Statistical Tests for Comparison of Daily Variability in Observed and Simulated Climates

T. ADRI BUISHAND AND JULES J. BEERSMA

Royal Netherlands Meteorological Institute (KNMI), De Bilt, The Netherlands

(Manuscript received 17 February 1995, in final form 1 April 1996)

ABSTRACT

Tests for differences in daily variability based on the jackknife are presented. These tests properly account for the effect of autocorrelation in the data and are reasonably robust against departures from normality. Three measures for the daily variability are considered: process, within-month, and innovation variance. The jackknife statistic compares the logarithm of these measures. The standard errors of this logarithm are obtained by recomputing the variance estimates for all subsamples wherein one month is omitted from the complete sample. A simple extension of the jackknife procedure is given to obtain a powerful multivariate test in situations that the differences in variance have the same sign across the region considered or over the year.

As an illustration the tests are applied to near-surface temperatures over Europe simulated by the coupled ECHAM/LSG model. It is shown that the control run of the model significantly overestimates the process variance in winter and spring and the within-month variance in all seasons. Significant differences are also found for the innovation variances of the daily temperatures, but the sign of the differences varies over the year. In a perturbed run with enhanced atmospheric greenhouse gas concentrations the daily temperature variability over Europe significantly decreases in winter and spring compared with the control run.

1. Introduction

Variability refers to the magnitude of the fluctuations about an average condition. For impact assessment of climate change on society the potential changes in variability are considered at least as important as changes of the mean. The frequency of extreme events, for example, can be more sensitive to changes in the variability than changes in the mean (Katz 1992). Validation of the variability within general circulation models (GCMs), which are the "state of the art" tools for predicting anthropogenic climate change, is despite its importance somewhat neglected. Knowledge of the deficiencies with respect to variability can be used to improve GCMs. This may in turn result in a further reduction of the uncertainties about anthropogenic climate change.

Despite the large interest of the impacts community in the potential changes of the characteristics of daily values, the number of studies comparing the daily variability of (GCM simulated) time series is rather limited. In most of these studies different approaches are followed. Measures that are alternatively taken for the daily variability are the process variance (the variation of daily values about the long-term mean) and the within-month or intramonthly variance (the variation of daily values about the individual monthly means). The process variance contains daily variability as well as variability on longer timescales.

Daily values of atmospheric variables are autocorrelated and their distribution generally deviates from the normal distribution. The basic assumptions for the traditional F test, independent and normally distributed data, are thus violated. Unfortunately, the results of the test are very sensitive to departures from these assumptions. Wilson and Mitchell (1987) and Mearns et al. (1990) removed the autocorrelation from the time series by a linear filter as advocated by Katz (1988). The test is then based on the variance of the filtered data, which is known as the innovation variance. A standardized fourth moment is also introduced in that method to account for nonnormality. Rind et al. (1989) and Cao et al. (1992) considered the within-month variance. The first authors applied an F test with a reduced number of degrees of freedom based on the persistence in the data. They did not investigate the performance of their modification of the F test nor did they consider the effect of nonnormality. Cao et al. did not give details about statistical tests.

The above tests refer to data at a single location. GCM data consist, however, of a large number of time series on a spatial grid. Wigley and Santer (1990) presented a number of tests to compare monthly variances of such multivariate data. Computer intensive resampling techniques are, however, required to decide whether a result is significant or not. Multivariate tests for daily variability have not been considered yet in the climatological literature.

Corresponding author address: Dr. T. Adri Buishand, Royal Netherlands Meteorological Institute, P.O. Box 201, 3730 AE De Bilt, The Netherlands.

This paper focuses on jackknife tests for comparison of daily variability. The critical values of these tests can generally be based on Student's t-distribution. The method bears strong resemblance with the use of the jackknife in Buishand and Beersma (1993) for a test on the autocorrelation coefficients. The jackknife procedure properly accounts for the effect of the autocorrelation on the standard error of variance estimates and the tests are reasonably robust against departures from normality. The jackknife not only provides a test for the variances of two different climate time series at a single location but is also a useful tool for implementing a simple multivariate test statistic whose critical values can be derived from the t distribution rather than empirically evaluating the statistical significance with a permutation or Monte Carlo method.

The paper is organized as follows. In section 2 the jackknife is introduced to obtain estimates of the logarithm of the process variance and its standard error. This leads to a simple statistic for testing for equality of process variances. In section 3 similar statistics are presented for comparing the within-month variances. Section 4 deals with the multivariate extension of the jackknife procedure. In section 5 the test of Katz (1988) on the innovation variances is reviewed and alternatives based on the jackknife are proposed, which make a multivariate extension possible. The use of the tests is illustrated in section 6 with near-surface temperature data over Europe from the analyses of the European Centre for Medium-Range Weather Forecasts (ECMWF) and two simulations of a coupled global atmosphere–ocean model (ECHAM/LSG) developed jointly by the Max-Planck-Institut für Meteorologie in Hamburg and the Meteorologisches Institut der Universität Hamburg.

2. A jackknife test for process variances

The mean μ and the process variance σ^2 of an atmospheric variable X generally depend on the time of the year. Estimates of these parameters are therefore generally presented on a monthly or seasonal basis. In this study all years of a certain calendar month are pooled to get an individual estimate for each of the 12 different calendar months first. For the pooling, a rather short period of one month is chosen because otherwise the variance estimate could be seriously biased as a result of the annual cycle in the mean. The monthly estimates are combined to seasonal and annual estimates in a later stage.

Let $x_{i,j}$ denote the value of X on day i ($i = 1, \cdots, n$) in year j ($j = 1, \cdots, J$) of the calendar month under consideration. Then the sample estimates $\bar{x}$ and s^2 of μ and σ^2 are given by

$$\bar{x} = \frac{1}{nJ}\sum_{i=1}^{n}\sum_{j=1}^{J} x_{i,j}, \quad s^2 = \frac{1}{nJ}\sum_{i=1}^{n}\sum_{j=1}^{J}(x_{i,j} - \bar{x})^2. \quad (1)$$

The estimate s^2 will be denoted as sample process variance.

In contrast with the F test, which is based on the ratio of sample variances, jackknife tests consider the differences of their logarithms (Miller 1968). This has two reasons: First, the distribution of $\hat{\theta} = \ln(s^2)$ is closer to the normal distribution than that of s^2 and, second, the variance of $\hat{\theta}$ is practically independent of the true process variance σ^2. There is, however, not a simple analytical expression for $\mathrm{var}(\hat{\theta})$ that properly accounts for the effects of autocorrelation and nonnormality. The jackknife provides a distribution-free estimate of $\mathrm{var}(\hat{\theta})$ by recomputing $\hat{\theta}$ a large number of times after successive deletion of one observation or a group of observations from the entire dataset (Efron 1982). The jackknife method requires that the data in different groups are independent. To accomplish this requirement a complete month is deleted each time.

For the description of jackknife tests it is convenient to introduce the pseudovalues θ_j^*:

$$\theta_j^* = \hat{\theta} + (J - 1)(\hat{\theta} - \hat{\theta}_{-j}), \quad (2)$$

with $\hat{\theta}_{-j}$ the estimate of $\theta = \ln(\sigma^2)$ after omitting the data for year j in the pooled sample of the calendar month under consideration. The term $(J - 1)(\hat{\theta} - \hat{\theta}_{-j})$ is mainly determined by the data in year j and can be regarded as a measure of the influence of these data on the estimate $\hat{\theta}$ (Hinkley 1983). The key idea behind a jackknife test is that the pseudovalues may often be treated as independent normal variables. This holds in particular when a group of observations is deleted each time. The assumptions underlying a jackknife test are generally also better satisfied when the number J of years is large. It is therefore necessary to check the validity of the test for small and moderate values of J by a simulation study.

The mean of the pseudovalues θ_j^* provides a bias-adjusted estimate of θ, which is known as the jackknife estimate:

$$\hat{\theta}_{\mathrm{jack}} = \frac{1}{J}\sum_{j=1}^{J}\theta_j^*. \quad (3)$$

The jackknife variance

$$\hat{V}_{\mathrm{jack}} = \frac{1}{J(J-1)}\sum_{j=1}^{J}(\theta_j^* - \hat{\theta}_{\mathrm{jack}})^2 \quad (4)$$

estimates both $\mathrm{var}(\hat{\theta})$ and $\mathrm{var}(\hat{\theta}_{\mathrm{jack}})$. Tests for equality of variances are therefore often based on the values of $\hat{\theta}_{\mathrm{jack}}$ rather than $\hat{\theta}$.

A number of Monte Carlo experiments were performed to judge the quality of the estimate $\hat{V}_{\mathrm{jack}}$. Different autocorrelation structures were considered by generating normal first-order AR(1) and second-order AR(2) autoregressive processes. The effect of nonnormality was examined by generating Laplace and exponential AR(1) processes as described by Lawrance

(1980). For generated 10-year samples Table 1 compares the mean of $\hat{V}_{\text{jack}}^{1/2}$ with the standard deviations $\sigma(\hat{\theta}_{\text{jack}})$ and $\sigma(\hat{\theta})$ of $\hat{\theta}_{\text{jack}}$ and $\hat{\theta}$, respectively. The statistic $\hat{V}_{\text{jack}}^{1/2}$ yields an almost unbiased estimate of $\sigma(\hat{\theta})$ and slightly underestimates $\sigma(\hat{\theta}_{\text{jack}})$. Different values for the lag 1 and lag 2 autocorrelation coefficients ρ_1 and ρ_2 are considered in Table 1. For the AR(1) processes $\rho_2 = \rho_1^2$, whereas for the given AR(2) process $\rho_2 < \rho_1^2$. For the normal AR processes in Table 1 the standard deviation of $\hat{\theta}$ increases with the strength of the autocorrelation. It is also seen from the table that $\sigma(\hat{\theta})$ is much larger when the data follow a Laplace or exponential distribution instead of a normal distribution. An important quantity that causes these differences in $\sigma(\hat{\theta})$ is the kurtosis γ_2 of the distribution of X. The kurtosis is usually defined as the standardized fourth cumulant:

$$\gamma_2 = \frac{E[(X-\mu)^4]}{\sigma^4} - 3. \tag{5}$$

For the normal distribution γ_2 is zero. Kurtosis tends to be positive (negative) for distributions with larger (smaller) relative frequencies in one or both tails and for more sharply peaked (flat topped) distributions than the normal distribution. For the Laplace and exponential distributions γ_2 is 3 and 6 respectively. From Table 1 it is seen that $\sigma(\hat{\theta})$ increases with γ_2.

A test for equality of the process variances $\sigma^2(\text{I})$ and $\sigma^2(\text{II})$ of two mutually independent climate time series, for example, observations over two different time periods or data from a GCM control run and a perturbed run, can be constructed on the basis of Eqs. (3) and (4). A suitable test statistic is

$$T = \frac{\hat{\theta}_{\text{jack}}(\text{II}) - \hat{\theta}_{\text{jack}}(\text{I})}{[\hat{V}_{\text{jack}}(\text{I}) + \hat{V}_{\text{jack}}(\text{II})]^{1/2}}, \tag{6}$$

with $\hat{\theta}_{\text{jack}}(\text{I})$ and $\hat{\theta}_{\text{jack}}(\text{II})$ the jackknife estimates of $\ln(\sigma^2)$ for climate I and climate II, respectively, and $\hat{V}_{\text{jack}}(\text{I})$, $\hat{V}_{\text{jack}}(\text{II})$ the jackknife variances of $\ln(s^2)$ for these two climates. The critical values of the test statistic can be derived from Student's t-distribution, with an effective number of degrees of freedom given by

$$d = \frac{[\hat{V}_{\text{jack}}(\text{I}) + \hat{V}_{\text{jack}}(\text{II})]^2}{\hat{V}_{\text{jack}}^2(\text{I})/(J-1) + \hat{V}_{\text{jack}}^2(\text{II})/(K-1)}, \tag{7}$$

with J and K the number of years for climate I and climate II, respectively. The value of d cannot be larger than the degrees of freedom $J + K - 2$ in the classical Student's t-test (Scheffé 1970). The test is similar to that of Welch (1938) for comparing means of two normal populations with possible unequal variances. This famous Behrens–Fisher problem comes up in tests for equality of process variances when the two climate time series have different kurtosis or autocorrelation coefficients. To allow for this possibility the expression for d differs somewhat from that in Buishand and Beersma (1993) for a jackknife test for equality of the lag 1 autocorrelation coefficients. The Monte Carlo results in appendix A show that the proposed t distribution works quite well. The test is reasonably robust against nonnormality. Some discrepancies between the empirical and nominal significance levels occur when a very short record (~5 years) is tested against a longer record, in particular when autocorrelation is stronger in the shorter record or when this record has a higher kurtosis than the longer record.

TABLE 1. Standard deviations of $\hat{\theta} = \ln(s^2)$ and $\hat{\theta}_{\text{jack}}$ (see main text) and the mean of the estimate $\hat{V}_{\text{jack}}^{1/2}$ of $\sigma(\hat{\theta})$ for different AR processes in a series of Monte Carlo experiments ($J = 10$, $n = 30$, $\sigma^2 = 1$, 5000 simulations): N refers to the normal distribution, L to the Laplace distribution, and E to the exponential distribution.

Process	Autocorrelation ρ_1	ρ_2	$\sigma(\hat{\theta})$	$\sigma(\hat{\theta}_{\text{jack}})$	$E(\hat{V}_{\text{jack}}^{1/2})$
N AR(1)	.50	.25	0.106	0.107	0.102
N AR(1)	.80	.64	0.169	0.176	0.165
N AR(2)	.80	.45	0.138	0.138	0.133
L AR(1)	.80	.64	0.261	0.270	0.256
E AR(1)	.80	.64	0.333	0.355	0.315

For noninteger d the critical values of the test can be obtained numerically (Gardiner and Bombay 1965; Bukač and Burstein 1980) or by interpolation in a table of Student's t-distribution. With the exception of very small sample sizes the critical values from this approximate solution are almost identical to those in the Welch–Aspin test (Lee and Gurland 1975).

3. Jackknife tests for within-month variances

The sample process variance in Eq. (1) can be decomposed as follows:

$$s^2 = \frac{1}{J}\sum_{j=1}^{J} s_{w,j}^2 + s_b^2. \tag{8}$$

In this decomposition $s_{w,j}^2$ represents the variation of the daily values within the calendar month under consideration in year j ($j = 1, \cdots, J$) and s_b^2 the interannual variation between the monthly means of the considered calendar month. These variance components are defined by

$$s_{w,j}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{i,j} - \bar{x}_{\cdot j})^2, \quad s_b^2 = \frac{1}{J}\sum_{j=1}^{J}(\bar{x}_{\cdot j} - \bar{x})^2, \tag{9}$$

where

$$\bar{x}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j} \tag{10}$$

are the individual monthly averages.

In the climatological literature the within-month variances $s_{w,j}^2$ are more popular than the sample process variance s^2. The within-month variance depends on the

autocorrelation of the process. A strong autocorrelation implies that the variation of the daily values within a month will be small. For an AR(1) process the mean of $s^2_{w,j}$ can be approximated as (appendix B)

$$E(s^2_{w,j}) \approx \sigma^2\left(1 - \frac{1}{n}\frac{1+\rho_1}{1-\rho_1}\right). \qquad (11)$$

With $n = 30$ it follows from this approximation that $E(s^2_{w,j}) \approx 0.7\sigma^2$ if $\rho_1 = 0.8$. Even for this rather strong autocorrelation, the within-month variance is still the dominant term in Eq. (8).

A test similar to that in the previous section is obtained by applying the jackknife procedure to the logarithm $\hat{\theta}_w$ of the average of the within-month variances:

$$\hat{\theta}_w = \ln\left(\frac{1}{J}\sum_{j=1}^{J} s^2_{w,j}\right). \qquad (12)$$

As an alternative one might consider the average θ_w of the logarithms of the within-month variances:

$$\bar{\theta}_w = \frac{1}{J}\sum_{j=1}^{J} \hat{\theta}_{w,j}, \qquad (13)$$

with $\hat{\theta}_{w,j} = \ln(s^2_{w,j})$. An estimate $\hat{V}_w$ of the variance of $\bar{\theta}_w$ can be obtained from the sample variance of the $\hat{\theta}_{w,j}$:

$$\hat{V}_w = \frac{1}{J(J-1)}\sum_{j=1}^{J} (\hat{\theta}_{w,j} - \bar{\theta}_w)^2. \qquad (14)$$

For this alternative it is easily verified that the pseudovalues θ^*_j in the jackknife procedure are equal to $\hat{\theta}_{w,j}$ and that $\hat{V}_w$ is the jackknife estimate of $\mathrm{var}(\bar{\theta}_w)$. The test for equality of within-month variances is based on the statistic:

$$T_w = \frac{\bar{\theta}_w(\mathrm{II}) - \bar{\theta}_w(\mathrm{I})}{[\hat{V}_w(\mathrm{I}) + \hat{V}_w(\mathrm{II})]^{1/2}}. \qquad (15)$$

This test is identical to Welch's (1938) test for comparing the means of two normal populations with possible unequal variances. The test is more robust against nonnormality than the jackknife test, which considers $\hat{\theta}_w$ in Eq. (11). It is, however, more sensitive to differences in kurtosis. Details are given in appendix C.

4. Multivariate extensions

The jackknife tests in the previous sections have been developed for data in a specific calendar month at a given location. Such tests can be applied simultaneously to a number of successive calendar months and to various points on a grid. An important problem is then the determination of the global significance of these tests. When the tests are independent the number of times that the null hypothesis is rejected can be compared with that expected from the binomial distribution (Livezey and Chen 1983). As an alternative a chi-square test can be done on the sum of the logarithms of the observed significance levels of the individual tests (Sneyers 1990). The independence assumption is, however, strongly violated for different climate time series on a grid as in GCM output. Computationally intensive permutation techniques are then required to determine the global significance (Wigley and Santer 1990).

The use of a single multivariate test statistic avoids the above multiplicity problem. The data for the individual (calendar) months can be combined into a single seasonal (e.g., DJF) or annual test statistic as follows. First, the pseudovalues $\theta^*_1, \theta^*_2, \ldots, \theta^*_J$ are calculated for each month separately. These pseudovalues are then averaged over the various months, giving $\overline{\theta^*_1}, \overline{\theta^*_2}, \ldots, \overline{\theta^*_J}$. The jackknife statistics T and T_w are finally obtained by applying Eqs. (3), (4), and (14) to these average pseudovalues. The use of Student's t-distribution as an approximation to the null distribution of T and T_w requires that these pseudovalues behave like independent normal variables. A beneficial result of averaging is that the distribution of the $\overline{\theta^*_j}$ will be closer to the normal distribution than that of the pseudovalues for an individual calendar month. Dependence is avoided because for each $\overline{\theta^*_j}$ a different season or year is deleted. A small difficulty arises when the whole year is taken into account because then the group of deleted observations contains a number of days that are directly adjacent to those left in. It is unlikely, however, that this will impair the use of the jackknife procedure because only a relatively small number of the daily values in two subsequent years are significantly correlated.

A combined test using average pseudovalues of various calendar months is more powerful than that for an individual calendar month when the sign of the differences in variances for the climates under consideration is the same in these months because of the larger sample size in the combined test. The combined test may, however, fail when the sign of the differences varies over the year. The derivation of a global significance level from the binomial distribution or the chi-square statistic in Sneyers (1990) does not have this disadvantage, but these methods require that the data in successive months are independent.

The data at several grid points in a region can be combined by a similar averaging procedure. The test then compares spatial averages of the logarithm of time variances. The resulting test is a simple multivariate test that does not require permutation techniques. The test will be most powerful when the differences between climate I and climate II have the same sign across the whole region. The averaging procedure is not suitable for very large regions (e.g., the whole globe or a hemisphere) where areas with negative differences may compensate those with positive differences. An alternative test would have been obtained by comparing the variances of the daily spatial averages of the data. Climate change impact assessment is, however, usually

concerned with the variances of local data. Changes in these local variances might be different from those in the variances of the spatial averages because the latter also depend on the spatial correlation. Variances of spatial averages are therefore not considered further.

The multivariate test compares the averages of the logarithm of a variance measure. This is equivalent with a comparison of the geometric mean rather than the arithmetic mean. It is possible that the differences between the geometric means conflict with those in the arithmetic means. Such a case is considered in the examples in section 6. An alternative would be to compare the averages of the variances without taking the logarithm first. There is less need for taking logarithms in the multivariate extension because temporal and spatial averaging already leads to practically normally distributed pseudovalues in the jackknife procedure.

5. Application of the jackknife to innovation variances

In the literature on time series analysis the variance of an underlying uncorrelated process, called the innovation variance, is often considered instead of the variance of the original process. A linear filter is used to obtain the innovations $a_{i,j}$ from the daily values $x_{i,j}$. This operation is sometimes called "prewhitening," with "white" referring to a purely random or white-noise process being produced. Davis (1977) studied robust methods for setting confidence intervals for the variances of the innovations. Significance tests for an abrupt change in the innovation variances were studied in a later paper (Davis 1979). The emphasis was on asymptotic properties of robust techniques. It was Katz (1988) who recognized the value of this work for statistical inference about climate variability.

In Katz (1988) it is assumed that the climate time series can be represented by an autoregressive process of order $p (p \geq 1)$, denoted by AR(p). The prewhitening filter then takes the form:

$$a_{i,j} = (x_{i,j} - \bar{x}) - \sum_{k=1}^{p} \phi_k (x_{i-k,j} - \bar{x}),$$

$$i = p + 1, \cdots, n; \quad j = 1, \cdots, J. \tag{16}$$

The use of $x_{i-k,j} = \bar{x}$ if $i \leq k$, as is sometimes proposed in the literature, can lead to anomalous values of $a_{1,j}$, $\ldots$, $a_{p,j}$. In contrast with Katz (1988) the $a_{i,j}$s for $i \leq p$ are therefore not considered. The autoregression coefficients ϕ_k, $k = 1, \cdots, p$ determine the autocorrelation coefficients ρ_k, $k > 0$ via the Yule–Walker equations:

$$\rho_k = \sum_{j=1}^{p} \phi_j \rho_{|k-j|}, \quad k = 1, 2, \cdots. \tag{17}$$

Estimates of the ϕ_k can be obtained from the first p estimated autocorrelation coefficients by solving Eq. (17) for $k = 1, \cdots, p$. In the examples in section 6 ρ_k is estimated as (Buishand and Beersma 1993):

$$r_k = \frac{1}{(n-k)J} \sum_{i=1}^{n-k} \sum_{j=1}^{J} (x_{i,j} - \bar{x})(x_{i+k,j} - \bar{x})/s^2,$$

$$k = 1, 2, \cdots, \tag{18}$$

where s^2 is the sample process variance given by Eq. (1).

For an AR(p) process the variance σ_a^2 of the innovations is related to the process variance σ^2 as (Box and Jenkins 1976)

$$\sigma_a^2 = (1 - R^2)\sigma^2, \tag{19}$$

where

$$R^2 = \rho_1\phi_1 + \rho_2\phi_2 + \cdots + \rho_p\phi_p. \tag{20}$$

The quantity R is comparable to the multiple correlation coefficient in a linear regression. For an AR(1) process Eq. (20) reduces to $R^2 = \rho_1^2$, from which it follows that the innovation variance is only about one-third of the process variance if $\rho_1 = 0.8$.

The average of the innovations in Eq. (16) is approximately zero. The innovation variance can therefore be estimated as

$$s_a^2 = \frac{1}{J^*} \sum_{i=p+1}^{n} \sum_{j=1}^{J} a_{i,j}^2, \tag{21}$$

with $J^* = (n - p)J$. Again the test for equality of variances is based on the logarithms of their estimates. Assuming that the $a_{i,j}$ are independent, the following approximation holds for the variance of $\hat{\theta}_a = \ln(s_a^2)$:

$$\text{var}(\hat{\theta}_a) = (2 + \gamma_2)/J^*, \tag{22}$$

where γ_2 is the kurtosis of the $a_{i,j}$. In Katz (1988) an estimate $\hat{V}_{\text{Katz}}$ of this variance is obtained by replacing γ_2 in Eq. (22) by the sample kurtosis:

$$\hat{\gamma}_2 = \frac{\sum_{i=p+1}^{n} \sum_{j=1}^{J} a_{i,j}^4}{J^* s_a^4} - 3. \tag{23}$$

A test for equality of the innovation variances of climate I and climate II can be based on the statistic:

$$Z = \frac{\hat{\theta}_a(\text{II}) - \hat{\theta}_a(\text{I})}{[\hat{V}_{\text{Katz}}(\text{I}) + \hat{V}_{\text{Katz}}(\text{II})]^{1/2}}. \tag{24}$$

Under the null hypothesis $[\sigma_a^2(\text{I}) = \sigma_a^2(\text{II})]$ the distribution of this statistic is approximately standard normal for large sample sizes.

As before, the jackknife procedure can be used to obtain estimates of $\theta_a = \ln(\sigma_a^2)$ and $\text{var}(\hat{\theta}_a)$. This alternative was already studied by Davis (1977, 1979). An advantage of the jackknife is that it allows for multivariate extensions as discussed in section 4. Here both the jackknife estimates $\hat{\theta}_{\text{jack1}}$ and $\hat{V}_{\text{jack1}}$ of θ_a and $\text{var}(\hat{\theta}_a)$ based on one-day deletion and the jackknife estimates $\hat{\theta}_{\text{jack30}}$ and $\hat{V}_{\text{jack30}}$ based on one-month deletion

TABLE 2. Standard deviations of $\hat{\theta} = \ln(s^2)$ and $\hat{\theta}_a = \ln(s_a^2)$ and the means of the estimates $\hat{V}^{1/2}_{\text{Katz}}$, $\hat{V}^{1/2}_{\text{jack1}}$, and $\hat{V}^{1/2}_{\text{jack30}}$ of $\sigma(\hat{\theta}_a)$ for different AR processes in a series of Monte Carlo experiments ($J = 10$, $n = 30$, 5000 simulations). The lag 1 autocorrelation coefficient ρ_1 is always 0.80; for the AR(1) processes $\rho_2 = 0.64$ and for the AR(2) process $\rho_2 = 0.45$. The coefficients ϕ_1 and ϕ_2 determine the prewhitening filter in Eq. (16): *N* refers to the normal distribution, *L* to the Laplace distribution, and *E* to the exponential distribution.

Process	ϕ_1	ϕ_2	$\sigma(\hat{\theta})$	$\sigma(\hat{\theta}_a)$	$E(\hat{V}^{1/2}_{\text{Katz}})$	$E(\hat{V}^{1/2}_{\text{jack1}})$	$E(\hat{V}^{1/2}_{\text{jack30}})$
N AR(1)	0.80	—	0.169	0.083	0.082	0.083	0.081
	0.60	—	0.169	0.097	0.082	0.083	0.092
N AR(2)	1.22	−0.53	0.138	0.085	0.084	0.085	0.082
	0.80	0	0.138	0.110	0.083	0.084	0.107
L AR(1)	0.80	—	0.267	0.236	0.216	0.229	0.233
E AR(1)	0.80	—	0.327	0.314	0.281	0.311	0.305

are considered. In the latter case the expressions for the pseudovalues and jackknife estimates are the same as those for the process variance in section 2. For the case of one-day deletion the pseudovalues are given by

$$\theta^*_{i,j} = \hat{\theta}_a + (J^* - 1)(\hat{\theta}_a - \hat{\theta}_{a,-(i,j)}), \quad (25)$$

with $\hat{\theta}_{a,-(i,j)}$ the estimate of $\theta_a = \ln(\sigma_a^2)$ after omitting $a_{i,j}$. The jackknife estimate $\hat{\theta}_{\text{jack1}}$ is the average of the $\theta^*_{i,j}$s and the jackknife variance $\hat{V}_{\text{jack1}}$ is obtained as

$$\hat{V}_{\text{jack1}} = \frac{1}{J^*(J^* - 1)} \sum_{i=p+1}^{n} \sum_{j=1}^{J} (\theta^*_{i,j} - \hat{\theta}_{\text{jack1}})^2. \quad (26)$$

Significance tests can be based on the standard normal distribution in case of one-day deletion.

Table 2 compares the means of $\hat{V}^{1/2}_{\text{Katz}}$, $\hat{V}^{1/2}_{\text{jack1}}$, and $\hat{V}^{1/2}_{\text{jack30}}$ with the standard deviation $\sigma(\hat{\theta}_a)$ of $\hat{\theta}_a$ for 10-year samples of different AR processes. For the normal AR processes the statistical properties of $\hat{V}^{1/2}_{\text{Katz}}$ and $\hat{V}^{1/2}_{\text{jack1}}$ are almost identical. A second case (lower rows) is considered for these processes where the coefficients ϕ_k in the prewhitening filter deviate from the original autoregression coefficients. In these two cases both $\hat{V}^{1/2}_{\text{Katz}}$ and $\hat{V}^{1/2}_{\text{jack1}}$ underestimate $\sigma(\hat{\theta}_a)$. A similar discrepancy will occur when the climate time series considered cannot be represented properly by a low-order AR process. For the exponential distribution, $\hat{V}^{1/2}_{\text{Katz}}$ underestimates $\sigma(\hat{\theta}_a)$. The estimate $\hat{V}^{1/2}_{\text{jack30}}$ using one-month deletion is not corrupted with a large bias when the filter coefficients ϕ_k deviate from the autoregression coefficients but has a larger standard error than $\hat{V}^{1/2}_{\text{Katz}}$ and $\hat{V}^{1/2}_{\text{jack1}}$. The test for equality of innovation variances should then be based on the *t* distribution. The critical values of the *t* distribution are greater than those from the normal distribution. The differences are substantial when the number *d* of degrees of freedom is less than 10. Because of the use of the *t* distribution the test on one-month deletion will be less powerful than that based on one-day deletion. This holds in particular when the sample sizes are small.

From Table 2 it is also seen that for the normal processes $\sigma(\hat{\theta}_a)$ is much smaller than $\sigma(\hat{\theta})$. Prewhitening

TABLE 3. Mean ($\bar{x}$), variances, kurtosis ($\hat{\gamma}_2$), lag 1 autocorrelation coefficient (r_1), and order (p) of fitted AR processes of near-surface temperature (°C) over Europe (averages of monthly estimates, see main text): E refers to the ECMWF analyses, C to the ECHAM control run, and A to the ECHAM scenario-A run.

			Variance			$\hat{\gamma}_2$			
Season	Data	$\bar{x}$	s^2	s_w^2	s_a^2	$\{x_{i,j}\}$	$\{a_{i,j}\}$	r_1	p
DJF	E	4.2	10.76	6.65	3.13	0.15	0.82	.80	1.8
	C	2.1	18.30	14.06	5.75	0.48	1.34	.81	2.7
	A	5.7	11.29	8.09	2.87	0.36	1.04	.83	2.6
MAM	E	8.1	7.42	5.14	1.92	0.51	1.16	.80	1.9
	C	7.8	9.45	7.79	1.99	0.14	1.18	.83	2.6
	A	10.5	6.56	5.35	1.33	0.05	0.78	.83	2.6
JJA	E	16.3	5.51	3.82	1.56	0.66	1.47	.82	1.9
	C	17.8	5.48	4.30	0.80	0.14	0.78	.86	2.7
	A	20.5	5.39	3.87	0.82	0.08	0.79	.85	2.8
SON	E	11.3	7.88	5.56	2.19	0.34	0.76	.80	1.7
	C	11.3	8.26	6.63	1.72	0.17	1.07	.84	2.7
	A	14.0	7.01	5.72	1.37	0.29	0.83	.84	2.5
Year	E	10.0	7.89	5.29	2.20	0.41	1.05	.80	1.8
	C	9.7	10.38	8.19	2.56	0.23	1.09	.83	2.7
	A	12.7	7.56	5.75	1.60	0.20	0.86	.84	2.6

TABLE 4. Ratios of the sample variances of near-surface temperature over Europe in the ECHAM control run to those in the ECMWF analyses and results of jackknife tests for equality of variances.

Variance	Statistic	DJF	MAM	JJA	SON	Year
s^2	ratio	1.70	1.27	0.90	1.05	1.32
	T	4.24[a]	2.90[a]	−0.25	1.49	3.49[a]
s_w^2	ratio	2.11	1.52	1.13	1.19	1.55
	T_w	6.60[a]	6.99[a]	2.56[a]	3.75[a]	8.07[a]
s_a^2	ratio	1.84	1.04	0.51	0.78	1.17
	$T_{a,1}$	11.21[a]	−4.02[a]	−20.90[a]	−9.48[a]	−11.20[a]
	$T_{a,30}$	3.73[a]	−1.87	−5.73[a]	−3.56[a]	−3.39[a]

[a] Differences significant at the 5% level (two-sided test).

then has a positive effect on the power of the test. For the AR processes from the Laplace and exponential distributions prewhitening leads to a strong increase in the kurtosis [for an AR(1) process with $\rho_1 = 0.8$ the kurtosis increases by a factor of 45/9]. The value of $\sigma(\hat{\theta}_a)$ is therefore relatively high for these processes.

It may be possible that differences between $\sigma_a^2(\mathrm{I})$ and $\sigma_a^2(\mathrm{II})$ are not only caused by differences between $\sigma^2(\mathrm{I})$ and $\sigma^2(\mathrm{II})$ but also by differences in the autocorrelation structure of the two climates. Equality of $\sigma_a^2(\mathrm{I})$ and $\sigma_a^2(\mathrm{II})$ implies equality of $\sigma^2(\mathrm{I})$ and $\sigma^2(\mathrm{II})$ only if the autocorrelation structure in both climates is the same (Wilson and Mitchell 1987). When the distribution of X is not too far from the normal distribution, a test based on the innovations is in that case preferable because of the larger power. In situations where there are differences in the autocorrelation structure a test on the process variances may lead to quite other conclusions than a test based on the innovation variances. When distributional properties of daily values are of interest (e.g., extremes) one could restrict oneself to a test on the process variances. For a time series description it seems, on the other hand, more relevant to perform tests on the innovation variances and the autocorrelation properties.

TABLE 5. Average jackknife standard deviations $\bar{V}_{\mathrm{jack1}}^{1/2}$ and $\bar{V}_{\mathrm{jack30}}^{1/2}$ and lag 1 autocorrelation coefficient $r_1(\bar{a})$ of daily area averages of the innovations. E: ECMWF analyses, C: ECHAM control run, and A: ECHAM scenario-A run.

Data	Statistic	DJF	MAM	JJA	SON	Year
E	$\bar{V}_{\mathrm{jack1}}^{1/2}$	0.108	0.108	0.116	0.103	0.109
	$\bar{V}_{\mathrm{jack30}}^{1/2}$	0.153	0.175	0.218	0.155	0.177
C	$\bar{V}_{\mathrm{jack1}}^{1/2}$	0.113	0.110	0.103	0.109	0.109
	$\bar{V}_{\mathrm{jack30}}^{1/2}$	0.179	0.147	0.131	0.146	0.152
A	$\bar{V}_{\mathrm{jack1}}^{1/2}$	0.109	0.103	0.104	0.104	0.105
	$\bar{V}_{\mathrm{jack30}}^{1/2}$	0.155	0.131	0.124	0.130	0.135
E	$r_1(\bar{a})$	.56	.57	.48	.61	.56
C	$r_1(\bar{a})$	.48	.57	.51	.61	.54
A	$r_1(\bar{a})$	.49	.65	.50	.67	.58

6. Examples

The various tests in the previous sections have been applied to time series of daily near-surface temperature over Europe. Temperature variances of ECMWF analyses and two simulations of the coupled ECHAM/LSG model (Cubasch et al. 1992) are considered and compared for the area extending from 34° to 67°N, 14°W to 26°E. The ECMWF analyses refer to the period 1984–1993 and were available on a 2.5° × 2.5° horizontal grid. The quality of these data and the impact of changes in the analysis through 1989 are discussed in Hurrell and Trenberth (1992). For the ECHAM GCM daily temperatures on a 5.6° × 5.6° horizontal grid are considered for 10 years (2075–2084) in a 100-year control run with constant 1985 equivalent CO_2 concentration and those in the scenario-A run for the same period. In this transient run the average CO_2 concentration in the 2075–2084 decade is about 2.8 times that of 1985.

a. Seasonal means, variances, and kurtosis

Table 3 summarizes some relevant sample statistics. The values in this table are averages of monthly estimates over a season or year and over the grid. The monthly estimates of the innovation variance were obtained by fitting AR processes of order 1, 2, and 3 to the daily data of each grid point and selecting the optimum order by the Bayesian information criterion (BIC) as recommended by Katz (1982). The values of the kurtosis were based on $\hat{\gamma}_2 = \sum_{i=1}^{n} \sum_{j=1}^{J} (x_{i,j} - \bar{x})^4 / (nJs^4) - 3$ for the daily temperatures and Eq. (23) for the innovations. To avoid rounding errors Spicer's (1972) algorithm was used to obtain the power sums about the mean. Table 3 shows that in the control run the average temperature is underestimated in winter (DJF) and overestimated in summer (JJA). As a consequence the annual cycle in the ECHAM model is about 3.5°C too large. Rather large differences are sometimes found between the variance estimates from the ECMWF analyses and those from the GCM data. In the scenario-A run the average temperature is about 3°C higher for all seasons and the (daily) variances are

TABLE 6. Ratios of the sample variances of near-surface temperature over Europe in the ECHAM scenario-A run to those in the control run and results of jackknife tests for equality of variances.

Variance	Statistic	DJF	MAM	JJA	SON	Year
s^2	ratio	0.62	0.69	0.98	0.85	0.73
	T	−6.26[a]	−5.97[a]	−0.48	−2.06	−8.01[a]
s_w^2	ratio	0.58	0.69	0.90	0.86	0.70
	T_w	−6.91[a]	−3.64[a]	−1.46	−1.91	−7.69[a]
s_a^2	ratio	0.50	0.67	1.03	0.80	0.62
	$T_{a,30}$	−9.03[a]	−4.82[a]	0.60	−1.42	−6.85[a]

[a] Differences significant at the 5% level (two-sided test).

generally lower than in the control run. The reduction in the latitudinal temperature gradient at enhanced atmospheric CO_2 concentration could be responsible for this decrease in temperature variability (Rind et al. 1989; Rind 1991; Cao et al. 1992). From the values of p in Table 3 it is seen that the ECMWF data are on average close to an AR(2) process, whereas the ECHAM data are closer to an AR(3) process. The kurtosis is positive and, as indicated in section 5, the innovations $a_{i,j}$ always have larger kurtosis than the original data $x_{i,j}$. For the three climates considered in Table 3 the differences in kurtosis and lag 1 autocorrelation coefficients are small. From the Monte Carlo results in appendices A and C it is clear that the jackknife tests are robust against these differences.

b. *Comparing observed daily variability with GCM-simulated daily variability*

Ratios of the sample variances in the ECHAM control run to those in the ECMWF analyses and results of jackknife tests for equality of variances over Europe are given in Table 4. The ECHAM model significantly (at the 5% level) overestimates the within-month variances in all seasons. The process variance is significantly overestimated in winter and spring in the ECHAM model. For the summer and autumn seasons the process variances in the ECHAM control run and the ECMWF analyses are almost equal. Two jackknife statistics for the innovation variance are given in Table 4: $T_{a,1}$ based on one-day deletion[1] and $T_{a,30}$ based on one-month deletion. The innovation variance is significantly overestimated by the ECHAM model in winter and significantly underestimated in summer and autumn. When all months in the year are averaged, the sample innovation variance ratio indicates an overestimation by the model of 17%, whereas the test statistics indicate a significant underestimation by the model. Such a situation can occur when the variance shows large seasonal or spatial variation. In this case the sample innovation variance of the ECHAM control simulation varies strongly over the seasons; in winter it is about seven times as large as in summer (Table 3). The arithmetic mean is then much larger than the geometric mean.

c. *One-month deletion versus one-day deletion*

A striking point in Table 4 is that the values of $T_{a,1}$ are much more significant than those of $T_{a,30}$. This is caused by a lower value of the variance estimate in the denominator of the test statistic in case of one-day deletion. In section 5 it was already mentioned that there is a risk that $\hat{V}^{1/2}_{\text{jack1}}$ systematically underestimates the standard deviation of $\ln(s_a^2)$. For the ECMWF analyses and the ECHAM data this can be verified by averaging $\hat{V}^{1/2}_{\text{jack1}}$ and $\hat{V}^{1/2}_{\text{jack30}}$ over months and grid points. Table 5 shows that for both the ECMWF analyses and the ECHAM data the average $\bar{V}^{1/2}_{\text{jack30}}$ is systematically larger than $\bar{V}^{1/2}_{\text{jack1}}$. For the ECMWF analyses the difference is about 60% and for the ECHAM simulations about 35%. These differences indicate that there is dependence between the innovations. This might be due to the presence of long-term persistence in the data, which cannot be described by a low-order AR process. Linear filtering may also fail to produce independent innovations in case of nonlinear dependence between $x_{i,j}$ and the values on preceding days or other systematic departures from the AR model. The differences between $\bar{V}^{1/2}_{\text{jack1}}$ and $\bar{V}^{1/2}_{\text{jack30}}$ are, however, not large enough to explain the differences between $T_{a,1}$ and $T_{a,30}$ in Table 4. In contrast to the average jackknife standard deviation $\bar{V}^{1/2}_{\text{jack1}}$ in Table 5, the value of the multivariate statistic $T_{a,1}$ is sensitive to spatial correlation. The derivation of a variance estimate from daily area-averaged pseudovalues assumes that there is no lagged cross-correlation between the innovations at different grid points. Pointwise filtering does not guarantee that this correlation is removed. A simple check on lag 1 cross-correlation of the innovations is possible by calculating the lag 1 autocorrelation co-

[1] In order to average the pseudovalues over successive calendar months, an equal number of days in each month was considered in case of one-day deletion. The first day was always day 4 to allow that p can be as large as 3 in Eq. (16), whereas the last day was day 28 for the periods containing February and day 30 for the other periods.

efficient, $r_1(\bar{a})$, of the daily area averages of the innovations. Table 5 shows that the seasonal and annual values of $r_1(\bar{a})$ are about 0.5 in all cases from which it must be concluded that there is a rather strong lag 1 cross-correlation between the innovations. This lagged cross-correlation leads to a serious underestimation of the variance of $\hat{\theta}_{\text{jack1}}$ in the multivariate extension and invalidates the use of the standard normal distribution in case of one-day deletion. It is therefore recommended to base the jackknife statistic for the equality of innovation variances on one-month deletion rather than on one-day deletion.

d. Comparing variances in the control and scenario-A run

The results of tests for equality of variances of the ECHAM control simulation and the scenario-A simulation are presented in Table 6 in a similar way as in Table 4. The earlier mentioned reduction in the variances as a result of enhanced greenhouse gas concentrations is statistically significant for all three components in winter and spring, but there is no statistical evidence of a change in the variances in the summer and autumn seasons. The reductions in the first two

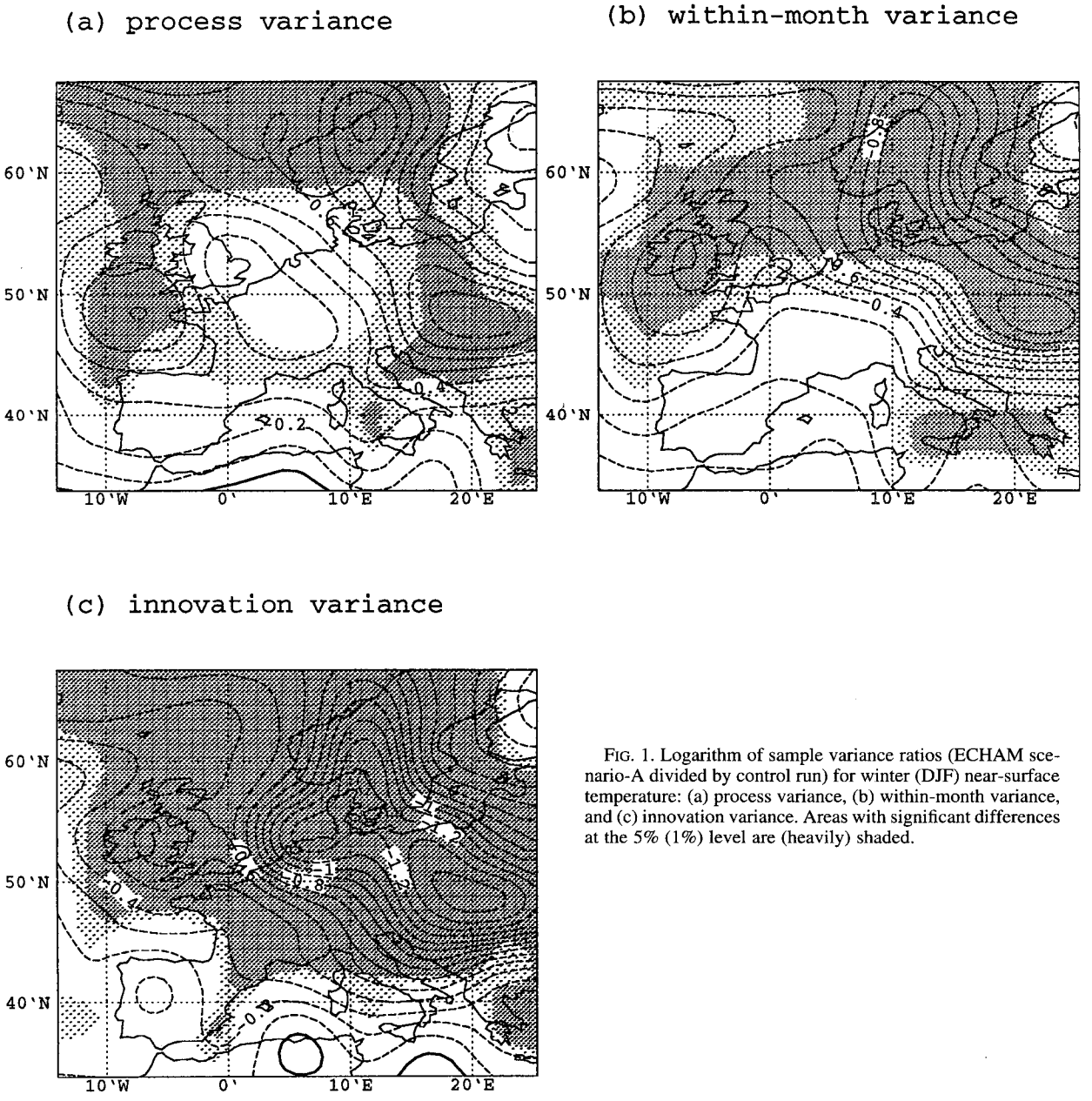

FIG. 1. Logarithm of sample variance ratios (ECHAM scenario-A divided by control run) for winter (DJF) near-surface temperature: (a) process variance, (b) within-month variance, and (c) innovation variance. Areas with significant differences at the 5% (1%) level are (heavily) shaded.

seasons are strong enough to yield a significant reduction over the year.

The magnitude of the differences varies over the study area. Figure 1 presents maps of the logarithm of the local sample variance ratios (ECHAM scenario-A divided by control run) for the winter season. Significant differences in the local variances, based on jackknifing the data of individual grid points, are shaded. The maps of process variance, within-month variance, and innovation variance ratios have a more or less similar structure. For all three components the number of grid points with a significant reduction is relatively large at high latitudes. The ratios of the innovation variances tend to be smaller and more significant than those of the process variances and the within-month variances. This is probably due to differences in the autocorrelation structure of the two model runs. The area-averaged lag 1 autocorrelation coefficient is 0.81 for the control run and 0.83 for the scenario-A run (Table 3). The jackknife test in Buishand and Beersma (1993) reveals that the difference between these two values is significant at the 5% level.

e. *Comparison with other GCM simulations*

The results for this version of the ECHAM model only partly agree with those reported elsewhere for other GCMs. The tendency to overestimate the daily temperature variability is consistent with comparisons by Reed (1986), Rind et al. (1989), and Mearns et al. (1990). Cao et al. (1992), however, found an underestimation of the within-month variance over the midlatitude continents during summer for a version of the U.K. high-resolution GCM. The causes for these model imperfections are often sought in the soil or hydrology related parameterizations.

The reduction of the temperature variability under enhanced greenhouse gas concentrations in the ECHAM model is also featured by other models (Rind et al. 1989; Wilson and Mitchell 1987). Mearns et al. (1990), however, found mixed results for the changes in the innovation variance of the daily temperature in a version of the NCAR Community Climate Model.

7. Conclusions

Tests for equality of daily variances based on the jackknife have a much wider applicability than the traditional F test. To avoid difficulties with autocorrelation it is necessary to delete one month of data simultaneously in the jackknife procedure. In most cases the null distribution of the jackknife statistic can then be approximated by Student's t-distribution. The effective number of degrees of freedom in that distribution follows easily from the jackknife variances.

A multivariate test statistic can be obtained by averaging the pseudovalues over successive months and/or grid points in a region. This avoids the multiple comparison problem. The examples indicate that the multivariate jackknife tests are able to detect differences of about 15% between the seasonal averages of the process and the within-month variances when 10-year samples over Europe are combined (significance level 5%).

Removing the autocorrelation in the daily data by a linear autoregressive filter does not guarantee the validity of a jackknife procedure with one-day deletion. There is a serious risk that the variance of $\ln(s_a^2)$ is underestimated, in particular in case of spatial averaging, because of lagged cross-correlations between the innovations. The jackknife tests on the innovation variances should therefore also be based on one-month deletion. For two single time series of normally or almost normally distributed data with the same autocorrelation structure, a test on the innovation variances is more powerful than a test on the process variances. In the multivariate extension the gain in power is reduced by remaining lagged cross-correlations.

The validation of temporal variability in climate models and the effects of enhanced greenhouse gas concentrations on variability need more attention. In previous studies different measures of daily variability have been used, which makes it difficult to compare the results. It is therefore desirable that simulated and observed variability are systematically analyzed in an equal manner. The multivariate jackknife test statistics presented here are simple, powerful, and nondemanding on computer resources.

Acknowledgments. We thank A. J. Coops and G. P. Können for their detailed comments on an earlier version of this paper. Helpful suggestions by two referees are also acknowledged. The GCM data were kindly provided by the Max-Planck-Institut für Meteorologie in Hamburg, Germany. This work was partly supported by the Dutch National Research Programme on Global Air Pollution and Climate Change (NRP) under Grant 850035.

APPENDIX A

Validity of the t Distribution in Tests on the Process Variances

The approximation of the null distribution of the statistic T in Eq. (6) by Student's t-distribution with d degrees of freedom was checked by a similar Monte Carlo study as in Buishand and Beersma (1993). Table A1 shows that the empirical significance levels are close to the nominal values when the data came from the normal distribution and have the same autocorrelation structure. The largest discrepancy occurs when a

TABLE A1. Actual rejection rates of the null hypothesis of equal process variances for two-sided tests based on the jackknife statistic T in a series of Monte Carlo experiments (n = 30; 2500 simulations for J = 10, K = 30; 5000 simulations in the other cases). The critical values of T are obtained from Student's t-distribution with d degrees of freedom. For the generated AR(1) processes the lag 1 autocorrelation coefficient ρ_1 = 0.8, whereas for the AR(2) processes ρ_1 = 0.8 and ρ_2 = 0.45: N refers to the normal distribution, L to the Laplace distribution, and E to the exponential distribution.

Process	Climate I J	Climate II K	Nominal significance level 0.100	0.050	0.010
N AR(1)	5	5	0.105	0.053	0.010
N AR(1)	10	10	0.109	0.058	0.013
N AR(1)	5	15	0.121	0.066	0.018
N AR(1)	10	30	0.108	0.058	0.022
N AR(2)	5	5	0.090	0.043	0.008
N AR(2)	10	10	0.106	0.055	0.010
L AR(1)	5	5	0.116	0.051	0.009
L AR(1)	5	15	0.120	0.069	0.021
E AR(1)	5	5	0.140	0.077	0.020
E AR(1)	5	15	0.149	0.088	0.031

very short record is tested against a longer record (J = 5, K = 15). As with the test on the lag 1 autocorrelation coefficients in Buishand and Beersma (1993), the critical values from the t approximation are too low for a test on the variances of autocorrelated exponential variables. Table A2 presents empirical significance levels for normal AR(1) processes with different lag 1 autocorrelation coefficients. These results indicate that our jackknife procedure is reasonably robust against this Behrens–Fisher type of problem. The largest differences between the empirical and nominal significance levels are found when a short record with a relatively high value of ρ_1 is tested against a longer record with a relatively small value of ρ_1 (case J = 15, K = 5 in Table A2). The higher autocorrelation in the short record then strengthens the differences in the variances of $\hat{\theta}_{\text{jack}}$(I) and $\hat{\theta}_{\text{jack}}$(II). The same occurs when the short record is generated from a Laplace AR(1) process and the long record from a normal AR(1) process (Table A3). For climate data the differences in kurtosis are,

TABLE A2. As in Table A1 except that the data came from normal AR(1) processes with ρ_1 = 0.7 in climate I and ρ_1 = 0.8 in climate II.

Climate I J	Climate II K	Nominal significance level 0.100	0.050	0.010
5	5	0.104	0.050	0.009
10	10	0.113	0.057	0.014
5	15	0.115	0.058	0.013
15	5	0.129	0.070	0.020

TABLE A3. As in Table A1 except that the data came from a normal AR(1) process in climate I and a Laplace AR(1) process in climate II with the same lag 1 autocorrelation coefficient (ρ_1 = 0.8).

Climate I J	Climate II K	Nominal significance level 0.100	0.050	0.010
5	5	0.102	0.047	0.010
10	10	0.107	0.059	0.016
5	15	0.106	0.059	0.016
15	5	0.124	0.073	0.024

however, much smaller than those between the normal and the Laplace distribution.

APPENDIX B

The Mean of the Within-Month Variances for an AR(1) Process

Each $s^2_{w,j}$ has the same mean. Taking expectations on both sides of Eq. (8), it follows

$$E(s^2_{w,j}) = E(s^2) - E(s^2_b). \tag{B1}$$

The sample process variance is an asymptotically unbiased estimate of the true process variance:

$$E(s^2) \approx \sigma^2. \tag{B2}$$

The interannual variance s^2_b is an almost unbiased estimate of the variance of the monthly averages $\bar{x}_{\cdot j}$. Using the expression for var($\bar{x}_{\cdot j}$) in Katz (1985) for an AR(1) process, the following approximation to the mean of s^2_b is obtained:

$$E(s^2_b) \approx \frac{\sigma^2}{n}\left(\frac{1+\rho_1}{1-\rho_1}\right). \tag{B3}$$

Substitution of Eqs. (B2) and (B3) into Eq. (B1) results in

$$E(s^2_{w,j}) \approx \sigma^2\left(1 - \frac{1}{n}\frac{1+\rho_1}{1-\rho_1}\right). \tag{B4}$$

APPENDIX C

Validity of the t Distribution in Tests on the Within-Month Variances

The Monte Carlo experiments in Table A1 also provide a check on the use of Student's t-distribution for testing equality of the within-month variances with the statistic T_w in Eq. (15). From the results in Table C1 it is seen that the t approximation of the null distribution of T_w performs at least as well as for the test on the process variances. Even for the samples from the exponential distribution the empirical significance levels are close to the nominal values. The t approximation performs, however, rather poorly when one climate time series is generated from a normal AR(1) process

TABLE C1. Actual rejection rates of the null hypothesis of equal within-month variances for two-sided tests based on the jackknife statistic T_w in a series of Monte Carlo experiments (n = 30; 2500 simulations for J = 10; K = 30; 5000 simulations in the other cases). The critical values of T are obtained from Student's t-distribution with d degrees of freedom. For the generated AR(1) processes the lag 1 autocorrelation coefficient ρ_1 = 0.8, whereas for the AR(2) processes ρ_1 = 0.8 and ρ_2 = 0.45: N refers to the normal distribution, L to the Laplace distribution, and E to the exponential distribution.

Process	Climate I J	Climate II K	Nominal significance level 0.100	0.050	0.010
N AR(1)	5	5	0.096	0.044	0.007
N AR(1)	10	10	0.107	0.053	0.012
N AR(1)	5	15	0.109	0.058	0.014
N AR(1)	10	30	0.106	0.058	0.013
N AR(2)	5	5	0.094	0.047	0.007
N AR(2)	10	10	0.103	0.049	0.009
L AR(1)	5	5	0.088	0.039	0.006
L AR(1)	5	15	0.096	0.049	0.014
E AR(1)	5	5	0.092	0.041	0.009
E AR(1)	5	15	0.103	0.052	0.013

and the other from a Laplace AR(1) process (Table C2). For this particular variant of the jackknife there are considerable differences in the expected values of $\bar{\theta}_w$(I) and $\bar{\theta}_w$(II) under the null hypothesis in that situation due to the large differences in kurtosis. Except for very short record lengths (case J = 5, K = 5 in Table B2) this leads to a progressive test, that is, a test that rejects the null hypothesis too frequently.

The use of Student's t-distribution in the jackknife test based on $\hat{\theta}_w$ in Eq. (11) has the same limitations as that in the test on the process variances. The Student approximation does not work for autocorrelated data from an exponential distribution. The test is rather robust against differences in kurtosis. For the situation that one climate time series is generated from a normal AR(1) process and the other from a Laplace AR(1) process, the results are comparable to those for the process variances in Table A3.

It should finally be noted that the Monte Carlo experiments Table A2 cannot be used to investigate the null distribution of T_w. The differences in ρ_1 for the two climates in that table lead to differences in the within-month variances.

TABLE C2. As Table C1 except that the data came from a normal AR(1) process in climate I and a Laplace AR(1) process in climate II with the same lag 1 autocorrelation coefficient (ρ_1 = 0.8).

Climate I J	Climate II K	Nominal significance level 0.100	0.050	0.010
5	5	0.101	0.051	0.010
10	10	0.137	0.072	0.015
5	15	0.150	0.083	0.022
15	5	0.117	0.067	0.021

REFERENCES

Box, G. E. P., and G. M. Jenkins, 1976: *Time Series Analysis: Forecasting and Control.* 2d ed. Holden-Day, 575 pp.

Buishand, T. A., and J. J. Beersma, 1993: Jackknife tests for differences in autocorrelation between climate time series. *J. Climate,* **6,** 2490–2495.

Bukač, J., and H. Burstein, 1980: Approximations of Student's t and chi-square percentage points. *Commun. Stat. B,* **9,** 665–672.

Cao, H. X., J. F. B. Mitchell, and J. R. Lavery, 1992: Simulated diurnal range and variability of surface temperature in a global climate model for present and doubled CO_2 climates. *J. Climate,* **5,** 920–943.

Cubasch, U., K. Hasselmann, H. Höck, E. Maier-Reimer, U. Mikolajewicz, B. D. Santer, and R. Sausen, 1992: Time-dependent greenhouse warming computations with a coupled ocean-atmosphere model. *Climate Dyn.,* **8,** 55–69.

Davis, W. W., 1977: Robust interval estimation of the innovation variance of an ARMA model. *Ann. Stat.,* **5,** 700–708.

——, 1979: Robust methods for detection of shifts of the innovation variance of a time series. *Technometrics,* **21,** 313–320.

Efron, B., 1982: *The Jackknife, the Bootstrap and other Resampling Plans.* Soc. Ind. Appl. Math., 92 pp.

Gardiner, D. A., and B. F. Bombay, 1965: An approximation to Student's t. *Technometrics,* **7,** 71–72.

Hinkley, D. V., 1983: Jackknife methods. *Encyclopedia of Statistical Sciences.* Vol. 4, S. Kotz, N. L. Johnson, and C. B. Read, Eds., Wiley and Sons, 280–287.

Hurrell, J. W., and K. E. Trenberth, 1992: An evaluation of monthly mean MSU and ECMWF global atmospheric temperatures for monitoring climate. *J. Climate,* **5,** 1424–1440.

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.

——, 1985: Probabilistic models. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 261–288.

——, 1988: Statistical procedures for making inferences about climate variability. *J. Climate,* **1,** 1057–1064.

——, 1992: Role of statistics in the validation of general circulation models. *Climate Res.,* **2,** 35–45.

Lawrance, A. J., 1980: Some autoregressive models for point processes. *Point Processes and Queuing Problems,* P. Bartfai and J. Tomko, Eds., Elsevier, 257–275.

Lee, A. F. S., and J. Gurland, 1975: Size and power of tests for equality of means of two normal populations with unequal variances. *J. Amer. Statist. Assoc.,* **70,** 933–941.

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

Mearns, L. O., S. H. Schneider, S. L. Thompson, and L. R. McDaniel, 1990: Analysis of climate variability in general circulation models: Comparison with observations and changes in variability in 2 × CO_2 experiments. *J. Geophys. Res.,* **95,** 20 469–20 490.

Miller, R. G., 1968: Jackknifing variances. *Ann. Math. Statist.,* **39,** 567–582.

Reed, D. N., 1986: Simulation of time series of temperature and precipitation over Eastern England by an atmospheric general circulation model. *J. Climatol.,* **6,** 233–253.

Rind, D., 1991: Climate variability and climate change. *Greenhouse-Gas-Induced Climate Change: A Critical Appraisal of Simulations and Observations,* M. E. Schlesinger, Ed., Elsevier, 111–128.

——, R. Goldberg, and R. Ruedy, 1989: Change in climate variability in the 21st century. *Clim. Change,* **14,** 5–37.

Scheffé, H., 1970: Practical solutions of the Behrens–Fisher problem. *J. Amer. Statist. Assoc.*, **65,** 1501–1508.

Sneyers, R., 1990: On the statistical analysis of series of observations. World Meteorological Organization Tech. Note No. 143, WMO No. 415, 192 pp. (original French version published in 1975). P.O. Box 2300, CH 1211, Genève 2, Switzerland.

Spicer, C. C., 1972: Calculation of power sums of deviations about the mean. *Appl. Statist.*, **21,** 226–227.

Welch, B. L., 1938: The significance of the difference between two means when the population variances are unequal. *Biometrika,* **29,** 350–361.

Wigley, T. M. L., and B. D. Santer, 1990: Statistical comparison of spatial fields in model validation, perturbation, and predictability experiments. *J. Geophys. Res.*, **95,** 851–865.

Wilson, C. A., and J. F. B. Mitchell, 1987: Simulated climate and CO_2-induced climate change over Western Europe. *Clim. Change,* **10,** 11–42.