A Simple Test for Equality of Variances in Monthly Climate Data

JULES J. BEERSMA AND T. ADRI BUISHAND

Royal Netherlands Meteorological Institute, De Bilt, the Netherlands

(Manuscript received 20 January 1998, in final form 12 June 1998)

ABSTRACT

Tests for equality of variances of monthly climate data using resampling techniques are discussed. The application of a jackknife test to spatially correlated time series is worked out in this paper. Besides this spatial extension, it is also possible to combine the data for the individual calendar months into a single seasonal or annual test statistic. The derivation of the critical values of the test statistic from Student's t-distribution in such multivariate applications is investigated. A modification to improve the use of the t-distribution is given for the case that the distribution of the data is close to the normal distribution. The power of the simple jackknife test is compared with that of a permutation test.

The test is illustrated with a comparison of the variances of monthly temperatures and precipitation amounts in the anomaly simulation, with enhanced greenhouse gas concentrations, and in the control simulation of the high-resolution transient experiment with the Hadley Centre coupled ocean–atmosphere general circulation model. Three regions are considered: central North America, southern Europe, and northern Europe. For a number of regions and seasons the differences between the variances of the two simulations are significant at the 5% level. In particular, a significant increase in the variance of monthly precipitation over northern Europe is found in the anomaly simulation for winter, summer, and autumn. Limitations of the use of the test to monthly precipitation time series containing a large proportion of zeros are identified.

1. Introduction

The study of changes in the variance of meteorological variables is of recent interest. It is now well recognized that climate change may not be restricted to changes in the mean alone. Several authors have compared the variances of monthly and seasonal values of observed data or simulated data from general circulation models (GCMs). The determination of the statistical significance of observed differences meets, however, difficulties. Rind et al. (1989), Mearns et al. (1990), Cao et al. (1992), and Gordon and Hunt (1994) used the F test for this purpose. The F test assumes that the data are independent and normally distributed. Furthermore, the test often fails to discover meaningful differences in the variances due to lack of degrees of freedom. Zwiers and Thiébaux (1987) tried to overcome the low power of the F test by deriving the interannual variability from the spectral density function of the daily values. Their test requires a careful elimination of the annual cycle in the mean. Moreover, the distribution of the test statistic has been studied only for (daily) samples from a normal distribution.

Corresponding author address: Jules J. Beersma, Royal Netherlands Meteorological Institute (KNMI), P.O. Box 201, 3730 AE De Bilt, the Netherlands.
E-mail: beersma@knmi.nl

The above tests refer to data at a single location. GCM data consist, however, of a large number of correlated time series on a spatial grid. Wigley and Santer (1990) presented a number of tests to compare the variances of such multivariate data. Resampling techniques using computer-intensive Monte Carlo methods were proposed to decide whether a result is significant or not.

Buishand and Beersma (1996) discussed the use of the jackknife for the comparison of daily variability in observed and simulated climates. The jackknife method is a resampling technique that does not require Monte Carlo methods. The resulting tests are reasonably robust against nonnormality of the data. The critical values can generally be based on Student's t-distribution both for univariate testing with data at a single location and for multivariate testing with data on a spatial grid.

The present paper focuses on the use of the jackknife for testing equality of variances of monthly values. Section 2 presents an overview of tests for equality of variances using resampling techniques. Particular attention is given to the jackknife method in the multivariate situation. The method is illustrated in section 3 with simulated monthly temperatures and precipitation amounts from the high-resolution transient experiment (UKTR) with the Hadley Centre coupled ocean–atmosphere GCM (Murphy 1995; Murphy and Mitchell 1995). Section 4 concludes the paper with a discussion.

2. Tests based on resampling

For estimating standard errors, resampling techniques are often good alternatives to analytic approximations. They also provide tests of significance in situations that the validity of the normal distribution is questionable. Several papers in the statistical literature have discussed the use of the jackknife and the bootstrap for testing equality of variances. An attractive property of these tests is that rather simple multivariate versions for samples on a spatial grid or samples of different seasons can be obtained. A correction of a standard jackknife test is proposed for such multivariate applications.

a. Univariate tests

In this section we confine ourselves to the monthly means (or totals) at a single location. A sample for J successive years (e.g., January average temperatures) is represented as $x_1, x_2, \ldots, x_J$. The sample mean is denoted as $\overline{x}$ and the unbiased sample variance s^2 is given by

$$s^2 = \frac{1}{J-1}\sum_{j=1}^{J}(x_j - \overline{x})^2. \tag{1}$$

The statistic s^2 is an unbiased estimate of the true variance σ^2 of the monthly values x_j if these data are independent, a quite common assumption for monthly data from different years. Tests for equality of variances are often based on $\hat{\theta} = \ln(s^2)$ rather than on s^2 itself, because the distribution of $\hat{\theta}$ is usually closer to the normal distribution than that of s^2. For independent data, $\text{var}(\hat{\theta})$ can be approximated as (O'Brien 1978)

$$\text{var}(\hat{\theta}) \approx \frac{2 + \gamma_2}{J}, \tag{2}$$

where γ_2 is the kurtosis (a standardized fourth-order moment). For the normal distribution $\gamma_2 = 0$. An estimate of $\text{var}(\hat{\theta})$ can be obtained by replacing γ_2 by the sample kurtosis:

$$\hat{\gamma}_2 = \frac{J\sum_{j=1}^{J}(x_j - \overline{x})^4}{\left[\sum_{j=1}^{J}(x_j - \overline{x})^2\right]^2} - 3. \tag{3}$$

Some caution is needed, however, because $\hat{\gamma}_2$ can be seriously biased (see appendix B). The jackknife provides a distribution-free alternative estimate of $\text{var}(\hat{\theta})$.

Although s^2 is an unbiased estimate of σ^2 for independent and identically distributed data, $\hat{\theta} = \ln(s^2)$ is a biased estimate of $\theta = \ln(\sigma^2)$. The bias is of order $1/J$ (O'Brien 1978):

$$\text{bias}(\hat{\theta}) \approx \frac{-1 - 2\gamma_2}{J}. \tag{4}$$

1) THE JACKKNIFE

In the jackknife method the statistic $\hat{\theta}$ is recomputed for each subsample of size $J - 1$. Let $\hat{\theta}_{-j}$ be the value of the statistic after omitting x_j. From $\hat{\theta}$ and $\hat{\theta}_{-j}$ a pseudovalue can be formed as

$$\theta_j^* = \hat{\theta} + (J-1)(\hat{\theta} - \hat{\theta}_{-j}). \tag{5}$$

Although the pseudovalues θ_j^* can be seen as estimates of θ, they have a much larger variance than $\hat{\theta}$. However, their mean,

$$\hat{\theta}_{\text{jack}} = \frac{1}{J}\sum_{j=1}^{J}\theta_j^*, \tag{6}$$

which is known as the jackknife estimate of θ (Miller 1968), can be a good alternative to $\hat{\theta}$. The jackknife estimate reduces the bias in estimating $\ln(\sigma^2)$ to order $1/J^2$.

Unlike the $\hat{\theta}_{-j}$ values, the pseudovalues exhibit little correlation (section 2c). Jackknife tests treat the pseudovalues as independent normal variables. Tests for equality of variances are then similar to those for equality of means in normal populations using Student's t-distribution. These tests need $\hat{\theta}_{\text{jack}}$ and its estimated variance:

$$\hat{V}_{\text{jack}} = \frac{1}{J(J-1)}\sum_{j=1}^{J}(\theta_j^* - \hat{\theta}_{\text{jack}})^2. \tag{7}$$

From the jackknife estimates a number of different statistics can be derived to test for equality of the variances $\sigma^2(\text{I})$ and $\sigma^2(\text{II})$ of two mutually independent time series of monthly climate data. Let $\hat{\theta}_{\text{jack}}(\text{I})$, $\hat{\theta}_{\text{jack}}(\text{II})$ be jackknife estimates of $\ln(\sigma^2)$ and $\hat{V}_{\text{jack}}(\text{I})$, $\hat{V}_{\text{jack}}(\text{II})$ their estimated variances, then the usual two-sample pooled t-statistic can be represented as

$$T_a = \left[\frac{JK(J+K-2)}{J+K}\right]^{1/2} \times \frac{\hat{\theta}_{\text{jack}}(\text{II}) - \hat{\theta}_{\text{jack}}(\text{I})}{[J(J-1)\hat{V}_{\text{jack}}(\text{I}) + K(K-1)\hat{V}_{\text{jack}}(\text{II})]^{1/2}}, \tag{8}$$

with J and K the number of years for climate I and climate II, respectively. Under the null hypothesis of equal variances, the distribution of T_a is approximated by Student's t-distribution with $J + K - 2$ degrees of freedom. For the case of equal sample sizes, Miller (1968) demonstrates that this approximation works well for sample sizes as small as $J = 10$. The test is also quite robust against nonnormality. However, for $J \neq K$, Monte Carlo experiments show that the critical values of the test should be larger than those obtained from Student's t-distribution, especially for long-tailed distributions (Brown and Forsythe 1974; Boos and Brownie 1989). Besides the earlier mentioned assumptions about the normality and correlation of the pseudovalues, there are two additional complications in the case of unequal sample sizes that limit the use of Student's t-

distribution with $J + K - 2$ degrees of freedom (O'Brien 1978). First, the pseudovalues have different variances in climate I and climate II if $J \neq K$. Second, the fact that bias(θ_{jack}) depends on J implies that the mean of the numerator of the test statistic slightly differs from zero under the null hypothesis. Furthermore, the two-sample Student test becomes less robust against nonnormality if the sample sizes are unequal (Kendall and Stuart 1973).

Keselman et al. (1979) suggested the use of Welch's t-statistic to cope with variance heterogeneity of the pseudovalues in case of unequal sample sizes. The test statistic reads

$$T_b = \frac{\hat{\theta}_{\text{jack}}(\text{II}) - \hat{\theta}_{\text{jack}}(\text{I})}{[\hat{V}_{\text{jack}}(\text{I}) + \hat{V}_{\text{jack}}(\text{II})]^{1/2}}. \quad (9)$$

Buishand and Beersma (1996) used a similar statistic to compare the daily variability in observed and simulated climates. The critical values of T_b are derived from Student's t-distribution with an effective number d^* of degrees of freedom:

$$d^* = \frac{[\hat{V}_{\text{jack}}(\text{I}) + \hat{V}_{\text{jack}}(\text{II})]^2}{\hat{V}^2_{\text{jack}}(\text{I})/(J-1) + \hat{V}^2_{\text{jack}}(\text{II})/(K-1)}. \quad (10)$$

For equal sample sizes, $T_b = T_a$, but d^* tends to be smaller than $J + K - 2$. Besides unequal sample sizes, differences in kurtosis γ_2 for the two climates also lead to variance heterogeneity. A correction of the test for correlation between pseudovalues is presented in section 2c.

2) Permutation and bootstrap procedures

Permutation procedures are computer-intensive techniques to determine the statistical significance of a result. The method is free of assumptions about the parametric form of the distribution of the data. A pooled permutation procedure can be used to test for equality of variances of two climate time series $\{x_1, \ldots, x_J\}$ and $\{y_1, \ldots, y_K\}$. The means $\overline{x}$ and $\overline{y}$ have to be subtracted first (Boos and Brownie 1989). The method then assumes that under the null hypothesis each permutation of the combined sample $\{x_1 - \overline{x}, \ldots, x_J - \overline{x}, y_1 - \overline{y}, \ldots, y_K - \overline{y}\}$ is equally likely. A permutation sample is obtained by taking a sample of size J *without* replacement to represent the centered data for climate I; the remaining K values represent the centered data for climate II. For each permutation sample the ratio of the sample variances s^2(I) and s^2(II) is computed. Comparing the distribution of this ratio in the permutation samples with the observed ratio gives the achieved significance level.

In contrast to this permutation test, the pooled bootstrap procedure in Boos and Brownie (1989) resamples *with* replacement from the combined sample of centered data. The two techniques are further identical. Downton and Katz (1993) applied the pooled bootstrap technique to test for discontinuities in the variance in long-term records of seasonal mean maximum temperatures. They observed that a test at the 10% level can detect changes of 25%–30% in the standard deviations of seasonal mean maximum temperatures in records of 10 yr or more and that such a test is generally not sensitive enough to be able to detect changes less than 20%.

For the bootstrap it makes sense to consider studentized statistics like T_a and T_b instead of the ratio of the sample variances (Boos and Brownie 1989). The actual rejection rate of the null hypothesis is then closer to the desired significance level. Boos and Brownie show that bootstrapping the jackknife statistic T_a results in an improvement compared with the use of Student's t-distribution with $J + K - 2$ degrees of freedom, in particular if $J \neq K$.

Pooled permutation and bootstrap procedures are not robust against unequal kurtoses. It is possible to achieve asymptotic correct significance levels in that case by bootstrapping the scaled samples $\{x_1/s(\text{I}), \ldots, x_J/s(\text{I})\}$, $\{y_1/s(\text{II}), \ldots, y_K/s(\text{II})\}$, separately (Boos et al. 1989). Because convergence is slow, the test cannot be applied to small and moderate samples (say $J, K \leq 50$).

b. Multivariate extensions

The jackknife procedure allows for a multivariate test for equality of variances in two different climates using data at several grid points in a region. Such a multivariate extension is presented in Buishand and Beersma (1996). First, the pseudovalues $\theta_1^*, \theta_2^*, \ldots, \theta_J^*$ are calculated for each grid point separately. These pseudovalues are then averaged over the various grid points, giving $\overline{\theta_1^*}, \overline{\theta_2^*}, \ldots, \overline{\theta_J^*}$. The jackknife statistic T_a or T_b is finally obtained by applying (6) and (7) to these average pseudovalues. This combined test will be more powerful than that for an individual grid point when the differences between climate I and climate II have the same sign across the whole region because of the larger sample size. The test is not suitable for very large regions (e.g., a hemisphere) where areas with negative differences may compensate those with positive differences.

The above multivariate extension compares spatial averages of the logarithms of the variances. This is equivalent with a comparison of the geometric means rather than the arithmetic means as in the SPRET1 statistic of Wigley and Santer (1990):

$$\text{SPRET1} = \overline{s^2}(\text{I})/\overline{s^2}(\text{II}), \quad (11)$$

where $\overline{s^2}$(I) and $\overline{s^2}$(II) are the spatial averages of the sample variances[1] for climate I and climate II, respec-

[1] In contrast to the unbiased estimate in Eq. (1), Wigley and Santer divide the sum of the squared deviations about the mean by J rather than $J - 1$. This choice does not influence the outcome of a permutation test and the value of the jackknife statistics T_a and T_b is affected only in case of unequal sample sizes.

TABLE 1. Actual rejection rates of the null hypothesis of equal variances for two-sided tests based on the jackknife statistic T_b (5000 simulations). The pseudovalues in the test statistic are averaged over N independent sequences. The critical values of T_b are obtained from Student's t-distribution with d^* degrees of freedom.

				Significance level		
Distribution	J	K	N	0.100	0.050	0.010
Normal	5	5	1	0.074	0.035	0.009
			3	0.055	0.025	0.005
			5	0.051	0.021	0.003
			9	0.055	0.021	0.002
	10	10	1	0.099	0.050	0.012
			3	0.074	0.037	0.008
			5	0.077	0.032	0.005
			9	0.074	0.031	0.005
Exponential	10	10	1	0.169	0.107	0.037
			3	0.129	0.071	0.017
			5	0.123	0.066	0.016
			9	0.119	0.062	0.012

tively, for a particular calendar month. There is, however, not a simple approximation to the distribution of SPRET1 under the null hypothesis. Wigley and Santer (1990) used the pooled permutation procedure of Preisendorfer and Barnett (1983) to determine the statistical significance of the observed value of SPRET1. The method is usually unnecessarily restricted to equal sample sizes only. As for the univariate tests in section 2a, it is also necessary here to adjust the monthly values for differences in the means of the two climates (Santer and Wigley 1990). Otherwise the kurtosis in each permutation would differ from that in the original series, resulting in an incorrect significance level.

The data for the individual (calendar) months can be combined into a single seasonal or annual test by averaging the monthly pseudovalues in a similar way. There is a gain in power when the sign of the differences in variance for the two climates is the same for the months under consideration. On the other hand the combined test may fail when the sign of the differences varies over the year.

c. A corrected jackknife test

In section 2a it was noted that in case of equal sample sizes the t-approximation of the null distribution did quite well in a jackknife test for sample sizes as small as 10. Correlation between the pseudovalues and the fact that their distribution deviates from the normal distribution, even if the data come from a normal distribution, limits the use of the t-distribution for smaller sample sizes. The situation is different in the multivariate extension of section 2b because spatial averaging influences the distribution of the pseudovalues. The effect of spatial averaging on the validity of the t-approximation for the test based on the statistic T_b has been investigated in a Monte Carlo experiment. Table 1 considers both the situation of two single climate time series and that of averaging the pseudovalues of N independent sequences. This averaging does not affect the correlation between the pseudovalues, while the effect of nonnormality of the pseudovalues decreases with increasing N. For N large enough the distribution of T_b therefore no longer depends on N. The empirical significance levels for samples from the normal distribution are for large N much lower than the nominal values because of the negative correlation between the pseudovalues (O'Brien 1978). The situation is in fact better if $N = 1$ because then the correlation effect is counteracted by the nonnormality of the pseudovalues. For the case $J = K = 10$ the two effects just compensate. In the generated samples from the exponential distribution the correlation between the pseudovalues is positive (O'Brien 1978). Because of this positive correlation and nonnormality of the pseudovalues the test is progressive, that is, the null hypothesis is rejected too frequently.

Like the F test our jackknife statistic T_b has little power to detect differences in the variances of two short independent climate time series at a single location. Averaging over successive months or grid points is therefore necessary to obtain a meaningful test. Through the averaging procedure the effect of nonnormality of the pseudovalues is small. Departures from the assumed t-distribution are then mainly due to correlation between the pseudovalues. These pseudovalues are equicorrelated, that is,

$$\mathrm{Corr}(\overline{\theta_i^*}, \overline{\theta_j^*}) = \rho \tag{12}$$

for all $i \neq j$. If ρ is known, the test statistic can easily be corrected for this type of correlation (Walsh 1947). The main point behind the correction is that $\hat{V}_{\mathrm{jack}}$ in Eq. (7) does not provide a purely unbiased estimate of $\mathrm{var}(\hat{\theta}_{\mathrm{jack}})$, but such an estimate is given by

$$\tilde{V}_{\mathrm{jack}} = \frac{1 + (J - 1)\rho}{1 - \rho}\hat{V}_{\mathrm{jack}}, \tag{13}$$

leading to the modified test statistic:

$$\tilde{T}_b = \frac{\hat{\theta}_{\mathrm{jack}}(\mathrm{II}) - \hat{\theta}_{\mathrm{jack}}(\mathrm{I})}{[\tilde{V}_{\mathrm{jack}}(\mathrm{I}) + \tilde{V}_{\mathrm{jack}}(\mathrm{II})]^{1/2}}. \tag{14}$$

From the Satterthwaite procedure in Welch (1938), it follows that the variance estimate $\tilde{V}_{\mathrm{jack}}$ should also be used in Eq. (10) for the degrees of freedom. Table 2 shows that the null distribution of the corrected statistic $\tilde{T}_b$ is generally much better approximated by Student's t-distribution than that of the jackknife statistic T_b. The corrected test even works well in case of unequal sample sizes despite the differences in the means of the jackknife estimates in the numerator of Eq. (14) under the null hypothesis.

Table 2 also presents estimates of ρ. Details about the derivation of these estimates are given in appendix A. The table shows that the values of $\hat{\rho}$ are rather small. Nevertheless this correlation may have a considerable effect on the distribution of the test statistic, because it

TABLE 2. Actual rejection rates of the null hypothesis of equal variances for two-sided tests based on the jackknife (2500 simulations for $J = 10$, $K = 30$; 5000 simulations in the other cases). The results in the first row refer to the jackknife statistic T_b and those in the second row to the corrected jackknife statistic $\tilde{T}_b$. The pseudovalues in the test statistics are averaged over $N = 9$ independent sequences. The estimated correlation coefficients between these pseudovalues in climate I and climate II are denoted as $\hat{\rho}$(I) and $\hat{\rho}$(II), respectively. The critical values of T_b and $\tilde{T}_b$ are obtained from Student's t-distribution with d^* degrees of freedom.

					Significance level		
Distribution	J	K	$\hat{\rho}$(I)	$\hat{\rho}$(II)	0.100	0.050	0.010
Normal	5	5	−0.063	−0.064	0.055	0.021	0.002
					0.095	0.048	0.008
	10	10	−0.017	−0.020	0.074	0.031	0.005
					0.098	0.050	0.010
	5	15	−0.066	−0.010	0.067	0.030	0.006
					0.111	0.054	0.015
	10	30	−0.015	−0.003	0.073	0.037	0.006
					0.100	0.049	0.013
Exponential	5	5	0.021	0.018	0.106	0.047	0.007
					0.090	0.037	0.006
	10	10	0.014	0.014	0.119	0.062	0.012
					0.098	0.050	0.008

does not decrease with increasing separation in time. Unfortunately, the procedure on which the estimates of ρ in Table 2 are based does not apply to a single realization. Moreover, the amount of data is generally not sufficient to obtain a sensible estimate of ρ directly. The value of ρ is determined by the sample size and the underlying distribution. This dependence was examined in order to obtain a suitable modification of the jackknife statistic T_b.

Table 3 presents estimates of ρ for $J = 5$, 10, and 30 for a number of distributions. These values increase with increasing kurtosis γ_2 of the distribution. Both for the symmetric Laplace distribution and the skewed χ^2_4 distribution the effect of correlation on the distribution of the test statistic can be neglected. The kurtosis of these distributions is, however, as large as 3. The monthly means of climatic data generally have kurtosis close to zero. It is therefore often sufficient to apply a correction to the test statistic valid for the normal distibution. The estimates of ρ for the normal distribution in Table 3 can be approximated as

$$\tilde{\rho} = -J^{-1.7}. \tag{15}$$

Substitution of $\tilde{\rho}$ in Eq. (13) gives the desired correction. Unfortunately, it is difficult to verify the validity of this correction. The sample kurtosis in Eq. (3) has a very strong bias in small samples from distributions with positive kurtosis, the so-called leptokurtic distributions (see appendix B). In the examples in section 3, the kurtosis for a single grid point was estimated as

$$\hat{\gamma}_2 = \frac{n_s J \sum_{i=1}^{n_s} \sum_{j=1}^{J} (x_{i,j} - \overline{x}_{i\cdot})^4}{\left[\sum_{i=1}^{n_s} \sum_{j=1}^{J} (x_{i,j} - \overline{x}_{i\cdot})^2\right]^2} - 3, \tag{16}$$

TABLE 3. Estimated correlation coefficients between the pseudovalues of sequences of J independent observations from the normal and other distributions (10 000 simulations for $J = 5$ and $J = 10$; 2500 simulations for $J = 30$). As in Table 2 the correlation coefficients are derived from average pseudovalues taken over $N = 9$ independent sequences.

			$\hat{\rho}$		
Distribution	Skewness	Kurtosis	$J = 5$	$J = 10$	$J = 30$
Uniform	0	−1.2	−0.101	−0.040	−0.005
Normal	0	0	−0.064	−0.019	−0.003
Laplace	0	3	−0.013	0.002	0.002
χ^2_4	$\sqrt{2}$	3	−0.002	0.009	0.002
Exponential	2	6	0.020	0.014	0.004

where $x_{i,j}$ is the value of the ith calendar month for year J, $\overline{x}_{i\cdot}$ is the average of that calendar month, and n_s is the number of calendar months in the season of interest. The pooling over successive months reduces the bias because of the larger sample size. The estimate in Eq. (16) is, however, sensitive to a systematic variation of the variance within the season of interest.

d. Power of tests for equality of variances

A Monte Carlo experiment was performed to study the performance of the proposed jackknife test. The SPRET1 statistic of Wigley and Santer (1990) was also considered in that experiment. To demonstrate the effect of spatial averaging, one set of data was generated for univariate tests on the variances at a single location, and another set was generated for multivariate tests on the variances of $N = 30$ sequences. In the latter case, vectors of length 30 were generated from a multivariate normal distribution analogous to a Monte Carlo experiment of Zwiers (1987), where the correlation coefficient between the ith and jth sequence was set equal to the lag k autocorrelation coefficient of a second-order autoregressive process:

$$\left.\begin{aligned} \rho_0 &= 1 \\ \rho_1 &= \phi_1/(1 - \phi_2) \\ \rho_k &= \phi_1 \rho_{k-1} + \phi_2 \rho_{k-2}, \qquad k \geq 2 \end{aligned}\right\}, \tag{17}$$

with $k = |i - j|$, $\phi_1 = 1.6$, and $\phi_2 = -0.8$. This correlation function represents a damped sine curve. It can be seen as the one-dimensional analog of the spatial correlation function of a climate variable exhibiting teleconnection patterns. From the Monte Carlo experiment it turns out that averaging the pseudovalues over $N = 30$ correlated sequences leads to a reduction in the standard error of $\hat{\theta}_{\text{jack}}$ of 66%, which is comparable to the effect of averaging over nine independent sequences. The standard deviations, σ(I) in climate I and σ(II) in climate II, were taken to be the same for every sequence.

Table 4 presents the results for two-sided tests at the 5% level in samples of size 10. The power is low in the univariate case ($N = 1$), in agreement with the dis-

TABLE 4. Power of tests for equality of variances for various ratios of the standard deviations in climate I and climate II (1000 simulations, and 1000 permutations of each combined sample for the two climates to determine the statistical significance of the SPRET1 statistic). The values of the power refer to a two-sided test at the 5% level for samples of size 10 ($J = K = 10$) from a univariate ($N = 1$) or a multivariate ($N = 30$) normal distribution.

	$N = 1$		$N = 30$	
$\sigma(\text{II})/\sigma(\text{I})$	T_b	SPRET1	$\tilde{T}_b$	SPRET1
1.2	0.077	0.076	0.273	0.262
1.5	0.176	0.175	0.808	0.842
2.0	0.434	0.399	0.996	0.994

cussion on the power of the F test in Zwiers and Thiébaux (1987). Even if $\sigma(\text{II})/\sigma(\text{I})$ is as large as 2 more than 50% of the cases passes the test. A large gain in power is achieved with the multivariate tests. About 80% of the cases is declared significant if $\sigma(\text{II})/\sigma(\text{I}) = 1.5$. It is further seen in Table 4 that the power of the simple jackknife test is comparable with that of a computer-intensive permutation test using the SPRET1 statistic.

3. Examples

The multivariate jackknife test in the previous section was applied to simulated time series of monthly mean near-surface temperature and precipitation from the UKTR climate change experiment with the Hadley Centre coupled ocean–atmosphere GCM (Murphy 1995; Murphy and Mitchell 1995). Data of the last 10 yr of the 75-yr integration from the control simulation (with constant CO_2 concentration) are compared with those from the anomaly simulation for the same decade (with an increase in CO_2 of 1% per year, resulting in an effective CO_2 doubling after 70 yr). The land areas of three regions are considered: central North America (CNA; 35°–50°N, 85°–105°W), southern Europe (SEU; 35°–50°N, 10°W–45°E), and northern Europe (NEU; 50°–70°N, 10°W–60°E). The first two regions were previously selected for analysis of regional climate change simulation by the Intergovernmental Panel on Climate Change (IPCC 1990, 1996). The latter region was introduced by Raisanen (1995) and later also considered by IPCC (1996). The monthly mean near-surface temperature is obtained by averaging the monthly mean maximum and minimum temperature. Results for monthly mean maximum and minimum temperature separately are similar to those for monthly mean temperature and are therefore not presented. The precipitation amounts considered are the sums of large-scale and convective precipitation.

a. Near-surface temperature

Table 5 summarizes some relevant sample statistics. The values in this table are averages of monthly estimates over a season or year and over the land grid points in the region. The kurtosis estimates are generally close to zero. Exceptions occur in spring and autumn. Values of $\hat{\gamma}_2 > 1$ in those transition seasons are rather due to systematic differences between the temperature variances of successive calendar months than to leptokurtic distributions. Equation (15) was therefore used to correct for correlation between the pseudovalues in the jackknife test for equality of variances.

TABLE 5. Mean, variance, and kurtosis of monthly near-surface temperature for central North America (CNA), southern Europe (SEU), and northern Europe (NEU). Here C refers to the control simulation and A to the anomaly simulation of the UKTR experiment: Dec–Feb (DJF), Mar–May (MAM), and etc.

Area	Data	DJF	MAM	JJA	SON	Year
			Mean (°C)			
CNA	C	−9.73	6.48	22.58	10.74	7.52
CNA	A	−4.96	8.95	27.00	15.55	11.63
SEU	C	−1.89	6.42	20.41	10.33	8.82
SEU	A	0.80	9.39	24.29	14.19	12.17
NEU	C	−22.04	−5.86	12.91	−2.70	−4.42
NEU	A	−17.34	−1.52	15.47	0.95	−0.61
			Variance (°C²)			
CNA	C	13.39	5.65	4.28	4.98	7.07
CNA	A	11.58	4.70	7.30	4.84	7.10
SEU	C	11.74	6.19	4.38	4.24	6.64
SEU	A	11.90	3.94	5.33	3.63	6.20
NEU	C	17.49	11.01	2.92	8.09	9.88
NEU	A	23.16	6.53	3.73	5.37	9.70
			Kurtosis			
CNA	C	0.16	1.66	0.34	0.64	0.70
CNA	A	−0.55	−0.21	0.29	−0.23	−0.18
SEU	C	−0.08	1.15	0.07	1.90	0.76
SEU	A	−0.11	−0.24	0.53	−0.08	0.02
NEU	C	−0.36	0.00	0.17	1.20	0.25
NEU	A	−0.54	−0.01	0.35	0.55	0.09

In the anomaly simulation the average temperature is about 4°C higher for most seasons. For the three regions the variances are larger in summer but smaller in spring and autumn. However, these changes in the variance resulting from enhanced greenhouse gas concentrations are statistically significant for only two cases; the 71% increase in summer for CNA and the 41% decrease in spring for NEU (Table 6). In winter the temperature variance changes have different signs.

Note that it is possible that the variance ratio indicates

TABLE 6. Ratios of the sample variances of monthly near-surface temperature in the anomaly simulation to those in the control simulation and results of jackknife tests for equality of variances.

Area	Statistic	DJF	MAM	JJA	SON	Year
CNA	ratio	0.86	0.83	1.71	0.97	1.00
CNA	$\tilde{T}_b$	−0.56	0.05	2.54*	−0.26	1.01
SEU	ratio	1.01	0.64	1.22	0.86	0.94
SEU	$\tilde{T}_b$	0.10	−1.41	0.96	−0.47	−0.43
NEU	ratio	1.32	0.59	1.28	0.66	0.98
NEU	$\tilde{T}_b$	1.84	−5.18*	0.74	−0.89	−0.75

* Differences significant at the 5% level (two-sided test).

TABLE 7. Mean, variance, and kurtosis of monthly precipitation for CNA, SEU, and NEU. Here C refers to the control simulation and A to the anomaly simulation of the UKTR experiment.

Area	Data	DJF	MAM	JJA	SON	Year
			Mean (mm day^{-1})			
CNA	C	1.16	3.00	3.07	1.52	2.19
CNA	A	1.30	3.05	2.73	1.44	2.13
SEU	C	2.98	2.31	2.01	1.93	2.30
SEU	A	3.01	2.32	1.61	1.76	2.18
NEU	C	1.34	1.72	2.34	2.24	1.91
NEU	A	1.68	1.95	2.51	2.58	2.18
			Variance (mm^2 day^{-2})			
CNA	C	0.31	1.03	1.63	0.76	0.93
CNA	A	0.34	1.47	1.79	0.92	1.13
SEU	C	0.91	0.82	1.17	0.85	0.93
SEU	A	1.37	0.76	1.11	0.81	1.01
NEU	C	0.29	0.44	0.67	0.46	0.46
NEU	A	0.47	0.48	0.92	0.68	0.64
			Kurtosis			
CNA	C	0.62	−0.03	−0.11	0.47	0.21
CNA	A	1.19	0.20	0.12	0.33	0.46
SEU	C	0.29	−0.08	0.60	0.26	0.27
SEU	A	0.16	−0.06	1.49	0.79	0.59
NEU	C	−0.19	−0.07	−0.03	−0.22	−0.13
NEU	A	−0.06	0.08	0.06	0.14	0.06

a decrease in variance whereas the test statistic indicates an increase in variance (see, e.g., CNA in spring) or the other way around. However, this usually happens only when the test statistic is close to zero (and thus far from the critical value). It generally requires that the arithmetic mean is much different from the geometric mean, which occurs if the variance shows large seasonal or spatial variation.

b. Precipitation

The distribution of monthly precipitation generally differs more from the normal distribution than that of monthly temperature. The largest departures from normality are found in areas or seasons where completely dry months frequently occur. If at a particular grid point the monthly mean preciptation is zero for the whole period considered, the sample variance is clearly also zero but the pseudovalues, since they involve $\ln(s^2)$, are undefined. Similarly, when only one of the monthly mean precipitation values in a time series is larger than zero, one of the pseudovalues is undefined. Both situations are found in a small number of the grid points in SEU in summer and autumn. Furthermore, time series containing many zeros have a strong effect on the spatial kurtosis estimate. To avoid problems related to such situations only those grid points are considered for which the monthly mean precipitation time series contains at least four values larger than zero.

It should further be noted that the precipitation in GCM simulations has often been regarded as being representative of the average over the grid box concerned (Reed 1986; Gregory and Mitchell 1995). The distribution of a spatial average of monthly precipitation is less skewed and has a lower kurtosis than that of monthly precipitation at a point.

TABLE 8. Ratios of the sample variances of monthly precipitation in the anomaly simulation to those in the control simulation and results of jackknife tests for equality of variances.

Area	Statistic	DJF	MAM	JJA	SON	Year
CNA	ratio	1.09	1.44	1.09	1.22	1.21
CNA	$\tilde{T}_b$	1.10	2.25*	0.50	1.15	2.36*
SEU	ratio	1.50	0.93	0.95	0.95	1.09
SEU	$\tilde{T}_b$	1.36	−1.85	−0.53	−0.48	0.04
NEU	ratio	1.61	1.09	1.37	1.49	1.37
NEU	$\tilde{T}_b$	3.72*	1.03	2.61*	3.01*	4.77*

* Differences significant at the 5% level (two-sided test).

Table 7 summarizes the sample statistics for precipitation in the same way as for temperature in section 3a. For NEU the monthly mean precipitation in the anomaly simulation is for all seasons 5%–25% higher than in the control simulation. This increase in the mean is accompanied by an increase in the variance. Table 8 shows that the changes in variance vary between 10% (spring) and 60% (winter), and are, except for spring, statistically significant. For CNA the anomaly simulation shows an increase in mean winter precipitation of about 10% and a decrease in mean summer precipitation of about 10%; for SEU there is a 20% decrease in mean precipitation in summer and an almost 10% decrease in autumn (Table 7). For these two regions the changes in the mean are not accompanied by similar changes in the variance. The largest changes in the variance are found in other seasons, namely a statistically significant increase of 44% in spring for CNA and an increase of 50% in winter for SEU (Table 8).

In the statistical tests above, a correction for correlation between pseudovalues was applied using Eq. (15) for the normal distribution. The kurtosis estimates in Table 7 support this correction for most cases. Exceptions are CNA in winter and SEU in summer and autumn. In particular for the SEU precipitation, the positive kurtosis cannot be attributed to within-season variations of the variance only. According to the simulation results in appendix B, a spatial average of $\hat{\gamma}_2$ in the range of 1–1.5 indicates that $\gamma_2 \approx 3$, so that a correction for correlation between pseudovalues would not be needed. For the cases mentioned above the correction had only a small effect; the values of the test statistic T_b without correction are 0.99 for CNA in winter, −0.47 for SEU in summer, and −0.43 for SEU in autumn.

c. Comparison with other GCM simulations

The results for the UKTR experiment only partly agree with those of Rind et al. (1989), Gordon and Hunt (1994), and Liang et al. (1995) for mixed layer models. In contrast to a coupled model, as used in the UKTR experiment, a mixed layer model cannot produce vari-

ability associated with dynamical ocean processes such as the Atlantic thermohaline circulation and the El Niño–Southern Oscillation. Since such processes contribute to the interannual variability, they should be included in experiments that investigate the response of atmospheric variability to enhanced greenhouse gas concentrations, as is demonstrated by Meehl et al. (1994). They found that the changes of (interannual) temperature variability in a mixed layer version of their model differed from those in a coupled version, particularly in the Tropics.

However, particular responses, that can be understood from physical relationships, seem quite robust. Examples are reduced temperature variability over areas where sea ice retreats (Gordon and Hunt 1994; Meehl et al. 1994; Liang et al. 1995), enhanced summer temperature variability in areas of reduced soil moisture (Meehl et al. 1994; Liang et al. 1995), and enhanced precipitation variability due to the enhanced hydrological cycle and greater atmospheric moisture content in the extratropics (Rind et al. 1989; Liang et al. 1995).

Liang et al. (1995), for example, found increased summer temperature variability over CNA, which they ascribe to reduced soil moisture. In UKTR there is an increased temperature variability over CNA in summer, which is accompanied by a reduction in mean precipitation, and this generally leads to reduced soil moisture in a warmer climate. With respect to enhanced precipitation variability, all substantial changes in precipitation variance (larger than 10%), in the three areas considered, are increases. Increases in precipitation variability over CNA in spring and summer similar to those in UKTR were also reported by Liang et al. (1995).

4. Discussion

A test for equality of variances based on the jackknife has been described that is suitable for correlated time series of monthly climate data on a spatial grid (e.g., those produced by GCMs). In contrast to other resampling techniques the method does not require computer-intensive simulation to derive the statistical significance of observed differences in variances. The null distribution of the test statistic can be approximated by Student's *t*-distribution with an effective number of degrees of freedom. For a test on multivariate data this approximation can be improved by a correction for correlation between the pseudovalues in the jackknife procedure. The proposed correction does, however, not apply if there are strong departures from the normal distribution as is for instance the case for monthly precipitation data containing a considerable fraction of zeros.

Besides the reported nonnormality of monthly precipitation during the dry season in SEU, more serious problems were encountered with the application of the jackknife procedure to monthly precipitation in Southeast Asia (5°–40°N, 60°–101°E). Even for the wet monsoon the distributions of the monthly precipitation at several grid points in the area appeared to be very leptokurtic. The area-average kurtosis can be reduced by excluding the relatively dry grid points from the analysis. Disregarding grid points with mean monthly precipitation smaller than 0.5 mm day^{-1} yields an 18% increase in monthly precipitation variance in summer (June–August) and a 25% increase in the monsoon season (June–September). Both increases are significant at the 5% level. These results are in line with the increase in interannual variability of the area-averaged south Asian or Indian monsoon precipitation reported by Meehl and Washington (1993) and Bhaskaran et al. (1995).

Like the traditional F test, the jackknife test in this paper assumes that the monthly values from different years are independent. If there is a positive correlation between the values in successive years, then the jackknife variance tends to underestimate the true variance, which results in a progressive test.

Tests for equality of variances are known to have little power for typical sample sizes encountered in climate change experiments. In a jackknife test the low power is due to variability of $\hat{\theta}_{\text{jack}}$. The averaging of the pseudovalues over calendar months and/or grid points in a region leads to a considerable reduction in the standard error of $\hat{\theta}_{\text{jack}}$. Because monthly data generally exhibit no or only weak autocorrelation, averaging over three successive calendar months reduces the standard error of $\hat{\theta}_{\text{jack}}$ by about 40%. In the application to the monthly values in the UKTR experiment, spatial averaging over the grid points in each of the three regions yields a reduction in standard error of about 50% for temperature and 65% for precipitation. For temperature, the total reduction in standard error is comparable with that in the Monte Carlo experiment in section 2d. Despite these reductions in standard error quite substantial differences in variances can pass the test. For instance, for the monthly temperatures of NEU the changes in variance for the four seasons are 32%, −41%, 28%, and −34%, respectively. Only the largest of these changes (corresponding to a change in standard deviation of about 20%) is significant at the 5% level. Furthermore, for precipitation the observed changes in the variance of 37% (NEU, summer), 44% (CNA, spring), and 49% (NEU, autumn) are statistically significant at the 5% level, but this is not the case for the observed increase of 50% in the variance of monthly precipitation in SEU during winter.

The paper focused on monthly values. The presented jackknife procedure can, of course, also be used to compare the variances of seasonal values. However, for nearly normally distributed data, a test on the seasonal values (e.g., winter temperatures) has only about the same power as a test on the values for a particular calendar month (e.g., January temperatures). This is because $\text{var}(\hat{\theta}) \approx 2/J$ for both the monthly and seasonal values. For leptokurtic data, a seasonal mean or total will have much smaller kurtosis than the individual monthly values. It

is therefore possible that the proposed correction for correlation between pseudovalues can be applied to the variances of seasonal values but not to the variances of monthly values. Furthermore, var($\hat{\theta}$) will be smaller for the seasonal values due to their reduced kurtosis. This is advantageous for the power of a jackknife test on the seasonal variances.

Although there is strong evidence of an increase in the variance of monthly precipitation over NEU in the anomaly simulation, the relative variability or coefficient of variation (standard deviation divided by the mean) shows much less change. In principle, a test for equality of variation coefficients can be developed along the same lines as that for the variance in this paper. In case of absence of zero values, a test on the relative variability can also be obtained by applying the jackknife procedure to the variance of the logarithms of the monthly precipitation amounts.

Acknowledgments. The GCM data were kindly provided by D. Viner through the Climate Impacts LINK Project (Climatic Research Unit, University of East Anglia, Norwich, United Kingdom). The authors would like to thank the reviewers for their useful comments.

APPENDIX A

Estimation of Correlation between Pseudovalues

The estimates of the correlation coefficients ρ in Tables 2 and 3 were obtained from the Monte Carlo experiment as follows. Let $\overline{\theta^*_{j,m}}$ be the average pseudovalue for year j in the mth simulation ($j = 1, \ldots, J$; $m = 1, \ldots, M$), taken over N independent sequences. A natural estimate of ρ is then

$$\hat{\rho} = \frac{2\sum_{m=1}^{M}\sum_{i=1}^{J}\sum_{j=1}^{i-1}(\overline{\theta^*_{i,m}} - \overline{\theta^*})(\overline{\theta^*_{j,m}} - \overline{\theta^*})}{MJ(J-1)\hat{v}}, \quad \text{(A1)}$$

where

$$\overline{\theta^*} = \frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}\overline{\theta^*_{j,m}} \quad \text{and}$$

$$\hat{v} = \frac{1}{JM}\sum_{j=1}^{J}\sum_{m=1}^{M}(\overline{\theta^*_{j,m}} - \overline{\theta^*})^2.$$

For computational purposes it is more convienient to obtain $\hat{\rho}$ by (Koch 1983)

$$\hat{\rho} = (J\hat{v}_b - \hat{v})/[(J-1)\hat{v}], \quad \text{(A2)}$$

where

$$\hat{v}_b = \frac{1}{M}\sum_{m=1}^{M}(\overline{\theta^*_{\cdot m}} - \overline{\theta^*})^2, \quad \text{(A3)}$$

with

$$\overline{\theta^*_{\cdot m}} = \frac{1}{J}\sum_{j=1}^{J}\overline{\theta^*_{j,m}} \quad \text{(A4)}$$

the mean of the pseudovalues in the mth simulation. Equation (A2) cannot be used to estimate ρ from a single record, because the result $\hat{\rho} = -1/(J-1)$ for $M = 1$ does not depend on the true value of ρ.

APPENDIX B

Properties of Kurtosis Estimates

In a Monte Carlo study, Pearson (1935) observed that kurtosis estimates can be heavily biased. For the normal distribution, it can be shown (Cramér 1946) that $E(\hat{\gamma}_2) = -6/(J+1)$. Table B1 compares the mean of $\hat{\gamma}_2$ with the true kurtosis γ_2 for sample sizes encountered in this paper. For leptokurtic distributions ($\gamma_2 > 0$) the true kurtosis is seriously underestimated. The bias grows with increasing γ_2. For $J = 5$, $E(\hat{\gamma}_2) \approx -1$ for all distributions considered in Table B1, no matter their true kurtosis. This bias is partly caused by the boundedness of $\hat{\gamma}_2$. Expressions for the bounds of standardized sample moments are given in Dalén (1987). The upper bound of $\hat{\gamma}_2$ is 0.25, 5.11, and 25.03 for $J = 5$, 10, and 30, respectively. The sample kurtosis of a sample of size 5 from a Laplace distribution is thus always smaller than the true kurtosis.

The last column in Table B1 gives the mean of the pooled estimate $\hat{\gamma}_2$ in Eq. (16) for $n_s = 3$ samples of size 10 from the same distribution. The bias is roughly of the same order as that in a single sample of size 30. Note that for this sample size, $E(\hat{\gamma}_2) \approx 1$ for the two distributions with $\gamma_2 = 3$.

Besides the bias, the large variability of kurtosis estimates is a point of concern. For the leptokurtic distributions in Table B1, the standard deviation of $\hat{\gamma}_2$ increases with increasing J as a result of the growth of its upper bound. It is only for larger samples than those in Table B1 that var($\hat{\gamma}_2$) becomes proportional to $1/J$. Further, the standard deviation of the pooled estimate of 3×10 observations is somewhat smaller than that of a single sample

TABLE B1. Mean (first row) and standard deviation (second row) of kurtosis estimates for sequences of independent observations from various distributions (5000 simulations).

		$\hat{\gamma}_2$, Eq. (3)			$\hat{\gamma}_2$, Eq. (16)
Distribution	Kurtosis	$J = 5$	$J = 10$	$J = 30$	$J = 10$, $n_s = 3$
Uniform	−1.2	−1.11	−1.01	−1.11	−0.97
		0.53	0.54	0.26	0.33
Normal	0	−1.00	−0.54	−0.19	−0.21
		0.50	0.76	0.71	0.69
χ^2_{12}	1	−0.98	−0.42	0.26	0.17
		0.53	0.95	1.35	1.21
χ^2_4	3	−0.94	−0.16	1.12	0.89
		0.56	1.21	2.12	1.86
Laplace	3	−0.87	0.08	1.41	1.15
		0.54	1.13	1.86	1.67
Exponential	6	−0.88	0.19	2.27	1.88
		0.62	1.46	2.94	2.56

of size 30. Spatial averaging over grid points will strongly reduce the standard deviation of $\hat{\gamma}_2$.

REFERENCES

Bhaskaran, B., J. F. B. Mitchell, J. R. Lavery, and M. Lal, 1995: Climatic response of the Indian subcontinent to doubled CO_2 concentrations. *Int. J. Climatol.,* **15,** 873–892.

Boos, D. D., and C. Brownie, 1989: Bootstrap methods for testing homogeneity of variances. *Technometrics,* **31,** 69–82.

——, P. Janssen, and N. Veraverbeke, 1989: Resampling from centered data in the two sample problem. *J. Stat. Plan. Inf.,* **21,** 327–345.

Brown, M. B., and A. B. Forsythe, 1974: Robust tests for the equality of variances. *J. Amer. Stat. Assoc.,* **69,** 364–367.

Buishand, T. A., and J. J. Beersma, 1996: Statistical tests for comparison of daily variability in observed and simulated climates. *J. Climate,* **9,** 2538–2550.

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.

Cramér, H., 1946: *Mathematical Methods of Statistics.* Princeton University Press, 575 pp.

Dalén, J., 1987: Algebraic bounds on standardized sample moments. *Stat. Prob. Lett.,* **5,** 329–331.

Downton, M. W., and R. W. Katz, 1993: A test for inhomogeneous variance in time-averaged temperature data. *J. Climate,* **6,** 2448–2464.

Gordon, H. B., and B. G. Hunt, 1994: Climate variability within an equilibrium greenhouse simulation. *Climate Dyn.,* **9,** 195–212.

Gregory, J. M., and J. F. B. Mitchell, 1995: Simulation of daily variability of surface temperature and precipitation over Europe in the current $2\times CO_2$ climates using the UKMO climate model. *Quart. J. Roy. Meteor. Soc.,* **121,** 1451–1476.

IPCC, 1990: *Climate Change: The IPCC Scientific Assessment.* Cambridge University Press, 365 pp.

——, 1996: *Climate Change 1995: The Science of Climate Change.* Cambridge University Press, 572 pp.

Kendall, M., and A. Stuart, 1973: *The Advanced Theory of Statistics.* Vol. 2. 3d ed. Charles Griffin, 723 pp.

Keselman, H. J., P. A. Games, and J. J. Clinch, 1979: Tests for homogeneity of variance. *Commun. Stat.-Simul. Comput.,* **B8,** 113–129.

Koch, G. G., 1983: Intraclass correlation coefficient. *Encyclopedia of Statistical Sciences,* S. Kotz, N. L. Johnson, and C. B. Read, Eds., Vol. 4, Wiley and Sons, 212–217.

Liang, X.-Z., W.-C. Wang, and M. P. Dudeck, 1995: Interannual variability of regional climate and its change due to the greenhouse effect. *Global Planet. Change,* **10,** 217–238.

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 $2\times CO_2$ experiments. *J. Geophys. Res.,* **95,** 20 469–20 490.

Meehl, G. A., and W. M. Washington, 1993: South Asian summer monsoon variability in a model with doubled atmospheric carbon dioxide concentration. *Science,* **260,** 1101–1104.

——, M. Wheeler, and W. M. Washington, 1994: Low-frequency variability and CO_2 transient climate change. Part 3: Intermonthly and interannual variability. *Climate Dyn.,* **10,** 277–303.

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

Murphy, J. M., 1995: Transient response of the Hadley Centre Coupled Ocean–Atmosphere Model to increasing carbon dioxide. Part I: Control climate and flux adjustment. *J. Climate,* **8,** 36–56.

——, and J. F. B. Mitchell, 1995: Transient response of the Hadley Centre Coupled Ocean–Atmosphere Model to increasing carbon dioxide. Part II: Spatial and temporal structure of response. *J. Climate,* **8,** 57–80.

O'Brien, R. G., 1978: Robust techniques for testing heterogeneity of variance effects in factorial designs. *Psychometrika,* **43,** 327–342.

Pearson, E. S., 1935: A comparison of β_2 and Mr. Geary's W_n criteria. *Biometrika,* **27,** 333–352.

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

Raisanen, J., 1995: A comparison of the results of seven GCM experiments in northern Europe. *Geophysica,* **30,** 3–30.

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

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

Walsh, J. E., 1947: Concerning the effect of intraclass correlation on certain significance tests. *Ann. Math. Stat.,* **18,** 88–96.

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.

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

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