 0.14°C and values frequently less than 0.5°C in low latitudes to a maximum of 1.81°C and values frequently above 1.0°C in high latitudes. In the summer these standard deviations range from a minimum of 0.16°C to a maximum of 1.04°C.

As in the case of an uncorrelated process, the variance of time averages [Eq. (10)] is inversely proportional to the sample size. Thus, the estimated standard deviations of three-year seasonal time averages ($n_c = 270$ for winter, $n_c = 276$ for summer) could be converted to the estimated standard deviations of time averages based on a different sample size, n'_c say, simply by multiplying by the factor d, where

$$d = (n_c/n'_c)^{1/2}. \tag{23}$$

Here the new sample size n'_c must also be large.

The standard deviations can be converted into at least approximate estimates of the magnitude of the climate change that could be detected in a GCM experiment. Under the assumption that the variance of the experiment time averages equals the variance of the control time averages, the denominator of the test statistic (18) equals $2^{1/2}$ times the estimated standard deviation of the control time averages. So, if three years of control and experiment runs were available, then we would expect that any climatic change larger than roughly the factor b times the estimated standard deviation, where

$$b = (2^{1/2})(1.96) = 2.77, \tag{24}$$

to be identified as statistically significant (i.e., $P < 0.05$). For example, the areas in Figs. 3 and 4 with standard deviations between 0.5 and 1.0°C corre-

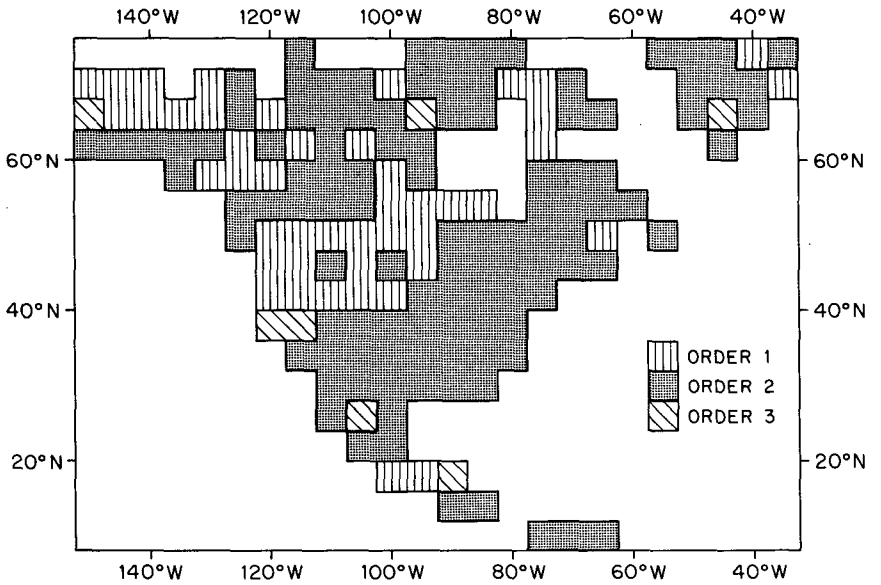

FIG. 1. Selection of order of autoregressive process for winter GCM control daily mean surface air temperature time series.

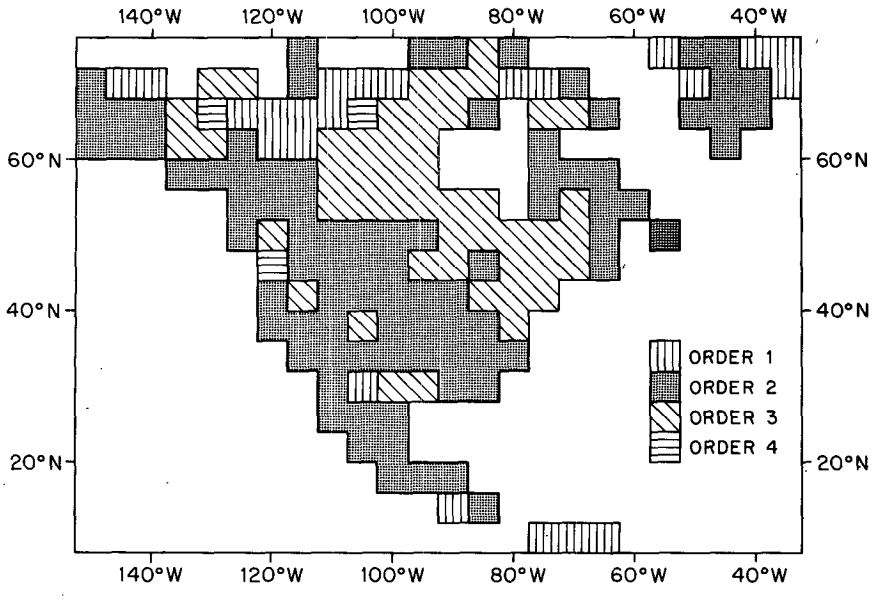

FIG. 2. Selection of order of autoregressive process for summer GCM control daily mean surface air temperature time series.

spond to areas for which climatic changes of ~1.4–2.8°C could be detected with such an experiment.

Although autoregressive processes of higher than first order were selected to model OSU GCM temperature time series at many grid points, it is still reasonable to question whether always fitting AR(1) processes instead would have any effect on the tests of significance. In this regard, which particular order autoregressive process is chosen only matters insofar as its effect on the estimated variance of time averages. To address this question, a subset of the grid points contained within the United States were considered in detail. For those cases, whether winter or summer, in which an autoregressive process of higher than first order was selected, the estimated standard deviation of time averages was compared with the estimate based on fitting an AR(1) process instead. The differences between the two estimates are relatively small but very systematic. In nearly every case, the estimated standard deviation using an

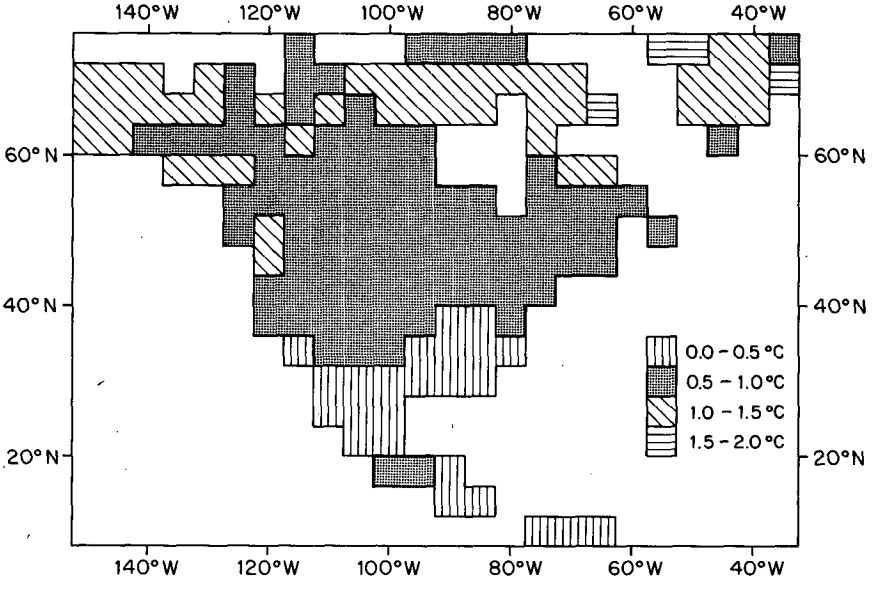

FIG. 3. Standard deviation of time averages for winter GCM control daily mean surface air temperature time series.

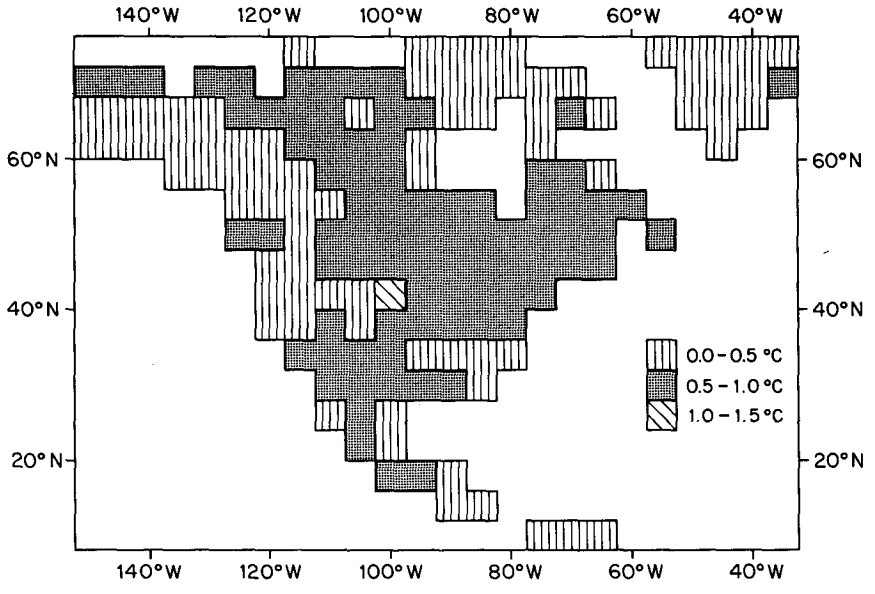

FIG. 4. Standard deviation of time averages for summer GCM control daily mean surface air temperature time series.

AR(1) process is greater than the estimated standard deviation using an autoregressive process of higher than first order. The typical difference in estimated standard deviations is ~0.15°C. This difference can be converted by (24) into a difference in the smallest climatic change that could be detected of ~0.4°C. The apparent reason for the differences is that the sample autocorrelation function decays toward zero at a rate faster than the geometric rate of the theoretical autocorrelation function for an AR(1) process. This sort of behavior was detected by Straus and Halem (1981) for both observed and GCM time series of several atmospheric variables including temperature.

c. *Tests of significance*

To demonstrate how tests for significant differences between time averages from GCM control and experiment time series may be conducted, the test statistic (18) was applied to winter and summer GCM control runs. Here, for illustrative purposes, we present the results for only one particular grid point, located at longitude 100°W and latitude 34°N. The time averages of winter and summer three-year seasonal daily mean surface air temperatures are 5.01 and 28.40°C, respectively. The problem at hand is to test whether this difference in time averages (summer minus winter) of 23.39°C is statistically significant. For both the winter and summer time series, second-order autoregressive processes were selected, with the estimated autoregression coefficients being $\hat{\phi}_1(2) = -0.853$ and $\hat{\phi}_2(2) = 0.294$ in the winter and $\hat{\phi}_1(2) = -1.114$ and $\hat{\phi}_2(2) = 0.271$ in the summer. The estimated white noise variances are $\hat{\sigma}^2(2) = 14.882$ $(°C)^2$ for the winter and $\hat{\sigma}^2(2) = 2.484$ $(°C)^2$ for the summer, resulting in estimated standard deviations of the time averages of 0.532 and 0.604°C, respectively. Substituting these values into (18) yields a test statistic value of $Z = 29.05$ ($P \approx 0$). Consequently, we conclude that this difference between summer and winter time averages is statistically significant. Using (20), an ~95% confidence interval for the difference between summer and winter population means is 23.39 ± 1.58°C.

5. Comparison with other procedures

Other methods for estimating the variance of time averages of atmospheric data have been proposed and applied to GCM climate experiments (Section 1). In this section we first compare the parametric time series modeling approach presented here with the technique proposed by Chervin and Schneider (1976). Both the test statistic (18) for the parametric time series modeling approach and the test statistic for the Chervin-Schneider procedure (Chervin and Schneider, 1976, Eq. (2)) can be written in the form

$$\text{Test statistic} = \frac{\Delta}{[\widehat{\text{var}}(\Delta)]^{1/2}}, \tag{25}$$

where

$$\Delta = \bar{X}_e - \bar{X}_c \tag{26}$$

and $\widehat{\text{var}}(\Delta)$ is an estimator of $\text{var}(\Delta)$.

These two procedures, of course, differ in the method of estimating $\text{var}(\Delta)$. Because the parametric time series modeling method of estimating $\text{var}(\Delta)$ is

based on a large number of daily samples, even though these samples are dependent, the test statistic (18) has an approximately Gaussian distribution. On the other hand, the Chervin-Schneider method of estimating var(Δ) is based on only a relatively small number of GCM runs, so that the associated test statistic has a t-distribution with degrees of freedom equal to the total number of control and experiment runs minus 2. For example, with three control runs (as in Section 4) and with, say, one additional experiment run, the Chervin-Schneider test statistic has a t-distribution with two degrees of freedom. In this case, the critical t-value equals 4.30 for a 5% level of significance as compared with 1.96 for the standard Gaussian distribution. Hence a difference in time averages Δ more than twice as small in magnitude would be statistically significant (i.e., $P < 0.05$) according to the parametric time series modeling approach.

The Chervin-Schneider technique is also based on the requirement that the population variances of the control and experiment time series are equal. The parametric time series modeling approach, as mentioned earlier, does not require this assumption. If more than one control run and more than one experiment run were available, the Chervin-Schneider t-statistic could be modified to avoid the equality of variances requirement (e.g., Laurmann and Gates, 1977).

Madden and Ramanathan (1980) estimate the variance of time averages of atmospheric variables using spectral analysis. A close correspondence exists between parametric time series modeling and spectral analysis. In particular, the expression (10) for the variance of time averages of an AR(p) process is proportional to the spectral density function for such a process evaluated at zero frequency. The spectral analysis approach directly estimates the spectral density function near zero frequency and involves the smoothing of raw spectral density estimates, rather than selecting a specific autoregressive process to model the data. These methods should provide estimates of the variance of time averages that are in reasonably close agreement. Nevertheless, when dealing with many data sets, as in GCM experiments, one procedure might have a computational advantage over the other. We note, in this regard, that the parametric time series modeling approach is completely automatic, whereas the spectral analysis approach may require some subjective assumptions concerning the manner in which the spectral density estimates are smoothed.

6. Concluding remarks

A procedure for making statistical inferences about population means from the output of GCM climate experiments has been presented. This parametric time series modeling approach, involving the fitting of low-order autoregressive processes to the data, yields a potentially more powerful technique for detecting climatic change than the simpler schemes used heretofore. The application of the procedure has been demonstrated through the use of control data generated by the OSU atmospheric GCM. The test application provides estimates of the variance of winter and summer time averages of daily mean surface air temperature, as well as estimates of the magnitude of climatic change that the procedure should be able to detect. A related result of the analysis is that autoregressive processes of higher than first order, rather than Markov processes, are needed to adequately model the majority of the GCM time series considered.

We have concentrated on the analysis of simulated time series from GCM climate experiments. The same methodology, of course, could be applied to real atmospheric measurements. As part of a comprehensive GCM diagnostic study, comparisons could be made with the results of fitting autoregressive processes in a similar manner to corresponding observed atmospheric data. In particular, figures (similar to Figs. 3 and 4) could be constructed for the estimated standard deviation of time averages of real daily mean surface air temperature measurements. We note that such figures have been determined for U.S. temperature data using spectral analysis (Madden and Shea, 1978) and using AR(1) processes (Madden, 1981).

Because a univariate statistical approach has been taken here, data at many different grid points will be examined to detect climatic change, making repeated significance tests a necessity. The question of multiplicity arises when repeated significance tests are required. Even if no climatic change has actually occurred, "climatic change" can be expected to be detected for at least a small proportion of the different significance tests. To avoid this multiplicity problem, an exploratory analysis could be performed for a first GCM climate experiment. This exploratory analysis would lead to the formulation of hypotheses requiring only a limited number of significance tests; for example, hypotheses in terms of averages over latitudinal zones or other regions. Then, as part of a confirmatory analysis, parametric time series models could be fit directly to the spatially averaged data from a second GCM climate experiment. Alternatively, a multivariate statistical approach could be taken, allowing a collection of adjacent grid points to be treated simultaneously. Some references that deal with the application of multivariate time series models to atmospheric data include Jones (1964) and Lemke *et al.* (1980).

The present study has been restricted to the problem of making inferences about the population means of GCM simulated atmospheric variables. An analogous procedure, still based on parametric time series modeling, could be devised to make statistical infer-

ences about variances, such as testing for a change in climatic variability. The technique proposed here also could not be applied to daily total precipitation time series at grid points. Because these time series consist of mixed continuous-discrete variables, a specialized approach is required to convert the data to a form that can be treated within the general statistical framework introduced here.

Acknowledgments. The author thanks W. Lawrence Gates for suggesting this topic, William McKie and Robert Mobley for providing programming assistance, and Robert Chervin and Allan Murphy for supplying comments on this work. An earlier version of this paper, entitled "Statistical Evaluation of Climate Experiments with General Circulation Models: Inferences about Means," appeared as Climatic Research Institute Report No. 15, Oregon State University (July 1980). This research was supported by the National Science Foundation under Grant ATM-8001702.

APPENDIX

Estimation of Autoregression Coefficients

The Yule-Walker recursive method of calculation of the estimated autoregression coefficients (Box and Jenkins, 1976, p. 82) is as follows:

(i) Initial and boundary conditions

$$\hat{\phi}_0(p) = 1, \quad p = 0, 1, \ldots, p_u, \tag{A1}$$

$$\hat{\sigma}^2(0) = c_0 . \tag{A2}$$

(ii) Recurrence relations for estimated autoregression coefficients

$$\hat{\phi}_p(p) = -[1/\hat{\sigma}^2(p-1)][\sum_{k=0}^{p-1} \hat{\phi}_k(p-1)c_{p-k}],$$

$$p = 1, 2, \ldots, p_u, \tag{A3}$$

$$\hat{\phi}_k(p) = \hat{\phi}_k(p-1) + \hat{\phi}_{p-k}(p-1)\hat{\phi}_p(p),$$

$$k = 1, 2, \ldots, p-1, \quad p = 2, 3, \ldots, p_u . \tag{A4}$$

(iii) Recurrence relation for estimated white noise variances

$$\hat{\sigma}^2(p) = \{1 - [\hat{\phi}_p(p)]^2\}\hat{\sigma}^2(p-1),$$

$$p = 1, 2, \ldots, p_u . \tag{A5}$$

REFERENCES

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

Albers, W., 1978: Testing the mean of a normal population under dependence. *Ann. Statist.*, **6,** 1337–1344.

Anderson, T. W., 1971: *The Statistical Analysis of Time Series.* Wiley, 704 pp.

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

Chervin, R. M., 1980: Estimates of first- and second-moment climate statistics in GCM simulated climate ensembles. *J. Atmos. Sci.*, **37,** 1889–1902.

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

Hannan, E. J., 1980: The estimation of the order of an ARMA process. *Ann. Statist.*, **8,** 1071–1081.

Hinkley, D. K., 1977: On quick choice of power transform. *Appl. Statist.*, **26,** 76–79.

Jones, R. H., 1964: Prediction of multivariate time series. *J. Appl. Meteor.*, **3,** 285–289.

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

——, 1976: On estimating the variance of time averages. *J. Appl. Meteor.*, **15,** 514–515.

Katz, R. W., 1981: On some criteria for estimating the order of a Markov chain. *Technometrics*, **23,** 243–249.

——, and R. H. Skaggs, 1981: On the use of autoregressive-moving average processes to model meteorological time series. *Mon. Wea. Rev.*, **109,** 479–484.

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.

Leith, C. E., 1973: The standard error of time-average estimates of climate means. *J. Appl. Meteor.*, **12,** 1066–1069.

Lemke, P., E. W. Trinkl and K. Hasselmann, 1980: Stochastic dynamic analysis of polar sea ice variability. *J. Phys. Oceanogr.*, **10,** 2100–2120.

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

——, 1981: A quantitative approach to long-range prediction. *J. Geophys. Res.*, **86,** 9817–9825.

——, and D. J. Shea, 1978: Estimates of the natural variability of time-averaged temperatures over the United States. *Mon. Wea. Rev.*, **106,** 1695–1703.

——, and V. Ramanathan, 1980: Detecting climate change due to increasing carbon dioxide. *Science*, **209,** 763–768.

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

Straus, D. M., and M. Halem, 1981: A stochastic-dynamical approach to the study of the natural variability of the climate. *Mon. Wea. Rev.*, **109,** 407–421.

Ulrych, T. J., and T. N. Bishop, 1975: Maximum entropy spectral analysis and autoregressive decomposition. *Rev. Geophys. Space Phys.*, **13,** 183–200.