The Effect of Serial Correlation on Statistical Inferences Made with Resampling Procedures

FRANCIS W. ZWIERS

Canadian Climate Center, Downsview, Ontario, Canada

(Manuscript received 20 February 1990, in final form 16 July 1990)

ABSTRACT

Resampling procedures include hypothesis testing methods based on permutation procedures and interval estimation methods based on bootstrap procedures. The former are widely used in the analysis of climate experiments conducted with general circulation models (GCMs) and in the comparison of the simulated and observed climates. The latter are used less frequently than their flexibility and utility warrants. Both resampling techniques are powerful tools, which provide elegant means of overcoming fundamental statistical difficulties encountered in the analysis of observed and simulated climate data. Unfortunately, inferences based on both resampling schemes are as sensitive to the effects of serial correlation as classical statistical methods. These tools must therefore be used with the same amount of caution as other statistical methods when it is suspected that the data might be serially correlated.

1. Introduction

Resampling procedures have been part of the statistical arsenal of climatologists for almost a decade. Such procedures are available for two (or more) sample hypothesis testing problems (the permutation procedure) and for parameter estimation problems (the bootstrap). The permutation procedure, which is often used as the basis of "field" significance tests, was first described by Mielke et al. (1981) and Preisendorfer and Barnett (1983). It has also been discussed by Livezey (1985), Zwiers (1987a), Santor and Wigley (1990) and others. Permutation procedures have found application in the comparison of simulated and observed climates (e.g., Wigley and Santor 1990), the analysis of climate sensitivity experiments (e.g., von Storch and Zwiers 1987; Barnett et al. 1989; and others), the comparison of simulated climates (e.g., Zwiers and Boer 1987), and other problems in which it is necessary to compare the properties of two or more samples of observations. Bootstrap procedures (Efron 1982, 1987; DiCiccio and Tibshirani 1987) have had less application in the climate literature to date. One application can be found in Labitzke and van Loon (1988). Solow (1985) discusses the bootstrap in the context of geophysical problems. He points out that it cannot be applied to serially correlated data as formulated by Efron (1982) and goes on to describe how to circumvent this difficulty in some situations.

Corresponding author address: Francis W. Zwiers, Canadian Climate Center/CCRN, Numerical Modeling Division, 4905 Dufferin Street, Downsview, Ontario, Canada M3H 5T4.

Statistical tests based on permutation procedures differ from classical "parametric" and "non-parametric" testing methods because the analyst is not required to make as many assumptions about the statistical processes which generated the samples, and because the analyst may use statistics that have unknown reference distributions. In classical parametric statistics the functional form of the sampled distributions is fully specified except for the value of a few unknown parameters such as the mean and the variance. The reference distribution (i.e., the distribution of the test statistic when the null hypothesis is true) is derived analytically from these functional representations of the sampled distributions and from various assumptions about the sampling mechanism using basic probability theory and the tools of integral calculus (e.g., Mood and Graybill 1963). Depending upon the complexity of the problem, either an exact or an asymptotic reference distribution is obtained. The former results in statistical tests that have exactly the specified significance level when all assumptions are satisfied. The latter results in tests that achieve the specified significance level asymptotically (i.e., as sample size increases) if all assumptions are satisfied. It is generally necessary to make the following sampling assumptions:

(a) All observations within a sample come from the same distribution;

(b) observations are taken independently of each other; and

(c) samples are taken independently of each other,

and to specify the functional form of the distributions which generated the observations. The sampling assumptions together with the distributional assumption

comprise the full statistical model which is imposed upon the problem by the analyst.

Classical nonparametric tests relax the requirement for explicit functional representations of the sampled distributions by either eliminating the need to make any assumptions about the distributions or by replacing distributional assumptions with less restrictive assumptions about the shape of the distributions. For example, when no assumption is made about the shape of the sampled distribution the Mann–Whitney test (see Conover 1980, p. 216) may be used to test the null hypothesis that the sampled distributions are identical. However, it is necessary to assume that the two distributions are identical in all respects except possibly the mean to test the more general hypothesis that the means are equal with the Mann–Whitney statistic. As with classical parametric statistics, the suite of classical nonparametric tests available to the analyst is limited by the constraint that the reference distribution be obtainable analytically. While the use of nonparametric procedures does reduce the amount of structure imposed upon the problem by the analyst, their use does not permit him/her to relax assumptions (a), (b), and (c) above, which pertain to the sampling mechanism. Indeed, these sampling assumptions play an even more important role in determining the reference distribution in nonparametric statistics than in parametric statistics precisely because there is less information available about the functional form of the sampled distributions.

Tests based on permutation procedures are nonparametric because it is not necessary to have any information about the form of the sampled distributions to use them. On the other hand, these methods are similar to classical parametric methods because the procedure that is used to derive the reference distribution follows the approach which is taken in classical statistics (see section 2). However, the classical constraint that the reference distribution be obtainable via analytical techniques is removed because the necessary integrations are performed with Monte Carlo techniques. Thus, the permutation procedure approach has the advantage that tests can be conducted in situations where it is impossible to derive reference distributions analytically. In particular, it is possible to conduct "field significance" tests in the comparison of two climate simulations where the dimensionality of the observations is always much larger than the number of observations. Permutation procedures also make it possible to use innovative test statistics for which the derivation of a reference distribution is analytically intractable.

Estimation procedures based on the bootstrap may be either parametric or nonparametric depending on whether or not the analyst is able to make a distributional assumption in addition to sampling assumptions (a)–(c). In both cases, procedures differ from classical approaches to interval estimation because information in the sample (together with sampling assumptions (a)–(c) and possibly a distributional assumption) is used to derive an estimate of the sampling distribution of the parameter of interest by means of Monte Carlo simulation. The estimated distribution is then used to construct a confidence interval for the parameter of interest (such as a mean, variance, or correlation coefficient). This differs from classical approaches to interval estimation, which rely either upon an analytic or an asymptotic derivation of the sampling distribution of the parameter of interest. Both the parametric and nonparametric bootstrap procedures have "percentile," "bias corrected" (BC) and "bias corrected with acceleration" (BC_a) (see DiCiccio and Tibshirani 1987) variants which differ in the way in which the confidence interval is computed.

In some problems the bootstrap procedure is conceptual only. That is, it is sometimes possible to construct the bootstrap confidence intervals analytically (just as it is sometimes possible to derive the critical values of a permutation procedure based test analytically). However, in general, the appeal of the technique, as that of the permutation procedure, is that it can be used to construct confidence intervals where it is not possible to specify the underlying distribution or where confidence intervals cannot be obtained via classical analytic techniques.

2. Resampling procedures

a. Hypothesis testing with permutation procedures

As noted above there are several formalisms for performing statistical hypothesis testing in multisample problems. They range from the full parametric model (in which the functional form of the sampled distributions is specified and a set of sampling assumptions are made) to permutation procedures in which no specification of the sampled distribution is required. The hypothesis testing paradigm works the same way in every formalism:

(1) the null hypothesis is assumed to be true;
(2) the distribution of the test statistic is derived using this assumption together with the details of the statistical model which was imposed by the analyst;
(3) the test statistic is computed from the observations; and
(4) an assessment of the rarity of the computed value is made by locating it on the reference distribution.

The null hypothesis is rejected if the computed value occurs relatively infrequently within the assumed statistical model when the null hypothesis is true (typically less than 5% or 1% of the time). The alternate hypothesis is used only to determine the direction of the departure from the null hypothesis which is important when making a decision; that is, to determine whether a test should be "one-sided" or "two-sided" (should

the null hypothesis be rejected when one observes extreme values of the test statistic in a particular direction or simply when one observes extreme values of the test statistic).

The key ingredient in this paradigm is the reference distribution. In the case of a parametric procedure, the reference distribution is obtained analytically from the functional form of the sampled distributions using the tools of integral calculus (e.g., the derivations of the Student's t- and Fisher's *F*-distributions in Mood and Graybill 1963). However, one might also use other integration tools such as numerical quadrature (Carnahan et al. 1969) or Monte Carlo integration (Rubinstein 1981) techniques to perform the necessary integrations and obtain equally correct, if less generally applicable, results. Such an approach can extend the parametric approach by allowing one to derive reference distributions that cannot be easily derived via analytic techniques (Zwiers and Storch 1989, contains such an example). Furthermore, the use of numerical or Monte Carlo integration techniques allows one to remove assumptions about the sampled distributions by replacing functional representations of the sampled distributions with corresponding empirical estimates.

Permutation procedures operate in precisely this manner. The complex integrations needed to obtain the reference distribution can, in simple cases, be performed analytically but are generally performed via a Monte Carlo integration technique. The Monte Carlo integration is conducted as follows. The two (or more) samples, which are to be compared, are pooled to form a single large sample. The pooled sample is resampled to form a "new" pair (or set) of samples, and the test statistic is recomputed from these new samples. The last two steps are repeated a large number of times and the collection of simulated realizations of the test statistic is used to produce an empirical estimate of the reference distribution.

b. *Interval estimation with the bootstrap*

There are also several formalisms with which one can construct interval estimates of parameters such as means, variances, and covariances. Again, these range from classical approaches in which a full parametric model is specified to nonparametric bootstrap procedures which require only sampling assumptions (a)–(c). In all cases the estimation paradigm is similar:

(1) an "efficient" parameter estimate is constructed;
(2) the sampling distribution of the estimator is derived either analytically using exact or approximate methods, or numerically; and
(3) the derived sampling distribution is used to construct a confidence interval for the parameter estimate.

Interval estimates are generally constructed at the 95% or 99% level—the probability that the estimated interval contains the true parameter value when all assumptions implicit in the procedure are satisfied.

In bootstrap procedures the sampling distribution of the parameter estimator is derived via a Monte Carlo integration similar to that used in permutation procedures. Nonparametric bootstrap procedures begin by estimating an empirical distribution function from each sample. Parametric procedures begin by fitting a specified distributional form (such as the Gaussian or Exponential distributions) to the observations in each sample by means of an efficient fitting technique (such as maximum likelihood). Once a distribution has been fitted to each sample, random samples of the same size as the observed samples are generated from the estimated distributions. In the nonparametric case, this is equivalent to selecting observations from each observed sample at random with replacement. The parameter of interest is then recomputed from the simulated samples. This process is repeated a large number of times to create an empirical distribution for the estimated parameter. In this way the information in the original samples is used to estimate the parameter and to estimate the sampling distribution of the estimator.

Once an estimate of the sampling distribution has been constructed it is used to obtain confidences intervals. In the percentile variant of the bootstrap, a $(1 - 2\alpha) \times 100\%$ confidence interval is constructed by using the $\alpha \times 100$th and $(1 - \alpha) \times 100$th percentiles of the empirical distribution function of the estimator as the interval's end points.

The bias correcting variants of the bootstrap (BC and BC_a) improve upon the percentile variant by assuming that some transformation exists which can transform the true sampling distribution of the estimator into the Gaussian distribution (Efron 1982, 1987). Confidence intervals are computed as follows. Let $\hat{\theta}$ be the value of the parameter estimator obtained from the observed samples ($\hat{\theta}$ could be a correlation coefficient or a difference of means). Locate $\hat{\theta}$ on the empirical distribution generated by the resampling and reestimation process. That is, let p_0 be the proportion of bootstrap realizations of $\hat{\theta}$ which are less than the observed $\hat{\theta}$. Let z_0 be the standard Gaussian deviate which corresponds to p_0. That is, z_0 is the solution of

$$p_0 = \Phi(z_0) \quad \text{or} \quad z_0 = \Phi^{-1}(p_0) \tag{1}$$

where Φ represents the standard Gaussian distribution function. Then compute tail probabilities p_l and p_u according to the formulas

$$p_l = \Phi\left\{z_0 + \frac{[z_0 + z^{(\alpha)}]}{1 - a[z_0 + z^{(\alpha)}]}\right\}$$

$$p_u = \Phi\left\{z_0 + \frac{[z_0 + z^{(1-\alpha)}]}{1 - a[z_0 + z^{(1-\alpha)}]}\right\} \tag{2}$$

where $z^{(\alpha)}$ and $z^{(1-\alpha)}$ are standard Gaussian deviates given by

$$z^{(\alpha)} = \Phi^{-1}(\alpha) \quad \text{and} \quad z^{(1-\alpha)} = \Phi^{-1}(1-\alpha), \quad (3)$$

the level of confidence is $(1 - 2\alpha) \times 100\%$, and a is the "acceleration" constant. The details for computing a can be found in Efron (1987) and DiCiccio and Tibshirani (1987). The lower and upper limits of the BC and BC_a confidence intervals are found by extracting the $p_l \times 100$th and $p_u \times 100$th percentiles of the empirical distribution of parameter estimates generated by the resampling process. The percentile variant of the bootstrap confidence interval is obtained by setting $z_0 = 0$ and $a = 0$. The BC variant is obtained by setting $a = 0$ and the BC_a version is obtained by using both z_0 and a.

3. The effect of serial correlation

The resampling process used in permutation and bootstrap procedures is meant to replicate the process that led to the observations which make up the original samples. However, implicit in both resampling schemes is the assumption that the observed climate processes which yielded the observations behave as white noise processes. Thus, resampling procedures may fail to replicate the sampling process, which generated the original samples, because they may not take into account the stochastic characteristics of the observed climate processes. In particular, when observations are serially correlated, inferences will be made relative to incorrectly derived reference or sampling distributions because the resampling process does not replicate the serial correlation structure of the observed climate processes. This is not likely to be much of a problem with data coming from atmospheric GCMs (AGCMs), which are not coupled to ocean and ice models because the interannual correlation of monthly, seasonal, and annual means simulated by such models is small [although there is evidence that AGCMs generate more long time scale variability than would be expected from daily variability (Zwiers 1987b)]. However, the effects of serial correlation may be a problem in the real atmosphere and in coupled models because monthly, seasonal, and annual means are affected by slowly varying boundary conditions.

Three simulation experiments were conducted to illustrate the possible effects of serial correlation on tests and interval estimates based on resampling schemes. The first example deals with a simple difference of means test of the type that might be employed in a comparison of two samples of climate data. The remaining examples deal with the estimation of confidence intervals for the difference of two means and for the lag-0 cross correlation between two variables.

a. A simple hypothesis testing problem

To see that serial correlation effects inferences made with permutation procedures, consider a pair of samples $X(1), \ldots, X(n)$ and $Y(1), \ldots, Y(n)$ obtained from identical autoregressive processes of order one [frequently referred to as AR(1), see Box and Jenkins (1976)] (or red noise processes) with lag-one correlation coefficients β. For convenience, assume that the sample size n is even. Now suppose that one wanted to test the null hypothesis that the samples come from climates with equal means. A natural test statistic is the difference of means statistic $\overline{X} - \overline{Y}$. It is easily shown that the variance this statistic is given by

$$\frac{2\sigma^2}{n}\left[1 + 2\sum_{\tau=1}^{n-1}\left(1 - \frac{\tau}{n}\right)\beta^{\tau}\right] = \frac{2\sigma^2}{n}\left\{\frac{1+\beta}{1-\beta} - \frac{2\beta}{n}\left[\frac{1-\beta^n}{(1-\beta)^2}\right]\right\} \quad (4)$$

where σ^2 is the variance of the AR(1) processes.

Randomly partition the data into two "new" samples and rederive the variance of the difference of means statistic, conditional upon that particular partitioning of the data. One resampling of the data might yield the samples $X(1), X(3), \ldots, X(n-1), Y(1), Y(3), \ldots, Y(n-1)$ and $X(2), X(4), \ldots, X(n), Y(2), Y(4), \ldots, Y(n)$. The variance of the difference of these sample means, say $\overline{X}' - \overline{Y}'$, conditional on this particular partition of the pool, is given by

$$\frac{\sigma^2}{n}\frac{(1-\beta)}{(1+\beta)}\left[2 + \beta\frac{2\beta(1-\beta^n)}{n(1-\beta^2)}\right]. \quad (5)$$

This conditional variance is less than the variance of $\overline{X} - \overline{Y}$ for all $\beta \in (0, 1)$. Moreover, this is true for every possible partitioning of the data. The conclusion is that the permutation procedure will generate a reference distribution, which on average is too narrow, and that the subsequent test for difference of means will have a significance level which is greater than the nominal level.

The following simulation experiment was conducted to illustrate the effect of serial correlation on the difference of means test:

(1) A pair of length n samples were generated from independent AR(1) processes with identical properties. Sample sizes of $n = 20$, 40, and 100 were used to represent sample sizes which are commonly available to climatologists.

(2) A permutation procedure (with 1000 permutations) was used to test the null hypothesis, that the two samples come from stochastic processes with equal means. The test statistic used was simply $\overline{X} - \overline{Y}$. Only one-sided tests were performed, meaning that the null hypothesis was rejected at the $\alpha \times 100$ percent level if the observed $\overline{X} - \overline{Y}$ was less than the $\alpha \times 100$th percentile of the simulated reference distribution. Tests were conducted at the 1%, 2.5%, 5%, and 10% significance levels.

(3) Steps (1) and (2) were replicated 2080 times.

The results obtained from this simple experiment are illustrated in Fig. 1. Each panel represents a different significance level, and each curve within a panel represents a different sample size. When serial correlation is absent (i.e., $\beta = 0$), the permutation procedure acts as it should and produces a test for which the actual significance level is equal to the nominal level to within the effects of sampling variation. However, when serial correlation is present (i.e., $\beta > 0$), its effects can be devastating, at least when $\overline{X}-\overline{Y}$ is used as a test statistic. Figure 1 shows that the likelihood of falsely rejecting the null hypothesis is not greatly affected by sample size, but that it is strongly affected by the presence of serial correlation. When the lag-1 correlation coefficient $\beta = 0.5$, a one-sided test conducted at the nominal 2.5% level will actually operate at a significance level of about 12.5% (i.e., a two-sided test conducted at the nominal 5% level will actually operate at a significance

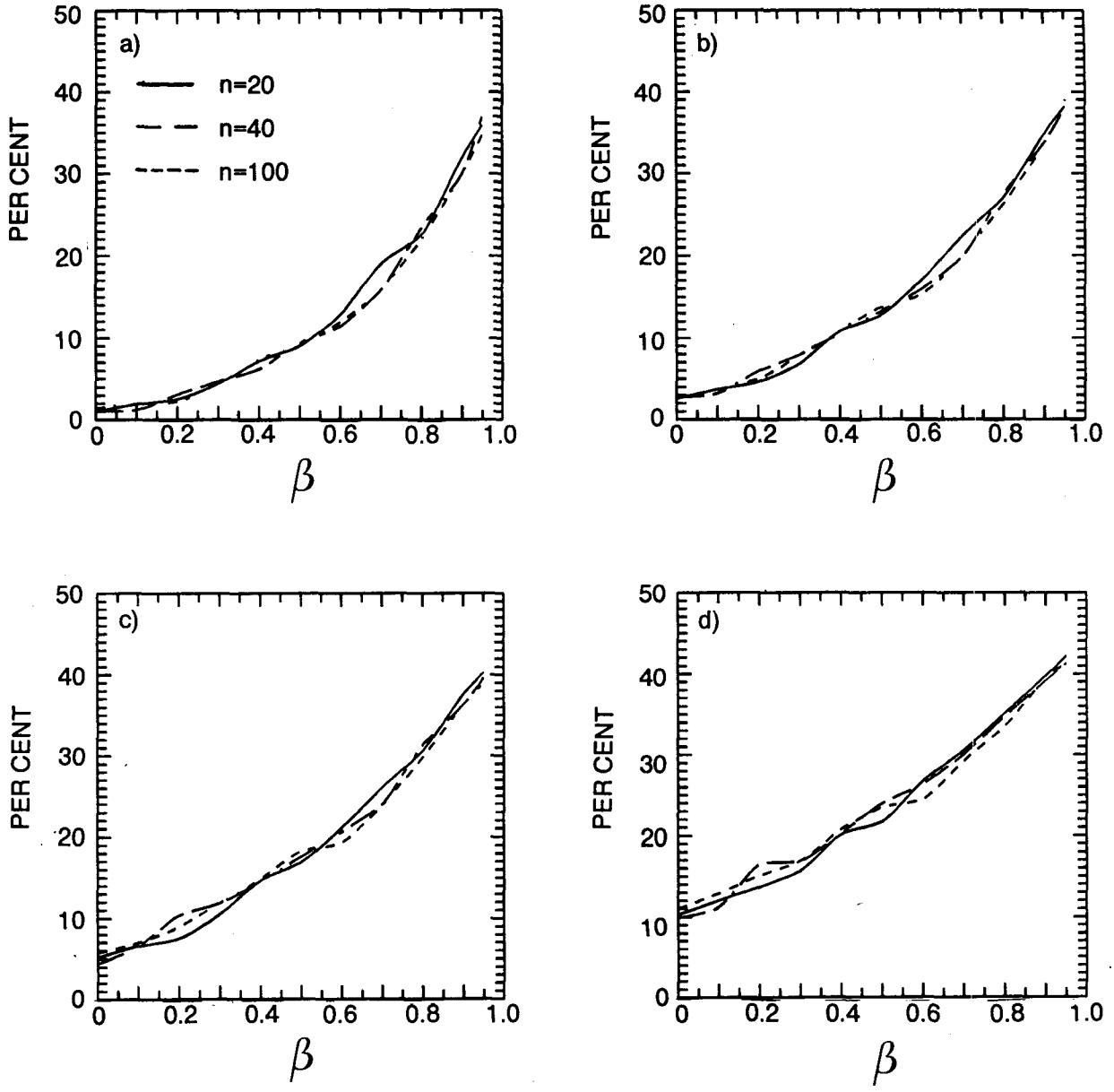

FIG. 1. The actual significance level of a permutation procedure based difference of means test applied to pairs of samples obtained from a simulated AR(1) process. The actual significance level was estimated by replicating the process of generating the samples and conducting the test 2080 times. The curves display the proportion of the 2080 tests which were rejected in each simulation experiment as a function of the serial correlation of the observations. Panels (a)–(d) correspond to tests conducted nominally at the 1%, 2.5%, 5% and 10% significance levels. Within each frame, the solid line, the long-dashed line, and the short-dashed line record the performance of the test with samples of size $n = 20$, 40, and 100, respectively.

level of approximately 25%!). One would expect to obtain similar results for other test statistics.

b. Two simple estimation problems

To illustrate the effects of serial correlation when interval estimates are made with the bootstrap, the problems of estimating a confidence interval for the difference of means, and of estimating a confidence interval for the contemporaneous (lag-0) correlation between two variables were considered.

1) A CONFIDENCE INTERVAL FOR THE DIFFERENCE OF MEANS

Again, consider a pair of samples $X(1), \ldots, X(n)$ and $Y(1), \ldots, Y(n)$ obtained from identical AR(1) processes with lag-1 correlation coefficients β. To gain some insight into the effects of serial correlation on the bootstrap distribution of the difference of means statistic, let us suppose that a nonparametric bootstrapping procedure is employed. Thus, bootstrap samples are obtained from the observed samples via simple random sampling with replacement. Let $X'(1), \ldots, X'(n)$ by a bootstrap sample taken from $X(1), \ldots, X(n)$ and let $Y'(1), \ldots, Y'(n)$ be a bootstrap sample taken from $Y(1), \ldots, Y(n)$. Then $\overline{X}'-\overline{Y}'$ is a bootstrap realization of the difference of means statistic. The variance of $\overline{X}'-\overline{Y}'$, conditional upon the particular observations $X(1), \ldots, X(n)$ and $Y(1), \ldots, Y(n)$, is given by

$$\mathrm{Var}[\overline{X}'-\overline{Y}' \mid X(1),\cdots, X(n), Y(1),\cdots, Y(n)] = (S_X^2 + S_Y^2)/n \quad (6)$$

where

$$S_X^2 = \sum_{i=1}^{n} [X(i) - \overline{X}]^2/n$$

and S_Y^2 is defined similarly. Therefore, the unconditional variance of $\overline{X}'-\overline{Y}'$ is given by

$$\mathrm{Var}(\overline{X}'-\overline{Y}') = E(S_X^2 + S_Y^2)/n \quad (7)$$

where E denotes the expectation operator. Taking expectations, it can be shown that the right-hand side of (7) reduces to

$$[\sigma_X^2 + \sigma_Y^2 - (\sigma_{\bar{X}}^2 + \sigma_{\bar{Y}}^2)]/n \quad (8)$$

which is less than $\mathrm{Var}(\overline{X}-\overline{Y})$ when the observations $X(1), \ldots, X(n)$ and $Y(1), \ldots, Y(n)$ come from identical AR(1) processes with $\beta > 0$. It follows that the nonparametric bootstrap procedure underestimates the width of the distribution of $\overline{X}-\overline{Y}$ in these circumstances just as the permutation procedure underestimates the width of the reference distribution. Consequently, one expects that confidence intervals will be narrower than they should be and that their coverage (i.e., probability that they contain the true difference of means) will be lower than it should be. This expectation is confirmed by the results of the simulation experiment that we conducted.

The experiment was conducted as follows:

(1) A pair of samples of length n (n = 20, 40, and 100) were generated from independent AR(1) processes with identical properties.

(2) Parametric and nonparametric BC bootstrap procedures (with 200 bootstrap samples) were used to estimate 95% confidence intervals for the difference of means.

(3) Steps (1) and (2) were replicated 1000 times and the number of times that the bootstrapped confidence interval contained the true difference of means was recorded.

The results of this experiment for the nonparametric bootstrap are displayed in Fig. 2. Results for the parametric bootstrap (not displayed) are virtually identical. There are only small differences between the coverage of the confidence intervals for the different sample sizes. When the data are not serially correlated the coverage is near the nominal value (approximately 92% as opposed to 95%). However, the coverage drops steadily as the persistence (i.e., β) of the simulated data increases. The bootstrap procedure for estimating con-

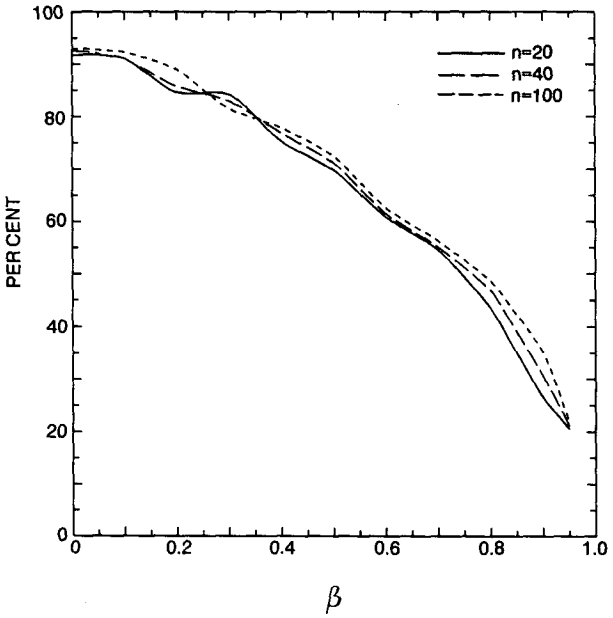

FIG. 2. The coverage of nominal 95% nonparametric BC bootstrap confidence intervals for the difference of means of a pair of samples obtained from identical simulated AR(1) processes. The coverage was estimated by replicating the process of obtaining a pair of samples and constructing the bootstrap confidence interval 1000 times. The curves display the proportion of the 1000 intervals containing the true difference of means in each simulation experiment as a function of the serial correlation of the observations. The solid line, the long-dashed line, and the short-dashed line record the coverage for samples of size n = 20, 40, and 100, respectively.

fidence intervals for the difference of means (or simply for a single mean) is clearly not very tolerant of departures from the assumption that observations are independent of each other (sampling assumption b). The coverage of the nominal 95% confidence interval drops to 90% for $\beta \approx 0.2$ (i.e., the previous observation explains 4% of the variance of the present observation) and falls to about 80% when $\beta \approx 0.35$ (in which case the previous observation explains approximately 12% of the variance of the present observation). The actual coverage of the bootstrap confidence interval is only about 50% when the lag-1 correlation coefficient $\beta = 0.75$.

2) A CONFIDENCE INTERVAL FOR THE CORRELATION COEFFICIENT

Considerations similar to those made previously for the difference of means problem will lead to the conclusion that the bootstrap also underestimates the width of confidence intervals for the correlation between two variables. This is because the products $X(t)Y(t)$, $t = 1, \cdots, n$, will be serially correlated when the individual observations $X(t)$ and $Y(t)$ are serially correlated, albeit in a more complicated manner than the $X(t)$ and $Y(t)$ themselves.

A simulation experiment was conducted to illustrate the effects of serial correlation on bootstrap interval estimates of the lag-0 cross-correlation coefficient. The experiment was conducted in a manner similar to that previously with the exception that pairs of observations $[X(t), Y(t)]$, $t = 1, \cdots, n$, were generated from a bivariate AR(1) process constructed so that the lag-0 cross correlation between the time series is 0.75, and so that the lag-1 serial correlation of the individual time series has values β between 0 and 0.95. This was done by generating samples from the bivariate AR(1) process

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \beta \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \vec{\epsilon}_t \tag{9}$$

where $\vec{\epsilon}_t$ is bivariate Gaussian white noise with variance/covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 0.75 \\ 0.75 & 1 \end{bmatrix}. \tag{10}$$

The results of this experiment are displayed in Fig. 3. Each panel in the figure represents a different sample size. Within each panel, the solid line represents the coverage of the nonparametric BC confidence interval and the dashed line represents the coverage of the corresponding parametric interval. Note that in this case the parametric bootstrap improves upon the nonparametric version. When the observations are not serially correlated the coverage of the parametric BC confidence intervals is very close to the nominal 95% level. The coverage of the nonparametric intervals is roughly 2% less than the nominal level. The parametric intervals also give slightly better coverage than the nonparametric intervals when the observations are serially correlated. The difference between the coverage of the parametric, and nonparametric intervals decreases with increasing sample size (as it should). The second thing that is apparent from these graphs is that the bootstrap confidence intervals for lag-0 cross correlation are more tolerant of serial correlation than those for the difference of means. Coverage of the nominal 95% interval falls to approximately 90% for values of β between 0.35 and 0.4. Coverage drops to approximately 80% for values of β between 0.6 and 0.7. The loss of coverage appears to increase with sample size for $\beta > 0.6$.

4. Conclusions and recommendations

The use of permutation and bootstrap procedures has allowed climatologists to conduct statistical tests, and to estimate uncertainty in many situations where it is not possible to derive parametric or nonparametric tests analytically. They have allowed analysts to relax critical assumptions about the nature of sampled distributions in the hypothesis testing and estimation paradigms, and have made the process more robust in the sense that the correct reference distribution is always used, at least asymptotically, provided the sampling mechanism used in the resampling procedure matches that used to obtain the data.

Permutation and bootstrap procedures, as they are currently formulated, implicitly assume that observations within samples represent realizations of independent and identically distributed observations. It is clear from the examples described above that neither procedure is particularly tolerant of departures from this sort of sampling assumption. Tests and estimates of uncertainty based on resampling schemes are apparently as sensitive to the effects of serial correlation as many other statistical procedures used in climatology. The latter have been discussed in detail by Livezey and Chen (1983), Thiébaux and Zwiers (1984), Trenberth (1984a,b), Zwiers and Thiébaux (1987), and others.

Resampling schemes should be used with caution when there is the possibility that data are serially correlated. Resampling procedures can be modified to take the presence of relatively simple stochastic structure into account in univariate inference problems. Solow (1985) describes one approach that adapts the nonparametric bootstrap to situations in which simple stochastic models can be used to describe the serial correlation structure of the data. Permutation procedures could be adapted in a similar manner. The parametric bootstrap can be adapted to these situations by including the serial correlation structure as part of the parametric model used by the bootstrap. On the other hand, it seems unlikely that it will be possible to make such adaptations for multivariate statistics given the complexities that are possible in multivariate lagged correlation behavior.

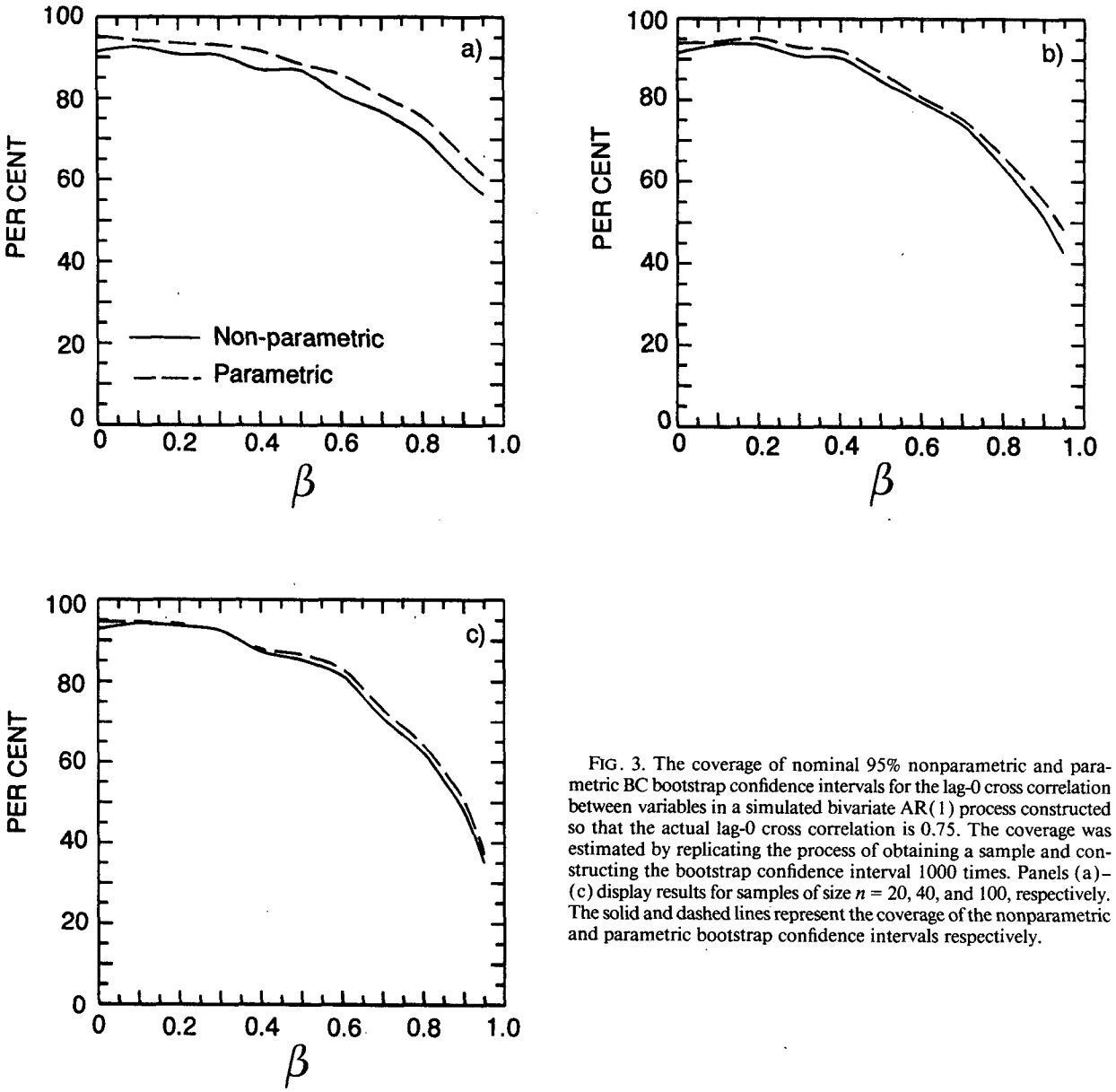

FIG. 3. The coverage of nominal 95% nonparametric and parametric BC bootstrap confidence intervals for the lag-0 cross correlation between variables in a simulated bivariate AR(1) process constructed so that the actual lag-0 cross correlation is 0.75. The coverage was estimated by replicating the process of obtaining a sample and constructing the bootstrap confidence interval 1000 times. Panels (a)–(c) display results for samples of size n = 20, 40, and 100, respectively. The solid and dashed lines represent the coverage of the nonparametric and parametric bootstrap confidence intervals respectively.

The applications of the permutation procedure that were previously cited should be relatively free from the effects of serial correlation because serial correlations are relatively small in these problems. On the other hand, the application of the bootstrap by Labitzke and van Loon (1988, hereafter LvL) may be quite strongly affected. Labitzke and van Loon use the bootstrap to determine the uncertainty of estimated correlations between 10.7-cm solar flux and 30-mb North Pole temperature observations after stratifying the data according to the phase of the QBO. Their Fig. 1b (for the west phase of the QBO) shows a pair of somewhat irregularly sampled time series which exhibit strong quasi-periodic behavior. The estimated correlation between the series is 0.76.

Simple low-order AR processes can display quasi-periodic behavior similar to that displayed in LvL's Figs. 1b,c. For example, Fig. 4 displays four realizations of a 32-yr time series generated from the AR(2) model

$$X_t = 1.60X_{t-1} - 0.91X_{t-2} + \epsilon_t. \qquad (11)$$

The model's coefficients were chosen so that the simulated process has a spectral peak at the 11-yr period.

Simulation experiments similar to those described in section 3b(2) were conducted to determine how well the bootstrap estimates confidence intervals for the correlation coefficient when observations are taken from correlated AR(2) processes such as (11). Observations were generated from the AR(2) equivalent of (9),

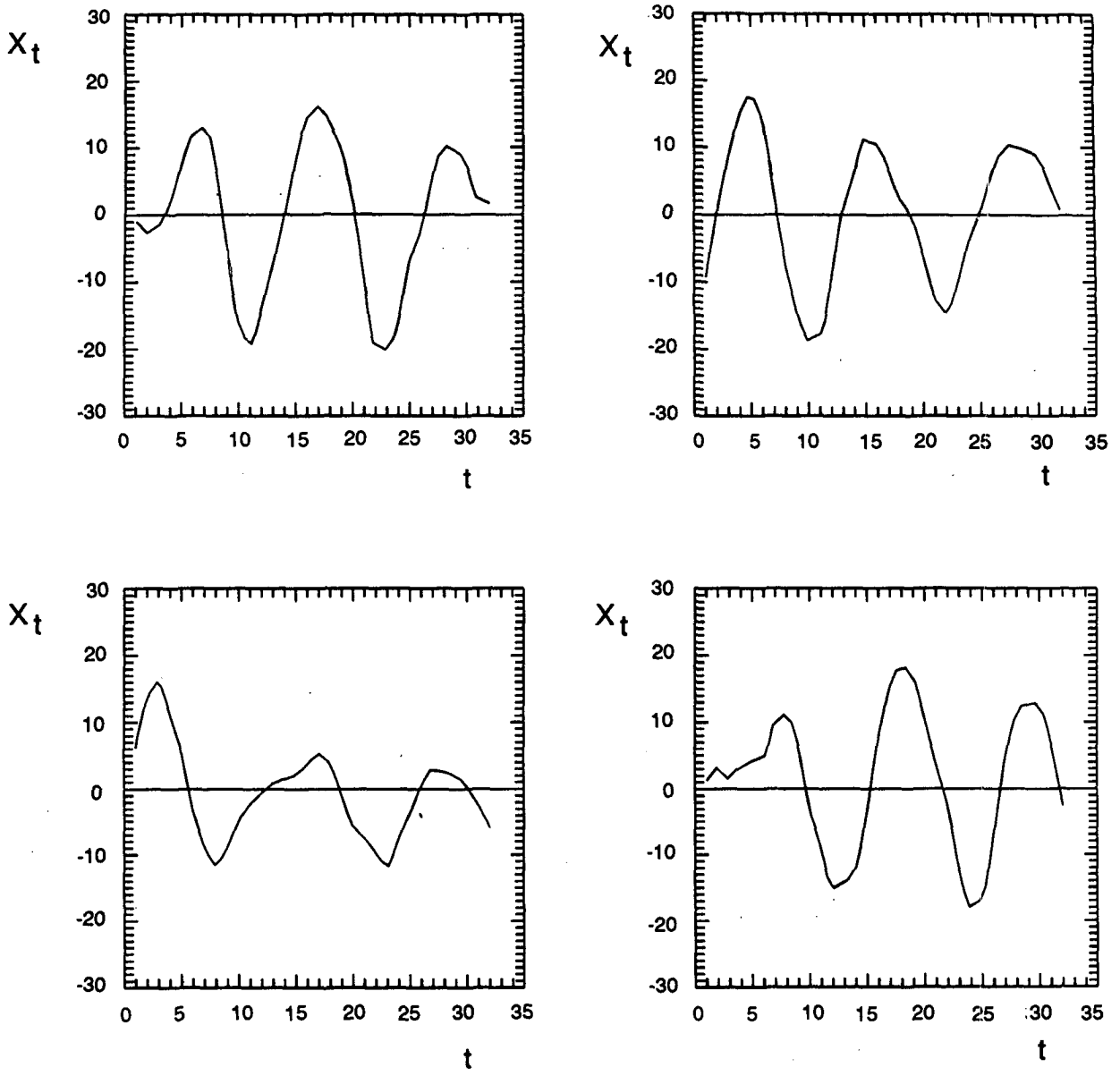

FIG. 4. Four samples of length 32 generated from AR(2) model (11).

$$\begin{bmatrix} X_t \\ Y_t \end{bmatrix} = \beta_1 \begin{bmatrix} X_{t-1} \\ Y_{t-1} \end{bmatrix} + \beta_2 \begin{bmatrix} X_{t-2} \\ Y_{t-2} \end{bmatrix} + \vec{\epsilon}_t \qquad (12)$$

where $\vec{\epsilon}_t$ is defined as in section 3b(2). Realizations of length 40 were generated from this model but observations were taken only every second time step to approximate the effect of the stratification carried out by LvL. Thus, the resulting samples contained 20 observations.

Coverage of the bootstrap confidence intervals was estimated for three bivariate AR(2) models like (12), all constructed so that the individual time series X_t and Y_t have spectral peaks at the 11-yr period. Model A ($\beta_1 = 1.40$ and $\beta_2 = -0.71$) does not produce time series realizations which are obviously periodic. In this case, coverage of the nominal 95% confidence interval for the lag-0 cross-correlation coefficient is close to 85%. Model B ($\beta_1 = 1.60$ and $\beta_2 = -0.91$) produces realizations like those in Fig. 4, which appear to be quasi-periodic. When samples were taken from this model coverage of the nominal 95% confidence interval dropped to about 60% and the average end points of the parametric BC bootstrap confidence interval were 0.44 and 0.87. Corresponding estimates of the true end points (obtained from 10 000 realizations of the sample correlation coefficient) are 0.10 and 0.96, respectively. Model C ($\beta_1 = 1.65$ and $\beta_2 = -0.96$) produces real-

izations that are often more regular than those shown in Fig. 4 (or LvL's Figs. 1b,c). In this last case coverage was only about 50%.

The purpose of these few examples is not to argue that the irregularly sampled solar flux and 30-mb North Pole temperature time series behave like nearly nonstationary AR(2) time series. It is virtually impossible to identify an appropriate stochastic model for these time series on the basis of the short realizations which are available because of their apparent quasi-periodic nature. Rather, the purpose is to suggest that the uncertainty of the correlation reported by LvL may be much greater than they have indicated. If these times series really do behave like model B then the 95% confidence interval for the west phase correlation should be considerably wider than the 0.54–0.91 interval which was reported. Also, the null hypothesis that the west phase correlation is zero may not be inconsistent with the observations. A Monte Carlo simulation was used to determine that the critical values for a 5% (2%) test are approximately ±0.74 (±0.81) when samples of length 20 are taken every second time step from uncorrelated time series of the form of (11). The correlation observed by LvL was 0.76.

Acknowledgments. The comments of the editor (R. E. Livezey) and Hans von Storch led to a clearer and more complete paper. I thank them both.

REFERENCES

Barnett, T. P., L. Dumenil, U. Schlese, E. Roeckner and M. Latif, 1989: The effects of Eurasian snow cover on regional and global climate variations. *J. Atmos. Sci.*, **46,** 661–685.

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

Carnahan, B., H. A. Luther and J. O. Wilks, 1969: *Applied Numerical Methods.* Wiley, 604 pp.

Conover, W. J., 1980: *Practical Nonparametric Statistics.* Second Edition. Wiley, 493 pp.

DiCiccio, T., and R. Tibshirani, 1987: Bootstrap confidence intervals and bootstrap approximations. *J. Amer. Statist. Assoc.*, **82,** 163–170.

Efron, B., 1982: *The Jackknife, the Bootstrap, and other Resampling Plans.* J. W. Arrowsmith, 92 pp.

——, 1987: Better bootstrap confidence intervals. *J. Amer. Statist. Assoc.*, **82,** 171–185.

Labitzke, K., and H. van Loon, 1988: Associations between the 11-year solar cycle, the QBO, and the atmosphere. Part I: The troposphere and stratosphere in the Northern Hemisphere in winter. *J. Atmos. Terr. Phys.*, **50,** 197–206.

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

——, and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.*, **111,** 46–59.

Mielke, P. W., K. J. Berry and G. W. Brier, 1981: Application of the multi-response permutation procedure for examining seasonal changes in monthly mean sea level pressure patterns. *Mon. Wea. Rev.*, **109,** 120–126.

Mood, A. M., and F. A. Graybill, 1963: *Introduction to the Theory of Statistics.* Second Edition. McGraw-Hill, 443 pp.

Preisendorfer, R. W., and T. P. Barnett, 1983: Numerical model-reality intercomparison tests using small-sample statistics. *J. Atmos. Sci.*, **40,** 1884–1896.

Rubinstein, R. Y., 1981: *Simulation and the Monte Carlo Method.* Wiley, 278 pp.

Santor, B. D., and T. M. L. Wigley, 1990: Regional validation of means, variances and spatial patterns in general circulation model control runs. *J. Geophys. Res.*, **95**(D), 829–850.

Solow, A. R., 1985: Bootstrapping correlated data. *Math. Geol.*, **17,** 769–775.

Thiébaux, H. J., and F. W. Zwiers, 1984: The interpretation and estimation of effective sample size. *J. Climate Appl. Meteor.*, **23,** 800–811.

Trenberth, K. E., 1984a: Some effects of finite sample size and persistence on meteorological statistics. Part I: Autocorrelations. *Mon. Wea. Rev.*, **112,** 2359–2368.

——, 1984b: Some effects of finite sample size and persistence on meteorological statistics. Part II: Potential predictability. *Mon. Wea. Rev.*, **112,** 2369–2379.

Wigley, T. M. L., and B. D. Santor, 1990: Statistical comparison of spatial fields in model validation, perturbation, and predictability experiments. *J. Geophys. Res.*, **95**(D), 851–866.

von Storch, H., and F. W. Zwiers, 1988: Recurrence analysis of climate sensitivity experiments. *J. Climate,* **1,** 157–171.

Zwiers, F. W., 1987a: Statistical considerations for climate experiments. Part II: Multivariate tests. *J. Climate Appl. Meteor.*, **26,** 477–487.

——, 1987b: A potential predictability study conducted with an atmospheric General Circulation Model. *Mon. Wea. Rev.*, **115,** 2957–2974.

——, and G. J. Boer, 1987: A comparison of climates simulated by a General Circulation Model when run in the Annual Cycle and Perpetual Modes. *Mon. Wea. Rev.*, **115,** 2626–2644.

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

——, and H. von Storch, 1989: Multiple recurrence analysis. *J. Climate,* **2,** 1538–1553.