Resampling Hypothesis Tests for Autocorrelated Fields

D. S. WILKS

Department of Soil, Crop and Atmospheric Sciences, Cornell University, Ithaca, New York

(Manuscript received 5 March 1996, in final form 10 May 1996)

ABSTRACT

Presently employed hypothesis tests for multivariate geophysical data (e.g., climatic fields) require the assumption that either the data are serially uncorrelated, or spatially uncorrelated, or both. Good methods have been developed to deal with temporal correlation, but generalization of these methods to multivariate problems involving spatial correlation has been problematic, particularly when (as is often the case) sample sizes are small relative to the dimension of the data vectors. Spatial correlation has been handled successfully by resampling methods when the temporal correlation can be neglected, at least according to the null hypothesis. This paper describes the construction of resampling tests for differences of means that account simultaneously for temporal and spatial correlation. First, univariate tests are derived that respect temporal correlation in the data, using the relatively new concept of ‘‘moving blocks’’ bootstrap resampling. These tests perform accurately for small samples and are nearly as powerful as existing alternatives. Simultaneous application of these univariate resampling tests to elements of data vectors (or fields) yields a powerful (i.e., sensitive) multivariate test in which the cross correlation between elements of the data vectors is successfully captured by the resampling, rather than through explicit modeling and estimation.

1. Introduction

It is often of interest to compare two datasets for differences with respect to particular attributes. For example, one might want to investigate whether the mean value under one set of conditions is different than the mean in some other circumstance. In order to make fair and reliable comparisons, it is necessary to account for sampling variability in this process. That is, even if there is no difference with respect to an attribute of interest (e.g., the mean) in the generating processes for two batches of data, the two sample estimates of the quantity (the pair of sample means) will rarely be exactly equal. Standard statistical tests (e.g., the familiar *t* test) have been devised to discern what magnitude of difference, in relation to the variability evident in the data samples, can be declared with high confidence to be real (i.e., statistically significant). Statistical tests have most commonly been applied to atmospheric datasets in the context of climate studies (e.g., Livezey 1985), although their range of potential applicability is much broader (e.g., Daley and Chervin 1985).

Two complications arise when attempting to apply standard hypothesis tests to climatic or other geophysical data. These are illustrated schematically in the middle portion of Fig. 1. First, standard tests, including the *t* test, rest on the assumption that the underlying data are composed of independent samples from their parent populations. Very often, atmospheric and other geophysical data do not satisfy this assumption even approximately. In particular, such data typically exhibit serial (or auto-) correlation. The effect of this autocorrelation on the sampling distribution of the mean is to increase its variance above the level that would be inferred under the assumption of independence. That is, sample means are less consistent from batch to batch than would be the case for independent data.

The estimated variance of the sampling distribution of the mean appears in the denominator of the test statistic for the *t* test:

$$t = \frac{\bar{x} - \mu_0}{[\mathrm{var}(\bar{x})]^{1/2}}. \tag{1}$$

Underestimation of the variance of the sample mean thus inflates the standard *t* statistic, leading to unwarranted rejections of the null hypothesis (e.g., Cressie 1980; Wilks 1995). However, a number of univariate (i.e., scalar) tests have been developed that perform properly, even when the data upon which they operate exhibit serial correlation (Albers 1978; Chervin and Schneider 1976; Jones 1975; Katz 1982; Zwiers and Thiébaux 1987; Zwiers and von Storch 1995). One approach, which is successful when the sample size *n* is sufficiently large, is to estimate the variance of the sampling distribution of the mean as

Corresponding author address: Dr. Daniel S. Wilks, Department of Soil, Crop and Atmospheric Sciences, Cornell University, 1123 Bradfield Hall, Ithaca, NY 14853-1901.
E-mail: dsw5@cornell.edu

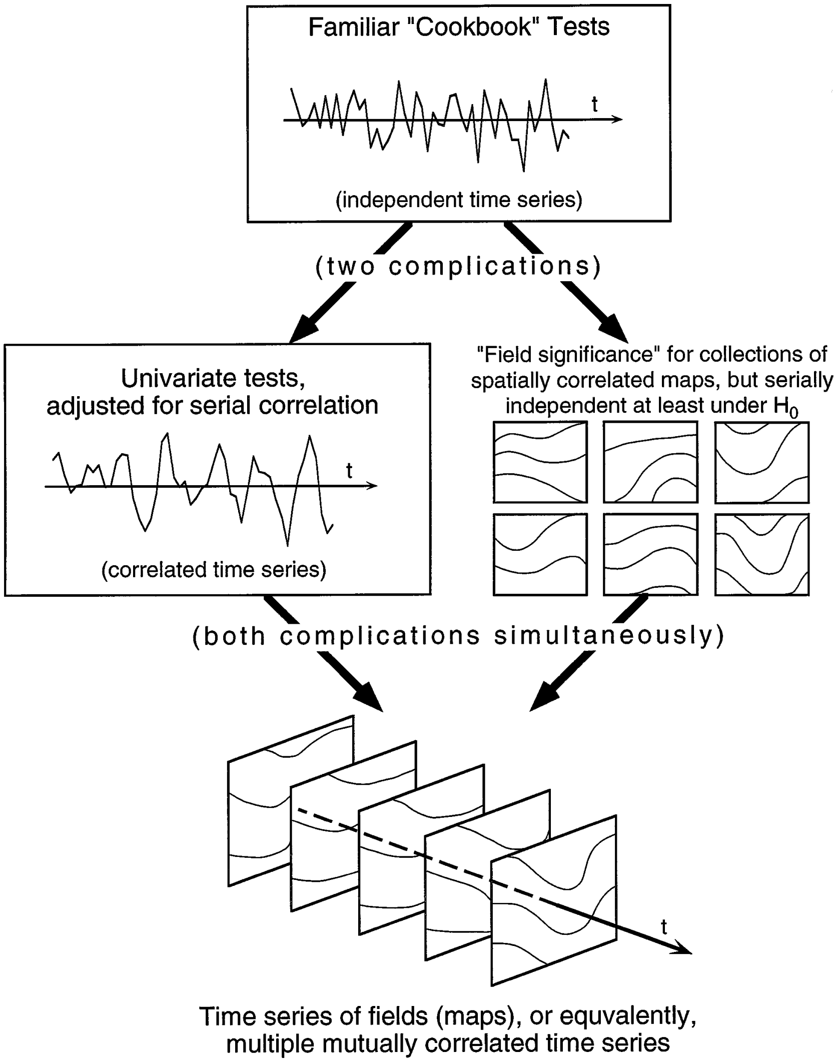
Familiar "Cookbook" Tests
t
(independent time series)
(two complications)
Univariate tests, adjusted for serial correlation
t
(correlated time series)
"Field significance" for collections of spatially correlated maps, but serially independent at least under H0
(both complications simultaneously)
t
Time series of fields (maps), or equvalently, multiple mutually correlated time series

FIG. 1. Conceptual illustration of therelationships between conventional univariate hypothesis tests (top), univariate tests modified to account for temporal correlation in the data (middle left), multivariate tests assuming serial independence (middle right), and tests respecting both types of correlations (bottom) that are the subject of this paper.

$$\mathrm{var}(\overline{x}) = V\frac{s_x^2}{n}, \tag{2}$$

where s_x^2 is the sample variance of the data and V is the "variance inflation factor," which depends on the autocorrelation in the data according to

$$V = 1 + 2\sum_{k=1}^{n-1}\left(1 - \frac{k}{n}\right)r_k. \tag{3}$$

Here, the r_k are estimates of the autocorrelations at lags k. For independent data, $r_k = 0$ for $k \neq 0$, in which case $V = 1$ and the conventional t test is recovered in (1). The variance inflation factor is sometimes called the "time between effectively independent samples," (Leith 1973; Trenberth 1984), or the "decorrelation time," although this terminology is only really applicable to (3) for problems involving estimation of the mean (Livezey 1995; Thiébaux and Zwiers 1984; von Storch 1995; Zwiers and von Storch 1995). Similarly, (3) is sometimes used to define the "effective sample size"

$$n_e = \frac{n}{V}, \tag{4}$$

which also pertains only to inferences concerning means.

The second complication indicated in Fig. 1 is that the data of interest may be vector valued, or multivariate. If the elements of the data vectors are mutually independent, it is straightforward to evaluate jointly the results of a collection of scalar hypothesis tests (the "field significance") using the binomial distribution (e.g., Livezey and Chen 1983; von Storch 1982). However, for problems of practical interest, it is often the case that the data exhibit substantial cross correlation. For example, it might be of interest to test whether the means of two sets of atmospheric fields (these could be values at a spatial array of points, or expansion coefficients for the data projected onto a different basis, such as selected Fourier harmonics) are significantly different between two time series of these fields. If the data are cross correlated but temporally independent, the classical Hotelling test [the multivariate analog of the simple *t* test, e.g., Johnson and Wichern (1982)] or related tests (e.g., Hasselmann 1979) can be used, provided the number of samples is much larger than the dimension of the vector data.

Alternatively, resampling tests (Chung and Fraser 1958; Livezey and Chen 1983; Mielke et al. 1981; Preisendorfer and Barnett 1983; Wilks 1996; Zwiers 1987) can be employed. These tests involve, in one form or another, developing sampling distributions for a test statistic (e.g., the length of the difference between vector means, according to a particular norm) by repeated random reordering the data vectors (fields). This procedure is attractive because the resampling procedure captures the cross correlation between elements of the data vectors without requiring that the covariance structure be explicitly modeled and estimated. However, conventional resampling procedures, which repeatedly rearrange the time ordering of individual data points, necessarily destroy important information regarding the temporal correlation that may be present in the data (e.g., Zwiers 1990). Such tests are thus applicable only to situations where either the temporal correlation is negligible, or temporal independence is implied by the null hypothesis.

Multivariate geophysical data exhibiting both time and space correlations, indicated in the lower portion of Fig. 1, are often of interest. In principle, multivariate parametric tests that account for serial correlation could be constructed by extending univariate tests that include serial correlation (Kabaila and Nelson 1985), or through likelihood methods (Hillmer and Tiao 1979). However, either of these approaches would require estimation of a prohibitively large number of parameters for many problems. Although some suggestions for dealing simultaneously with auto- and cross correlation in resampling tests applied to multivariate data have been offered by Livezey (1995), an adequate resampling test respecting both the temporal and spatial correlation typically found in atmospheric data has not previously been devised. This paper develops such a test for the difference of vector means, using a relatively new idea called the ''moving blocks'' bootstrap, which is a resampling procedure that can capture time dependence in autocorrelated data. The simpler univariate (i.e., for scalar series) bootstrap tests for differences of the mean are developed in section 2. In section 3 these tests are extended to multivariate data (i.e., vector series), where the resampling procedure captures the effects of cross correlation on the test statistic without the need to explicitly model those correlations. Section 4 summarizes and concludes with suggestions for future directions.

2. Univariate tests

a. Bootstrap tests and the moving blocks bootstrap

The bootstrap is a relatively recent computer-based statistical technique (Efron 1982; Efron and Tibshirani 1993). The philosophical basis of the bootstrap is the idea that the sampling characteristics (i.e., batch-to-batch variations exhibited by different samples from a parent population) of a statistic of interest can be simulated by repeatedly treating a single available batch of data in a way that mimics the process of sampling from the parent population. As a nonparametric approach, it is often useful when the validity of assumptions underlying more traditional theoretical approaches is questionable and when the more traditional approaches are either unavailable or intractable.

Bootstrapping is used most frequently to estimate sampling distributions, or particular aspects of sampling distributions, such as standard errors and confidence intervals. Although computationally intensive, the procedure is algorithmically simple. The bootstrap approximation to the sampling distribution of a statistic of interest is constructed by repeatedly resampling the available data *with replacement* to yield multiple synthetic samples of the same size as the original set of observations. For example, consider an arbitrary statistic $S(\mathbf{x})$, which is some function of a sample of n data values $\mathbf{x} = \{x_1, x_2, x_3, \cdots, x_n\}$. A bootstrap sample from $\mathbf{x}$, say $\mathbf{x}^*$, is constructed by treating $\mathbf{x}$ as nearly as possible as if it were the full parent population. Each bootstrap sample consists of n values drawn independently and with replacement from $\mathbf{x}$ and in general will contain multiple copies of some of the original data values while omitting others. Some large number n_B of bootstrap samples is drawn, and the statistic of interest, $S(\mathbf{x}^*)$, is computed for each. Remarkably, it has been found that the distribution of the resulting n_B values of $S(\mathbf{x}^*)$ is often a reasonable representation of the sampling distribution of $S(\mathbf{x})$. Consequently, one can, for example, estimate a confidence interval for $S(\mathbf{x})$ on the basis of the dispersion of the distribution of $S(\mathbf{x}^*)$.

There is a close relationship between confidence intervals and hypothesis tests. In particular, bootstrap confidence intervals can be used as the basis for nonparametric hypothesis tests. If a bootstrap resampling procedure is designed in a way that is consistent with a specified null hypothesis regarding a statistic of interest, the central $(1 - \alpha)100\%$ confidence interval for that statistic, as estimated through its bootstrap distribution, comprises the acceptance region for the corresponding two-sided hypothesis test. When the observed statistic lies outside of this interval, the null hypothesis can be rejected at the α100% level. A corresponding one-sided test would be significant at the $(\alpha/2)$100% level.

An alternative approach to the construction of nonparametric resampling hypothesis tests is through permutation procedures (e.g., Mielke et al. 1981; Preisendorfer and Barnett 1983). A two-sample permutation test operates by assuming that, under the null hypothesis, the two batches of data at hand have been drawn from the same parent population. Accordingly, their labelling as belonging to one batch or the other is arbitrary, and all the data can be pooled into a single body from which pairs of synthetic samples can be drawn repeatedly, without replacement. This process leads, in a manner analogous to the bootstrap procedure, to the construction of a synthetic sampling distribution for the statistic of interest. A limitation of permutation tests is the implication that the null hypothesis specifies that the underlying distributions for the two data samples are the same in all respects. By contrast, bootstrap tests can investigate much more focused null hypotheses. In the bootstrap tests described below, the null hypothesis will be that pairs of means (only) are equal, without assuming equality of variances, autocorrelations, or other aspects of the joint distributions of the data.

In their basic forms, neither bootstrap nor permutation procedures are appropriate tools for resampling time series or other autocorrelated data because independent resampling destroys the correlation structure. For positively correlated time series, a hypothesis test based on the conventional bootstrap would be a nonparametric analog of the t test (1), but with independent bootstrap resampling effectively yielding $V = 1$ in (2). As is the case for the t test, the resulting bootstrap test would be permissive: the null hypothesis would be rejected more easily and frequently than warranted. One approach to modifying the bootstrap to account for correlated data is to use a fitted parametric model (e.g., an autoregressive scheme for time series data) to produce approximately uncorrelated residuals, bootstrap these ''prewhitened'' values, and then reconstruct simulated data series by inverting the parametric model (Efron and Tibshirani 1993; Solow 1985). However, extending this approach to multivariate settings would be problematic.

A nonparametric alternative for bootstrapping correlated data is the recently proposed moving blocks bootstrap (Künsch 1989; Liu and Singh 1992). This technique preserves much of the correlation structure in bootstrapped time series by resampling ''blocks,'' or sets of fixed length of consecutive data values, rather than single data points. Figure 2 illustrates the construction of a single bootstrap sample from a data record of length $n = 12$, using blocks of length $l = 4$. The basic set of objects operated upon by the moving blocks bootstrap consists of the $n - l + 1$ distinct blocks of length l, rather than the n individual data values in $\mathbf{x}$. Defining b as the number of blocks comprising each bootstrap sample ($b = 3$ in Fig. 2), the number of distinct bootstrap samples that can be drawn from a given time series of length n, assuming all the possible blocks are distinct, is $(n - l + 1)^b$. For $l = 1$ the moving blocks bootstrap reduces to the conventional bootstrap.

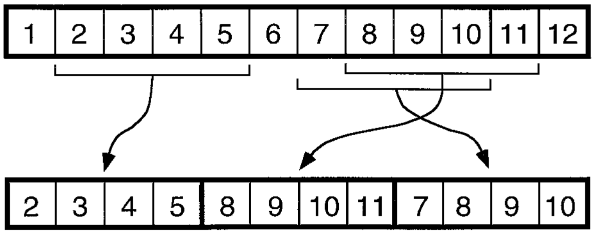

FIG. 2. Schematic illustration of the moving blocks bootstrap. Beginning with a time series of length $n = 12$ (above), $b = 3$ blocks of length $l = 4$ are drawn with replacement. The resulting time series (below) is one of $(n - l + 1)^b = 729$ equally likely bootstrap samples.

An outstanding problem for implementation of the moving blocks bootstrap is the selection of the block length l. Limited theoretical work on block length selection done to date has indicated that as $n \rightarrow \infty$, the block length l should also tend to infinity, but slowly enough that $l/n \rightarrow 0$ (Künsch 1989; Liu and Singh 1992). In addition, it is expected that, if other factors are equal, time series exhibiting stronger autocorrelation will require longer block lengths to capture adequately the effects of the dependence (Carlstein 1986). However, as a practical matter, choices for l in applied work have been ad hoc and qualitative (Leger et al. 1992). Prescriptions offered below for the choice of l have been developed specifically for the problem of testing sample means of time series.

b. Bootstrap test for AR(1) data

One of the simplest possible time series models is the first-order autoregressive [AR(1)] process (e.g., Box and Jenkins 1976), defined by

$$x'_t = \phi x'_{t-1} + \epsilon_t \qquad (5)$$

Here the primes indicate ''anomalies'' constructed by subtraction of the mean (i.e., $x'_t = x_t - \mu$, with $\mu = E[x]$), ϕ is the autoregressive parameter (which is equal to the population value of the lag-1 autocorrelation), and the ϵ_t series consists of independent random variates with expectation $E[\epsilon] = 0$ and standard deviation σ_ϵ. Often it is assumed that the ϵ_t values follow a Gaussian distribution. Zwiers and von Storch (1995) have suggested that this simple model is a reasonable approximation to the behavior of many atmospheric variables not exhibiting quasi-periodic behavior, in that its spectrum is maximum at zero frequency and decreases monotonically for higher frequencies. Accordingly, they devised a test (hereafter the ZvS test), based on simulations using (5), for inferences concerning means of time series. For AR(1) data the ZvS test yields accurate results for smaller samples than does the test based on (1)–(3) or its two-sample counterpart.

The tests described here will be based on the modified one-sample *t*-test statistic, (1)–(3), or the corresponding two-sample test described below. Both of these require estimation of the variance inflation factor V (3). Direct substitution of sample estimates of the $n - 1$ sample autocorrelations into (3) leads to erratic results for V (Thiébaux and Zwiers 1984). More stable estimates for V can be obtained by fitting a time series model to the data series $\mathbf{x}$ and using the properties of the time series model to extrapolate to lagged correlations beyond the first few (Katz 1982; Thiébaux and Zwiers 1984). When assuming an AR(1) model for the data, only the lag-1 autocorrelation needs to be directly estimated from the data:

$$r_1 = \frac{\sum_{t=1}^{n-1} (x_t - \overline{x})(x_{t+1} - \overline{x})}{\sum_{t=1}^{n} (x_t - \overline{x})^2}. \tag{6}$$

Then, the properties of the AR(1) process are exploited to compute

$$r_k = r_1^k, \tag{7}$$

which are used for the remaining $n - 2$ correlation estimates in (3).

Although they are more stable (i.e., lower variance), estimates of V thus obtained are still substantially biased for samples that are not large. The nature of this bias is that the estimates are on average too low, and the tendency for the test in (1)–(3) to be permissive for small samples (e.g., Thiébaux and Zwiers 1984; Zwiers and von Storch 1995) may be attributable in part to this bias. For the variance estimate in (2) to be compatible with the corresponding bootstrap estimates developed below, it is necessary to correct the bias in the estimation of V. Linear regression equations specifying the logarithm of the ratio of the (known) variance inflation factor to the (biased) V (3), for synthetic series from a variety of AR(1) models and sample sizes, lead to the adjusted variance inflation factor

$$V' = V \exp\left(\frac{2V}{n}\right). \tag{8}$$

These V' approach V for large n and are very nearly unbiased for all but very small sample sizes in combination with very strong correlations (i.e., $n_e < 2$), for which none of the available tests operate correctly.

The methods and results here will be presented mainly for two-sample tests for differences of mean. In this setting, there are two mutually independent time series, $\mathbf{x}_1$ and $\mathbf{x}_2$, of lengths n_1 and n_2, respectively, to be compared. However, the methods and most of the results are also valid for, and apply in an obvious way to, the corresponding one-sample tests, the statistic for which is given in (1). One-sample tests are appropriate for problems where one is interested in whether the mean of a single series could be a particular value, perhaps zero. For example, if the available data consisted of two *paired* (and possibly mutually correlated) series $\mathbf{y}$ and $\mathbf{z}$, equality of means for the two could be investigated through a one-sample test operating on $\mathbf{x} = \mathbf{y} - \mathbf{z}$, whose null hypothesis would be that the mean $\mu_X = 0$.

The basic statistic of interest for the two-sample tests developed below is the difference between the two sample means,

$$S(\mathbf{x}_1, \mathbf{x}_2) = \overline{x}_1 - \overline{x}_2. \tag{9}$$

Usually, the null hypothesis (H_0) will be that the means of the two populations from which the samples were drawn are equal. That is,

$$S_0 = \mu_1 - \mu_2 = 0, \tag{10}$$

although investigation of other null hypotheses presents no additional difficulty. Regardless of the specific H_0 of interest, the test statistic used to investigate the plausibility of that null hypothesis will be

$$d = \frac{S(\mathbf{x}_1, \mathbf{x}_2) - S_0}{\hat{\sigma}_S}, \tag{11}$$

in which the denominator is the square root of the estimated sampling variance of $S(\mathbf{x}_1, \mathbf{x}_2)$,

$$\begin{aligned}\hat{\sigma}_S &= [\mathrm{var}(\overline{x}_1) + \mathrm{var}(\overline{x}_2)]^{1/2} \\ &= \left(V_1' \frac{s_{x_1}^2}{n_1} + V_2' \frac{s_{x_2}^2}{n_2}\right)^{1/2}.\end{aligned} \tag{12}$$

Equation (11) is the two-sample version of (1). Except for the bias-corrected variance-inflation factors, this is the same test statistic used by Katz (1982) and is the basis of the two-sample ‘‘usual’’ test in the terminology of Zwiers and von Storch (1995). Asymptotically (i.e., for sufficiently large sample size), the test statistic d in (11) follows the standard Gaussian distribution when H_0 is true. For the nonparametric tests derived in this section, the form of the distribution of d is unimportant, but the point of scaling the test statistic (11) by the standard deviation in (12) is to improve the accuracy of the resulting tests, by removing the dependence of its distribution on unknown quantities (e.g., Hall and Wilson 1991).

The bootstrap test for the difference of means is based on evaluation of the unusualness of the observed $S(\mathbf{x}_1, \mathbf{x}_2)$, in terms of the test statistic d, in the context of an estimated sampling distribution for d derived by bootstrapping $\mathbf{x}_1$ and $\mathbf{x}_2$. The moving blocks bootstrap is applied repeatedly to produce n_B realizations of the bootstrap analog of (11),

$$d^* = \frac{S(\mathbf{x}_1^*, \mathbf{x}_2^*) - S(\mathbf{x}_1, \mathbf{x}_2)}{\hat{\sigma}_S^*}. \tag{13}$$

For a two-tailed test, the null hypothesis is then rejected at the $\alpha 100\%$ level if the observed value of d would place in the most extreme $\alpha 100\%$ of the distribution of d^*. That is, the test rejects H_0 if

$$\frac{\#\{d^* \geq d\}}{n_B + 1} = \frac{\sum_{i=1}^{n_B} I(d_i^* \geq d)}{n_B + 1} \leq \frac{\alpha}{2} \qquad (14a)$$

for the upper tail, or

$$\frac{\#\{d^* \leq d\}}{n_B + 1} = \frac{\sum_{i=1}^{n_B} I(d_i^* \leq d)}{n_B + 1} \leq \frac{\alpha}{2} \qquad (14b)$$

for the lower tail. Here, $I(\cdot)$ is the indicator function, which equals 1 if its argument is true and is zero otherwise. For example, the results presented below will be based on the sampling distribution of d being estimated using $n_B = 1999$ bootstrap realizations of d^*. These two-tailed tests will then reject H_0 at the 5% level either if the observed value of d is within the range of or more extreme than the largest 50, or the smallest 50, of these d^* values. The corresponding one-tailed test would reject H_0 at the $(\alpha/2)100\%$ level if either (14a) or (14b), as appropriate, were satisfied. Equations (14a) and (14b) constitute the simple ''percentile method'' (Efron and Tibshirani 1993) for bootstrap confidence interval estimation. More elaborate refinements to the percentile method exist (Efron 1987; Efron and Tibshirani 1993), but these were not found to materially improve the performance of the difference-of-mean tests described here.

The difference of bootstrap means $S(\mathbf{x}_1^*, \mathbf{x}_2^*)$ is compared in the numerator of (13) to the original statistic $S(\mathbf{x}_1, \mathbf{x}_2)$—*not* to S_0, which is the corresponding term in (11). This substitution is necessary to ensure that the resulting realizations of d^* reflect H_0, considering that from the perspective of the bootstrap samples $\mathbf{x}_1$ and $\mathbf{x}_2$ constitute the population, and the difference of their sample means $S(\mathbf{x}_1, \mathbf{x}_2)$ is the bootstrap model for the population counterpart S_0 (10). The resulting bootstrap distribution of d^* will be centered near zero. Failure to formulate d^* in this way leads to tests of low power (i.e., poor sensitivity to violations of H_0). Further discussion on this point can be found in Hall and Wilson (1991).

In (13), the term $S(\mathbf{x}_1^*, \mathbf{x}_2^*)$ is computed by separately applying the moving blocks bootstrap to $\mathbf{x}_1$ and $\mathbf{x}_2$, and then computing and differencing the sample means of the resulting pair of bootstrap samples. Note that the two samples $\mathbf{x}_1$ and $\mathbf{x}_2$ are not mixed in the resampling process. Maintaining the separateness of the two samples in the resampling is an important difference between the present test and, for example, the ''BP'' bootstrap test described by Zwiers (1987). A practical consequence is that the null hypothesis can pertain exclusively to the means of the two samples and that equality of other aspects of the two distributions need not be assumed.

The denominator in (13) is a resampling counterpart of the denominator in (11). However, application of a second pass of moving blocks bootstrapping to the bootstrap samples $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ would introduce further rifts in the time sequences of the original series $\mathbf{x}_1$ and $\mathbf{x}_2$. The approach taken here is to estimate the sampling variability of $S(\mathbf{x}_1^*, \mathbf{x}_2^*)$ through the jackknife variance estimates for $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$ (e.g., Efron 1982; Efron and Tibshirani 1993), but to compute them by operating on the bootstrap blocks rather than on the individual data values. That is, the denominator in (13) is computed using jackknife-after-bootstrap variance estimates from the two bootstrap samples,

$$\hat{\sigma}_S^* = (\hat{\sigma}_{\mathrm{JAB},1}^2 + \hat{\sigma}_{\mathrm{JAB},2}^2)^{1/2}. \qquad (15)$$

The two jackknife-after-bootstrap variance estimates are based on b sample means derived from the bootstrap samples $\mathbf{x}^*$, each computed by leaving out a different one of the b data blocks. Define

$$\bar{x}_{(i)}^* = \frac{\sum \mathbf{x}_{(1)}^* + \sum \mathbf{x}_{(2)}^* + \sum \mathbf{x}_{(i-1)}^* + \sum \mathbf{x}_{(i+1)}^* + \cdots + \sum \mathbf{x}_{(b)}^*}{n' - l} \qquad (16)$$

as the ith of these b averages, where $\sum \mathbf{x}_{(i)}^*$ is the sum of the data values over the ith block and $n' = bl$ is the length of the full bootstrap sample $\mathbf{x}^*$. (It may happen that $n' \neq n$ if b and l do not divide n evenly.) Define also

$$\bar{x}_{(\cdot)}^* = \frac{1}{b} \sum_{i=1}^{b} \bar{x}_{(i)}^* \qquad (17)$$

as the average of these averages. Then the jackknife-after-bootstrap estimate of the variance of the mean can be written as

$$\hat{\sigma}_{\mathrm{JAB}}^2 = \left(\frac{n'}{n}\right)\left(\frac{b-1}{b}\right) \sum_{i=1}^{b} (\bar{x}_{(i)}^* - \bar{x}_{(\cdot)}^*)^2, \qquad (18)$$

in which the factor (n'/n) is included for cases where $n' \neq n$. As a practical consideration, particularly for small b, it sometimes happens that all the blocks drawn in a particular bootstrap sample are the same. If so, this variance estimate is zero, and a new $\mathbf{x}^*$ must be drawn for the test to proceed.

It remains to specify the block length l. A practical requirement for the rule used to choose the block length is that it must depend only on sample statistics and not on unknown population quantities. If l is too small, the resulting test will be permissive (H_0 rejected too frequently), with the limit of $l = 1$ corresponding to the test that ignores the serial correlation altogether (cf. Zwiers 1990). It is also found that the test is stringent (H_0 is rejected too rarely) if l is too large. A workable rule for choice of the block length was developed here by trial and error evaluation of test results for synthetic AR(1) series in cases where H_0 was true, respecting the constraints mentioned in the last paragraph of section 2a. This process led to the rule: select the largest integer no greater than

$$l = (n - l + 1)^{(2/3)(1-1/V')}, \qquad (19)$$

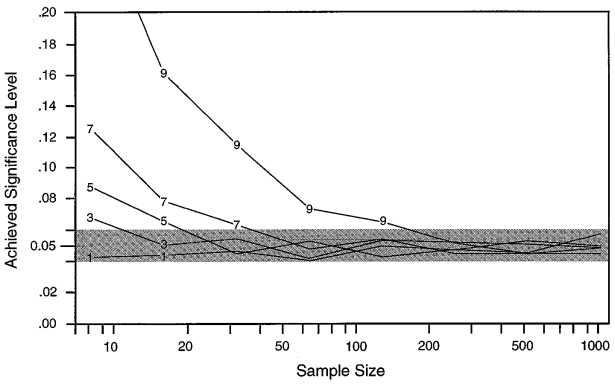

FIG. 3. Achieved significance levels for two-sided, two-sample AR(1) tests using sample sizes ranging from $n = 8$ to $n = 1024$ and autoregressive parameter $\phi = 0.1, 0.3, 0.5, 0.7$, and 0.9. Shaded region indicates the 95% confidence interval around the nominal significance level of 5%.

which must be evaluated iteratively, but converges quickly (a reasonable initial guess for l is $\sqrt{n}$). The specified block length thus increases with increasing sample size, and for a given sample size is larger for larger values of the bias-adjusted variance inflation factor V'. Having chosen the block length, the number of blocks b in the bootstrap sample is determined as the nearest integer to n/l. The two samples $\mathbf{x}_1$ and $\mathbf{x}_2$ may yield different values for b, l, and $n' = bl$.

The criterion used for the choice of (19) for specification of l was that the resulting tests would exhibit accurate rejection probabilities for the smallest possible effective sample sizes n_e, without compromising test performance for large n_e. Figure 3 shows the achieved significance levels (estimated probability of rejecting H_0 when it is true) for the resulting tests, for selected positive values of the autoregressive parameter and for sample sizes ranging from $n = 8$ to $n = 1024$. These probabilities are based on two-sided tests at the 5% level, estimated as relative frequencies with which H_0 was rejected among tests on 2000 replications of pairs of AR(1) series having equal means. These results are for synthetic series independent of those used to arrive at (19). The shaded region in Fig. 3 indicates the 95% confidence interval around the nominal significance level of 5% for this number of replications. The figure indicates that the test operates correctly when n_e is greater than approximately 10 [or greater than about 8, according to (8)], which compares favorably to the ZvS test and to the tests described in Zwiers and Thiébaux (1987). For cases with inadequate sample size, the tests are uniformly permissive.

Equation (19) for the block length is also valid for the corresponding one-sample test of the mean. In that case, the statistic $S(\mathbf{x})$ is simply the sample mean of the single sample $\mathbf{x}$, the test statistic d is given by (1) [using also (2) and (3)], and the denominator in (13) is the square root of the jackknife-after-bootstrap variance in (18).

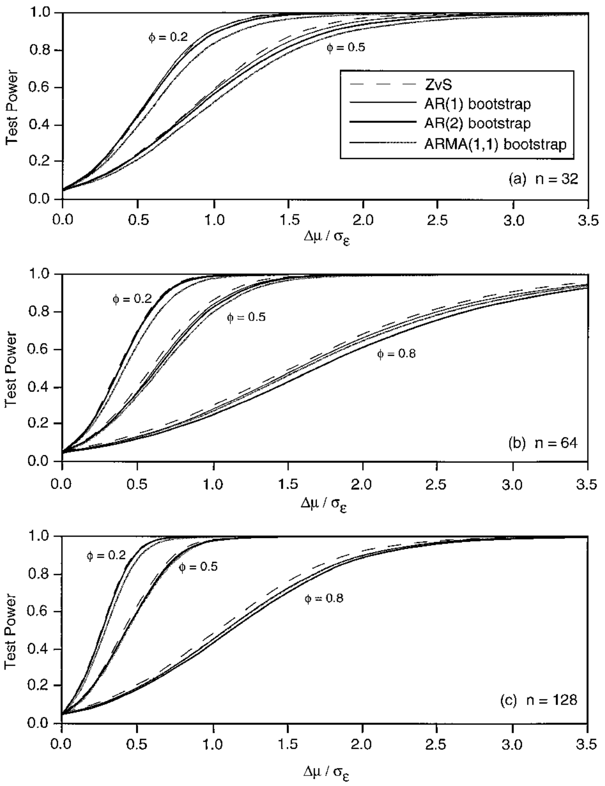

FIG. 4. Comparison of power functions for two-sample tests for differences of mean, using AR(1) data and conducted at the 5% level. Horizontal axis is difference in population means between the two samples, standardized by the square root of the white-noise variance.

In addition to accurate rejection probabilities when H_0 is true, one is interested in the power, or ability to detect violations of H_0, of these bootstrap tests. The power of a hypothesis test generally increases with the magnitude of the violation of H_0, and the probability with which H_0 will be rejected, as a function of the (true) alternative hypothesis, is known as the power function for that test. Figure 4 compares power functions for the AR(1) bootstrap tests described in this section (light solid curves) to those for the ZvS test (light dashed curves) for selected sample sizes and values of the autoregressive parameter. The power of the bootstrap tests is generally slightly less than, but often indistinguishable from, that of the ZvS test. Zwiers and von Storch (1995) show that the power of the ZvS test is nearly as great as the corresponding likelihood ratio test, which provides a theoretical maximum on test power.

Usually it is assumed that the white-noise series ϵ_t in (5) consists of independent Gaussian variates. There is no guarantee that real data will be Gaussian, however, so that the robustness of the bootstrap test to violations of this assumption is also of interest. Figure 5 shows achieved significance levels, again estimated from the results of 2000 tests of pairs of series for which H_0 is true, generated using $\phi = 0.5$ and independent white noise following uniform, Gaussian, exponential, and

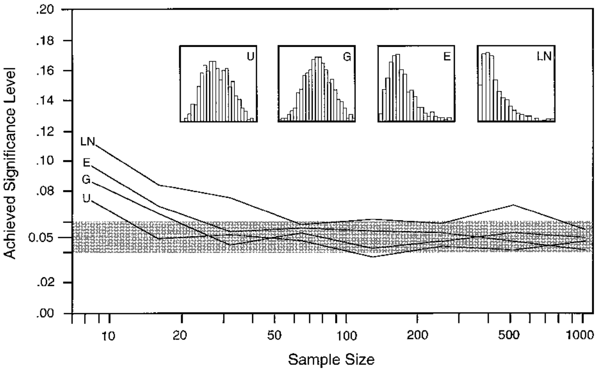

FIG. 5. Achieved significance levels for two-sided, two-sample AR(1) tests using sample sizes ranging from $n = 8$ to $n = 1024$ and autoregressive parameter $\phi = 0.5$ for data series forced with white noise following four different distributions: uniform (U), Gaussian (G), exponential (E), and lognormal (LN) Insets show histograms of samples of the resulting series values. Shaded region indicates the 95% confidence interval around the nominal significance level of 5%.

lognormal distributions with $E[\epsilon] = 0$. The insets show histograms of samples of the resulting **x** values. As in Fig. 3, the shaded region indicates the 95% confidence band around the nominal significance level of 5%. The results for Gaussian variates are the same as those for $\phi = 0.5$ in Fig. 3. Figure 5 shows that the test performance is even better (i.e., the test is accurate for smaller n) for autoregressive series forced with uniform variates. While somewhat larger sample sizes are required for the test to operate properly when the ϵ_t values are strongly skewed (exponential and lognormal distributions), the two-sample test is quite robust overall to these violations of the Gaussian assumption. The two-sample ZvS test (results not shown) is similarly robust to these deviations from Gaussian data. Overall achieved significance levels are comparable to those in Fig. 5 for the corresponding one-sample bootstrap tests, but for the strongly skewed time series the resulting bootstrap distributions are asymmetric, and the probabilities of rejection are higher for the left tail than for the right tail.

Finally, it is of interest to investigate the robustness of this test to violations of the assumed AR(1) time structure of the data. Zwiers and von Storch (1995) speculated that, while atmospheric and other geophysical time series do not necessarily follow an AR(1) model, data not clearly influenced by quasi-periodic processes such as ENSO or the quasi-biennial oscillation have power spectra sufficiently similar to the AR(1) model that the ZvS test should be applicable. Figure 6 shows achieved significance levels for two-sample, two-tailed bootstrap tests, which assume AR(1) time dependence, operating on data with more complicated autocorrelation structures. The results in Fig. 6a are for data generated according to the AR(2) model (section 2c) over that portion of the parameter space corresponding to stationary series with positive lag-1 autocorrelation. Here, the tests operate correctly only for generating processes that are very close to the AR(1)—that is, for very small $|\phi_2|$. The tests are permissive for $\phi_2 > 0$ and stringent for $\phi_2 < 0$, with very large deviations from accurate test performance evident for ϕ_2 far from zero. For ϕ_2 more negative than about $-2/3$, none of the 2000 test replications rejected H_0. Corresponding results (Fig. 6b) for the autoregressive–moving average [ARMA(1,1)] generating process (section 2d) are only slightly better, with accurate rejection probabilities only for small $|\theta_1|$, although the wild deviations from the nominal significance level seen in Fig. 6a are absent. Neither of the

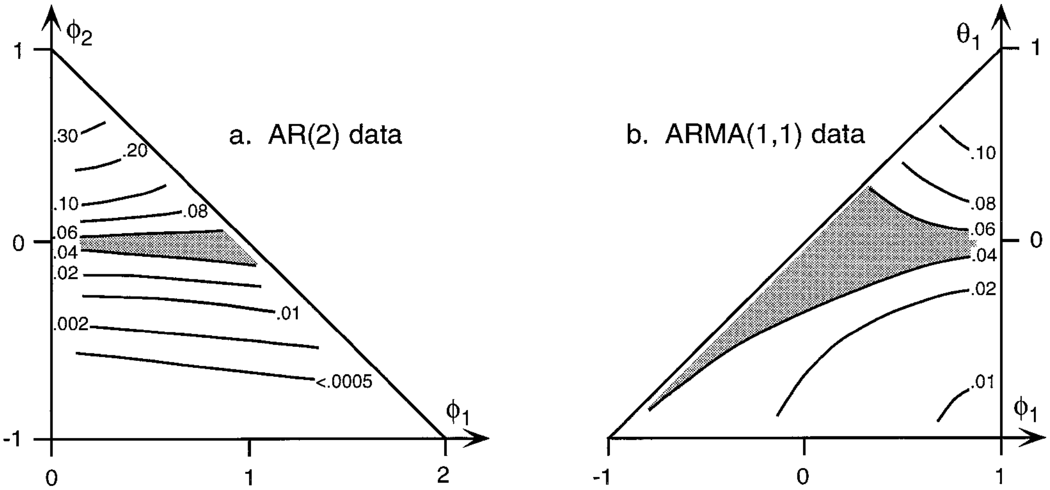

FIG. 6. Achieved significance levels for two-sided, two-sample AR(1) tests when applied to (a) AR(2) and (b) ARMA(1,1) data series for which $r_1 > 0$, with $n = 128$. Shading indicates approximate regions in the respective parameter spaces for which these are not different from the nominal test level of 5% with 95% confidence. Increasing sample size does not improve test performances. Corresponding results for the ZvS test are indistinguishable.

results in these two panels is improved by increasing the sample size. The corresponding results for the ZvS test are indistinguishable from those shown in Fig. 6.

c. Bootstrap test for AR(2) data

The second-order autoregressive [AR(2)] model is an extension of the AR(1) process in (5) to dependence on two time lags. It is defined by the equation

$$x'_t = \phi_1 x'_{t-1} + \phi_2 x'_{t-2} + \epsilon_t. \qquad (20)$$

The AR(2) model is considerably more flexible than the AR(1), which is obtained as a special case of (20) when $\phi_2 = 0$. This more general model can exhibit autocorrelation functions decreasing either more quickly or less quickly than the exponential decay of the AR(1) process given in (7) and also has the capacity to exhibit pseudoperiodic behavior (e.g., Box and Jenkins 1976; Wilks 1995; Zwiers 1990).

Figure 6a shows very strongly that the AR(1) test described in the previous section (as well as the ZvS test) is not robust to the more general case of AR(2) data, either (as expected) for regions of the parameter space where pseudoperiodicities are produced or for the many parameter combinations yielding spectra that are monotonically decreasing functions of frequency. This section describes a bootstrap test analogous to that in section 2b, but that operates correctly when applied to data generated by the more general AR(2) model.

Computing the test statistic d proceeds as in section 2b, with the following modifications. To estimate the variance inflation factor [(3)] for AR(2) data, the first two sample autocorrelations are required. Equation (6) is used for the first of these. The sample lag-2 autocorrelation is estimated here using

$$r_2 = \frac{\sum_{t=1}^{n-2} (x_t - \overline{x}_-)(x_{t+2} - \overline{x}_+)}{\left[\sum_{t=1}^{n-2} (x_t - \overline{x}_-)^2 \sum_{t=3}^{n} (x_t - \overline{x}_+)^2\right]^{1/2}}, \qquad (21)$$

which gives more stable results for small samples than does the simple extension of the form of (6) to two lags. In (21), the subscripts "−" and "+" denote sample means over the first and last $n - 2$ series values, respectively. The properties of the AR(2) process are then used to estimate the remaining $n - 3$ autocorrelations in (3), according to

$$r_k = \hat{\phi}_1 r_{k-1} + \hat{\phi}_2 r_{k-2}, \qquad k \geq 3, \qquad (22a)$$

where

$$\hat{\phi}_1 = r_1 \frac{1 - r_2}{1 - r_1^2} \qquad (22b)$$

and

$$\hat{\phi}_2 = \frac{r_2 - r_1^2}{1 - r_1^2}. \qquad (22c)$$

As is the case for the AR(1) test, the values of V thus obtained are biased. Empirical analysis of samples of synthetic AR(2) series, of the same kind used to develop (8), lead to the bias-adjusted variance inflation factor

$$V' = V \exp\left(\frac{3V}{n}\right), \qquad (23)$$

which again is very nearly unbiased except for very small n and strong time dependence.

Bootstrapping for the AR(2) test to generate a distribution of d^* (13) also proceeds as described in the previous section, excepting only that the block lengths are chosen using

$$l = (n - l + 1)^{(2/3)(1-1\sqrt{4V'})}. \qquad (24)$$

As before, this equation has been developed through trial and error, using the guidelines in the last paragraph of section 2a and according to the criterion that the resulting test operates correctly for the smallest possible sample sizes without compromising performance for larger samples. Figure 7 illustrates this performance for two-sample, two-tailed tests, again over that portion of the AR(2) parameter space corresponding to stationary models with positive lag-1 autocorrelation. The shaded regions indicate parameter combinations for which the achieved significance level is within the 95% confidence band of the nominal significance level of 5%, as estimated from 2000 replications for which H_0 is true. For $n = 16$, the test operates accurately only for very weak dependence, where $V' < 1$. The portion of the parameter space exhibiting accurate tests steadily increases with increasing sample size, until for $n = 64$ only tests operating on data exhibiting very strong dependence ($\phi_2 \gg \phi_1$) are permissive. For $n \geq 128$ the test yields accurate rejection probabilities for all AR(2) parameter combinations investigated (also indicated in Fig. 7). Overall, the AR(2) bootstrap test is accurate for n_e larger than about 12 to 15.

Since the AR(1) process can be regarded as a special case of the AR(2), the test described in this section can also be used, and performs accurately, for AR(1) data. Given the results in Fig. 6a, it is worth considering what is lost when the slightly more elaborate AR(2) test is applied to AR(1) series. The answer is that estimation of the second sample autocorrelation in (21) results in the AR(2) bootstrap test having slightly less power. The heavy solid lines in Fig. 4 show the power functions for the AR(2) test when applied in selected AR(1) settings. The AR(2) test is uniformly less powerful, but the loss of power is quite small except for small n_e. The AR(2) test also exhibits robustness to non-Gaussian data that are comparable to those shown in Fig. 5 for the AR(1) test. Unless one can be quite confident that a data series can be adequately modeled as AR(1), the serious nonrobustness of the AR(1) test to AR(2) data illustrated in Fig. 6a suggests that the AR(2) test should be preferred to either the AR(1) test or the ZvS test.

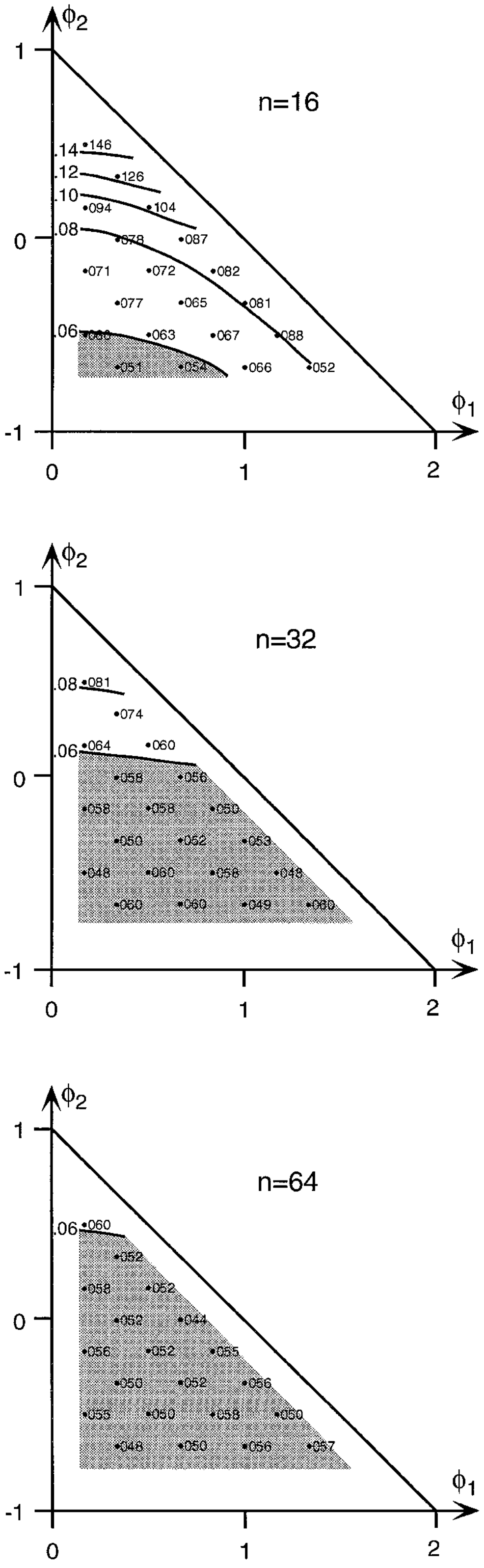

FIG. 7. Achieved significance levels for two-sided, two-sample AR(2) tests. Shading indicates approximate regions in the parameter space for which these are not different from the nominal test level of 5% with 95% confidence. For $n = 128$ and greater, the test operates correctly for all AR(2) parameter combinations considered.

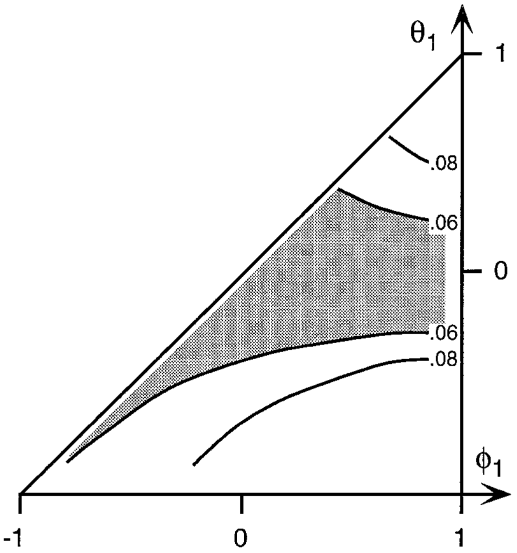

FIG. 8. Achieved significance levels for two-sided, two-sample AR(2) tests when applied to ARMA(1,1) data series with $n = 128$. Shading indicates approximate region for which these are not different from the nominal test level of 5% with 95% confidence. Test performance improves only very slightly with sample size.

Finally, Fig. 8 indicates the performance of the AR(2) test when applied to ARMA(1,1) data (section 2d). Accurate tests are achieved over a larger portion of this parameter space than was seen for the AR(1) test (compare Fig. 6b), and the achieved significance levels agree with the nominal level to within a factor of 2 throughout. It is always permissive where it does not operate correctly. Figure 8 shows results for $n = 128$, but improvements with increasing sample sizes are slight. An improvement with respect to the AR(1) test is not surprising, given that the AR(2) model is very much more flexible than the AR(1) model.

d. Bootstrap test for ARMA(1,1) data

Another simple but plausible time series model for geophysical data is the autoregressive–moving average process, of order 1 in both the autoregression and the moving average. This is the ARMA(1,1) process, defined by

$$x'_t = \phi_1 x'_{t-1} + \epsilon_t - \theta_1 \epsilon_{t-1}, \qquad (25)$$

where ϕ_1 is the autoregressive parameter, θ_1 is the moving average parameter, and the remaining symbols have the same definitions as in (5) and (20). The ARMA(1,1) process can be viewed as an AR(1) process forced by noise ϵ_t that is itself autocorrelated.

Figures 6b and 8 indicate that the tests in the previous two sections, based on the assumptions that the data conform to the AR(1) and AR(2) models, respectively, perform accurately over only limited portions of ARMA(1,1) parameter space. An analogous test can be developed assuming that the data have come from an

ARMA(1,1) generating process. For this test, the variance inflation factor V is computed using (3), (6), (21), and

$$r_k = \hat{\phi}_1 r_{k-1},\ k \geq 3, \tag{26a}$$

where

$$\hat{\phi}_1 = r_2/r_1. \tag{26b}$$

That is, autocorrelation functions for ARMA(1,1) processes decay exponentially from their values at the first lag, at a rate specified by the autoregressive parameter.

As before, the estimate of the variance inflation factor so obtained is biased, and a correction is necessary for the bootstrap test to perform properly for relatively small sample sizes. Following again the regression approach used to develop (8), it was found that a large part of this bias for ARMA(1,1) series can be removed by applying

$$V' = V \exp\left(\frac{(2 + \hat{\theta}_1)V}{n}\right), \tag{27a}$$

where $\hat{\theta}_1$ is obtained as the root of

$$(r_1 - \hat{\phi}_1)\theta_1^2 + (1 - 2\hat{\phi}_1 r_1 + \hat{\phi}_1^2)\theta_1 + (r_1 - \hat{\phi}_1) = 0 \tag{27b}$$

that statisfies $|\hat{\theta}_1| < 1$. The bias correction (27a) is somewhat less satisfying than its counterparts (8) and (23) in that it does not reflect the slight tendency for positive bias (i.e, V is too large, on average) for ARMA(1,1) processes for which θ_1 approaches ϕ_1, both of these parameters are relatively large, and the sample size is moderately large. Somewhat better test performance under these conditions might be obtained if a more refined bias correction were to be found.

The mechanics of the bootstrapping proceed as described before, using (24) to choose the block length. The accuracy of this test over the portion of the ARMA(1,1) parameter space corresponding to stationary series with positive lag-1 autocorrelation is shown as a function of sample size in Fig. 9. For $n = 16$ (Fig. 9a) the test is accurate over only a relatively small portion of the parameter space, and the inaccurate tests are all permissive. As the sample size increases to $n = 64$ (Fig. 9c), the test performs accurately over nearly all of the parameter space. Figure 9d illustrates the effect of the deficiency in the bias correction (27a), which, since the small positive bias of V is not corrected, produces tests that are stringent to a small degree in the upper portion of the parameter space. However, as a practical matter, the tests are quite workable even with this small problem, and the ARMA(1,1) test performs essentially correctly for $n > 64$ for all of the parameter combinations investigated here.

The power of these ARMA(1,1) tests is illustrated by the thick gray lines in Fig. 4. For most of these cases it is somewhat less sensitive to violations of H_0 than the other tests considered. However, for the more strongly autocorrelated data series ($\phi = 0.8$), the ARMA(1,1) test is slightly more powerful than the AR(2) test and is comparable to the AR(1) test.

Finally, the robustness of the ARMA(1,1) test to AR(2) data is shown in Fig. 10, for $n = 128$. While these results are better than the corresponding results for the AR(1) test (compare Fig. 6a), the ARMA(1,1) tests perform accurately only for relatively minor deviations from the AR(1) process, which is the special case of (25), with $\theta_1 = 0$. For AR(2) parameter combinations where the ARMA(1,1) test does not perform accurately, increasing the sample size yields tests that are increasingly stringent in relation to the results shown in Fig. 10. For smaller n the tests represented by the area below the stippling are still stringent, while tests in the area in and above the stippling are permissive for small samples.

3. Multivariate tests

The univariate (scalar data) tests developed in section 2 can be extended to multivariate (vector-valued data) tests by considering now that the data in each of the two samples are matrices [**x**] with elements $x_{t,m}$, where $t = 1, \cdots, n$ is the time index as before, and $m = 1, \cdots, M$ indexes the elements of the data vectors $\mathbf{x}_t$. The vector elements might correspond to points on a spatial grid, in which case each $\mathbf{x}_t$ could be a single realization of a field, or coefficients for these data projected onto a space of smaller dimension. The multivariate tests described below reduce to the univariate tests described in section 2 for $M = 1$.

The basic idea behind the multivariate tests is to simultaneously carry out the corresponding univariate tests for all data dimensions. That is, instead of block resampling scalar series, the same block resampling procedure will be applied to the vector data. Simultaneous application of the same resampling patterns to all dimensions of the data vectors will yield resampled statistics reflecting the cross correlations in the underlying data, without the necessity of explicitly modeling those cross correlations. Meanwhile, appropriate choice of the block length will allow the influence of the temporal correlation on the distribution of the test statistic to be represented as well.

For each dimension of the data, there will be a corresponding distance d_m computed using (11). The ''global,'' or field, significance relates to the size of the vector **d** of these ''local'' test statistics. If the null hypothesis of no real local difference in (vector) mean is true, the size of the vector **d**, having been drawn from a population with mean **0**, should be small. However, there are many choices for judging the size of the test vector **d**. The performance, with respect to test accuracy and power, of the following four vector norms will be investigated:

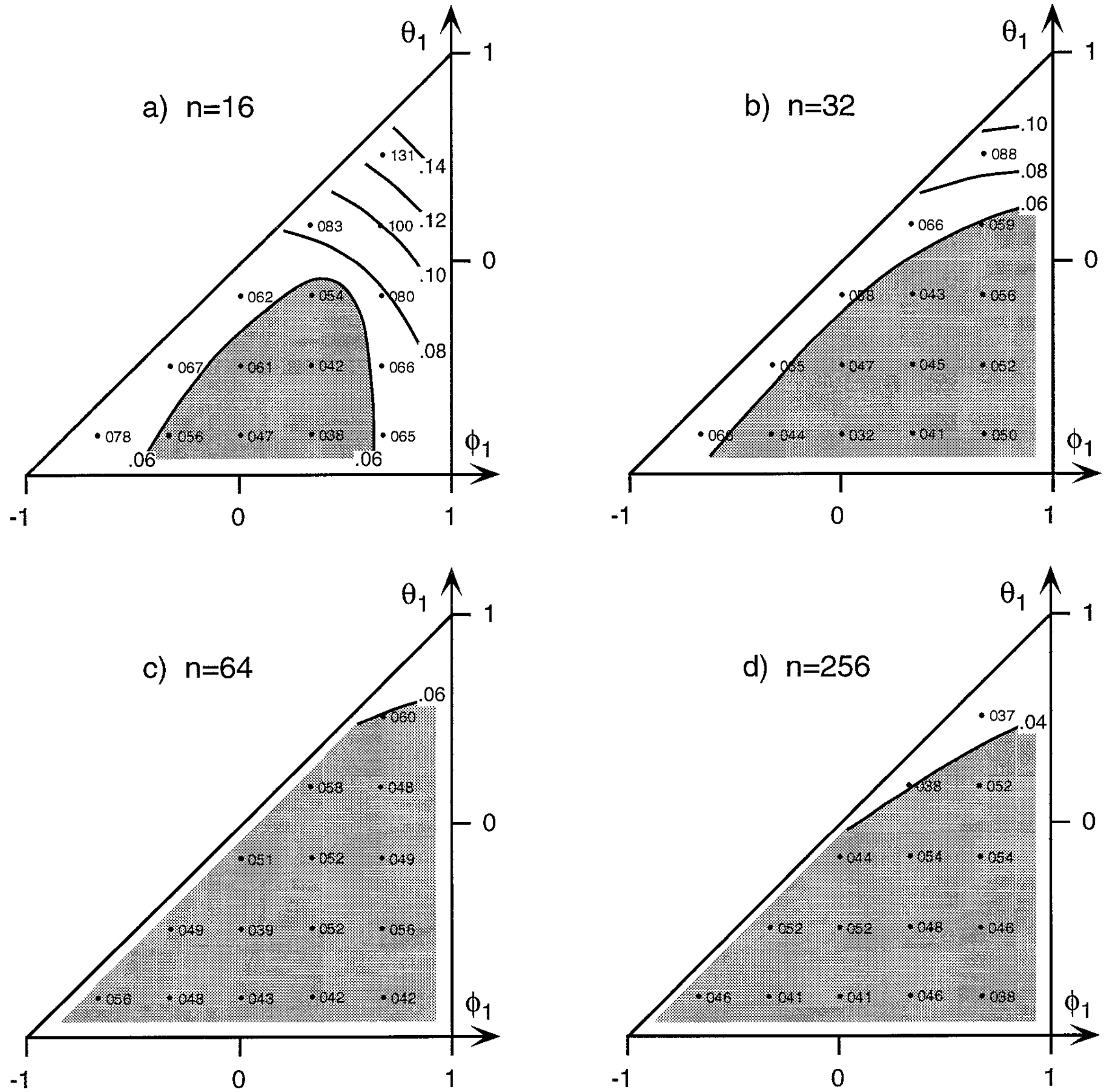

FIG. 9. Achieved significance levels for two-sided, two-sample ARMA(1,1) tests. Shading indicates approximate regions for which these are not different from the nominal test level of 5% with 95% confidence.

$$D_1 = \sum_{m=1}^{M} |d_m|, \tag{28a}$$

$$D_2 = \sum_{m=1}^{M} d_m^2, \tag{28b}$$

$$D_3 = \max_{m} |d_m|, \tag{28c}$$

and

$$D_4 = \#\{p_m \leq \alpha\} = \sum_{m=1}^{M} I[p_m \leq \alpha]. \tag{28d}$$

The first of these norms, D_1 (28a), measures the length of **d** as the sum of the absolute values of its elements, which has been recommended by Mielke (1985). The vector norm D_2 (28b) is the square of the traditional Euclidean distance in M-space, and this (or, equivalently, its square root) would be the conventional choice. Here, D_3 (28c) is the "max norm," or largest element of **d**, which typically corresponds to the smallest p value among the local tests, as suggested, for example, by Westfall and Young (1993). Finally, D_4 (28d) is the "counting norm" (Zwiers 1987), or the number of the M local tests that are significant at the α100% level, which has been employed frequently for testing climate data (e.g., Livezey and Chen 1983; Wilks 1996). In order to employ D_4, a specific local test must be chosen to compute the p value p_m, which in the following will

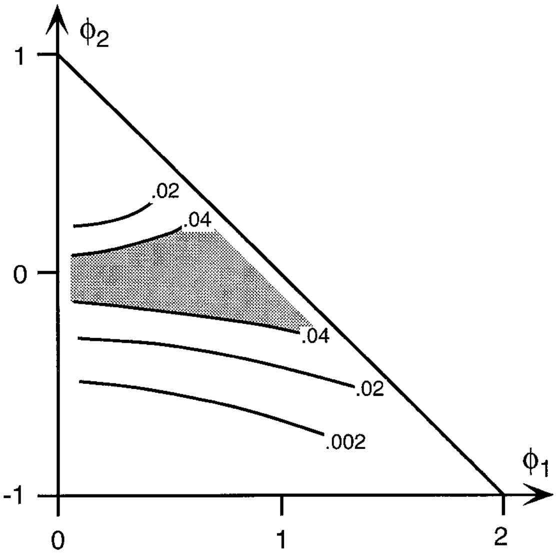

FIG. 10. Achieved significance levels for two-sided, two-sample ARMA(1,1) tests when applied to AR(2) data, with $n = 128$. Shading indicates approximate region in which these are not different from the nominal test level of 5% with 95% confidence.

be the z test assuming the asymptotic Gaussian distribution of (11).

Sampling distributions under H_0 for each of the norms in (28) are developed by bootstrapping the data [**x**] in the same way as described in section 2 for scalar series, with the same time blocks being chosen in each bootstrap sample for each of the M dimensions. Note that the two data batches [$\mathbf{x}_1$] and [$\mathbf{x}_2$] need not be resampled using the same block length. For each of the n_B bootstrap samples generated in this way, (13) is applied to each of the M elements to yield the vector $\mathbf{d}^*$, the length of which, D^*, is in turn computed according to the various norms in (28). The (univariate) distribution of each of these collections of n_B realizations of D^* is then used to assess the unusualness of the corresponding observed D by applying the percentile method (14), using D and D^* rather than d and d^*. That is, **d** is declared to be significantly different from **0** at the α100% level if D is among either the (α/2)100% smallest or the (α/2)100% largest of the n_B values of D^*.

The multivariate data used to demonstrate that the vector hypothesis tests will be realizations from the simplified vector autoregressive process

$$\mathbf{x}'_t = \phi\, \mathbf{x}'_{t-1} + [\mathbf{C}]\, \boldsymbol{\epsilon}_t, \qquad (29)$$

where $\boldsymbol{\epsilon}_t$ is an M-dimensional vector of independent standard Gaussian variates. The matrix [**C**] reflects the correlation among the elements of $\mathbf{x}_t$ and is related to their (unlagged) correlation matrix according to $[\mathbf{C}]\,[\mathbf{C}]^T = [\mathbf{R}_0]$. For simplicity, a scalar autoregressive coefficient ϕ is used in (29), which implies that the matrix of lag-1 correlations [$\mathbf{R}_1$] is symmetric and proportional to [$\mathbf{R}_0$].

Multivariate data with two forms of cross correlation will be investigated, which are the same as those used by Zwiers (1987). These correlations among the M elements of **x** are introduced through the matrix of unlagged correlations [$\mathbf{R}_0$], whose elements are

$$\rho_{i,j} = \begin{cases} 1, & |i - j| = 0 \\ s_1/(1 - s_2), & |i - j| = 1 \\ s_1\rho_{|i-j|-1} + s_2\rho_{|i-j|-2}, & |i - j| \geq 2. \end{cases} \qquad (30)$$

The two spatial correlation structures to be considered are a spatial analog of an AR(1) process, for which $s_1 = 0.9$ and $s_2 = 0$, and a spatial analog of an AR(2) process, for which $s_1 = 1.6$ and $s_2 = -0.8$. The (spatial) lag-1 autocorrelation for these two models is nearly identical. However, at larger spatial separations, the autocorrelation structures are quite different. The spatial AR(1) autocorrelation function decreases monotonically toward zero with increasing lag, while the spatial AR(2) autocorrelation function exhibits damped oscillations around zero, with local maxima and minima reminiscent of teleconnection patterns.

In order for the vector block bootstrapping to capture the cross correlation in the data produced by (30), it is essential that the same block length l be applied to each of the vector elements. In the following, this common block length is chosen by separately computing l for each vector element, as described in section 2, and then averaging these M specified block lengths. This procedure is a quick and ad hoc choice, which could very likely be improved upon. Note that it is not necessary for the two data vectors in a two-sample test to be resampled with the same block length.

Figures 11 and 12 show the achieved significance levels of these vector tests for data generated using the spatial AR(2) process and spatial AR(1) process from (30), respectively. The tests are two-sample, two-sided tests, and the shading again indicates 95% confidence intervals around the nominal significance level of 5%. The four panels in each figure show results for tests constructed using each of the four test statistic norms in (28). These results are for sample sizes ranging from $n = 16$ to $n = 2048$ and for $M = 4$ to $M = 64$ dimensional data. The symbols indicate data generated using $\phi = 0.2$, 0.5, and 0.8 in (29), and only points for which $n_e > 10$ are plotted.

Achieved significance levels for the multivariate tests based on both D_1 and D_2 are quite close to the nominal level of 5% when the effective sample size n_e is relatively large with respect to the dimension M of the data vectors, and the performances of these two norms in this respect are nearly indistinguishable. These observations apply as well to tests based on the max norm D_3, except when the temporal dependence is strong, in which case the tests tend to be stringent. Tests based on the counting norm D_4 appear to be consistently stringent even for large n_e/M, although not grossly so. However, for weak temporal dependence ($\phi = 0.2$), tests based on D_3 perform well even for small sample sizes in relation to M. For other situations in which $n_e < M$, all

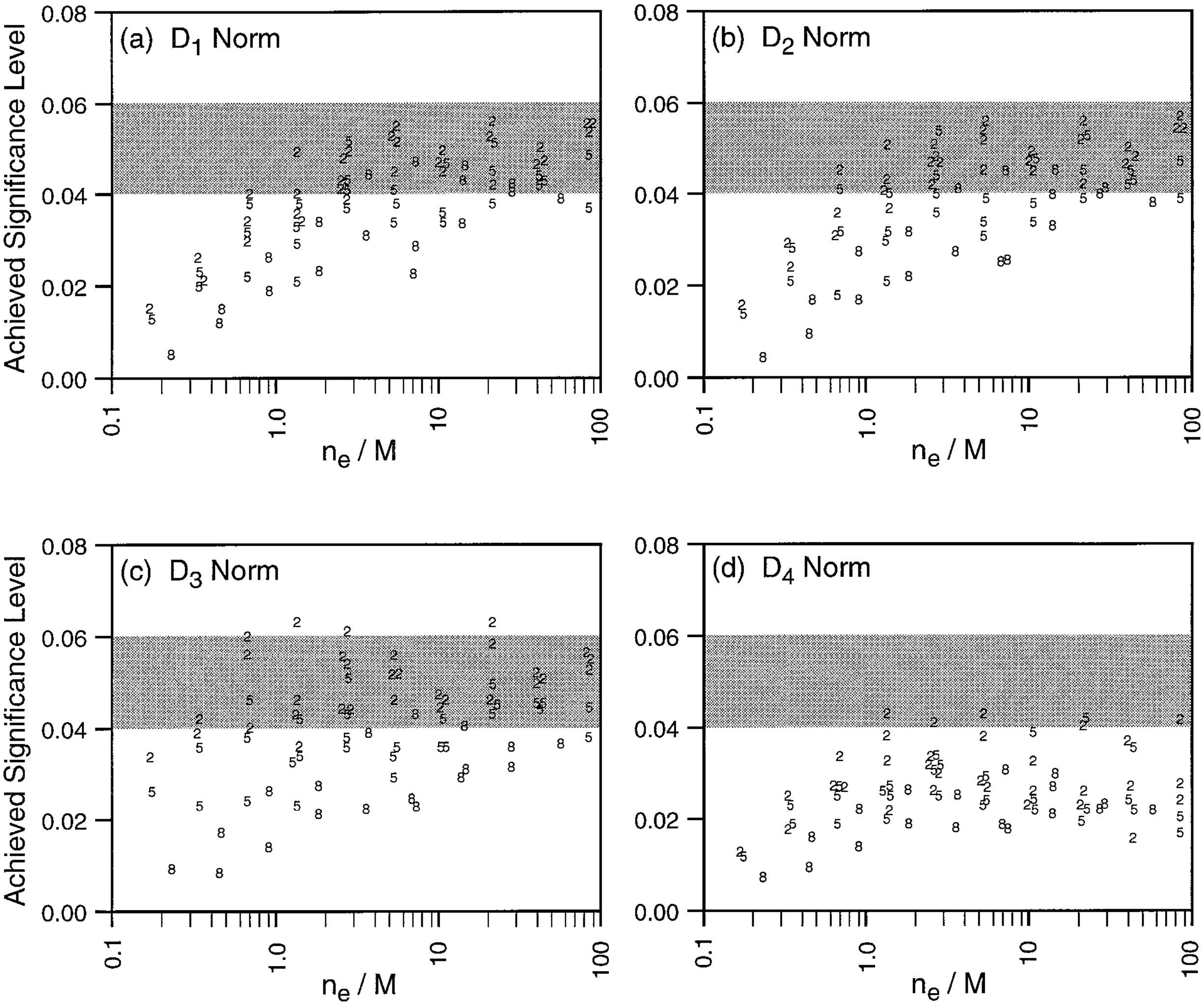

FIG. 11. Achieved significance levels for multivariate two-sample, two-sided tests. Shaded regions indicate 95% confidence intervals around the nominal significance level of 5%. Sample sizes range from $n = 16$ to $n = 2048$, and the dimension of the data vectors ranges from $M = 4$ to $M = 64$. Dependence between vector elements is specified by (30), with $s_1 = 1.6$ and $s_2 = -0.8$. Symbols indicate autoregressive parameter $\phi_1 = 0.2$, 0.5, or 0.8. Only points for which $n_e > 10$ are plotted.

of the tests are stringent, especially for the spatial AR(2) data in Fig. 11, but could still be used in practice given the understanding that the actual test level is somewhat smaller than the nominal one. Tests for data with $n_e < 10$ (not shown) are progressively more permissive, but exhibit gross inaccuracies (achieved significance levels larger than 0.10) only for n_e smaller than about 2. Results similar to those in Figs. 11 and 12 are obtained for extensions of (29) used to generate vector AR(2) data (not shown), with series having weaker or stronger autocorrelation (as measured by V) exhibiting results similar to those in Figs. 11 and 12, with smaller or larger values of ϕ, respectively.

The most pronounced differences in the performance of tests based on the four norms in (28) are with respect to their power. Figure 13 illustrates the sensitivity to violations of H_0 of the multivariate tests constructed using these four vector norms. The specific case presented pertains to two-sided, two-sample tests at the nominal 5% level, applied to vector AR(1) data with $n = M = 64$ and cross correlation described by (30), with $s_1 = 1.6$ and $s_2 = -0.8$. The six panels in Fig. 13 show power functions for different numbers m_Δ of the M vector elements of the population mean for one of the two samples being increased. The power functions in Fig. 13a pertain to the mean of only $m_\Delta = 1$ of the $M = 64$ elements being increased, with the identity of that element being chosen randomly for each replication. Since one would expect that violations of H_0 would tend to occur with spatial coherency, changes to the vector means in Figs. 13b–f are made randomly, but with the constraint that the changes are made to adjacent vector elements, in contiguous blocks of length m_Δ.

Clearly, the power of the tests increases as m_Δ increases, since increasing the means of larger numbers of vector elements by a fixed amount provides stronger

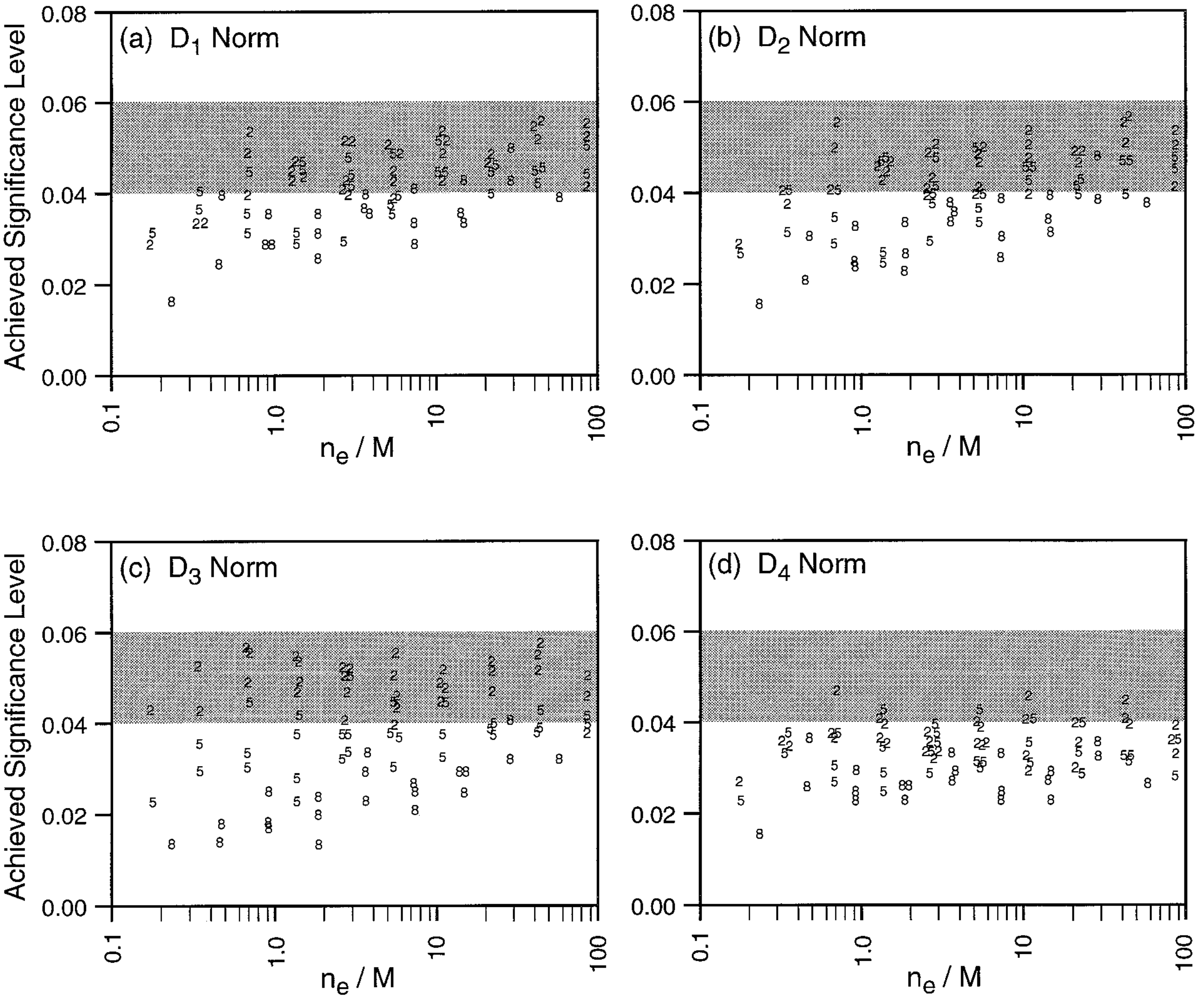

FIG. 12. As in Fig. 11 but for dependence between vector elements specified by (30) with $s_1 = 0.9$ and $s_2 = 0$.

signals for the tests to detect. As expected by extension of Fig. 4, power is greater for tests operating on less strongly correlated data ($\phi = 0.5$) than on more strongly correlated data ($\phi = 0.8$). For clarity, only results for $\phi = 0.5$ are shown in Figs. 13a,b. Power for tests with $\phi = 0.2$ (also not shown) bears the same qualitative relationship to the $\phi = 0.5$ and $\phi = 0.8$ results as in Fig. 4. The levels of the power functions for $\Delta\mu/\sigma_\epsilon = 0$ indicate that the tests are stringent for this sample size, particularly for $\phi = 0.8$. This is consistent with Fig. 11 since Fig. 13 pertains to the rather demanding cases of $n_e/M = 0.34$ and 0.12 for $\phi = 0.5$ and $\phi = 0.8$, respectively.

For a given value of m_Δ, there are very major differences in test sensitivity among the tests based on the four vector norms in (28). For the most challenging cases of small m_Δ, the max norm D_3 (heavy solid lines) is clearly superior to the others. The squared-distance norm D_2 (light solid lines), absolute-distance norm D_1 (light dashed lines), and the counting norm D_4 (heavy dashed lines) show progressively decreasing sensitivity, with D_4 exhibiting especially low power for small m_Δ. As m_Δ increases, the power of the D_3 test increases only slightly, while the power of the other tests increases substantially, until $m_\Delta = 16$, where power for most of the tests is fairly comparable. For $m_\Delta = 32$ the power of the D_3 tests is noticeably less than for the other three norms, which are now practically identical in power.

The same basic patterns shown in Fig. 13 occur for other sample sizes as well, although with less power for smaller n and greater power for larger n. Results for vector data generated using (30) with $s_1 = 0.9$ and $s_2 = 0$ are also very similar, but with a tendency for the tests to exhibit slightly less power. Power functions for analogous tests performed on temporal AR(2) data, with the component scalar tests as described in section 2c, are also basically the same but with (analogously to Fig. 4) slightly less power when applied to these temporal AR(1) data. Power functions for the tests applied to mutually independent multivariate data (i.e., $s_1 = s_2 =$

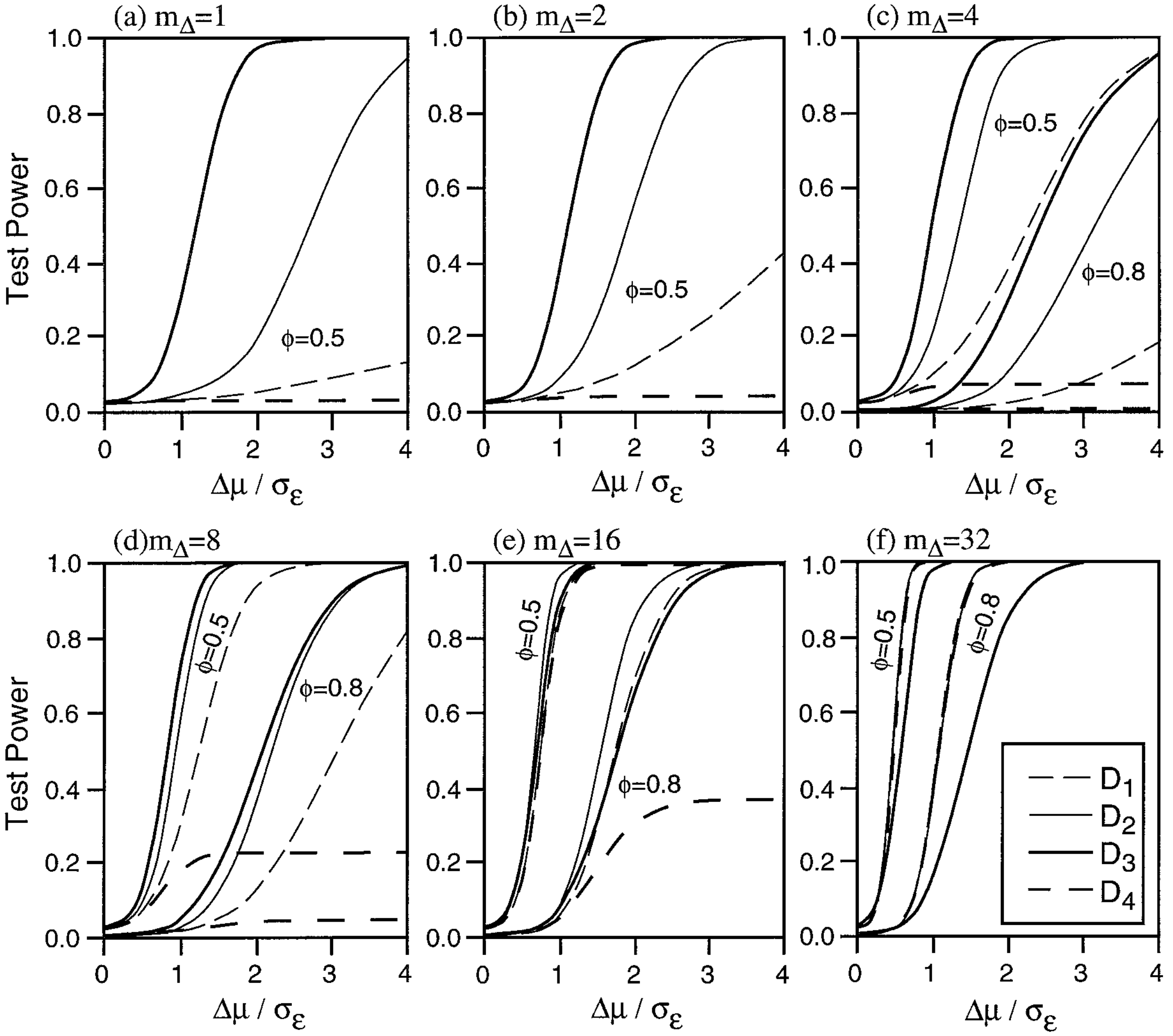

FIG. 13. Comparison of power functions for multivariate two-sample tests, using vector AR(1) data with $n = 64$ and $M = 64$, conducted at the nominal 5% level. Horizontal axis is difference between individual elements of the population means, standardized by the square root of the white-noise variance. (a)–(f) Results for differences between the mean vectors applied to $m_\Delta = 1, 2, 4, 8, 16$, and 32 randomly selected adjacent vector elements, respectively. Dependence between vector elements is specified by (30), with $s_1 = 1.6$ and $s_2 = -0.8$. Results for $\phi_1 = 0.8$ are omitted from (a) and (b) for clarity.

0) are also qualitatively similar, but show noticeably greater power.

The power functions shown in Fig. 13 are also representative of those for other values of the vector dimension M, except for the performance of tests based on the counting norm D_4. The ratio m_Δ/M required for tests based on the counting norm to exhibit good power must be larger for smaller M and can be smaller for larger M. As a rule of thumb, for the tests described in this paper at least, those based on the counting norm exhibited good power when the ratio m_Δ/M was larger than approximately $2(M)^{-1/2}$. Intuitively, it is reasonable that tests based on the D_4 would have low power for small m_Δ since this norm does not take account of how strongly each local scalar H_0 is rejected.

Finally, it is remarkable that the power of these multivariate tests based on the max norm D_3 is so great for the difficult cases where the "signal" is confined to only a very small number m_Δ of elements embedded in M-dimensional "noise." Comparison of Fig. 4b to Fig. 13a shows that for $m_\Delta = 1$ the test suffers surprisingly little degradation in power as the dimension of the data vectors increases from $M = 1$ to $M = 64$ for $\phi = 0.5$ and the relatively small sample size of $n = 64$. For $M = 1$ in Fig. 4b, the test achieves 90% power at $\Delta\mu/\sigma_\epsilon = 1.1$, while for $M = 64$ in Fig. 13a, the scaled difference in the mean (of only one of the M vector elements) necessary to achieve 90% power increases to only 1.7. For the intermediate values $M = 4$ and 16, the corresponding $\Delta\mu/\sigma_\epsilon$ are 1.2 and 1.4, respectively. Larger sample sizes yield the same qualitative patterns, but with decreasing $\Delta\mu/\sigma_\epsilon$. By focusing on the scalar test indicating violation of H_0 most strongly, the max norm D_3 appears to filter much of the mere sampling variability in the remaining $M - m_\Delta$ vector elements. Conversely, however, when m_Δ is large this focus leads D_3 to ignore important information, leading to tests with lower power.

4. Summary and conclusions

This paper has developed bootstrap tests that are based on a nonparametric analog of Katz's (1982) test for selected simple time-dependence structures. The tests are robust to deviations from Gaussian data and perform well for relatively small sample sizes. The univariate bootstrap tests developed in section 2 are nearly as powerful as the corresponding likelihood ratio tests. Proper choice of the bootstrap block length is important for accurate test performance. The prescriptions offered here for the block length work well for hypothesis tests of the mean, but may not be universally applicable.

The primary motivation for this paper has been construction of nonparametric multivariate tests, by operating the scalar tests of section 2 in parallel and thus resampling directly the cross correlation in vector data. The results in section 3 show that this can be a successful strategy, although in most cases the actual test levels are smaller than the nominal significance levels when the effective sample size n_e is comparable in magnitude or smaller than the dimension M of the data vectors. The condition $n_e/M < 1$ occurs frequently enough in practice to be important and yet is not handled well by existing tests. The tests described here can be applied successfully in these instances as well, provided the analyst accounts for their moderate conservatism.

The univariate test presented in section 2b, which assumes that the data arise from a first-order autoregressive process, performs very badly indeed when this assumption is violated. It was found here that this shortcoming applies equally to the test recently developed by Zwiers and von Storch (1995). The test presented in section 2c, which is based on the assumption that the data follow a second-order autoregressive process, is much more robust and widely applicable. If it can be assumed in a univariate test setting that the data are AR(1), then the ZvS test will be preferable to the scalar bootstrap tests developed here, from the standpoints of test power and computational simplicity. More generally, however, the AR(2) bootstrap test developed in-section 2c should be preferred. In practice, one could investigate formally which of the three time series models used in sections 2b–d best fits the data series at hand (e.g., Katz and Skaggs 1981).

If a dataset of interest clearly exhibits behavior more elaborate that any of the three models used in section 2, the performance of these tests can be readily investigated through Monte Carlo simulations of the type used here, if an adequate time series model for the data at hand can be identified. Following such an analysis, judgments regarding the outcome of the tests can be made using the resulting estimate of the achieved significance level, rather than the nominal one. It would clearly also be possible to extend the procedures described above to construct tests oriented toward specific higher-order or more complex models for the time dependence in the data.

The power of the multivariate tests developed in section 3 depends very strongly on the vector norm chosen to summarize the individual scalar test statistics. When differences in vector means are confined to a small number of vector elements, the tests based on the counting norm exhibit extremely poor power. Limited ability to discern violations of the null hypothesis in such ‘‘needle in a haystack’’ situations is expected, as has been pointed out by Hasselmann (1979). However, it is found here that use of the max norm, or largest of the M scalar statistics, yields surprisingly powerful multivariate tests even when the signal is confined to a small number of the vector elements and that the test power degrades relatively little as M increases. Unless it is known or strongly suspected that violations of H_0 will occur for a large proportion of the elements of the data vectors, the max norm D_3 is probably the best choice among those presented in (28).

For the multivariate tests it is essential that the vector bootstrapping operate with the same block length for each element of a data vector, although the block lengths for the two vectors in two-sample problems can be different. It is not necessary, however, that the same scalar time series model be assumed for each element of a data vector. For purposes of computing the denominator of (11), some of the scalar series could be modeled as ARMA(1,1), some as AR(2), etc., so long as comparable block lengths would result.

Although attention has been confined here to tests for differences in means, there is no reason why analogous tests addressing other aspects of the data could not be constructed as well. For example, tests involving measures of the variance of data fields, or functions of the spatial correlation patterns, could be of interest. In these cases, the more refined bootstrap confidence intervals described in Efron (1987) might produce significant improvements in test performance.

How the tests should proceed if the elements of a data vector exhibit grossly different degrees of serial correlation is unclear, although for large sample sizes the underlying scalar tests perform well over a fairly wide range of choices for l. The consequences of misspecifying the block length for some of the data elements could be explored more fully. This potential problem is relevant because of the clear need to adapt the block length to the strength of the autocorrelation in the data. This requirement has also motivated the use of parametric time series models to compute V' in (12), even though a wholly nonparametric approach might be preferable. Future work might also address more general and robust alternatives to this aspect of the construction of the tests.

Acknowledgments. I thank Rick Katz for telling me of the existence of block bootstrapping and Francis Zwiers for helpful discussions. This research has been funded in part by the Northeast Regional Climate Center under National Oceanic and Atmospheric Administration Grant NA16CP-0220-02.

REFERENCES

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

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

Carlstein, E., 1986: The use of subseries values for estimating the variance of a general statistic from a stationary sequence. *Ann. Stat.,* **14,** 1171–1179.

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

Chung, J. H., and D. A. S. Fraser, 1958: Randomization tests for a multivariate two-sample problem. *J. Amer. Stat. Assoc.,* **53,** 729–735.

Cressie, N., 1980: Relaxing assumptions in the one sample *t* test. *Aust. J. Stat.,* **22,** 143–153.

Daley, R., and R. M. Chervin, 1985: Statistical significance testing in numerical weather prediction. *Mon. Wea. Rev.,* **113,** 814–826.

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

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

——, and R. J. Tibshirani 1993: *An Introduction to the Bootstrap.* Chapman and Hall, 436 pp.

Hall, P., and S. R. Wilson, 1991: Two guidelines for bootstrap hypothesis testing. *Biometrics,* **47,** 757–762.

Hasselmann, K., 1979: On the signal-to-noise problem in atmospheric response studies. *Meteorology over the Tropical Oceans,* D. R. Shaw, Ed., Roy. Meteor. Soc., 251–259.

Hillmer, S. C., and G. C. Tiao, 1979: Likelihood function of stationary multiple autoregressive moving average models. *J. Amer. Stat. Assoc.,* **74,** 652–660.

Johnson, R. A., and D. W. Wichern, 1982: *Applied Multivariate Statistical Analysis.* Prentice-Hall, 594 pp.

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

Kabaila, P., and G. Nelson, 1985: On confidence regions for the mean of a multivariate time series. *Commun. Stat. Theory Methods,* **14,** 735–753.

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

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

Künsch, H. R., 1989: The jackknife and the bootstrap for general stationary observations. *Ann. Stat.,* **17,** 1217–1241.

Leger, C., D. N. Politis, and J. P. Romano, 1992: Bootstrap technology and applications. *Technometrics,* **34,** 378–398.

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

Liu, R. Y., and K. Singh, 1992: Moving blocks bootstrap captures weak dependence. *Exploring the Limits of Bootstrap,* R. Lepage and L. Billard, Eds., John Wiley, 225–248.

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

——, 1995: Field intercomparison. *Analysis of Climate Variability,* H. von Storch and A. Navarra, Eds., Springer, 159–176.

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

Mielke, P. W., 1985: Geometric concerns pertaining to applications of statistical tests in the atmospheric sciences. *J. Atmos. Sci.,* **42,** 1209–1212.

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

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

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

von Storch, H., 1982: A remark on Chervin-Schneider's algorithm to test significance of climate experiments with GCMs. *J. Atmos. Sci.,* **39,** 187–189.

——, 1995: Misuses of statistical analysis in climate research. *Analysis of Climate Variability,* H. von Storch and A. Navarra, Eds., Springer, 11–26.

Westfall, P. H., and S. S. Young, 1993: *Resampling-Based Multiple Testing.* John Wiley, 340 pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 464 pp.

——, 1996: Statistical significance of long-range ‘‘optimal climate normal’’ temperature and precipitation forecasts. *J. Climate,* **9,** 827–839.

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

——, 1990: The effect of serial correlation on statistical inferences made with resampling procedures. *J. Climate,* **3,** 1452–1461.

——, 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, 1995: Taking serial correlation into account in tests of the mean. *J. Climate,* **8,** 336–351.