A Method to Estimate the Statistical Significance of a Correlation When the Data Are Serially Correlated

WESLEY EBISUZAKI

Research and Data Systems Corp., Greenbelt, Maryland

(Manuscript received 13 May 1996, in final form 14 February 1997)

ABSTRACT

When analyzing pairs of time series, one often needs to know whether a correlation is statistically significant. If the data are Gaussian distributed and not serially correlated, one can use the results of classical statistics to estimate the significance. While some techniques can handle non-Gaussian distributions, few methods are available for data with nonzero autocorrelation (i.e., serially correlated). In this paper, a nonparametric method is suggested to estimate the statistical significance of a computed correlation coefficient when serial correlation is a concern. This method compares favorably with conventional methods.

1. Introduction

The lag correlation of two time series is a common statistic used to determine whether two time series have similar variations. It is also a quantity often calculated when building statistical models. Since a lagged correlation can be written as a zero-lag correlation by a suitable transformation ($a_i' = a_{i+\tau}$), we will consider only the case of a zero-lag sample correlation, which is given by

$$\hat{\rho}_{AB} = \frac{1}{N}\sum_{i=0}^{N-1}(a_i - \bar{a})(b_i - \bar{b})/(\hat{\sigma}_A\hat{\sigma}_B),$$

where N is length of the time series, and for $x = a, b$, $\bar{x}$ is defined by

$$\bar{x} = \frac{1}{N}\sum_{i=0}^{N-1} x_i,$$

and $\hat{\sigma}_X$ is the positive root of

$$\hat{\sigma}_X^2 = \frac{1}{N}\sum_{i=0}^{N-1}(x_i - \bar{x})^2.$$

Estimating the statistical significance of a correlation is simple when the data are Gaussian distributed and have no serial correlation (i.e., the autocorrelation is zero for nonzero lags). However, serial correlation is present in most geophysical time series. In addition, the serial correlation is often enhanced by the common practice of applying a low-pass filter to the data. (A low-pass filter reduces the temporal variability, thus increasing the lag-1 autocorrelation.) If unaccounted for, serial correlation would make the statistical tests less stringent. That is, one could incorrectly reject the null hypothesis, $\rho_{AB} = 0$, with a probability greater than the nominal significance level. Consequently, one needs a test that is relatively insensitive to the serial correlation.

There are three common methods for dealing with time series that are serially correlated. One method is to subsample the data; that is, use only some of the data in order to reduce the temporal resolution. For example, the surface temperature shows a strong day-to-day correlation. However, if the temperature measurements were separated by one month, the autocorrelation would be much smaller and perhaps we could assume that the data are independent.

Subsampling the data has been used in meteorology; for example, one often assumes that data separated by one month are statistically independent. However, phenomena such as the Southern Oscillation or the quasi-biennial oscillation have much longer timescales. If we want to eliminate the autocorrelation due to the Southern Oscillation, then we may have to separate the samples by many years. Given the length of many meteorological time series, this approach is often impractical.

Another method to deal with serial correlation is to estimate an effective sample size. The effective sample sizes for means and correlations (e.g., Davis 1976; Trenberth 1984; Thiébaux and Zwiers 1984) are given by

$$N_{eff,\bar{a}} = N\left[\sum_{\tau=-N+1}^{N-1}\left(1 - \frac{|\tau|}{N}\right)\rho_{AA}(\tau)\right]^{-1}, \quad (1)$$

$$N_{eff,\rho_{AB}} = N\left[\sum_{\tau=-N+1}^{N-1}\left(1 - \frac{|\tau|}{N}\right)\rho_{AA}(\tau)\rho_{BB}(\tau)\right]^{-1}, \quad (2)$$

where the autocorrelation for $|\tau| < N$ can be estimated by

Corresponding author address: Dr. Wesley Ebisuzaki, Climate Prediction Center, NOAA/NCEP,W/NP52, Washington, DC 20233.

$$\hat{\rho}_{XX}(\tau) = \frac{1}{N - |\tau|} \sum_{i=0}^{N-1-|\tau|} (x_i - \bar{x})(x_{i+|\tau|} - \bar{x})/\hat{\sigma}_X^2. \quad (3)$$

Thiébaux and Zwiers (1984) estimated $N_{eff,\bar{a}}$ using Eqs. (1) and (3) for a large number of random series and found that effective sample size could not be estimated consistently. The variability in the sample autocorrelation [Eq. (3)] caused the standard deviation of the estimated effective sample size to be larger than the effective sample size. These conclusions were consistent with Trenberth (1984), who also found difficulties in reliably estimating the autocorrelation. In addition, Thiébaux and Zwiers (1984) point out that the effective sample size should not be used in Student's t-test to estimate the significance of the difference of two means. Consequently, they suggest that the effective sample size should be considered only as a diagnostic quantity.

Yet another method to deal with serial correlation is to model the data statistically and then to determine the distribution of the desired quantities from the statistical model. Katz (1982), for example, fit some climate-simulation data to an autoregressive model whose order was determined by a Bayesian information criterion. Katz then found the variance of the time means from the statistical model. Thus, he was able to determine if the time means had changed significantly in his simulations. This approach involves identifying the appropriate statistical model and estimating that model's parameters.

The previous methods dealt with significance tests for the hypothesis that the correlation is zero. A related hypothesis is that the two series are statistically independent. Obviously a nonzero correlation implies statistical dependence but not vice versa. Nevertheless, the introduction would be incomplete without some mention of the two commonly used techniques.

The first method is to "prewhiten" the time series before computing the lag correlations. Prewhitening involves filtering the time series to remove the serial correlation. For example, suppose x_i was generated by an AR(1) process; that is, $x_i = \beta x_{i-1} + \epsilon_i$, where ϵ_i is a white noise forcing. Then applying a simple filter ($x_i' = x_i - \beta x_{i-1}$) gives a new time series, x_i', with a zero-expected serial correlation. Without the serial correlation, one can then make statements about the correlation between the two prewhitened time series. For example, if one finds that the computed correlation between the prewhitened series is statistically significant, one can then reject the hypothesis that the original series are statistically independent.

Another method is to examine the significance of the squared coherence from a cross-spectral (bivariate spectral) analysis. The procedure gives the relationship between the two time series as a function of the frequency. However, sample size can be an issue as one would probably not attempt a cross-spectral analysis with 8, 16, or 32 samples. Even when N is large, such as in time series of 20 yr of global atmospheric analyses, one would expect fewer than 10 cycles of the quasi-biennial oscillation and even fewer warm El Niño events during that period. While N may be large, the phenomenon of interest may be undersampled. Both of these techniques are discussed in textbooks (e.g., Brockwell and Davis 1991).

2. "Random-phase" test

Consider two time series, A and B, which were chosen from population PA and population PB, respectively. The significance of the computed correlation between A and B, $\hat{\rho}_{AB}$, could be determined from the distribution of computed correlations between randomly chosen members of PA and PB. Suppose only $\alpha \times 100\%$ of these randomly generated correlations had a magnitude greater than ρ_{crit}, then one could reject the null hypothesis at the α significance level whenever $|\hat{\rho}_{AB}| > \rho_{\text{crit}}$. However, this test is usually impractical as PA and PB are often unknown.

When the distributions of PA and PB are unknown, resampling methods attempt to generate random time series that have similar properties to the members of PA and PB. For example, the "bootstrap" can be used to evaluate the significance of $\hat{\rho}_{AB}$ by first generating a large number of random series (R^k), which are formed by taking elements of A at random; that is, $r_i^k = a_j$, where $j(i, k)$ is random integer between 0 and N $-$ 1, and k ranges between 1 and number of resampled time series. The correlations between R^k and B can be used to determine ρ_{crit} as described in the previous paragraph.

The idea behind resampling techniques is to generate random series that have properties similar to the original series, and to derive a confidence interval using these random series. Resampling depends on preserving the important properties of the original series. Zwiers (1990) tested the procedure described in the previous paragraph and found that it worked well when the sampling assumptions were applied ["(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."] However, when the series were serially correlated, the observations were no longer independent of each other, which violated a sampling assumption. The serial correlation decreased the effective sample size [Eq. (2)] and consequently increased ρ_{crit}. The bootstrap test, however, was not influenced by the serial correlation, and Zwiers found the test became more biased with increasing serial correlation. This illustrates a problem with the bootstrap—the generated random series are not serially correlated unlike many time series.

This paper considers another technique, which could be considered "resampling" in the frequency domain. This procedure will not preserve the distribution of values but rather the power spectrum (periodogram). The advantage of preserving the power spectrum is that resampled series retains the same autocorrelation as the

original series. The steps involved in this "random-phase" test are as follows.

1) Compute the discrete Fourier transform of the first series, a_i, by,

$$\tilde{a}_k = \frac{2 - \delta_k}{N} \sum_{j=0}^{N-1} a_j e^{2\pi ijk/N},$$

where $\delta_k = 0$ except for $k = 0$, $N/2$ (even N) in which case $\delta_k = 1$.

2) Let $\tilde{r}_0 = 0$, $\tilde{r}_k = |\tilde{a}_k|\exp(i\theta_k)$ for $0 < k < N/2$, and $\tilde{r}_{N/2} = 2^{1/2}|\tilde{a}_{N/2}|\cos(\theta_{N/2})$ for even N, where θ_k is a uniform random variable from $[0, 2\pi)$.

3) Compute the inverse Fourier transform of $\tilde{r}_k$ by

$$r_j = \mathrm{Re} \sum_{k=0}^{n} \tilde{r}_k e^{-2\pi ijk/N},$$

where $n = N/2$ or $\frac{N-1}{2}$ for even and odd N, respectively.

Step 2 creates a Fourier series with random phases and the same power spectrum as the original series. The two special cases are $k = 0$ and $k = N/2$. Setting $\tilde{r}_0 = 0$ is simply setting the mean of r_i to zero, which is inconsequential for a correlation. For $k = N/2$, we are dealing with the Nyquist frequency. At this frequency, one can sample only Re $(\tilde{a}_{N/2})$. If we assume $\mathrm{Im}(\tilde{a}_{N/2})^2$ has the same expected value as $\mathrm{Re}(\tilde{a}_{N/2})^2$, then the expected value of $|\tilde{a}_{N/2}|^2$ is $2\mathrm{Re}(\tilde{a}_{N/2})^2$, and the real part of the random-phase spectral component is $\tilde{r}_{N/2} = 2^{1/2}|\tilde{a}_{N/2}|\cos(\theta_{N/2})$.

In step (3), the inverse Fourier transform gives r_i, a random series with the same autocorrelation as the original series. The main difference is that r_i has random phases in each of its Fourier modes.

It should be noted that this procedure is very similar to the process of creating "surrogate" data (Fraser 1989; Kaplan 1993; Theiler et al. 1993; Gershenfeld and Weigend 1993), which has been used to test for nonlinearities in time series. However, these studies used $\tilde{r}_{N/2} = |\tilde{a}_{N/2}|\cos(\theta_{N/2})$ rather than $2^{1/2}|\tilde{a}_{N/2}|\cos(\theta_{N/2})$.

Once we have created a large number of random-phase series, we can calculate the significance in a manner similar to the bootstrap method. We take these random series and correlate them with the second series. (This step can be computed in the spectral domain for computational efficiency.) For example, if fewer than $\alpha \times 100\%$ of the correlations from the random-phase series have a magnitude greater than ρ_{crit}, then one can reject the null hypothesis, $\rho_{AR} = 0$, at the α significance level whenever $|\hat{\rho}_{AB}| > \rho_{\mathrm{crit}}$.

The random-phase test is a nonparametric test, which attempts to create random time series that could have come from PA, the population from which A is a member. The random-phase test is inappropriate for data with strong seasonality. In such a situation, the "random" time series are obviously not from PA as the phases of "plausible" time series are not arbitrary.

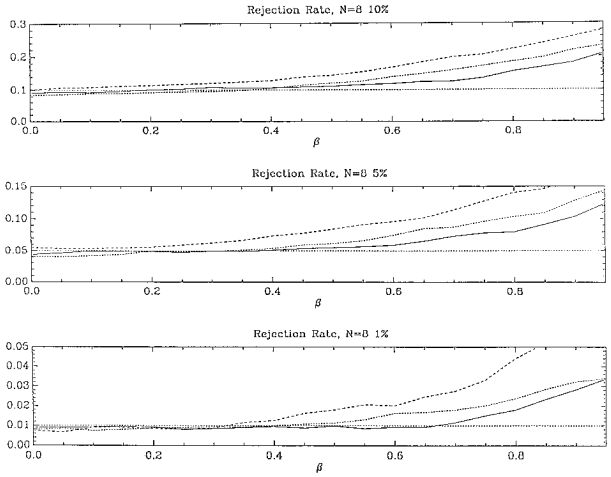

FIG. 1. The rejection rate (fraction of times that the $\rho_{AB} = 0$ hypothesis was rejected) for pairs of independent AR(1) series at the 0.1 (Fig. 1a), 0.05 (Fig. 1b), and 0.01 (Fig. 1c) significance levels. Shown are the rejection rates for the "bootstrap" (dashed line), conventional (dotted line), and "random-phase" (solid line) tests. Here, $N = 8$. Ideally the rejection rate should equal the significance level.

3. Tests using AR(1) test data

The performance of the random-phase test needs to be evaluated. At a minimum, one would like an unbiased test; that is, the null hypothesis is rejected at a rate equal to the nominal significance level whenever the null hypothesis is true. To evaluate the bias of the test, random time series were created using AR(1) and AR(2) models (first- and second-order autoregressive models).

a. Random-phase test

To evaluate the performance of this test, we generated 2000 pairs of time series using two independent AR(1) models given by

$$x_{i+1} = \beta x_i + \epsilon_i,$$

where β is the specified lag-1 autocorrelation, and ϵ_i is a normally distributed random variable with a zero mean and a variance of one. Since the two models are independent, the correlation between the time series has an expected value of zero.

We then subjected each pair of time series to a random-phase test and found the fraction of times the null hypothesis was rejected, that is, the rejection rate. Ideally, for the α significance level, we would reject only the null hypothesis α 100% of the time. For each random-phase test, 2000 random-phase time series were used to generate ρ_{crit}.

Figures 1–4 show the rejection rates for the correlation of independent AR(1) series at the 0.1, 0.05, and 0.01 significance levels using the random-phase test. The rejection rates are close to expectation and are much better than those for the bootstrap method. It should be noted that the curves have some "sampling" error, and

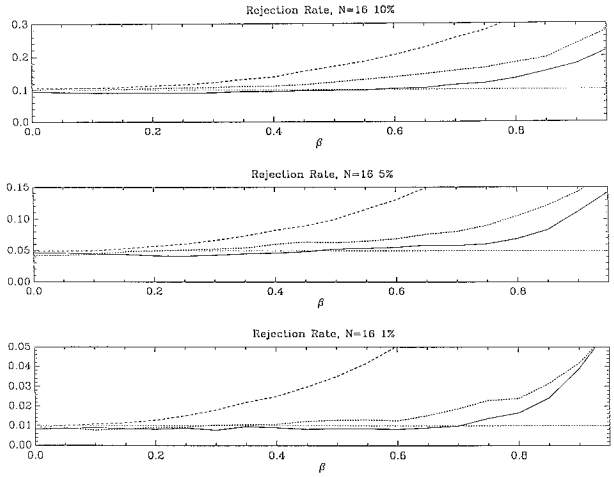

FIG. 2. Similar to Fig. 1 except $N = 16$.

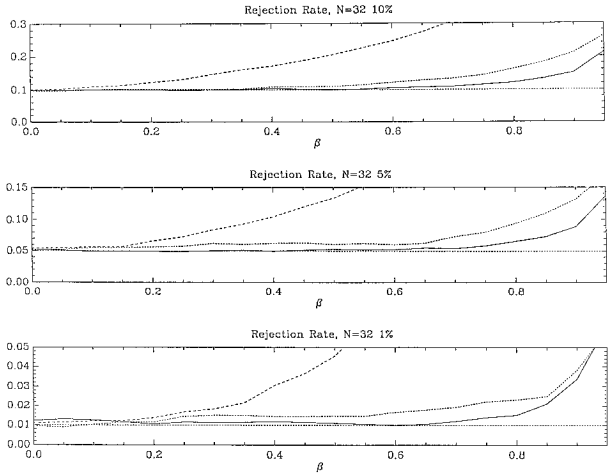

FIG. 3. Similar to Fig. 1 except $N = 32$.

the errors for different values of β are correlated because the random number generator used the same initial seed.

In Figs. 1–4, the random-phase test is too liberal as β approaches one. This is caused by poorly resolving the low frequencies (periods greater than or on the order of the length of the time series). As β approaches one, the AR(1) series have more power in these poorly or unresolved low frequencies. These unresolved low frequencies often appear as trends in the time series, and trends tend to (anti-) correlate with each other. The random-phase series, however, are periodic by construction and do not have trends except for very specific phases. Consequently, the magnitude of the correlation between randomly selected AR(1) series would tend to be larger than for the random-phase series. Thus, testing against random-phase series will produce a test with a too large rejection rate. This hypothesis was supported by tests using sample data that had its linear tread removed. As β approached one, the test did not show as large a liberal bias (not shown).

b. Conventional test

The conventional test that we are going to use is similar to the resampling techniques except we are going to use a statistical model to generate the random series. The statistical model should, in principle, be determined by fitting the data to a statistical model. Often data are modeled as from an AR(n) process where 'n' may be determined by "Akaike information criterion," the "final prediction error," etc. (Brockwell and Davis 1991). Since we are using data that came from an AR(1) model, we did not need to determine the order of the test data. This experiment was designed to be advantageous for the conventional test as the test data and statistical model are both AR(1). A more demanding situation would be the situation in which the data could not be statistically modeled by a low-order AR(n) process. Here, β was estimated by

$$\hat{\beta} = \frac{1}{\hat{\sigma}_x^2(N-1)} \sum_{i=0}^{N-2} (x_i - \bar{x})(x_{i+1} - \bar{x}). \qquad (4)$$

As seen in Fig. 1–4, the rejection rates for the random-phase and conventional tests are close to the nominal rates for large N and small to moderate β. While both tests suffer from a liberal bias for β approaching one, the bias was worse for the conventional test. The random-phase test appears to be relatively unbiased for a greater parameter range, especially for small N.

Both the random-phase and conventional tests had difficulties when N was small or β approached one. The problems with the conventional test were caused by underestimating β (Eq. 4), which can be approximated as

$$\hat{\beta} \approx \frac{\frac{1}{N-1}\sum_{i=0}^{N-2} x_i x_{i+1} - \bar{x}^2}{\overline{x^2} - \bar{x}^2},$$

where the overbar denotes the mean value. Since x_{i+1}

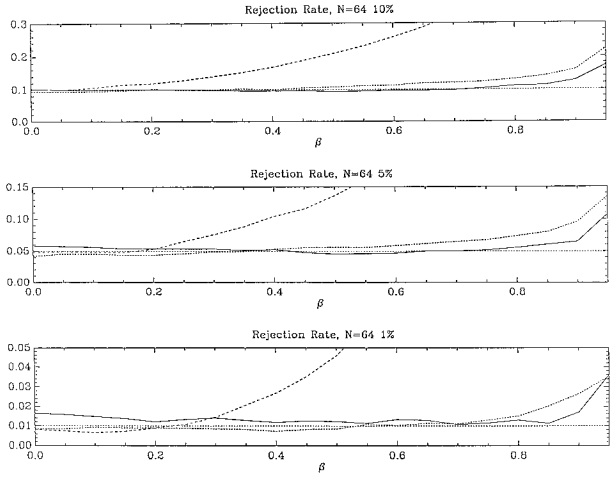

FIG. 4. Similar to Fig. 1 except $N = 64$.

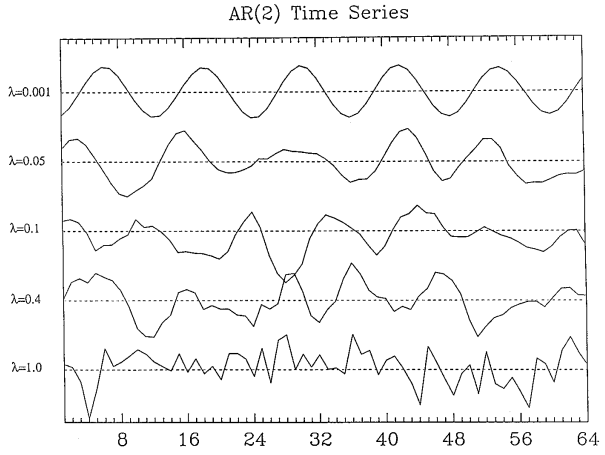

FIG. 5. Typical solutions of Eq. (5) using $\omega = \pi/6$ and λ equal to 0.001, 0.05, 0.1, 0.4, and 1.0 (top to bottom). The solutions have been normalized and offset in the vertical. The dashed lines are the zero line for the corresponding normalized time series.

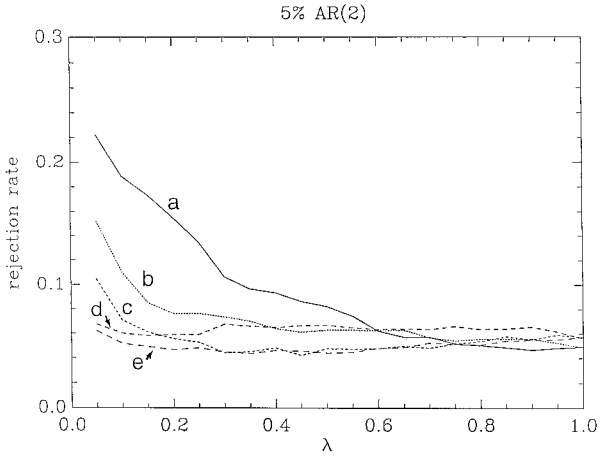

FIG. 6. The rejection rate for pairs of independent AR(2) series at the 0.05 significance level using random-phase test with (a) $N = 8$, (b) $N = 16$, (c) $N = 32$, (d) $N = 64$. Curve "e" is similar to "c" except for using AR(1) and AR(2) series: $\omega = \pi/6$ and $\beta = 0.5$.

$= \beta x_i + \epsilon_i$, the rhs of the previous equation can be approximated as

$$\hat{\beta} \approx \frac{\beta\overline{x^2} - \bar{x}^2}{\overline{x^2} - \bar{x}^2} \leq \beta.$$

The previous equation indicates that β is underestimated when the sample mean ($\bar{x}$) differs from the true mean (0). This difference of between the sample and true means is expected to increase for smaller N or for β closer to one that is consistent with the cases that the conventional test did poorly.

The underestimation of β caused the random series to have a smaller autocorrelation than the original data. Consequently, the test rejected the null hypothesis, $\rho_{AB} = 0$, too frequently. Both the random-phase and conventional tests became more liberal as β approached one; however, the random-phase test did perform better in this situation.

4. Tests using AR(2) test data

The AR(1) process is often used in statistical modeling. However, it cannot simulate many geophysical processes that are quasi-oscillatory (e.g., Thiébaux 1987). For this type of time series, we considered the equation for a damped oscillator forced by white noise (ϵ):

$$\frac{d^2}{dt^2}\psi + 2\lambda\frac{d}{dt}\psi + (\omega^2 + \lambda^2)\psi = \epsilon.$$

When $\epsilon = 0$, the solution is $\psi(t) = a\exp((i\omega - \lambda)t) + b\exp((-i\omega - \lambda)t)$, where a and b are constants determined from the initial conditions. If we make a finite-difference approximation by replacing $d^2\psi/dt^2$ by $(\psi_{i+2} - 2\psi_i + \psi_{i-1})/\Delta t^2$, $d\psi/dt$ by $(\psi_{i+1} - \psi_{i-1})/2\Delta t$, and letting $\Delta t = 1$, we get

$$\psi_i = \frac{2 - \omega^2 - \lambda^2}{1 + \lambda}\psi_{i-1} - \frac{1 - \lambda}{1 + \lambda}\psi_{i-2} + \epsilon_i'. \quad (5)$$

In Eq. (5), ω is the change in phase angle, and λ is the decay factor between adjacent data points. For our tests, ϵ_i' is a random variable that has a normal distribution, zero mean, and a variance of one. We will refer to Eq. (5) as the AR(2) model even though it is only one specific example.

In Fig. 5, we show some typical AR(2) time series with $\omega = \pi/6$ (30° phase shift) and various values of λ. For presentation, the values have been normalized and offset in the vertical. For small values of λ, the time series are almost sinusoidal. For moderate values, the time series has a preferred timescale between adjacent extrema, and for large values, the series has no apparent preferred timescale.

To calculate the rejection rate, 1000 pairs of time series were generated by two independent AR(2) models and the correlation was calculated for each pair. Each of these 1000 correlations was subject to a random-phase test that used 1000 random-phase time series. The critical correlation was computed and the fraction of times the null hypothesis, $\rho_{AB} = 0$, was rejected is shown in Figs. 6 and 7. These two figures show the rejection rate as a function of λ and N (curves a–d) for $\omega = \pi/6$. The rejection rate is close to the nominal significance except when λ, the damping factor, is small. When λ is too small, the rejection rate is too large, and this effect is more pronounced when N is also small ($\lambda < 0.1$ for $N = 64$, and $\lambda < 0.4$ for $N = 16$). This large rejection rate may be caused by spectral leakage. When λ is near zero, the sample time series is nearly sinusoidal and the true spectrum is a single narrow peak. The discrete Fourier transform, however, broadens the spectral peak (spectral leakage), which is more of a problem when N is small and the time series does not contain an integral

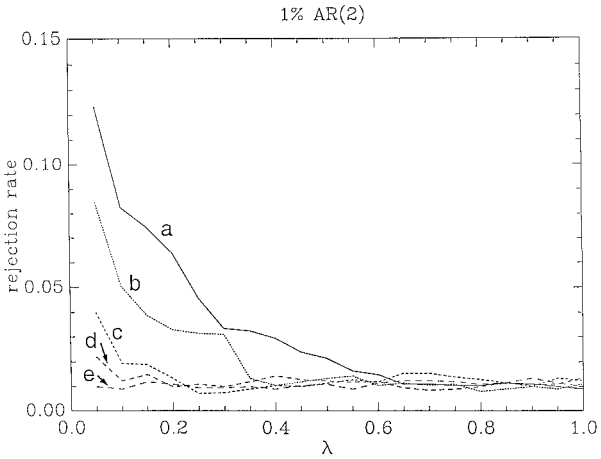

FIG. 7. Similar to Fig. 6 except for the 0.01 significance level.

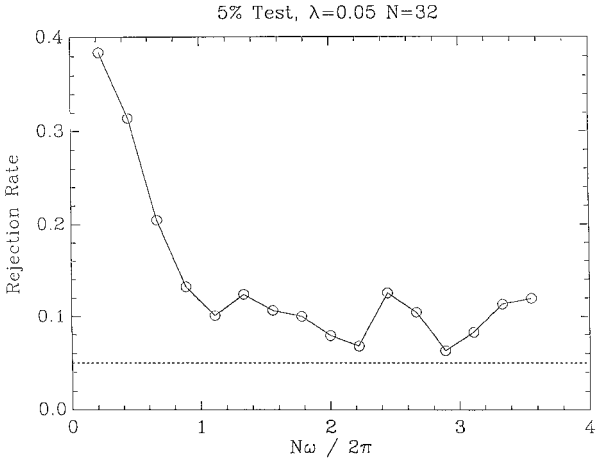

FIG. 8. The rejection rate for pairs of independent AR(2) series at the 0.05 significance level: $\lambda = 0.05$, $N = 32$, and the abscissa is $N\omega/2\pi$.

number of periods (i.e., $N\omega/2\pi$ is not an integer). This hypothesis is supported by Fig. 8, which shows the rejection rate is close to the nominal value when $N\omega/2\pi$ is an integer. When $N\omega/2\pi < 1$, the rejection rate was too high, which is the result of the period of the sinusoid being longer than the time series. As discussed in the AR(1) tests, unresolved low frequencies can increase the rejection rate.

In Figs. 6 and 7, we also show the rejection rate for the correlation of an independent AR(1) ($\beta = 0.5$) and AR(2) series (curve e). In both cases, the rejection rate is close to the nominal value. The test is unaffected by the spectral leakage in estimating the AR(2) power spectrum because we randomized the phase of the AR(1) series. It should be mentioned that this test is symmetric, the test for a_i and b_i is the same as the test for b_i and a_i except for the variability introduced by different random phases. This result is apparent as the correlation between b_i and the random-phase series, r_i, can be written as

$$\hat{\rho}_{RB} = \frac{1}{2\hat{\sigma}_R\hat{\sigma}_B}\sum_{k=1}^{n}(1 + \delta_k)^{3/2}|\tilde{r}_k||\tilde{b}_k|\cos(\epsilon_k'),$$

where ϵ_k' is a uniform random variable from $[0, 2\pi)$. From the above equation, we can infer that the test is sensitive to spectral leakage only if it affects both time series.

5. Summary

In summary, we suggested a nonparametric test for the significance of a nonzero correlation between two time series when serial correlation is a concern. The statistical test is based on generating a large number of random series with the same power spectra (periodogram) as the first series but with random phases in the Fourier modes. The statistical significance of the original correlation was found by comparing this correlation to the distribution of correlations between the random-phase series and the second original series.

The performance of this test was evaluated by generating independent time series and comparing the rejection rate, the fraction of the time that the $\rho_{AB} = 0$ hypothesis was rejected, to the nominal significance level. With time series created by an AR(1) model, the rejection rates were close to the expected values and was relatively insensitive to the length of the time series. The rejection rate was too large when trends were dominant, and significant power was in the unresolved low frequencies. These results compared favorably to a conventional test where the data was statistically modeled as from an AR(1) process.

In the AR(2) tests (stochastically forced, damped oscillator), the rejection rates for the random-phase test were close to the expected rates except when λ, the damping, became small. The effect of small λ was more pronounced on the smaller N cases. We attributed this problem to spectral leakage, which made the periodogram a poor estimate of the spectral power.

The advantage of a random-phase test is its simplicity. It is similar to the bootstrap in that we create random samples that have similar properties to the original data. Instead of preserving the distribution of the original data, we preserve the autocorrelation of the data. The remainder of the test is similar to the bootstrap. Compared with a conventional test using AR(1) data, the random-phase test showed less bias in the rejection rates and avoided the task of identifying the appropriate statistical model for the test.

Acknowledgments. This work was inspired by a pair of lectures given by Dr. R. Livezey on statistical evaluation of forecasts and general circulation models. Discussions with Dr. F. Zwiers and comments from two anonymous reviewers were also very helpful.

REFERENCES

Brockwell, P. J., and R. A. Davis, 1991: *Time Series: Theory and Methods*. Springer-Verlag, 577 pp.

Davis, R. E., 1976: Predictability of sea surface temperatures and sea level pressure anomalies over the North Pacific Ocean. *J. Phys. Oceanogr.,* **6,** 249–266.

Fraser, A. M., 1989: Reconstructing attractors from scalar time series: A comparison of singular system and redundancy criteria. *Physica D,* **34,** 391–404.

Gershenfeld, N. A., and A. S. Weigend, 1993: The future of time series: Learning and understanding. *Time Series Prediction: Forecasting the Future and Understanding the Past,* A. S. Weigend and N. A. Gershenfeld, Eds., Addison-Wesley, 1–70.

Kaplan, D. T., 1993: A geometrical statistic for detecting deterministic dynamics (data sets A, B, C, D). *Time Series Prediction: Forecasting the Future and Understanding the Past,* A. S. Weigend and N. A. Gershenfeld, Eds., Addison-Wesley, 415–428.

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

Theiler, J., P. S. Linsay, and D. M. Rubin, 1993: Detecting nonlinearity in data with long coherence times (data set E). *Time Series Prediction: Forecasting the Future and Understanding the Past,* A. S. Weigend and N. A. Gershenfeld, Eds., Addison-Wesley, 429–455.

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

——, and M. A. Pedder, 1987: *Spatial Objective Analysis: With Applications in Atmospheric Science*. Academic Press, 299 pp.

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

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