# Taking Serial Correlation into Account in Tests of the Mean

FRANCIS W. ZWIERS

*Canadian Centre for Climate Modelling and Analysis, Victoria, British Columbia, Canada*

HANS VON STORCH

*Max Planck Institute for Meteorology, Hamburg, Germany*

(Manuscript received 23 December 1993, in final form 10 June 1994)

ABSTRACT

The comparison of means derived from samples of noisy data is a standard part of climatology. When the data are not serially correlated the appropriate statistical tool for this task is usually the conventional Student's t-test. However, frequently data are serially correlated in climatological applications with the result that the $t$ test in its standard form is not applicable. The usual solution to this problem is to scale the $t$ statistic by a factor that depends upon the equivalent sample size $n_e$.

It is shown, by means of simulations, that the revised $t$ test is often conservative (the actual significance level is smaller than the specified significance level) when the equivalent sample size is known. However, in most practical cases the equivalent sample size is not known. Then the test becomes liberal (the actual significance level is greater than the specified significance level). This systematic error becomes small when the true equivalent sample size is large (greater than approximately 30).

The difficulties inherent in difference of means tests when there is serial dependence are reexamined. Guidelines for the application of the "usual" $t$ test are provided and two alternative tests are proposed that substantially improve upon the "usual" $t$ test when samples are small.

## 1. Introduction

Statistical comparisons of means are frequently conducted in climatology to intercompare observed and/or simulated climates among themselves or against fixed reference values. These comparisons are conducted by employing a paradigm in which (i) a statistical model is imposed upon the samples of climate data, (ii) a *null hypothesis* $H_0$ that is to be tested is specified, (iii) an *alternate hypothesis* $H_a$ that guides the interpretation of the test statistic is specified, and (iv) a test statistic is computed to determine how unusual the observed difference of means is in the context of the model and the null hypothesis.

It is well known that the classical method, which employs the Student's t-test (see, e.g., Mood and Graybill 1963) and assumes a statistical model in which climate observations are statistically independent and Gaussian, is sensitive to serial correlation within the samples. The effect of serial correlation is, usually, to make comparisons of means liberal. That is, "significant" differences are found more frequently than expected when there is no difference. This occurs because the $t$ test scales the difference of time averages with an estimate of the standard error of the difference, that relies explicitly upon the independence assumption. When the data are serially correlated the estimate of the standard error underestimates the sampling variability of the difference of time averages.

The purpose of this paper is to review existing methods for dealing with serial correlation, propose a new test for the difference of means that better takes the effects of serial correlation into account, and provide guidelines for application of comparison of means procedures.

We will operate by replacing the independence assumption referred to above with a red noise assumption. That is, it will be assumed that the evolution of an observed climate process is governed, approximately, by a difference equation of the form[1]

$$X_t - \mu_X = \rho_1(X_{t-1} - \mu_X) + \epsilon_t, \qquad (1)$$

where $\{\epsilon_t\}$ is a Gaussian white noise process, $\rho_1$ is the lag-1 correlation coefficient, and $\mu_X$ is the long-term mean. This simple difference equation model approximates observed climate behavior in many instances when $0 \leq \rho_1 < 1$. In this case the power spectrum, an

*Corresponding author address:* Dr. Francis W. Zwiers, Canadian Centre for Climate Modelling and Analysis, Atmospheric Environment Service, University of Victoria, P.O. Box 1700, MS 3339 Victoria, BC V8W 2Y2 Canada.

[1] Stochastic processes that satisfy (1) are frequently referred to as being autoregressive of order 1 [AR(1)].

example of which is illustrated in Fig. 1, has maximum energy at zero frequency and decreases smoothly with increasing frequency. This characteristic roughly approximates the behavior of many thermodynamic variables in the free atmosphere, even if accurate representation of their stochastic behavior requires the use of higher-order models. The AR(1) model also arises naturally when one employs stochastic climate models (e.g., Hasselmann 1976) to explain how the climate system can exhibit low-frequency variability without resorting to either normal modes, which operate at low frequencies; complex nonlinear feedbacks; or external forcing. We therefore feel that the difficulties with the standard $t$ test can be usefully alleviated in many situations by approximating the stochastic structure of climate observations with an AR(1) model.

While the Gaussian and red noise assumptions are reasonable in many instances, they should not be made blindly. Departures from both assumptions may cause difficulties with test reliability that are as serious as those that we are striving to correct. This will especially be the case when the observed processes have spectral peaks at greater than zero frequency. For example, any variable that exhibits an ENSO, QBO, or MJO signal is suspect. In general, it is imperative that the assumptions that are necessary for the application of a statistical procedure are checked. Relatively simple descriptive methods, such as plotting observations as a function of time; plotting estimates of the power spectrum, autocorrelation, and partial autocorrelation functions; and plotting frequency histograms of the data should be adequate to detect gross departures from the Gaussian and red noise assumptions.

The remainder of this paper is organized with two groups of readers in mind. Those interested in the practical aspects of tests of the mean when the data are approximately AR(1) and Gaussian will find our recommendations for one- and two-sample tests of the mean in section 2. Readers interested in a detailed description of the testing problem should read the remainder of this section, skip section 2, and return to it after reading the rest of the paper.

### a. *A pedagogical example*

A parochial and naive but nonetheless instructive question is whether long-term mean winter temperatures at two locations such as Hamburg and Victoria are equal. To attempt to answer this question, suppose that we have at our disposal the daily observations at both locations for the winter of 1992/93. We treat the winter temperatures at both locations as random variables, say $T_H$ and $T_V$. The "long-term mean" winter temperatures at the two locations, denoted as $\mu_H$ and $\mu_V$, respectively, are parameters of the probability distributions of these random variables.

The statistical question we pose is: Do the two samples of temperature observations contain sufficient evidence to reject the null hypothesis

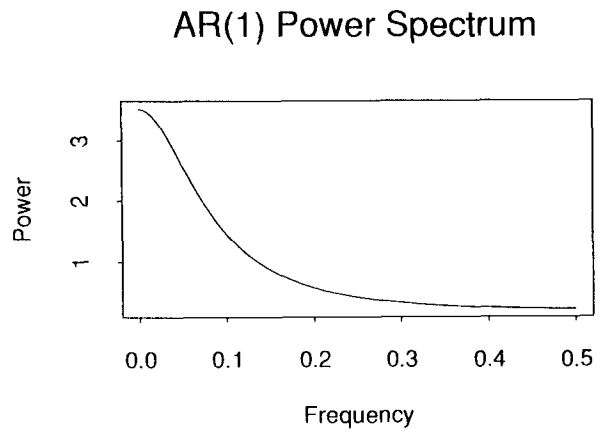


FIG. 1. The power spectrum of an AR(1) process with lag-1 correlation coefficient $\rho_1 = 0.75$ and innovation variance $\sigma_\epsilon^2 = 1$.

$$H_0 : \mu_H - \mu_V = 0? \tag{2}$$

In this example, and in many applications in climate research, the assumption that the observations are statistically independent is not satisfied. Consequently, the Student's t-test tends to reject that null hypothesis on weaker evidence than is implied by the significance level[2] that is specified for the test. Consequently, the Student's t-test will reject the null hypothesis more frequently than expected when the null hypothesis is true.

### b. *Subsampling the data*

A relatively clean and simple-minded solution to this problem is to form subsamples of independent observations. In the case of daily temperature data, one might argue that observations that are separated by, say, 5 days, are effectively independent of each other.[3] Let the number of observations, the sample means, and standard deviations of these reduced datasets be denoted by $n^*$, $\bar{T}_H^*$, $\bar{T}_V^*$, $S_H^*$, and $S_V^*$, respectively. Then the usual $t$ statistic

$$t = \frac{\bar{T}_H^* - \bar{T}_V^*}{((S_H^{*2} + S_V^{*2})/n^*]^{1/2}} \tag{3}$$

has a Student's $t$ distribution with $2n^* - 2$ degrees of freedom provided that the null hypothesis is true.[4] A test can be conducted at the specified significance level by comparing the value of (3) with the appropriate percentiles of this distribution.

---

[2] The significance level indicates the probability with which the null hypothesis will be rejected when it is true.

[3] The choice of the sampling interval is a nontrivial problem. It is determined either from physical considerations or by examination of the autocovariance function. See section 1c for a related discussion of the concept of effectively independent observations.

[4] Strictly speaking, this is true only if the standard deviations of $T_H$ and $T_V$ are equal.

Thus, appropriately subsampling the data will result in a test that operates as specified by the user. Unfortunately, this is achieved by throwing away much of the data and, presumably, at least part of the information contained in the data.

### c. *The equivalent sample size*

The main reason for the liberal behavior of the $t$ test when the data are not subsampled is that the denominator of (3) underestimates the sampling variance of the numerator [see, e.g., Laurmann and Gates 1974; Chervin and Schneider 1976; Jones 1975; Thiebaux and Zwiers 1984 (hereafter TZ)]. To correct this problem, the sum of sample variances $S_H^2 + S_V^2$ must be scaled by the *equivalent sample size* $n_e$. The equivalent sample size is given by

$$n_e = n \Big/ \left[ 1 + 2 \sum_{\tau=1}^{n-1} \left( 1 - \frac{\tau}{n} \right) \rho(\tau) \right], \tag{4}$$

where $\rho(\tau)$ is the correlation between temperature at time $t$ and temperature at time $t + \tau$. When the observed processes are AR(1) with lag-1 correlation coefficient $\rho_1$, $\rho(\tau) = \rho_1^{|\tau|}$. Thus

$$n_e = n \Big/ \left[ 1 + 2 \sum_{\tau=1}^{n-1} \left( 1 - \frac{\tau}{n} \right) \rho_1^{\tau} \right], \tag{5}$$

which may be approximated by

$$n_e \approx n \frac{(1 - \tau_1)}{(1 + \rho_1)} \tag{6}$$

when $n$ is large.

As an (important) aside, we reiterate the point made in TZ that the equivalent sample size is not uniquely defined. The equivalent sample size arises because we interpret $t$ as the distance between the sample means expressed in units of standard deviations of sample means. Many statistics have this basic form:

$$T = D/S_D, \tag{7}$$

where $D$ is some characteristic of the difference between two samples and $S_D$ is an estimate of the standard error of $D$. In many cases, the standard error estimator that is appropriate when observations are independent needs to be scaled by some function of the sample size $n$ when observations are serially correlated. The resulting expression for the equivalent sample size $n_e$ or *integral time scale* $n/n_e$ depends upon the definition of $D$.

Another way to think about the effective sample size is that it is a diagnostic that tells us something about the information loss due to serial correlation in a sample of size $n$. When $n_e$ is defined as in (5), the interpretation is that samples of independent observations of size $n_e$ contain as much information about the *difference of means* as samples of serially correlated observations of size $n$. In that context, we can think of $n_e$ as the *number of effectively independent observations.* However, when our interest is in some other feature of the difference between two samples, the definition of information changes. Consequently, the definition of the equivalent sample size, or number of effectively independent observations, also changes. It is therefore impossible to interpret $n_e$ as the number of effectively independent observations in an absolute sense.

### d. *Adjusting the t statistic*

When samples are sufficiently large, the adjusted $t$ statistic,

$$t = \frac{\bar{T}_H - \bar{T}_V}{[(S_H^2 + S_V^2)/n_e]^{1/2}} \tag{8}$$

has a standard Gaussian, or normal, distribution $N(0, 1)$ with mean 0 and standard deviation 1 (Albers 1978) under the null hypothesis. Thus one can conduct a test by comparing (8) to the percentiles of the standard Gaussian distribution. When samples are small it is often assumed that (8) will behave as Student's $t$ with $n_e - 1$ degrees of freedom under the null hypothesis. While this assumption, which appears to have a heuristic basis, is asymptotically correct (Albers 1978, see Lemma 2.1), it is not correct for small samples (see Katz 1982; TZ).

The imprecision of the assumption that (8) is distributed Student's $t$ is demonstrated with the following example. Suppose samples are obtained from a specified zero mean Gaussian AR(1) process. The exact equivalent sample size $n_e$ is known because the AR(1) parameters are known. Samples of length $n$ are randomly generated for various choices of $n$ and AR(1) parameters and each sample is used to test the null hypothesis that $\mu_X = 0$ with $t$ statistic (8) at the 5% significance level. We find that the actual rejection rate (Fig. 2) is notably smaller than the expected rate of 5% for $n_e \leqslant 30$. This finding is further supported by theoretical arguments in the appendix.

Additional difficulties arise when the equivalent sample size must be estimated from the data. In this case the actual rejection rate of the $t$ test tends to be greater than the nominal rate (see section 3). In some instances the actual significance level can be several times greater than the nominal significance level.

Although the effects of dependence on tests of the mean are well known in the statistical literature, it is relatively void of advice on how to counteract these effects. Albers (1978) considers the cost of making a large sample version of the $t$ test robust against some kinds of dependence. Cressie (1980) contains a survey of the effects of various departures from the assumptions that are implicit in the $t$ test. Tubbs (1980) describes the effects of serial correlation on multivariate versions of the $t$ test. Katz (1982) describes an asymptotic test. Kabaila and Nelson (1985) describe uni-

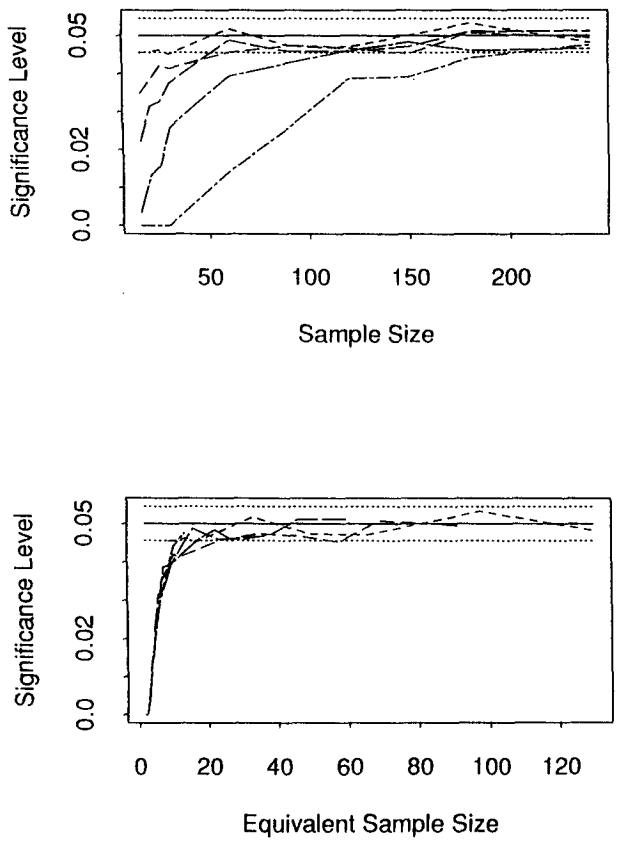


FIG. 2. Actual rejection rates of the $t$ test with $n_e$ known in a series of simulation experiments. The upper panel shows the rejection rate as a function of sample size, while the lower panel shows it as a function of equivalent sample size. Each panel contains five curves corresponding to data-generated first-order autoregressive processes $X_t$ with zero mean and lag-1 correlation coefficients of 0.3 (short dashes), 0.45 (medium dashes), 0.6 (long dashes), 0.75 (long broken dashes), and 0.9 (medium broken dashes), respectively. The null hypothesis $\mathcal{E}(X_t) = 0$ was tested at the 5% significance level in each of 1000 Monte Carlo trials conducted for each combination of sample size and lag-1 correlation coefficient. The solid horizontal line represents the nominal (5%) rejection rate and the dashed horizontal lines represent critical points at which the null hypothesis that the test operates at the nominal significance level is rejected.

variate and multivariate versions of a competing asymptotic test. Both approaches were examined in TZ. Sutrahdar et al. (1987) discuss one-way analysis of variance of experimental designs in which each "treatment" results in a sample taken from a time series. The comparison of the means of two relatively large samples is considered as a special case. None of these authors discuss the small sample case.

The remainder of this paper is organized as follows. Our recommendations for one- and two-sample tests of the mean for data from Gaussian red noise processes are contained in section 2. Readers interested in a detailed description of the testing problem should skip this section and return to it after reading the rest of the paper. The "usual" $t$ test, which is adjusted for serial correlation using the equivalent sample size, is discussed in section 3. The likelihood ratio (LR) test, a competing asymptotic test that is based on rigorous statistical principles, is described in section 4. An empirically developed table lookup test that has superior small sample properties is described in section 5. A summary is presented in section 6.

## 2. Recommended test procedures

The following procedures are appropriate for use when the observations are obtained from processes that are approximately Gaussian red noise processes. Unless otherwise noted, procedures for the two-sample case rely on the additional assumption that both samples come from red noise processes that have the same variance and lag-1 autocorrelation. The reliability of any statistical inference procedure will be compromised if the assumptions that are implicit in the procedure are not satisfied.

### *a. Large samples ($n_e \geq 30$)*

We recommend the use of the "usual" $t$ test when the equivalent sample size $n_e$ (one sample) or the sum of the equivalent sample sizes (two sample) is *known* to be greater than 30. The procedures are as follows.

- The one-sample case:

To test $H_0 : \mu = \mu_0$ using a sample of size $n$ compute

$$t = \frac{(\bar{x} - \mu_0)}{s/(\hat{n}'_e)^{1/2}}, \tag{9}$$

where $\bar{x}$ is the sample mean and $s^2$ is the sample variance. Compute the estimated equivalent sample size $\hat{n}'_e$ using

$$\hat{n}'_e = \begin{cases} 2 & \text{if } \hat{n}_e \leq 2, \\ \hat{n}_e & \text{if } 2 < \hat{n}_e \leq n, \\ n & \text{otherwise,} \end{cases} \tag{10}$$

where $\hat{n}_e = n(1 - r_1)/(1 + r_1)$ and $r_1$ is the sample lag-1 correlation coefficient that is given by

$$r_1 = \frac{\sum_{t=2}^{n} (x_t - \bar{x})(x_{t-1} - \bar{x})}{\sum_{t=1}^{n} (x_t - \bar{x})^2}. \tag{11}$$

Compare the computed $t$ value with the appropriate critical values of the standard Gaussian distribution.

- The two-sample case:

To test $H_0 : \mu_y = \mu_x$ using $Y$ and $X$ samples of size $m$ and $n$, respectively, compute

$$t = \frac{\bar{y} - \bar{x}}{s[1/(\hat{m}'_e)^{1/2} + 1/(\hat{n}'_e)^{1/2}]}, \tag{12}$$

where $\bar{y}$ and $\bar{x}$ are sample means and $s^2$ is the pooled sample variance. The latter is given by

$$s^2 = [\sum_{t=1}^{m} (x_t - \bar{x})^2 + \sum_{t=1}^{n} (y_t - \bar{y})^2]/(m + n - 2). \quad (13)$$

The equivalent sample size estimates $\hat{m}'_e$ and $\hat{n}'_e$ are obtained by substituting a pooled estimate of the lag-1 correlation coefficient

$$r_1 = \frac{\sum_{t=2}^{m} (x_t - \bar{x})(x_{t-1} - \bar{x}) + \sum_{t=2}^{n} (y_t - \bar{y})(y_{t-1} - \bar{y})}{\sum_{t=1}^{m} (x_t - \bar{x})^2 + \sum_{t=1}^{n} (y_t - \bar{y})^2} \quad (14)$$

into (10) for sample sizes $m$ and $n$, respectively. Compare the computed $t$ value with the appropriate critical values of the standard Gaussian distribution.

*b. Small samples ($n_e < 30$)*

When $n_e$ is *known* to be less than 30, a two-stage procedure is required.

- First make a subjective estimate of $n_e$, which you know to be no greater than the actual value. Use this estimate with the "usual" $t$ test [based on (9) or (12)] to make a preliminary decision. If the decision is to reject $H_0$ then the analysis can stop because evidence against $H_0$ has been found with a conservative test. Otherwise the analysis should continue with either the LR test [see section 2b(1)] or table lookup test [see section 2b(2)].
- The LR test is suggested for the second stage when it is known that $n_e > 15$ (or $n_e > 15$ for both samples in the two-sample case). The LR test is preferable to the table lookup test from a practical point of view because the required software is widely available and because the test does not rely upon specialized tables that are difficult to derive.
- The table lookup test is suggested for the second stage when it is not known if $n_e > 15$ (or $n_e > 15$ for both samples in the two-sample case). In this case it is prudent to utilize the protection against spurious reject decisions that is offered by the superior small sample properties of the table lookup test.

Outlines of the LR and table lookup tests follow.

1) THE LR TEST

Our recommendations for the LR test are made under the assumption that the user has access to computer codes that are able to make *exact* Gaussian maximum likelihood estimates of autoregressive time series models. We recommend Ansley's algorithm (Ansley 1979), which is contained in the *Splus arima.mle* function (Splus 1992; Becker et al. 1988). The Splus function also makes use of a transformation of the autoregressive parameters (in this case only $\rho_1$), which insures stationarity of the fitted model (Jones 1980). Similar functions are available in other statistical packages. For reasons of computational convenience, the recommended two-sample procedure does not utilize the assumption of common variance and lag-1 correlation. See section 4 for further details.

- The one-sample case:

1. Assume that $H_0 : \mu = \mu_0$ is true, compute the deviations $x'_t = x_t - \mu_0$, and choose parameters $\rho_1$ and $\sigma_\epsilon$ to maximize the log likelihood $l_r$ of the model $X'_t = \rho_1 X'_{t-1} + \epsilon_t$.
2. Assume the $H_0 : \mu = \mu_0$ is false, choose parameters $\rho_1$, $\mu$, and $\sigma_\epsilon$ to maximize the log likelihood $l_f$ of the model $(X_t - \mu) = \rho_1(X_{t-1} - \mu) + \epsilon_t$.
3. Compute the difference between the log likelihood of the *full* model ($l_f$) and that of the $H_0$ restricted model ($l_r$). Compare ($l_f - l_r$) against the critical values of the $\chi^2$ distribution with 1-df.

- The two-sample case:

1. Assume that $H_0 : \mu_X = \mu_Y$ is true, compute the common mean $\hat{\mu} = (\sum_{t=1}^{m} x_t + \sum_{t=1}^{n} y_t)/(m + n)$ from both samples, compute the deviations $x'_t = x_t - \hat{\mu}$ and $y'_s = y_s - \hat{\mu}$, choose parameters $\rho_{x,1}$, $\sigma_\epsilon$, $\rho_{Y,1}$ and $\sigma_\delta$ to maximize the log likelihoods of the models $X'_t = \rho_1 X'_{t-1} + \epsilon_t$ and $Y'_s = \beta Y'_{s-1} + \delta_s$, respectively, and set $l_r$ to the sum of the log likelihoods.
2. Assume that $H_0 : \mu_X = \mu_Y$ is false, choose parameters $\rho_{X,1}$, $\mu_X$, $\sigma_\epsilon$, $\rho_{y,1}$, $\mu_Y$, and $\sigma_\delta$ to maximize the log likelihoods of the models $x_t - \mu_x = \rho_1(X_{t-1} - \mu_x) + \epsilon_t$ and $Y_s - \sigma_y = \beta(Y_{s-1} - \mu_y) + \delta_s$, respectively, and set $l_f$ to the sum of the log likelihoods.
3. Compute the difference between the log likelihood of the *full* model ($l_f$) and that of the $H_0$ restricted model ($l_r$). Compare ($l_f - l_r$) with the critical values of the $\chi^2$ distribution with 1-df.

2) THE TABLE LOOKUP TEST

- The one-sample case:

To test $H_0 : \mu = \mu_0$ using a sample of size $n$ compute

$$t = \frac{(\bar{x} - \mu_0)}{s/(n)^{1/2}}, \quad (15)$$

where $\bar{x}$ is the sample mean and $s^2$ is the sample variance. Compute the sample lag-1 correlation coefficient $r_1$ using (11). Use Tables 6–10[5] in section 5 to determine the critical value for $t$ that is appropriate to a sample of size $n$ that has a lag-1 correlation coefficient $r_1$.

[5] Electronic versions of the tables are available from the authors, either on DOS compatible floppy disk or via e-mail (fzwiers@uvic.bc.doe.ca).

• The two-sample case:

To test $H_0 : \mu_y = \mu_x$ using $Y$ and $X$ samples of size $m$ and $n$, respectively, compute

$$t = \frac{\bar{y} - \bar{x}}{s\left(\frac{1}{n} + \frac{1}{m}\right)^{1/2}}, \tag{16}$$

where $\bar{y}$ and $\bar{x}$ are sample means and $s^2$ is the pooled sample variance (13). Compute the pooled sample lag-1 correlation coefficient $r_1$ using (14). Use Tables 6–10 to determine the critical value for $t$ that is appropriate to a sample of size $m + n$ that has a lag-1 correlation coefficient $r_1$.

### 3. The usual test

In this section we discuss the "usual" $t$ test in which the ordinary Student's $t$ statistic is scaled using an estimate of the equivalent sample size. For simplicity, the discussion will focus primarily on the one-sample test. We indicate how the results apply in the more usual two-sample context at the end of this section. As discussed in section 1, the common approach used in climatology for testing the null hypothesis

$$H_0 : \mu = \mu_0 \tag{17}$$

is to

• compute a $t$ statistic as

$$t = \frac{(\bar{x} - \mu_0)}{s/(\hat{n}_e)^{1/2}}, \tag{18}$$

where $\bar{x}$ is the sample mean, $s^2$ is the sample variance, and $\hat{n}_e$ is an estimate of the equivalent sample size, and

• compare the computed $t$ with critical values from the Student's $t$ distribution with $\hat{n}_e - 1$ df.

Variations in this approach are distinguished by the method used to estimate $\hat{n}_e$. In general, parsimonious methods (i.e., methods that rely upon the estimation of a small number of parameters) perform best.

The extreme antithesis of a parsimonious method is that which TZ called "DIRECT." In that method $n_e$ is estimated by substituting the sample autocorrelation function into (4) directly. This results in highly variable estimates of $n_e$ and by inference, very poor test performance. The variability is caused by sampling variability in the $n - 1$ estimated parameters that enter the calculation.

Substantial improvement can be obtained by truncating the sample autocorrelation function after a fixed number of lags (the method TZ call "DIRECT2"). Further improvement can be obtained by deriving the autocorrelation function from a parametric time series model fitted to the observations (the method TZ call "ARMA"; see also Katz 1982).

We modify the ARMA estimator in two ways:

1) We fit only AR(1) models to the observations. We do this by estimating the lag-1 correlation coefficient of the observed time series with (11) and substituting this estimate of $\rho_1$ into (5) to provide an estimate $\hat{n}_e$ of $n_e$. Note that (6) may also be used when samples are large.

2) We note that sampling variability sometimes results in unrealistic values of $\hat{n}_e$. We therefore constrain the estimates to realistic values using (10).

The performance of this equivalent sample size estimator and the corresponding approximate $t$ test of the mean were examined by means of a simulation experiment. One thousand samples of length $n$ = 15, 30, 60, 90, 120, and 240 were taken from simulated AR(1) stochastic processes with mean zero and lag-1 correlation coefficients $\rho_1$ = 0.3, 0.6, and 0.9. Thus a total of 18 combinations of sample length and persistence were considered. Each sample was used to estimate the equivalent sample size using the modified ARMA method described above, and each was used to conduct an approximate test of (17) with $\mu_0 = 0$.

The results for $\hat{n}'_e$ are summarized in Table 1. The modified ARMA estimator of the equivalent sample size shows improved performance compared to that reported by TZ. Both the bias and the variability of the estimates are reduced.

The results for a two-sided test of (17) with $\mu_0 = 0$ using the $t$ statistic (18) with $n_e$ estimator $\hat{n}'_e$ and critical values appropriate to the 5% significance level are summarized in Table 2. Because $H_0$ is true, we would expect about 5% of the 1000 simulated tests to result in reject decisions if the test operates as anticipated.

We need to take the sampling variability of the observed rejection rates into account to determine whether they are significantly different from the anticipated rate. Test decisions are independent of each other because the samples were generated in such a way as to insure their mutual independence. Consequently, the number of reject decisions in each 1000-trial experiment should have a binomial distribution in which the probability of a "success" (a rejection) on any trial will be 0.05 (von Storch 1982). We therefore expect the observed proportion of reject decision to lie in the interval (0.0365, 0.0635) with probability 0.95.

Table 2 shows that the test generally rejects the null hypothesis too frequently. The effect is particularly dramatic when samples are small and the sampled time series is persistent. The test operates as designed only when samples are "large," that is, roughly when $n_e > 30$. Our conjecture is that at these sample sizes the variance of $\bar{x}$ is well enough estimated by $s^2/\hat{n}'_e$ that $t$ has approximately a Gaussian distribution.

The fact that the test rejects $H_0$ too frequently at small sample sizes is due to the sampling variability of $\hat{n}'_e$. Table 3 and Fig. 2 show that $H_0$ is not rejected

TABLE 1. Each cell contains a summary of 1000 simulated realizations of estimator $\hat{n}'_e$. Each estimate was computed from a sample of length $n$ generated from an AR (1) stochastic process with lag-1 correlation $\alpha$. The first entry in each cell is the known equivalent sample size obtained from (11). The second entry is the mean of 1000 realizations of $\hat{n}'_e$. The entries in parentheses indicate the interquartile range (IQR) of the 1000 realizations of $\hat{n}'_e$. The IQR contains the middle 50% of all $\hat{n}'_e$ realizations.

| | Lag-1 correlation $\alpha$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0.3 | | | 0.6 | | | 0.9 | | |
| n | $n_e$ | $\hat{n}'_e$ | IQR | $n_e$ | $\hat{n}'_e$ | IQR | $n_e$ | $\hat{n}'_e$ | IQR |
| 15 | 8.4 | 10.9 | (8, 15) | 4.3 | 7.8 | (5, 10) | 1.6 | 5.1 | (3, 6) |
| 30 | 16.5 | 19.7 | (15, 24) | 8.0 | 11.2 | (8, 14) | 2.3 | 5.7 | (4, 7) |
| 60 | 32.7 | 36.7 | (30, 43) | 15.5 | 18.6 | (14, 22) | 3.7 | 6.9 | (5, 8) |
| 90 | 48.8 | 51.7 | (44, 58) | 23.0 | 26.3 | (21, 30) | 5.3 | 8.2 | (6, 10) |
| 120 | 65.0 | 68.3 | (59, 76) | 30.5 | 33.6 | (28, 38) | 6.9 | 9.7 | (7, 12) |
| 240 | 129.6 | 134 | (121, 144) | 60.5 | 63.3 | (57, 69) | 13.2 | 16.1 | (13, 19) |

frequently enough when the simulation is repeated with the known rather than estimated value of $n_e$.

Our advice then is that this simple and intuitively appealing test of the mean should be used either when samples are very large or when other knowledge about the problem can be used to infer something about the true value of $n_e$. As a very rough guideline, when $n_e$ is estimated users should expect the actual level of significance of this test (when conducted at the nominal 5% level) to be about 10% when $n_e \approx 15$. The significance level will approach the nominal level when $n_e > 30$. If $n_e$ cannot be reliably estimated a conservative *guesstimate* (which is known to be no greater than the true $n_e$) can be used in place of $\hat{n}'_e$. This results in a safe (i.e., conservative) test that requires stronger evidence to reject $H_0$ than would be required by an optimal test.

A two-sample test version of the test is constructed by analogy with the standard difference of means test that is described in section 1. To apply the test, it is necessary to assume that the two samples come from processes with the same variance and lag-1 correlation coefficient. The characteristics of the resulting two-sample test, and our recommendations for its use, are the same as those for the one-sample test except that they apply to the sum of the equivalent sample sizes for the two samples.

## 4. The likelihood ratio test

A more formal approach to the problem of testing the mean of a time series is to base the inference on the LR test (see, e.g., Kalbfleisch 1979; Breiman 1973; Cox and Hinkley 1974). The idea here is that the *likelihood* of the observations is maximized under two scenarios: one in which the null hypothesis is true and the other in which it is false. These likelihoods are compared by computing their ratio. Asymptotic theory provides a large sample reference distribution for the natural logarithm of the likelihood ratio and demonstrates that LR tests are asymptotically optimal.

Suppose that the sample $x_1, \ldots, x_n$ represents a realization of the random variables $X_1, \ldots, X_n$. Suppose also that our assumptions about the observed process $(X_t)$ together with either the null of alternate hypotheses allows us to determine the joint density function $f(X_1, \ldots, X_n|\boldsymbol{\theta})$ of $X_1, \ldots, X_n$. The vector $\boldsymbol{\theta}$ represents parameters, such as the lag-1 correlation coefficient,

TABLE 2. Each cell summarizes the results of 1000 tests of $H_0: \mu = 0$. The test was conducted by computing t statistic (15) and comparing it with the 5% critical points of the Student's t distribution with $\hat{n}'_e - 1$ df. Boldface type indicates observed rejection rates under $H_0$ that are different from the nominal 5% at the 5% significance level.

| | lag-1 correlation $\rho_1$ | | |
|---|---|---|---|
| n | 0.3 | 0.6 | 0.9 |
| 15 | **0.077** | **0.169** | **0.308** |
| 30 | **0.083** | **0.116** | **0.249** |
| 60 | **0.070** | **0.084** | **0.170** |
| 90 | 0.060 | **0.079** | **0.130** |
| 120 | 0.051 | 0.062 | **0.118** |
| 240 | 0.052 | 0.060 | **0.090** |

TABLE 3. Each cell summarizes the results of 1000 tests of $H_0: \mu = 0$. The test was conducted by computing t statistic (15) using the known value of $n_e$ and comparing it with the 5% critical points of the Student's t distribution with $n_e - 1$ df. Boldface type indicates observed rejection rates under $H_0$ that are different from the nominal 5% at the 5% significance level.

| | lag-1 correlation $\rho_1$ | | |
|---|---|---|---|
| n | 0.3 | 0.6 | 0.9 |
| 15 | 0.046 | **0.022** | **0.000** |
| 30 | 0.045 | 0.038 | **0.000** |
| 60 | 0.051 | 0.049 | **0.014** |
| 90 | 0.047 | 0.046 | **0.026** |
| 120 | 0.047 | 0.046 | 0.039 |
| 240 | 0.048 | 0.051 | 0.048 |

TABLE 4. Each cell summarizes the results of 1000 LR tests of $H_0 : \mu = 0$. The test was conducted by computing the LR statistic and comparing it with the 5% critical point of the $\chi^2$ distribution with 1 df. Boldface type indicates observed rejection rates under $H_0$ that are different from the nominal 5% rate at the 5% significance level.

| | lag-1 correlation $\alpha$ | | |
|---|---|---|---|
| n | 0.3 | 0.6 | 0.9 |
| 15 | **0.088** | **0.118** | **0.214** |
| 30 | 0.057 | **0.085** | **0.150** |
| 60 | 0.045 | **0.077** | **0.096** |
| 90 | 0.063 | 0.062 | **0.094** |
| 120 | 0.049 | 0.056 | **0.092** |
| 240 | 0.055 | 0.056 | 0.064 |

which must be estimated. The *likelihood function* is then defined as

$$L(\boldsymbol{\theta} | x_1, \ldots, x_n) = f(x_1, \ldots, x_n | \boldsymbol{\theta}) \quad (19)$$

and the *maximum likelihood* estimate $\theta$ of $\boldsymbol{\theta}$ is obtained determining the value of $\boldsymbol{\theta}$, which maximizes $L$.

In the present context the two models that are fitted to the observed time series via the method of maximum likelihood are

$$(X_t - \mu_0) = \rho_1(X_{t-1} - \mu_0) + \epsilon_t \quad \text{when } H_0 \text{ true,} \quad (20)$$

$$(X_t - \mu) = \rho_1(X_{t-1} - \mu) + \epsilon_t \quad \text{when } H_0 \text{ false.} \quad (21)$$

The noise, $(\epsilon_t)$, is Gaussian and white. When $H_0$ is true the likelihood function depends upon $\rho_1$ and $\sigma_\epsilon^2$, the variance of the noise. When $H_0$ is false, it depends upon $\rho_1$, $\sigma_\epsilon^2$, and $\mu$.

The likelihood functions for the *full* model (i.e., $H_0$ false) is given by

$$L(\mu, \rho_1, \sigma_\epsilon | \mathbf{x}) = (2\pi\sigma_\epsilon^2)^{-n/2} |M|^{1/2} \times \exp - \frac{(\mathbf{x} - \mu)' M (\mathbf{x} - \mu)}{2\sigma_\epsilon^2}, \quad (22)$$

where $\boldsymbol{\mu}$ is the $n \times 1$ vector $(\mu, \mu, \ldots, \mu)'$ and $\mathbf{x}$ is the $n \times 1$ vector of observations $(x_1, \ldots, x_n)$. The matrix $M$ is given by $\Sigma = \sigma_\epsilon^2 M^{-1}$, where $\Sigma$ is the variance–covariance matrix of the vector of observations $\mathbf{x}$. Here $\sigma_\epsilon^2$ is the white noise variance. The $(i, j)$ element of $\Sigma$ is given by $\sigma_{i,j} = \sigma_\epsilon^2 \rho^{|i-j|}/(1 - \rho_1^2)$. The likelihood

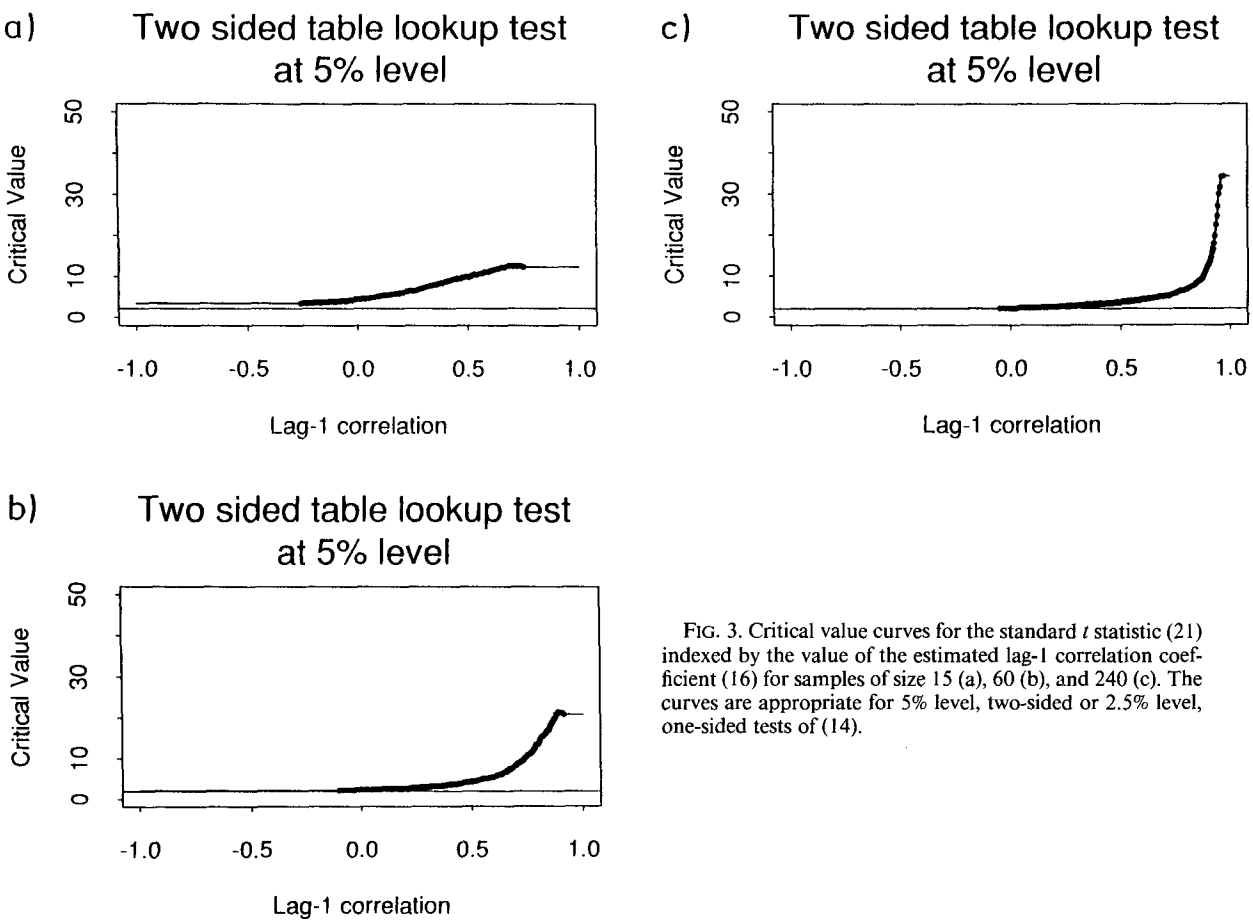


FIG. 3. Critical value curves for the standard $t$ statistic (21) indexed by the value of the estimated lag-1 correlation coefficient (16) for samples of size 15 (a), 60 (b), and 240 (c). The curves are appropriate for 5% level, two-sided or 2.5% level, one-sided tests of (14).

TABLE 5. Each cell summarizes the results of 1000 table lookup tests of $H_0: \mu = 0$. The test was conducted by computing the ordinary t statistic and comparing it with an empirical distribution that is indexed by the sample lag-1 correlation coefficient (see text). Boldface type indicates observed rejection rates under $H_0$ that are different from the nominal 5% at the 5% significance level.

| | lag-1 correlation $\alpha$ | | |
|---|---|---|---|
| n | 0.3 | 0.6 | 0.9 |
| 15 | 0.003 | 0.030 | 0.128 |
| 30 | 0.027 | 0.035 | 0.085 |
| 60 | 0.043 | 0.043 | 0.061 |
| 90 | 0.037 | 0.043 | 0.034 |
| 120 | 0.042 | 0.047 | 0.051 |
| 240 | 0.053 | 0.050 | 0.054 |

function for the $H_0$ restricted model is identical except that $\boldsymbol{\mu}_0 = (\mu_0, \mu_0, \ldots, \mu_0)'$ is substituted for $\boldsymbol{\mu}$. The intricacies of the exact likelihood function are described in detail in Box and Jenkins (1976, see section A7.4).

The LR test is conducted by computing the difference between the maximum log likelihood of the full (21) and partial (20) models. The resulting LR statistic is asymptotically distributed $\chi^2$ with 1 df under the null hypothesis.

We conducted an experiment identical to that described in section 3 to determine the actual significance level of the LR test when decisions are made at the nominal 5% significance level. Results of this experiment are summarized in Table 4.

The LR test clearly improves upon the "usual" test of the mean. Our advice is that this test should be used in preference to the "usual" test. Its mathematical and computational complexity should not be a deterrent because modern maximum likelihood estimation routines for Box–Jenkins models are readily available in a number of statistical packages. The LR test approach has desirable optimality properties and also has the advantage that it can be expanded to incorporate more sophisticated statistical representations of the stochastic nature of the observed climate. As a rough guideline, users should expect the actual level of significance of the LR test (when conducted at the nominal 5% level) to be about 10% for $n_e \approx 8$. The significance level will approach the nominal level when $n_e > 15$.

A two-sample version of the LR test is constructed by analogy to the one-sample test. The procedure is detailed in section 2b(1). Two-sample LR tests can be constructed with or without the assumption of section 3 that the two samples come from processes with the same variance and lag-1 correlation coefficient. We recommend the somewhat more general version that does not depend upon this assumption because the necessary computations can be done with existing software when the test is cast in this way. This generality comes at a cost in that larger samples are needed for the test to attain its asymptotic properties. The concept behind the test is that the *joint* log likelihood of samples $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ is maximized both with and without the restrictions imposed by the null hypotheses. The difference between the two log likelihoods is then computed and compared against the critical values of a $\chi^2$ distribution. The number of degrees of freedom is found by computing the difference between the number of free parameters in the full model and the $H_0$ restricted model.

Our guidelines for use of the two-sample test are the same as those for the one-sample test except that they should be satisfied individually by the equivalent sample size of each sample. This cautious approach, which differs from that for the two-sample test discussed in section 3, is required because of the generality of the two-sample LR test we advocate.

## 5. The table lookup test

The "usual" $t$ test described in section 3 works poorly because

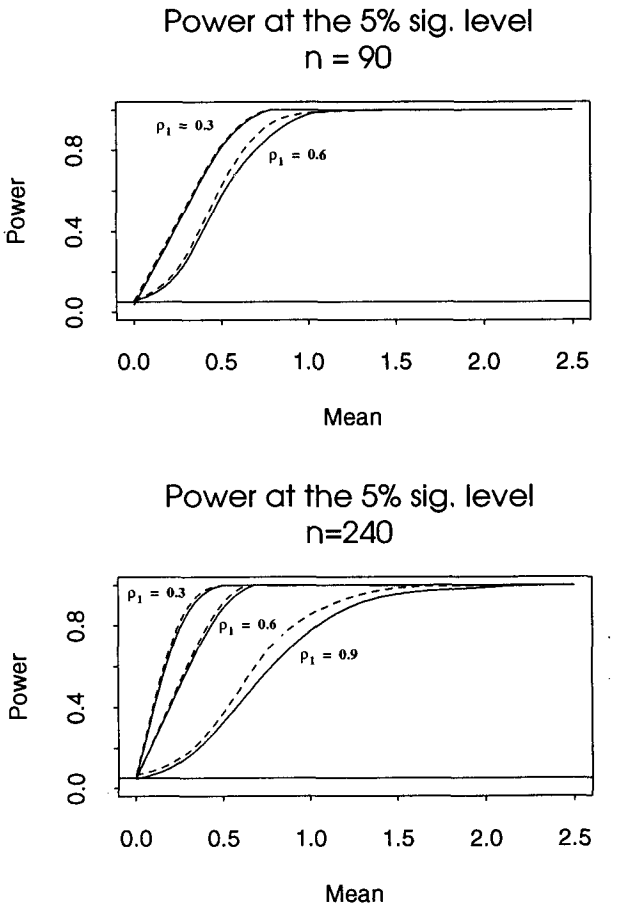


FIG. 4. The estimated power of the table lookup test (solid line) and the LR test (dashed line) for a range of alternatives to the null hypothesis. The power was estimated by generating 1000 samples from an AR(1) process with the specified mean and testing the null hypothesis on each sample. The upper panel shows the power of the test for samples of size 90 obtained from AR(1) processes with $\rho_1 = 0.3$ and 0.6. The lower panel displays corresponding results for samples of size 240 $\rho_1 = 0.3$, 0.6, and 0.9.

TABLE 6. Critical values for the table lookup test appropriate for a two- (one-) sided test conducted at the 20% (10%) significance level. Dashes indicate that sample correlations of this magnitude were not observed in the simulations used to create the table.

| | n | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $r_1$ | 10 | 15 | 20 | 25 | 30 | 45 | 60 | 75 | 90 | 120 | 180 | 240 |
| −0.35 | 2.30 | — | — | — | — | — | — | — | — | — | — | — |
| −0.30 | 2.34 | 1.87 | — | — | — | — | — | — | — | — | — | — |
| −0.25 | 2.41 | 1.88 | 1.70 | — | — | — | — | — | — | — | — | — |
| −0.20 | 2.46 | 1.92 | 1.71 | 1.62 | 1.54 | — | — | — | — | — | — | — |
| −0.15 | 2.52 | 1.99 | 1.75 | 1.64 | 1.55 | 1.45 | — | — | — | — | — | — |
| −0.10 | 2.61 | 2.04 | 1.80 | 1.67 | 1.59 | 1.47 | 1.44 | 1.42 | 1.39 | 1.38 | — | — |
| −0.05 | 2.67 | 2.09 | 1.84 | 1.68 | 1.64 | 1.50 | 1.47 | 1.44 | 1.39 | 1.39 | 1.36 | 1.33 |
| 0.00 | 2.75 | 2.18 | 1.88 | 1.72 | 1.66 | 1.53 | 1.49 | 1.46 | 1.43 | 1.40 | 1.38 | 1.35 |
| 0.05 | 2.82 | 2.26 | 1.97 | 1.79 | 1.71 | 1.60 | 1.55 | 1.47 | 1.50 | 1.46 | 1.43 | 1.42 |
| 0.10 | 2.94 | 2.35 | 2.08 | 1.87 | 1.76 | 1.64 | 1.58 | 1.55 | 1.50 | 1.48 | 1.46 | 1.40 |
| 0.15 | 3.10 | 2.48 | 2.18 | 1.96 | 1.82 | 1.70 | 1.66 | 1.60 | 1.58 | 1.57 | 1.53 | 1.51 |
| 0.20 | 3.23 | 2.59 | 2.27 | 2.08 | 1.91 | 1.77 | 1.70 | 1.66 | 1.64 | 1.63 | 1.59 | 1.57 |
| 0.25 | 3.36 | 2.75 | 2.42 | 2.22 | 2.05 | 1.87 | 1.78 | 1.75 | 1.77 | 1.69 | 1.69 | 1.70 |
| 0.30 | 3.48 | 2.96 | 2.57 | 2.36 | 2.22 | 1.96 | 1.91 | 1.91 | 1.86 | 1.86 | 1.79 | 1.77 |
| 0.35 | 3.61 | 3.20 | 2.79 | 2.56 | 2.38 | 2.12 | 2.02 | 2.03 | 1.96 | 1.88 | 1.92 | 1.90 |
| 0.40 | 3.77 | 3.46 | 3.02 | 2.74 | 2.61 | 2.31 | 2.22 | 2.13 | 2.11 | 2.07 | 2.00 | 2.02 |
| 0.45 | 3.95 | 3.66 | 3.38 | 3.05 | 2.79 | 2.55 | 2.36 | 2.25 | 2.23 | 2.19 | 2.14 | 2.13 |
| 0.50 | 4.13 | 3.92 | 3.71 | 3.41 | 3.24 | 2.73 | 2.57 | 2.47 | 2.46 | 2.39 | 2.31 | 2.30 |
| 0.55 | 4.27 | 4.23 | 4.13 | 3.77 | 3.55 | 3.03 | 2.86 | 2.65 | 2.61 | 2.56 | 2.50 | 2.43 |
| 0.60 | 4.45 | 4.59 | 4.47 | 4.26 | 3.94 | 3.45 | 3.14 | 3.03 | 2.83 | 2.75 | 2.65 | 2.72 |
| 0.65 | 4.55 | 4.83 | 4.85 | 4.72 | 4.49 | 3.89 | 3.60 | 3.26 | 3.16 | 3.02 | 2.91 | 2.94 |
| 0.70 | 4.56 | 5.17 | 5.37 | 5.28 | 5.25 | 4.49 | 4.01 | 3.74 | 3.59 | 3.36 | 3.27 | 3.33 |
| 0.75 | — | 5.37 | 5.71 | 5.82 | 5.85 | 5.35 | 4.91 | 4.40 | 4.14 | 3.92 | 3.69 | 3.55 |
| 0.80 | — | — | 5.99 | 6.49 | 6.49 | 6.42 | 6.00 | 5.51 | 5.18 | 4.67 | 4.28 | 4.25 |
| 0.85 | — | — | — | 6.66 | 7.26 | 7.82 | 7.33 | 7.28 | 6.76 | 6.09 | 5.41 | 5.05 |
| 0.90 | — | — | — | — | 7.31 | 8.77 | 9.47 | 9.45 | 9.01 | 8.59 | 7.55 | 6.97 |
| 0.95 | — | — | — | — | — | — | 9.93 | 10.7 | 11.3 | 13.2 | 13.3 | 12.5 |

• The test statistic does not have a Student's $t$ distribution under the null hypothesis.

• The denominator of the test statistic is highly variable because the equivalent sample size is poorly estimated.

• The critical value to which the test statistic is compared is subject to variability because the reference distribution is indexed by the same equivalent sample size estimator.

We reasoned that these problems could be ameliorated with an empirical test procedure based on the following ideas:

1. Base the test on a statistic that is not affected by the sampling variability that is present in equivalent sample size estimators. We chose to use the ordinary $t$ statistic, which does not take serial correlation into account. The statistic (15) is given by

$$t = \frac{\bar{x} - \mu_0}{s/(n)^{1/2}}. \tag{23}$$

2. Index the critical values that are used to make test decisions with an indicator of the serial correlation which has low variability. We chose to use the sample correlation coefficient $r_1$ rather than $\hat{n}'_e$. The latter is highly variable because it is a nonlinear function of the former.

3. Derive critical values appropriate to a particular value of $r_1$ via Monte Carlo simulation.

The last point deserves some comment.

When the true lag-1 correlation coefficient $\rho_1$ is known, very good estimates of the critical values for the test of (17) based on (23) are easily obtained via Monte Carlo simulation. The approach is to generate a large number of samples of size $n$ from the appropriate AR(1) process, compute $t$ for each sample, compute the appropriate percentiles from the ensemble of simulated $t$'s, and compare the observed $t$ with the derived percentiles. The operating significance level of such as test will be very close to the nominal level provided one derives the critical values (i.e., percentiles) from a large ensemble of simulated $t$ values.

When the true lag-1 correlation coefficient is not known, critical values are more difficult to obtain because a wide range of true lag-1 correlation coefficients $\rho_1$ will be consistent with the sample coefficient $r_1$. The Bayesian approach to this problem is to use probability distributions to express the likelihood that $\rho_1$ has a particular value.

The *prior distribution* expresses our subjective beliefs about likely values of $\rho_1$ before observations are taken. In the absence of any other information, a reasonable choice of prior distribution on $\rho_1$ is the uninformative prior, which places equal weight on all values in the

TABLE 7. Critical values for the table lookup test appropriate for a two- (one-) sided test conducted at the 10% (5%) significance level. Dashes indicate that sample correlations of this magnitude were not observed in the simulations used to create the table.

| | n | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $r_1$ | 10 | 15 | 20 | 25 | 30 | 45 | 60 | 75 | 90 | 120 | 180 | 240 |
| −0.35 | 3.46 | — | — | — | — | — | — | — | — | — | — | — |
| −0.30 | 3.55 | 2.57 | — | — | — | — | — | — | — | — | — | — |
| −0.25 | 3.68 | 2.59 | 2.26 | — | — | — | — | — | — | — | — | — |
| −0.20 | 3.80 | 2.66 | 2.28 | 2.15 | 2.03 | — | — | — | — | — | — | — |
| −0.15 | 3.85 | 2.76 | 2.35 | 2.17 | 2.04 | 1.89 | — | — | — | — | — | — |
| −0.10 | 4.03 | 2.89 | 2.44 | 2.22 | 2.09 | 1.92 | 1.85 | 1.83 | 1.81 | 1.78 | — | — |
| −0.05 | 4.21 | 2.98 | 2.52 | 2.24 | 2.13 | 1.96 | 1.90 | 1.88 | 1.81 | 1.80 | 1.73 | 1.72 |
| 0.00 | 4.34 | 3.15 | 2.59 | 2.31 | 2.18 | 2.02 | 1.92 | 1.89 | 1.84 | 1.83 | 1.75 | 1.74 |
| 0.05 | 4.48 | 3.30 | 2.76 | 2.40 | 2.26 | 2.09 | 1.98 | 1.89 | 1.94 | 1.86 | 1.84 | 1.84 |
| 0.10 | 4.75 | 3.47 | 2.88 | 2.56 | 2.38 | 2.15 | 2.07 | 2.01 | 1.96 | 1.91 | 1.89 | 1.85 |
| 0.15 | 5.07 | 3.71 | 3.06 | 2.69 | 2.43 | 2.24 | 2.14 | 2.09 | 2.04 | 2.02 | 1.96 | 1.96 |
| 0.20 | 5.27 | 3.96 | 3.26 | 2.85 | 2.58 | 2.33 | 2.21 | 2.17 | 2.14 | 2.12 | 2.05 | 2.03 |
| 0.25 | 5.42 | 4.25 | 3.63 | 3.07 | 2.79 | 2.48 | 2.35 | 2.28 | 2.29 | 2.21 | 2.19 | 2.20 |
| 0.30 | 5.63 | 4.60 | 3.82 | 3.32 | 3.07 | 2.66 | 2.53 | 2.51 | 2.42 | 2.37 | 2.33 | 2.26 |
| 0.35 | 5.90 | 5.10 | 4.20 | 3.72 | 3.35 | 2.85 | 2.63 | 2.65 | 2.56 | 2.47 | 2.44 | 2.43 |
| 0.40 | 6.17 | 5.56 | 4.74 | 4.07 | 3.67 | 3.11 | 2.94 | 2.81 | 2.76 | 2.68 | 2.56 | 2.56 |
| 0.45 | 6.44 | 5.60 | 5.28 | 4.59 | 4.09 | 3.53 | 3.19 | 2.97 | 2.94 | 2.84 | 2.79 | 2.76 |
| 0.50 | 6.80 | 6.54 | 5.93 | 5.23 | 4.80 | 3.79 | 3.52 | 3.26 | 3.19 | 3.11 | 3.02 | 2.95 |
| 0.55 | 7.00 | 6.93 | 6.64 | 5.96 | 5.31 | 4.39 | 3.85 | 3.52 | 3.49 | 3.33 | 3.25 | 3.13 |
| 0.60 | 7.15 | 7.45 | 7.36 | 6.96 | 6.14 | 5.05 | 4.29 | 4.10 | 3.76 | 3.61 | 3.44 | 3.54 |
| 0.65 | 7.20 | 7.91 | 8.03 | 7.67 | 7.15 | 5.80 | 5.03 | 4.49 | 4.24 | 4.06 | 3.79 | 3.78 |
| 0.70 | 7.21 | 8.39 | 8.70 | 8.72 | 8.50 | 6.87 | 5.74 | 5.30 | 4.85 | 4.53 | 4.34 | 4.21 |
| 0.75 | — | 8.57 | 9.13 | 9.44 | 9.22 | 8.43 | 7.19 | 6.43 | 5.80 | 5.23 | 4.89 | 4.67 |
| 0.80 | — | — | 9.65 | 10.3 | 10.5 | 10.4 | 9.11 | 8.28 | 7.64 | 6.55 | 5.79 | 5.50 |
| 0.85 | — | — | 9.67 | 10.4 | 11.4 | 12.8 | 11.8 | 11.4 | 10.3 | 9.02 | 7.40 | 6.72 |
| 0.90 | — | — | — | — | 11.4 | 13.8 | 14.4 | 14.9 | 14.9 | 13.4 | 10.9 | 9.58 |
| 0.95 | — | — | — | — | — | — | 14.9 | 16.4 | 17.3 | 20.3 | 20.7 | 19.8 |

interval [0, 1). That is, we think any value of $\rho_1$ consistent with a red spectrum is as likely as any other. The *posterior distribution* expresses our revised opinion about likely values of $\rho_1$ after sampling is complete.

Given the posterior distribution, Monte Carlo methods can be used to derive critical values for the test of (17) based on (23). The approach is to generate a large number of $\rho_1$'s from the posterior distribution, generate a sample of length $n$ from each corresponding AR(1) process, compute $t$'s from these samples, and then compute the appropriate percentiles from the ensemble of simulated $t$'s.

We emulate the Bayesian approach by deriving critical values as follows:

1. Generate an ensemble of 240 000 lag-1 correlation coefficients $\rho_1$ randomly on the interval [0, 1).
2. For each $\rho_1$ generate a sample of length $n$ from the corresponding AR(1) process.
3. Compute $r_1$ and $t$ from each sample.
4. Sort the resulting ensemble of 240 000 $(r_1, t)$ pairs in order of increasing $r_1$.
5. Select 200 equally spaced points $r_{1,i}$, $i = 1, \ldots,$ 200 between the minimum and maximum of the simulated $r_1$'s.
6. At each of these 200 "base" points select the $m$ $(r_1, t)$ pairs with $r_1$ nearest the base value. We used $m = 4800(240/n)^{1/2}$. This choice for $m$ will be explained below.
7. Compute the 80th, 90th, 95th, 98th and 99th quantiles of $|t|$ from each subset of selected $t$ values. We will refer to these quantiles as $t_{0.90,i}$, $t_{0.95,i}$, . . . , $t_{0.995,i}$.[6]

The result is five critical value tables (Tables 6–10), which are indexed by $r_1$ and are suitable for two-sided (one-sided) tests of (17) at the 20%, 10%, 5%, 2%, or 1% (10%, 5%, 2.5%, 1%, or 0.5%) level, respectively.

These tables form the basis of our empirical test procedure. To conduct a test at a given significance level we first compute $r_1$ and $t$ from the sample. The sample correlation $r_1$ is used to enter the appropriate critical value table. It may be necessary to interpolate between table entries that are nearest to $r_1$.

Examples of the tables of the quantiles appropriate for a 5% level two-sided test with samples of size $n$ = 15, 60, and 240 are illustrated in Fig. 3. Note that for the larger sample sizes the crossing point at $r_1 = 0$ corresponds closely to the critical value that would be

[6] Note that because of symmetry, the 80th (90th, . . .) percentile of the simulated $|t|$'s is a better estimator of the 90th (95th, . . .) percentile of $t$ than the 90th (95th, . . .) percentile of the simulated $t$'s.

TABLE 8. Critical values for the table lookup test appropriate for a two- (one-) sided test conducted at the 5% (2.5%) significance level. Dashes indicate that sample correlations of this magnitude were not observed in the simulations used to create the table.

| | n | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $r_1$ | 10 | 15 | 20 | 25 | 30 | 45 | 60 | 75 | 90 | 120 | 180 | 240 |
| −0.35 | 5.18 | — | — | — | — | — | — | — | — | — | — | — |
| −0.30 | 5.32 | 3.44 | — | — | — | — | — | — | — | — | — | — |
| −0.25 | 5.56 | 3.47 | 2.84 | — | — | — | — | — | — | — | — | — |
| −0.20 | 5.71 | 3.61 | 2.86 | 2.65 | 2.50 | — | — | — | — | — | — | — |
| −0.15 | 5.76 | 3.75 | 2.98 | 2.70 | 2.50 | 2.34 | — | — | — | — | — | — |
| −0.10 | 6.25 | 3.88 | 3.11 | 2.76 | 2.56 | 2.34 | 2.22 | 2.23 | 2.19 | 2.10 | — | — |
| −0.05 | 6.48 | 4.05 | 3.19 | 2.81 | 2.62 | 2.40 | 2.28 | 2.28 | 2.19 | 2.12 | 2.08 | 2.05 |
| 0.00 | 6.75 | 4.44 | 3.34 | 2.90 | 2.71 | 2.46 | 2.33 | 2.28 | 2.31 | 2.19 | 2.11 | 2.20 |
| 0.05 | 7.17 | 4.68 | 3.59 | 3.05 | 2.82 | 2.53 | 2.37 | 2.31 | 2.31 | 2.25 | 2.18 | 2.20 |
| 0.10 | 7.49 | 5.16 | 3.82 | 3.29 | 2.99 | 2.65 | 2.53 | 2.42 | 2.34 | 2.32 | 2.21 | 2.21 |
| 0.15 | 7.91 | 5.55 | 4.15 | 3.55 | 3.12 | 2.77 | 2.63 | 2.52 | 2.46 | 2.43 | 2.38 | 2.35 |
| 0.20 | 8.38 | 6.00 | 4.61 | 3.74 | 3.34 | 2.89 | 2.70 | 2.64 | 2.60 | 2.59 | 2.45 | 2.43 |
| 0.25 | 8.52 | 6.50 | 5.20 | 4.13 | 3.52 | 3.07 | 2.89 | 2.78 | 2.76 | 2.69 | 2.66 | 2.66 |
| 0.30 | 8.81 | 7.17 | 5.57 | 4.52 | 4.05 | 3.32 | 3.10 | 3.05 | 2.91 | 2.86 | 2.76 | 2.70 |
| 0.35 | 9.11 | 7.93 | 6.25 | 5.23 | 4.48 | 3.60 | 3.28 | 3.21 | 3.11 | 2.92 | 2.93 | 3.00 |
| 0.40 | 9.55 | 8.74 | 7.19 | 5.88 | 5.05 | 3.96 | 3.61 | 3.45 | 3.34 | 3.28 | 3.13 | 3.11 |
| 0.45 | 9.91 | 9.36 | 8.25 | 6.72 | 5.92 | 4.50 | 3.95 | 3.75 | 3.51 | 3.48 | 3.36 | 3.29 |
| 0.50 | 10.4 | 9.84 | 9.10 | 8.00 | 7.01 | 4.99 | 4.36 | 3.97 | 3.90 | 3.78 | 3.70 | 3.54 |
| 0.55 | 10.6 | 10.6 | 10.2 | 9.21 | 7.79 | 5.82 | 5.04 | 4.41 | 4.28 | 4.09 | 4.00 | 3.85 |
| 0.60 | 10.7 | 11.2 | 11.4 | 11.0 | 9.22 | 6.98 | 5.49 | 5.31 | 4.69 | 4.47 | 4.18 | 4.25 |
| 0.65 | 10.6 | 12.1 | 12.1 | 11.9 | 11.0 | 8.33 | 6.68 | 5.81 | 5.36 | 5.04 | 4.67 | 4.57 |
| 0.70 | 10.6 | 12.5 | 13.2 | 13.5 | 13.0 | 10.1 | 8.20 | 7.07 | 6.25 | 5.71 | 5.25 | 5.04 |
| 0.75 | — | 12.3 | 13.8 | 14.2 | 14.2 | 13.1 | 10.7 | 9.12 | 7.77 | 6.58 | 5.98 | 5.62 |
| 0.80 | — | — | 14.2 | 15.2 | 15.6 | 15.2 | 13.6 | 12.1 | 11.0 | 8.60 | 7.22 | 6.64 |
| 0.85 | — | — | — | 15.1 | 16.5 | 18.8 | 17.3 | 16.9 | 15.4 | 12.9 | 9.50 | 8.37 |
| 0.90 | — | — | — | — | 16.5 | 20.0 | 21.2 | 22.0 | 21.2 | 20.0 | 15.1 | 12.6 |
| 0.95 | — | — | — | — | — | — | 20.9 | 23.4 | 24.3 | 27.4 | 28.8 | 29.5 |

used if we could assume that the data are not serially correlated. The latter (percentiles of the Student's $t$ distribution with $n - 1$ degrees of freedom) are illustrated as horizontal lines. The discrepancy between the crossing point at $r_1 = 0$ and the Student's $t$ critical value is large for small samples because the distribution of $r_1$ is widely dispersed in this case. Note that quantiles gradually increase as $r_1$ increases. Critical values for tests conducted at higher levels of significance, such as the 1% level, actually peak for $r_1$ near 1 and then decrease as $r_1$ approaches 1. We conjecture that the numerator and denominator of the $t$ statistic (23) become less dependent as the true correlation coefficient $\rho_1$ approaches 1 (see the appendix for details). This would imply lower sampling variability for $t$ as $\rho_1$ (and hence $r_1$) approaches 1 and consequently smaller critical values.

This empirical approach emulates the Bayesian approach in the following sense. The critical value, which is referenced by a sample lag-1 correlation coefficient, was derived as the corresponding quantile of a collection of $m$ realizations of $t$. Each of these realizations was obtained from an AR(1) process with a different lag-1 correlation coefficient $\rho_1$. This collection of $\rho_1$'s constitutes an empirical posterior distribution on $\rho_1$. The prior distribution in this instance is the uniform distribution on the interval [0, 1).

The total number of simulated $(r_1, t)$ pairs used to obtain the critical value table is large to reduce noise in the table. The total number of pairs (240 000) and the number of pairs used to determine an individual table entry [$m = 4800(240/n)^{1/2}$] was chosen so that $(r_1, t)$ pairs spanning only 2% of the $r_1$ range could be used to accurately determine the critical values for the largest sample size considered (240). A narrow span is necessary when the sample size is large because the critical value curve changes quickly in this case for values of $r_1$ near 1. The number of simulated $(r_1, t)$ pairs used to determine critical value table entries is a function of $n^{-1/2}$ to compensate for the effects of sampling variability on $r_1$.

The operation of the test when the null hypothesis is true was examined in a simulation experiment analogous to those described in sections 3 and 4. The rejection rate for the test under $H_0$ in a 1000 trial experiment is reported in Table 5. Note that there is some imprecision when sample sizes are small: the test is conservative when the lag-1 correlation coefficient is small and somewhat liberal when it is large. Otherwise, the test operates more or less as advertised—and it generally does so for smaller samples than either of the competing tests described previously. Our experimentation has shown us that the imprecision in the test at small sample sizes is caused by sampling variability in the lag-1 correlation coefficient.

TABLE 9. Critical values for the table lookup test appropriate for a two- (one-) sided test conducted at the 2% (1%) significance level. Dashes indicate that sample correlations of this magnitude were not observed in the simulations used to create the table.

| | n | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $r_1$ | 10 | 15 | 20 | 25 | 30 | 45 | 60 | 75 | 90 | 120 | 180 | 240 |
| −0.35 | 8.76 | — | — | — | — | — | — | — | — | — | — | — |
| −0.30 | 9.03 | 5.02 | — | — | — | — | — | — | — | — | — | — |
| −0.25 | 9.26 | 5.12 | 3.80 | — | — | — | — | — | — | — | — | — |
| −0.20 | 9.98 | 5.43 | 3.89 | 3.34 | 3.07 | — | — | — | — | — | — | — |
| −0.15 | 10.1 | 5.67 | 4.03 | 3.42 | 3.10 | 2.79 | — | — | — | — | — | — |
| −0.10 | 10.9 | 5.87 | 4.21 | 3.52 | 3.19 | 2.82 | 2.68 | 2.68 | 2.63 | 2.51 | — | — |
| −0.05 | 11.1 | 6.49 | 4.38 | 3.67 | 3.30 | 2.97 | 2.77 | 2.69 | 2.61 | 2.51 | 2.52 | 2.45 |
| 0.00 | 11.9 | 7.35 | 4.71 | 3.91 | 3.44 | 3.04 | 2.85 | 2.76 | 2.69 | 2.61 | 2.50 | 2.41 |
| 0.05 | 12.4 | 7.97 | 5.30 | 4.07 | 3.60 | 3.08 | 2.89 | 2.82 | 2.76 | 2.69 | 2.61 | 2.59 |
| 0.10 | 12.8 | 8.92 | 5.72 | 4.49 | 3.95 | 3.28 | 3.03 | 2.88 | 2.78 | 2.71 | 2.66 | 2.64 |
| 0.15 | 14.0 | 10.1 | 6.51 | 4.78 | 4.26 | 3.45 | 3.24 | 3.08 | 3.01 | 2.92 | 2.83 | 2.86 |
| 0.20 | 14.6 | 10.5 | 7.18 | 5.28 | 4.40 | 3.60 | 3.39 | 3.19 | 3.06 | 3.09 | 2.94 | 2.96 |
| 0.25 | 14.7 | 11.1 | 8.59 | 6.16 | 4.81 | 3.81 | 3.55 | 3.43 | 3.27 | 3.21 | 3.19 | 3.23 |
| 0.30 | 15.2 | 12.8 | 8.94 | 6.88 | 5.68 | 4.26 | 3.97 | 3.72 | 3.57 | 3.42 | 3.31 | 3.21 |
| 0.35 | 15.4 | 13.7 | 10.5 | 8.26 | 6.58 | 4.70 | 4.05 | 3.99 | 3.82 | 3.51 | 3.49 | 3.56 |
| 0.40 | 15.7 | 14.9 | 12.8 | 9.54 | 7.85 | 5.26 | 4.60 | 4.37 | 4.07 | 3.85 | 3.84 | 3.79 |
| 0.45 | 16.4 | 16.3 | 14.9 | 11.4 | 8.92 | 6.32 | 5.14 | 4.73 | 4.33 | 4.28 | 4.07 | 3.92 |
| 0.50 | 17.1 | 17.4 | 16.3 | 13.8 | 11.4 | 6.99 | 5.81 | 5.18 | 4.79 | 4.58 | 4.39 | 4.35 |
| 0.55 | 16.9 | 17.9 | 17.3 | 15.6 | 12.6 | 8.73 | 6.91 | 5.65 | 5.34 | 5.05 | 4.85 | 4.57 |
| 0.60 | 17.2 | 18.6 | 18.7 | 18.9 | 14.8 | 10.6 | 7.53 | 6.78 | 5.94 | 5.63 | 4.94 | 5.10 |
| 0.65 | 16.9 | 19.8 | 19.2 | 20.6 | 18.9 | 14.2 | 10.1 | 7.76 | 7.05 | 6.24 | 5.62 | 5.37 |
| 0.70 | 16.9 | 19.3 | 21.3 | 22.0 | 22.2 | 17.6 | 12.4 | 9.77 | 8.35 | 7.32 | 6.51 | 6.12 |
| 0.75 | — | 18.8 | 21.4 | 22.1 | 23.0 | 21.9 | 18.1 | 14.2 | 11.0 | 8.56 | 7.64 | 6.98 |
| 0.80 | — | — | 20.7 | 22.9 | 24.6 | 24.6 | 21.3 | 20.0 | 16.6 | 12.6 | 9.15 | 8.18 |
| 0.85 | — | — | — | 22.7 | 25.1 | 28.6 | 28.4 | 30.3 | 24.6 | 19.1 | 12.9 | 11.1 |
| 0.90 | — | — | — | — | 25.0 | 27.9 | 31.0 | 33.7 | 33.1 | 31.1 | 23.8 | 17.8 |
| 0.95 | — | — | — | — | — | — | 29.7 | 32.2 | 33.4 | 37.4 | 42.3 | 45.8 |

The power of the table lookup test is contrasted with that of the LR test in Fig. 4 for samples of size 90 for which the true lag-1 correlation is 0.3 and 0.6 and for samples of size 240 for which the true lag-1 correlation is 0.3, 0.6, and 0.9. The comparison can be made fairly for these combinations of sample size and lag-1 correlation because both tests appear to operate at the nominal 5% significance level in this circumstance. The power curves were obtained from 1000 trial experiments in which departures from the null hypothesis ranging between $\sigma_x/4$ and $5\sigma_x$ were prescribed. What we see is that there is little difference between the power of the LR test (which is known to be asymptotically optimal) and the table lookup test.

Therefore, the table lookup test is an attractive competitor to the "usual" $t$ test discussed in section 3 and the LR test. A significant efficiency penalty is apparently not imposed through the use of the table lookup test. Moreover, the table lookup test operates at near the nominal significance level with smaller samples than either the "usual" test or the likelihood ratio test.

A two-sample version of the table lookup test is detailed in section 2b(2). The test is developed using the assumption that both sampled processes have the same variance and lag-1 correlation. The ingredients are virtually identical to those used in the one-sample test: the ordinary Student's $t$ statistic for the difference of means is computed and an estimate is made of the common lag-1 correlation coefficient. As in the one-sample case, the properties of the $t$ statistic depend upon the true lag-1 correlation coefficient and the number of observations used to compute the standard deviation in the denominator of the $t$ statistic. The uncertainty in the estimated lag-1 correlation coefficient also depends on the number of observations used in the estimate. Thus, the appropriate critical values are obtained by entering the critical value tables with the estimated lag-1 correlation coefficient and the sum of the two-sample sizes. Our guidelines for the use of the two-sample test are the same as those for the one-sample test except that they apply to the sum of the sample sizes.

## 6. Summary

We described the ordinary $t$ test and the usual way in which it is adapted to climatological inference problems in section 1. The usual test, which is adjusted by an estimate of the equivalent sample size, performs poorly because the equivalent sample size is poorly estimated and because it is incorrectly assumed that the adjusted statistic has a Student's $t$ distribution under the null hypothesis.

We also reiterated the observation that the equivalent sample size (or equivalently, the integral timescale) is

TABLE 10. Critical values for the table lookup test appropriate for a two- (one-) sided test conducted at the 1% (0.5%) significance level. Dashes indicate that sample correlations of this magnitude were not observed in the simulations used to create the table.

| | n | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $r_1$ | 10 | 15 | 20 | 25 | 30 | 45 | 60 | 75 | 90 | 120 | 180 | 240 |
| −0.35 | 13.0 | — | — | — | — | — | — | — | — | — | — | — |
| −0.30 | 13.3 | 6.82 | — | — | — | — | — | — | — | — | — | — |
| −0.25 | 14.0 | 7.01 | 4.70 | — | — | — | — | — | — | — | — | — |
| −0.20 | 15.1 | 7.42 | 4.83 | 3.87 | 3.58 | — | — | — | — | — | — | — |
| −0.15 | 15.5 | 7.70 | 5.12 | 3.98 | 3.58 | 3.20 | — | — | — | — | — | — |
| −0.10 | 16.2 | 8.41 | 5.26 | 4.16 | 3.69 | 3.23 | 3.02 | 2.96 | 2.97 | 2.82 | — | — |
| −0.05 | 16.2 | 9.41 | 5.45 | 4.47 | 3.80 | 3.34 | 3.20 | 2.98 | 2.94 | 2.83 | 2.80 | 2.79 |
| 0.00 | 18.0 | 10.6 | 5.91 | 4.77 | 4.01 | 3.48 | 3.24 | 3.05 | 2.99 | 2.93 | 2.82 | 2.67 |
| 0.05 | 19.0 | 11.6 | 7.22 | 5.11 | 4.21 | 3.56 | 3.20 | 3.20 | 3.13 | 3.00 | 2.92 | 2.88 |
| 0.10 | 19.2 | 13.7 | 8.28 | 5.80 | 4.77 | 3.66 | 3.43 | 3.31 | 3.16 | 2.99 | 2.95 | 2.93 |
| 0.15 | 20.5 | 14.9 | 9.44 | 6.20 | 5.27 | 3.96 | 3.66 | 3.44 | 3.37 | 3.30 | 3.16 | 3.13 |
| 0.20 | 20.6 | 15.5 | 11.0 | 7.08 | 5.37 | 4.25 | 3.93 | 3.60 | 3.47 | 3.55 | 3.29 | 3.28 |
| 0.25 | 20.5 | 16.6 | 12.1 | 8.80 | 6.17 | 4.43 | 4.04 | 3.85 | 3.70 | 3.65 | 3.50 | 3.49 |
| 0.30 | 21.2 | 18.4 | 12.7 | 10.1 | 7.51 | 5.25 | 4.57 | 4.29 | 3.93 | 3.95 | 3.76 | 3.59 |
| 0.35 | 21.5 | 19.8 | 15.9 | 12.0 | 8.80 | 5.71 | 4.63 | 4.50 | 4.26 | 3.89 | 3.90 | 3.93 |
| 0.40 | 21.8 | 20.9 | 19.7 | 14.1 | 11.3 | 6.32 | 5.44 | 4.96 | 4.58 | 4.36 | 4.25 | 4.22 |
| 0.45 | 22.3 | 23.0 | 23.0 | 17.6 | 13.6 | 7.65 | 6.20 | 5.49 | 4.92 | 4.83 | 4.61 | 4.41 |
| 0.50 | 23.3 | 23.7 | 24.9 | 19.6 | 16.4 | 9.56 | 7.13 | 6.00 | 5.49 | 5.17 | 4.90 | 4.75 |
| 0.55 | 23.3 | 25.0 | 25.7 | 22.6 | 19.2 | 12.0 | 8.75 | 6.60 | 6.25 | 5.82 | 5.56 | 5.07 |
| 0.60 | 22.9 | 25.9 | 26.3 | 26.1 | 22.0 | 14.5 | 9.83 | 8.50 | 7.01 | 6.43 | 5.63 | 5.62 |
| 0.65 | 22.1 | 26.8 | 26.3 | 28.6 | 27.6 | 20.6 | 14.4 | 9.76 | 8.43 | 7.29 | 6.43 | 5.88 |
| 0.70 | 22.0 | 26.3 | 27.8 | 30.8 | 30.4 | 26.5 | 17.8 | 12.8 | 10.7 | 8.69 | 7.72 | 7.09 |
| 0.75 | — | 24.3 | 27.3 | 29.9 | 29.7 | 29.8 | 25.0 | 20.6 | 14.1 | 10.6 | 8.97 | 7.96 |
| 0.80 | — | — | 26.8 | 29.8 | 31.3 | 33.5 | 32.0 | 28.5 | 26.8 | 15.7 | 10.7 | 9.92 |
| 0.85 | — | — | — | 29.3 | 30.9 | 36.3 | 38.4 | 41.2 | 35.1 | 26.5 | 15.8 | 13.2 |
| 0.90 | — | — | — | — | 30.8 | 33.8 | 39.1 | 42.9 | 42.9 | 41.1 | 35.8 | 21.4 |
| 0.95 | — | — | — | — | — | — | 37.6 | 39.5 | 41.4 | 46.3 | 54.3 | 61.5 |

obtained simply by asking how a measure of the difference between two samples (or a sample and reference point) should be scaled so that the difference can be expressed in units of standard deviations. Different measures of discrepancy will therefore result in different scaling factors. Thus the equivalent sample size is only one of many scaling factors and has no intrinsic physical interpretation of its own.

Next, we carefully revisited the subject of equivalent sample size estimation and examined the properties of the "usual" serial correlation compensated $t$ test in section 3. We were able to improve the ARMA $n_e$ estimator of TZ but found that this improvement had little effect on the operation of the test. We showed that the test is conservative when $n_e$ is known or subjectively estimated conservatively. We also showed that the test is liberal when $n_e$ is estimated objectively with the improved ARMA estimator. In both cases ($n_e$ "known" or objectively estimated), we found that the test has actual significance levels that are practically indistinguishable from the nominal levels when the true $n_e > 30$.

We then considered two competing tests of the mean: the LR test and a table lookup test. Both are computationally intensive. The LR ratio test requires numerical minimization of a complex quadratic form while the table lookup test requires extensive simulation to develop the conditional reference distributions that are used in that test. The LR test, which is an asymptotic test, operates at significance levels that are similar to the nominal levels when $n_e > 15$. The table lookup test operates at significance levels that are close to the nominal levels for all but the smallest sample sizes tested. Moreover, the power of the two tests is virtually indistinguishable for samples large enough so that the actual significance level of the LR test is equal to its nominal level.

Comparison and analysis of the three competing tests was performed in the one-sample setting in which comparisons are made between a sample mean and a fixed standard. The more usual setting requires the comparison of the means of two samples. Procedures for the application of the three tests in both the one- and two-sample setting were detailed in section 2. Recommendations for when to use the various tests were also given in section 2.

The working assumption in this paper has been that the stochastic behavior of the atmosphere can be approximated by a Gaussian autoregressive process of order 1 [AR(1) process—see (1)]. The Gaussian part of the assumption is important. The tests discussed above may be compromised to a considerable extent if applied to non-Gaussian data (such as daily precipitation accumulations). For processes that are Gauss-

ian, or nearly so (such as most thermodynamic variables of the free atmosphere that are not affected by long timescale quasi-periodic signals) the details of the stochastic behavior may be somewhat more complex than that of an AR(1) process without compromising the tests. Most such processes will exhibit power spectra that have maxima at the origin and have decreasing power with increasing frequency. Such spectra can be reasonably well approximated by AR(1) processes.

Finally, we reiterate a point made strongly in sections 1 and 2. The assumptions required to conduct a given statistical test impose a statistical model upon the observations. The reliability of the test will be compromised if the model does not have reasonable fidelity. It is therefore important to explore the validity of the assumptions before conducting the statistical test.

## APPENDIX

### *T* test with $n_e$ Known

We present corroborating evidence in this appendix that the $t$ test with $n_e$ known is conservative when data come from a Gaussian AR(1) processes with positive lag-1 correlation coefficient. For simplicity, we consider only the one-sample problem. We assume, without loss of generality, that the AR(1) process has zero mean and unit variance. Then we begin by writing the one-sample $t$ statistic with $n_e$ known as

$$t = \frac{\bar{x}(n_e)^{1/2}}{s} = \frac{\bar{x}(n_e)^{1/2}}{\left[\sum_{t=1}^{n} (x_t - \bar{x})^2/(n-1)\right]^{1/2}}$$

$$= \frac{N(x_1, \ldots, x_n)}{D(x_1, \ldots, x_n)}.$$

Next, we write

$$F = t^2 = N^2/D^2,$$

suppressing the arguments of $N(x_1, \ldots, x_n)$ and $D(x_1, \ldots, x_n)$ for clarity.

When observations are independent and identically distributed Gaussian, it can be shown that $N^2$ is a $\chi^2$ random variable with 1 df, that $N^2$ and $D^2$ are independent, and that $(n-1)D^2$ is a $\chi^2$ random variable with $n-1$ df. Consequently, $F$ has an $F$ distribution with $(1, n-1)$ df.

The derivation of this result (or the equivalent result that $t$ is distributed as Student's $t$ with $n-1$ df) depends critically upon the assumption of the independence of the observations.

The $t$ test with $n_e$ known attempts to accommodate serial correlation by comparing realizations of $t$ against critical values for the Student's $t$ distribution with $n_e - 1$ df. When the test is two-sided, this is equivalent to comparing realizations of $F$ to the critical values of the $F$ distribution with $(1, n_e - 1)$ df. Implicitly then, this accommodation for serial correlation attempts to approximate the true distribution of $F$ with that of a ratio of independent $\chi^2$ random variables with 1 and $n_e - 1$ df, respectively.

The approximation will work reasonably well if

(i) $N^2$ has a $\chi^2$ distribution with 1 df,
(ii) $N^2$ and $D^2$ are independent of each other, and
(iii) $(n_e - 1)D^2$ has a distribution that is well approximated by the $\chi^2$ distribution with $n_e - 1$ df.

We examine the extent to which this approximation works.

The first condition is completely satisfied. Here $N$ is Gaussian with mean 0 and variance 1 because the sample mean $\bar{x}$ is a linear combination of zero mean Gaussian random variables and because the factor $(n_e)^{1/2}$ is defined as the constant that scales the variance of $\bar{x}$ to unity. It therefore follows that $N^2$ is a $\chi^2$ random variable with 1 df.

The second condition is not satisfied. However, we will demonstrate below that $N^2$ and $D^2$ are not strongly related, at least when sample sizes are small. It is our opinion that the weak dependence between $N^2$ and $D^2$ is not an important factor in developing an understanding of why the distribution of $F$ is not well approximated by the $F$ distribution with $(1, n_e - 1)$ df.

The third condition is also not satisfied. We will demonstrate that the mean of $(n_e - 1)D^2$ is very close to that of a $\chi^2$ random variable with $n_e - 1$ df, but that the variance of $(n_e - 1)D^2$ is considerably less than anticipated. It follows that extreme values of $F$ (or equivalently $t$) will occur less frequently than predicted by the approximation and therefore that the $F$ or $t$ test with $n_e$ known will be conservative.

We begin by re-expressing $N^2$ and $D^2$ as quadratic forms in vector-matrix notation. Let $X = (x_1, \ldots, x_n)^t$ and let $e = (1/n, \ldots, 1/n)^t$. Then

$$N^2 = n_e\left(\frac{1}{n}\sum_{t=1}^{n} x_t\right)^2 = n_e(X^t e)^2$$

$$= n_e X^t(ee^t)X = n_e X^t A X,$$

where $A$ is the $n \times n$ array $ee^t$. Similarly, $D^2$ may be written as

$$D^2 = \frac{1}{(n-1)}\sum_{t=1}^{n} (x_t - \bar{x})^2 = \frac{1}{(n-1)} X^t(I - A)X.$$

Note that $X$ has a multivariate Gaussian distribution with mean 0 and variance–covariance matrix $\Sigma$. The $(i, j)$ element of $\Sigma$ is given by $\sigma_{i,j} = \rho_1^{|i-j|}$, where $\rho_1$ is the lag-1 correlation coefficient of the AR(1) process generating the data. Graybill (1983, Theorem 10.9.10) provides results concerning the variances and covariances of quadratic forms. In particular, if $X$ is an $n$-dimensional zero mean Gaussian random vector with variance–covariance matrix $\Sigma$, and if $C$ and $D$ are any symmetric $n \times n$ matrices of constants, then

$$\text{cov}(X^tCX, X^tDX) = 2 \text{ tr } (C\Sigma D\Sigma) \quad \text{(A.1)}$$

$$\text{var}(X^tCX) = 2 \text{ tr } (C\Sigma C\Sigma), \quad \text{(A.2)}$$

where tr denotes the trace of a matrix.

Equations (A.1) and (A.2) can be applied to obtain general expressions for the variances of $N^2$ and $D^2$ and their correlation. In practice, these expressions, complicated ratios of polynomials in $\rho_1$, are virtually impossible to derive reliably by hand. We therefore used a symbolic manipulator [*Maple V* (Char et al. 1992)] to derive expressions for specific sample sizes that are general in $\rho_1$ and used these expressions to make inferences about the behavior of $N^2$ and $D^2$. Space limitations prevent us from displaying these expressions here.

Calculations of the correlation between $N^2$ and $D^2$ show that they are not independent but also not strongly correlated, at least for the moderate sample sizes ($n$ = 5, 10, 20, 30) considered. The correlation is 0 when $\rho_1 = 0$, grows to a maximum at a value of $\rho_1$ near 1, and then falls to 0 as $\rho_1$ approaches 1. The value of the maximum increases slowly with sample size and its location approaches 1 with increasing sample size. For $n = 30$, the maximum correlation between $N^2$ and $D^2$ is about 0.038 and occurs at $\rho_1 \approx 0.95$. Dependence between $N^2$ and $D^2$ would not appear to have a substantial effect on the behavior of the $F$ ratio under serial correlation.

It can easily be shown that the mean of $(n_e - 1)D^2$ is given by

$$E[(n_e - 1)D^2] = \frac{n_e - 1}{n - 1} \text{ tr } [(I - A)\Sigma]$$

$$= (n_e - 1)\left(\frac{n}{n - 1}\frac{n_e - 1}{n_e}\right)$$

Thus $(n_e - 1)D^2$ has only a small negative bias as an estimator of the mean of a $\chi^2$ random variable with $n_e - 1$ df.[7] However, calculations of the variance of $(n_e - 1)D^2$ for samples sizes $n$ = 5, 10, 20, 30 indicate that the variance is substantially less than that of a $\chi^2$ random variable with $n_e - 1$ df except when $\rho_1 = 0$.[8] For the sample sizes considered, $\text{Var}[(n_e - 1)D^2]/(2n_e - 2)$ decreases monotonically from 1 to 0 as $\rho_1$ varies from 0 to 1. We conjecture that this is generally true for all sample sizes.

*Acknowledgments.* Hans von Storch contributed to this work during a visit to the Canadian Climate Centre in 1992 and thanks the Centre for its financial support of the visit. Both authors are appreciative of comments and discussion provided by George Boer, Bob Livezey, Jean Thiebaux, and an anonymous reviewer.

[7] The mean of a $\chi^2$ random variable with $\nu$ df is $\nu$.

[8] The variance of a $\chi^2$ random variable with $\nu$ df is $2\nu$.

REFERENCES

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

Ansley, C. F., 1979: An algorithm for the exact likelihood of a mixed autoregressive-moving average process. *Biometrika,* **66,** 59–65.

Becker, R. A., J. M. Chambers, and A. R. Wilks, 1988: *The New S Language: A Programming Environment for Data Analysis and Graphics.* Wadsworth, 702 pp.

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

Breiman, L., 1973: *Statistics with a View Towards Applications.* Houghton Mifflin, 399 pp.

Char, B. W., K. O. Geddes, G. H. Gonnet, B. L. Leong, and M. B. Monagan, 1992: *First Leaves: A Tutorial Introduction to Maple V.* Springer-Verlag, 253 pp.

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.

Cox, D. R., and D. V. Hinkley, 1974: *Theoretical Statistics.* Chapman and Hall, 511 pp.

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

Graybill, F. A., 1983: *Matrices with Applications in Statistics.* 2d ed., Wadsworth, 461 pp.

Hasselmann, K., 1976: Stochastic climate models. Part I. Theory. *Tellus,* **28,** 474–485.

Jones, R. H., 1975: Estimating the variance of time averages. *J. Appl. Meteor.,* **15,** 159–163.

——, 1980: Maximum likelihood fitting of ARMA models to time series with missing observations. *Technometrics,* **22,** 389–395.

Kabaila, P., and G. Nelson, 1985: On confidence regions for the mean of a multivariate time series. *Commun. Statist.-Theor. Meth.,* **14,** 735–753.

Kalbfleisch, J. G., 1979: *Probability and Statistical Inference II.* Springer-Verlag, 316 pp.

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

Laurmann, J. A., and W. L. Gates, 1974: Statistical considerations in the evaluation of climatic experiments with atmospheric general circulation models. *J. Atmos. Sci.,* **34,** 1187–1199.

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

Splus, 1992: Splus Version 3.1. Statistical Sciences Inc., Seattle.

Sutrahdar, B. C., I. B. MacNeill, and H. F. Sahrmann, 1987: Time series valued experimental designs: One-way analysis of variance with autocorrelated errors. *Time Series and Econometric Modelling,* I. B. MacNeill and G. J. Umphrey, Eds., D. Reidel, 113–129.

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

Tubbs, J. D., 1980: The effect of serial correlation on confidence regions for the parameters of a multivariate normal population. *Commun. Statist.-Theor. Meth.,* **9,** 1341–1351.

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