JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 95, NO. D1, PAGES 851–865, JANUARY 20, 1990

Statistical Comparison of Spatial Fields in Model Validation, Perturbation, and Predictability Experiments

T. M. L. WIGLEY

Climatic Research Unit, University of East Anglia, Norwich, England

B. D. SANTER

Max-Planck-Institut für Meteorologie, Hamburg, Federal Republic of Germany

The comparison of spatial fields of meteorological variables is an essential component of model validation studies and is central in assessing the significance of any change between a perturbed and control run of a general circulation model. Comparisons may be made of statistics which define the time-mean state, the temporal variability about this state, and/or spatial variability. Comparisons may also be made of the two time-mean spatial patterns, or of the temporal evolutions of spatial patterns. We consider here a suite of univariate and multivariate statistics which may be used to make these comparisons. Some of these statistics have been used previously, while others are either new or have not previously been used in the present context. The use of these statistics, their differences and similarities, and their relative performances are illustrated by considering mean sea level pressure changes between the decades 1951–1960 and 1971–1980 over an area covering North America, the North Atlantic Ocean, and Europe. Significance levels are assessed using the pool-permutation procedure of Preisendorfer and Barnett (1983) (henceforth P+B). This overcomes problems arising from nonideal behavior of the data (particularly spatial autocorrelation), unknown sampling distributions, and multiplicity in the case of univariate statistics. A subset of statistics is identified as most useful. For tests of differences in means these are the grid point by grid point t-test, a test comparing the overall means, and P+B's SITES statistic. For tests of differences in temporal variability they are the grid point by grid point F-test, and SPRET1 (the ratio of the spatial means of the time variances). SPRET1 is a modification of P+B's SPRED statistic designed to identify the direction of any variance difference. As a test of spatial variability differences, we identify SPREX1 (the ratio of the time means of the spatial variances), and for comparing spatial patterns the best statistic is the (spatial) correlation coefficient between the time-mean fields. For comparing the temporal evolution of spatial patterns, we recommend using the time-mean anomaly field correlation which is a more easily interpreted equivalent to P+B's SHAPE statistic.

1. INTRODUCTION

Many aspects of climatology and meteorology involve the use of general circulation models (GCMs) and the need to assess the differences between model-generated spatial fields and/or to compare model and observed fields. In model validation the output of a "control run" simulation of the present-day climate is compared with the observed climate in order to establish the degree of similarity of the two fields. In perturbation experiments, a control run and a "perturbed run" (in which an external forcing or boundary condition change has been made) are compared with a view to identifying a significant difference between the two fields. Model studies of climate predictability may involve similar comparisons, for example, between model variability in the presence and absence of temporally varying air-sea interactions and the observed variability [*Chervin*, 1986]. In addition to cases which involve GCM output, spatial field (or equivalent multivariate) comparisons are used in observational studies of climatic change. An example is the so-called "fingerprint method" used in seeking to positively identify the multivariate effects of atmospheric CO_2 concentration changes on the observed climate [*Wigley et al.*, 1985; *Barnett and Schlesinger*, 1987].

In all of these spatial field comparisons, the central issue is to assess the statistical significance of any differences between the fields. While straightforward in principle, the problem is made difficult in practice because the sample sizes involved are generally small (few model runs longer than 20 years have been performed). The magnitude of the task is increased by the fact that most field comparisons involve large spatial arrays, often many hundreds or even thousands of grid points. These issues have been recognized and addressed in important papers by *Hasselmann* [1979] and *Preisendorfer and Barnett* [1983] (henceforth P+B).

Three types of spatial field comparison can be identified: comparison of the means, comparison of the variances, and comparison of the spatial patterns. P+B have attempted to cover these by introducing a "trinity" of statistics. They describe the two multivariate space-time fields being compared as "n-point swarms in a common p-dimensional space" and define statistical measures of the swarms' relative centers of mass (their SITES statistic), relative sizes (SPRED, a measure of variability), and relative pattern evolutions (SHAPE). A novel feature of their method is that these statistics are themselves parts of a single p-dimensional separation statistic (L^2), a wholeness that is intellectually appealing but somewhat restrictive.

There are, of course, other ways that means, variances, and patterns can be and have been compared, some of which are rather more obvious and conceptually simpler than the statistics invented by P+B. For example, means and vari-

Paper number 89JD01617.
0148-0227/90/89JD-01617$05.00

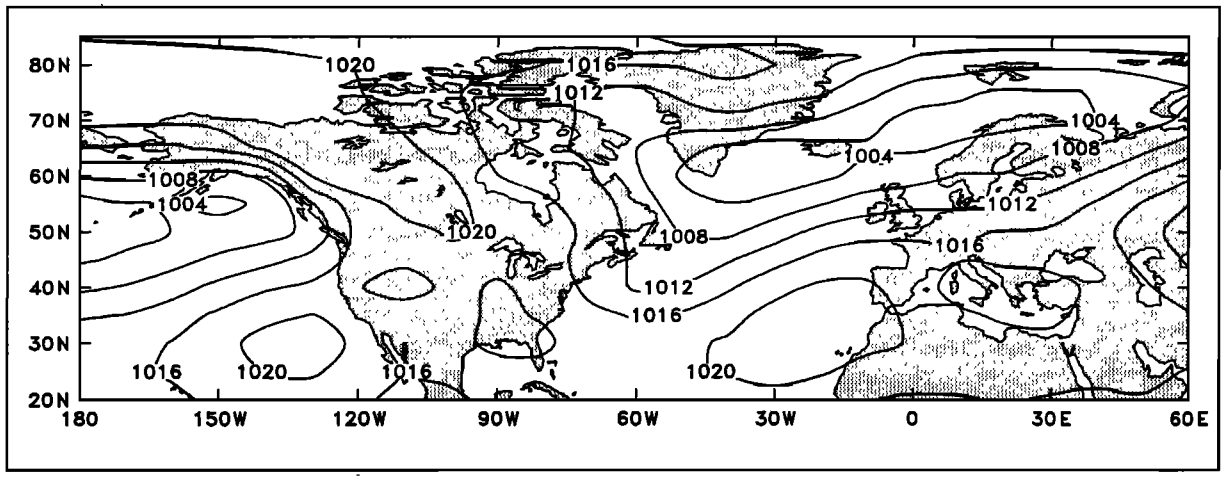

Fig. 1*a*

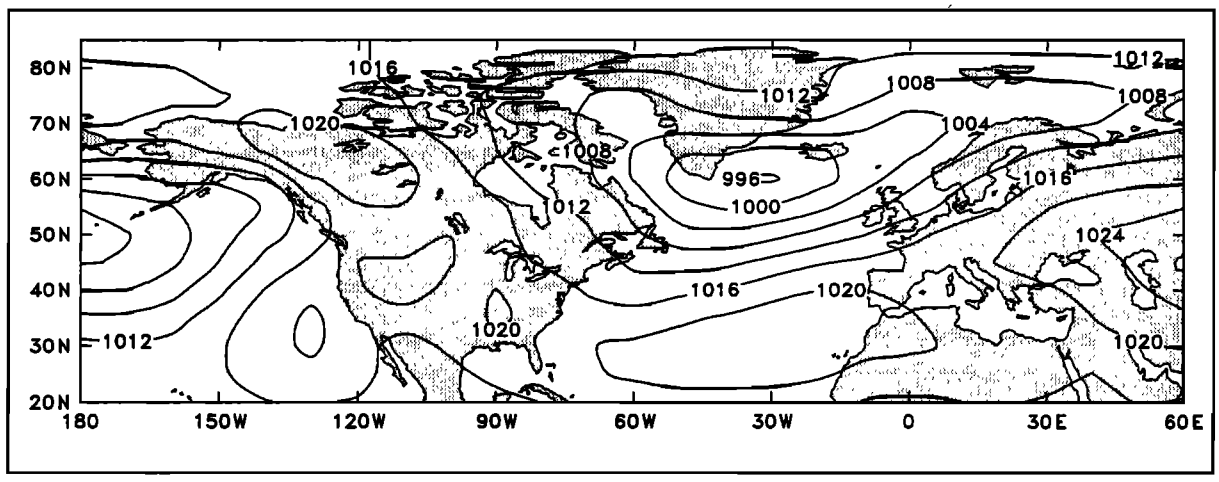

Fig. 1*b*

Fig. 1. Mean sea level pressures (millibars) for (*a*) January 1951–1960, (*b*) July 1951–1960, (*c*) January 1971–1980, and (*d*) July 1971–1980.

ances can be compared by applying conventional *t*-tests for the differences in means and *F*-tests for the variance ratios at each grid point, and spatial patterns can be compared by calculating the spatial correlation coefficient between the time-mean fields.

The difficulty with such methods (as indeed with P+B's trinity) lies in determining the statistical significance of the results. Only in very special circumstances is the sampling distribution of a test statistic known a priori, and the assumptions on which conventional *t*-tests and *F*-tests are based are, more often than not, invalid. In addition, in many multivariate cases, as noted by *von Storch* [1982] and *Livezey and Chen* [1983], there is a need to account for multiplicity (i.e., if many tests are performed, then a certain number would be "significant" at any prescribed level, purely by chance). Test results may also be affected by temporal and spatial autocorrelation. One way that significance can be assessed in such cases is to generate the sampling distribution from the available data using a permutation or Monte Carlo method [*Edgington*, 1987]. Examples in the meteorological literature include the PPP (pool permutation procedure) or APP (autocross permutation procedure) methods advocated by P+B.

Within each of the three types of field comparison noted above, there exists a number of different methods. What are the relationships between these different methods? For example, does the SITES statistic convey the same information as the accumulated results of individual grid point *t*-tests? If not, what are the differences? Which are the more sensitive indicators of the differences between spatial fields? Are there any other statistics that might be used? It is these questions that we seek to answer in this paper by comparing the significance results of a variety of different test statistics.

Our goal here therefore is to provide a review of some of the different methods that may be used to compare the means, variances, and patterns of spatial fields. Apart from the statistical measures already mentioned, we will introduce a number of other statistics that may be used for spatial field comparisons. The statistics described will be illustrated with examples which allow us to compare their practical value. This intercomparison, and our focus on a set of specific statistics, distinguishes our review from more descriptive and historically more comprehensive reviews such as that of *Livezey* [1985]. Further details of the methods described below are given by *Santer* [1988*a*].

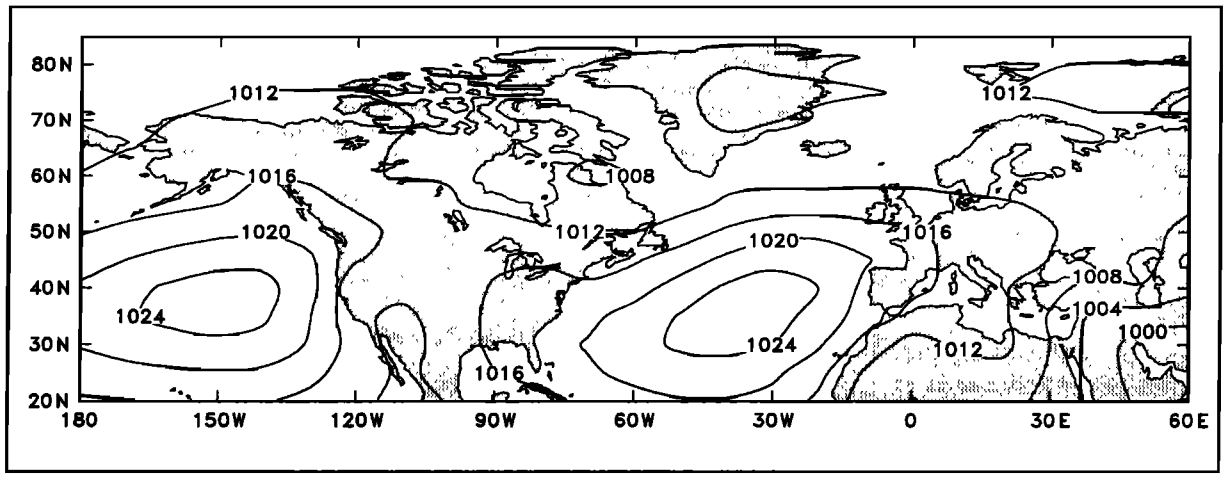

Fig. 1c

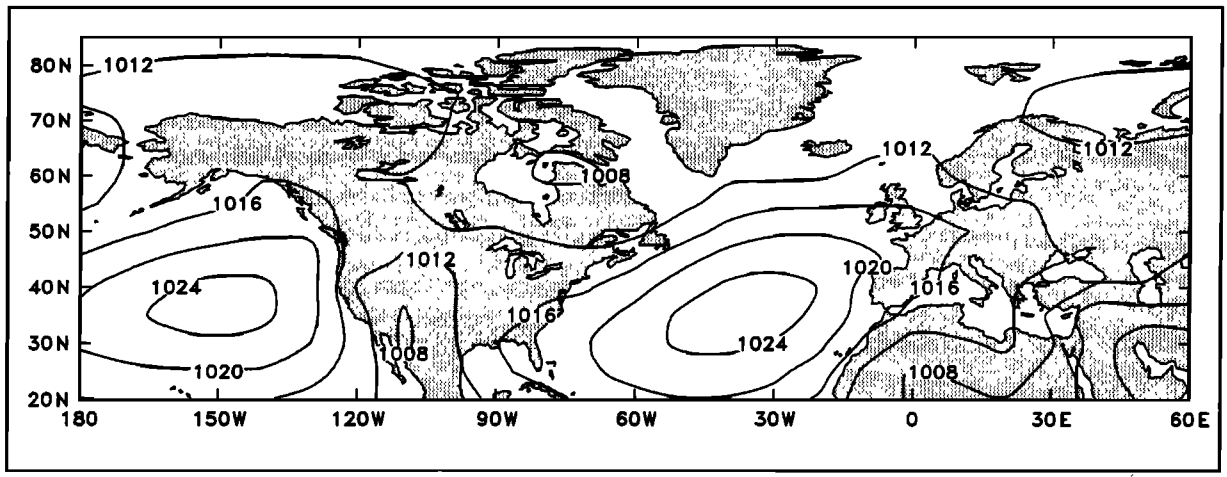

Fig. 1d

2. TERMINOLOGY

To clarify later calculations and to avoid confusion, we begin by summarizing the main terms. As far as is possible, we have employed the notation used by P+B.

We suppose that the two space-time fields to be compared are **D** and **M** (the letters may be identified with observed data and model output, as would be the case in a model validation experiment). Both fields may be multivariate of the form **D** = [**D1**(x, t), **D2**(x, t), $\cdots$] and comprised of three-dimensional spatial arrays of a set of variables **D1**, **D2**, etc., at various times t. To simplify the notation, we suppose the spatial information for all variables to be ordered sequentially with a one-to-one correspondence between **D** and **M**. **D** and **M** then become two-dimensional (x, t) arrays, and we can write the elements as d_{xt} and m_{xt}, where x and t are the independent discrete variables representing space and time ($x = 1, n_x$; $t = l, n_t$). (Note that for simplicity, we assume the temporal sample size, n_t, to be the same for **D** and **M**.) We are interested in comparing the spatial aspects (i.e., the x dependence) of **D** and **M**, so the x ordering is of prime importance, whereas the time ordering, in most cases, is of no direct consequence. Although the x dependence may involve more than one variable and more than two dimensions, it is convenient to imagine the **D** and **M** fields as two-dimensional spatial arrays, such as a mean sea level pressure field with values specified at a number of grid points (see Figure 1).

The main quantities that occur in our analysis are averages and variances over space and time, and various sums of squares. The sums of squares are the same as those which appear in an analysis of variance (of either **D** or **M**), and the use of these quantities considerably simplifies the computational aspects of the work. These items are defined in Table 1.

3. FIELD COMPARISON METHODS

In making spatial field comparisons, the underlying null hypothesis is that the two sample fields, **D** and **M**, come from the same population, i.e., that they have the same multivariate distributions. To test this hypothesis is clearly impractical; indeed, even in the univariate case, one would not normally attempt to test such a general hypothesis. Instead, we must break the problem down into a set of simpler and more restricted hypotheses.

In the univariate case, for example, we might formulate a set of hypotheses concerning specific moments of the distribution: means, variances, etc. Of course, these restricted hypotheses need not sum to the general hypothesis, so they represent a compromise solution, less general in that they

TABLE 1. Definitions of Statistical Quantities

Quantity	Definition
d_{xt}	value of d at point x and time t ($x = 1, n_x$; $t = 1, n_t$)
$\bar{d}_{.t}$	spatial average of d at time $t = \Sigma_x\ d_{xt}/n_x$
$\bar{d}_{x.}$	time average of d at point $x = \Sigma_t\ d_{xt}/n_t$
$s^2_{d,t}$	spatial variance of d at time $t = \Sigma_x\ (d_{xt} - \bar{d}_{.t})^2/n_x$
$s^2_{d,x}$	time variance of d at point $x = \Sigma_t\ (d_{xt} - \bar{d}_{x.})^2/n_t$
$\langle d \rangle$	grand mean $= (\Sigma_x\ \Sigma_t\ d_{xt})/n_x n_t$
GSSD	total (grand) sum of squares $= \Sigma_x\ \Sigma_t\ (d_{xt} - \langle d \rangle)^2 = (\Sigma_x\ \Sigma_t\ d^2_{xt}) - n_x n_t \langle d \rangle^2$
SSTD	within-x sum of squares $= n_x\ \Sigma_t\ (\bar{d}_{.t} - \langle d \rangle)^2 = n_x\ (\Sigma_t\ \bar{d}^2_{.t}) - n_x n_t \langle d \rangle^2$
SSXD	between-x sum of squares $= n_t\ \Sigma_x\ (\bar{d}_{x.} - \langle d \rangle)^2 = n_t(\Sigma_x\ \bar{d}^2_{x.}) - n_x n_t \langle d \rangle^2$
SSED	error sum of squares = GSSD − SSTD − SSXD
$V(\bar{d}_{x.})$	spatial variance of $\bar{d}_{x.} = \Sigma_x\ (\bar{d}_{x.} - \langle d \rangle)^2/n_x =$ SSXD$/n_x n_t$
$V(\bar{d}_{.t})$	time variance of $\bar{d}_{.t} = \Sigma_t\ (\bar{d}_{.t} - \langle d \rangle)^2/n_t =$ SSTD$/n_x n_t$
$\overline{s^2_{d,t}}$	time average of $s^2_{d,t}$ = (GSSD − SSTD)$/n_x n_t$
$\overline{s^2_{d,x}}$	spatial average of $s^2_{d,x}$ = (GSSD − SSXD)$/n_x n_t$
σ^2_D	$n_x n_t \overline{s^2_{d,x}} = \Sigma_x\ \Sigma_t\ (d_{xt} - \bar{d}_{x.})^2$ = (GSSD − SSXD) (used by P+B)

Definitions are for **D**; **M** items are defined similarly.

cover only low-order moments, but less demanding as well. Testing the equality of means, for instance, may or may not require equality of variances, depending on the method used for hypothesis testing; and standard tests for the equality of variances do not require equality of means (although the results do depend on the relative values of the means).

In the multivariate case we can also break the problem down into a set of restricted hypotheses. This is useful, partly because it makes it easier to formulate hypotheses, but also because it reflects the assumption that the relevant underlying distributions can be described by a small number of moments. Our primary concerns tend to be with means and variances. However, in multivariate problems, these moments can be defined in a number of different ways. In other words, we do not have a single mean or variance to consider, but a number of different possible means and variances. In addition, in multivariate problems we may wish to consider other data properties (such as spatial patterns) that are not simply described by the moments of individual variables.

Let us consider the means first. There are different levels of complexity in the way we can characterize the mean conditions. The simplest mean value is the grand mean, obtained by averaging over both space and time. Other single statistics that characterize the whole spatial field include P+B's SITES statistic, which is related to the mean of the squared differences in the time means at each spatial point, and *Mielke*'s [1985] δ statistic. At the next level of complexity we might examine the individual differences in time means at each point. (Such an analysis may be compressed to a lower level of complexity by considering only a summary statistic, such as the fraction of points that show significant differences at some prescribed significance level.) Finally, we may examine the n_x-dimensional vector of differences in means and ask whether the end of this vector lies close enough to the origin for one to accept the null hypothesis that it is drawn from a population with a zero mean vector (e.g., using Hotelling's T^2 statistic). This last possibility presents some practical difficulties because, in the present applications, n_x is generally many times larger than n_t, and it is usually necessary to reduce the dimensionality using spatial averaging, principal component analysis, harmonic analysis, etc.

In principle, the same hierarchy of tests can be formulated for an examination of time variances, but in this case, fewer of the possibilities are of interest. The overall space-time variance is more difficult to interpret than the space-time mean, so the lowest level of useful test statistic corresponds to those statistics which describe variance properties representative of the whole spatial field. Examples are P+B's SPRED statistic or variations on this theme, and the fraction of points that show locally significant differences in variance.

In the multivariate case, we can define spatial as well as temporal variances. Analogs of SPRED can easily be defined, but these have never been considered in the literature.

Although we may not be interested in spatial variability per se, we are interested in the spatial character of the two fields being compared. How does one compare the spatial patterns of two data sets? Some information can be gained by examining the spatial distribution of significance levels for local t-tests of differences in means or F-tests of variance ratios. Spatial patterns can also be compared more directly by correlating the time-mean fields. P+B have introduced a statistic, SHAPE, which is also a measure of spatial pattern similarity. However, SHAPE depends not only on the time-mean spatial patterns, but also on similarities in the temporal evolutions of the two data sets being compared. When time evolution is important, other statistics may also be used, for example, the mean over time of the spatial anomaly correlations.

3.1. *Comparison of Means*

The obvious (and usual) method is to compare the time averages ($\bar{d}_{x.}$ and $\bar{m}_{x.}$) grid point by grid point, assessing

significance by using a test for the difference in means [e.g., *Chervin and Schneider*, 1976]. However, while local significance is easily calculated, field significance is affected by spatial autocorrelation and is not easy to evaluate [e.g., *Livezey and Chen*, 1983]. A second method is to use P+B's SITES statistic. A third method would be to compare the grand means of d_{xt} and m_{xt} averaged over time and space. Finally, we could use a multivariate response permutation procedure (MRPP), a technique applied in the evaluation of weather modification experiments (e.g., *Mielke* [1985] and earlier references cited therein) but not previously applied in the present context. We will consider the grand means method first. Note that in a number of the statistics given in this section, one could equally well use an area-weighted mean.

Grand means. The statistical significance of the difference in grand means (i.e., $\langle d \rangle - \langle m \rangle$) can be estimated using an appropriate t-test. The test statistic is

$$t = (\langle d \rangle - \langle m \rangle)/S \tag{1}$$

where S^2 is a measure of the variance of the sampling distribution of $\langle d \rangle - \langle m \rangle$ and contains contributions from d and m. There are (at least) three ways that S^2 can be calculated: by ignoring the distinction between x and t and considering the overall d and m variances, by considering $\langle d \rangle$ and $\langle m \rangle$ as spatial averages of $\bar{d}_{x.}$ and $\bar{m}_{x.}$ and using the spatial variances of these time means, or by considering $\langle d \rangle$ and $\langle m \rangle$ as time averages of $\bar{d}_{.t}$ and $\bar{m}_{.t}$ and using the time variances of these spatial means. In the first case it can be shown that

$$S^2 = S_1^2 = (\mathrm{GSSD} + \mathrm{GSSM})/[n_x n_t (n_x n_t - 1)] \tag{2}$$

in the second case

$$S^2 = S_2^2 = \left[\mathrm{SSXD} - \frac{\mathrm{SSED}}{n_t - 1} + \mathrm{SSXM} - \frac{\mathrm{SSEM}}{n_t - 1} \right] \Big/ [n_x n_t (n_x - 1)] \tag{3}$$

and in the third case

$$S^2 = S_3^2 = \left[\mathrm{SSTD} - \frac{\mathrm{SSED}}{n_x - 1} + \mathrm{SSTM} - \frac{\mathrm{SSEM}}{n_x - 1} \right] \Big/ [n_x n_t (n_t - 1)] \tag{4}$$

The SSE terms appear because the SSX and SST values are not unbiased estimators of their corresponding population parameters, a standard result. Tests using the above test statistics assume that the d and m variances come from the same population. Modifications are required if the d and m variances are significantly different. Alternatively, one could use a paired t-test.

With the above test statistics, one might possibly use Student's distribution with $2n_x n_t - 2$, $2n_x - 2$, or $2n_t - 2$ degrees of freedom (for S_1^2, S_2^2, and S_3^2, respectively) to assess significance. However, the appropriateness of this distribution depends on a number of assumptions which may or may not be valid. The most important of these in the present context is that the data have no serial correlation, i.e., spatial autocorrelation in the case of (2) and (3) and temporal autocorrelation of the spatial means in (4).

A more serious practical disadvantage with these tests arises because different parts of the time-space fields may have large, significant, but compensating differences which could lead to no overall difference between $\langle d \rangle$ and $\langle m \rangle$. Because of this, the grand mean tests may be misleading. Nevertheless, they are important, since any overall spatial bias will affect the results of the other tests. The value of $\langle d \rangle - \langle m \rangle$, and its significance or otherwise, is useful in interpreting the results of other tests of means.

The SITES statistic. SITES is proportional to the spatial mean of the squared differences between the individual grid point time averages, standardized using the spatial means of the individual grid point time variances, i.e.,

$$\mathrm{SITES} = n_t \sum_x (\bar{d}_{x.} - \bar{m}_{x.})^2 / \sigma_D \sigma_M \tag{5}$$

The notation follows P+B, and σ is defined in Table 1.

SITES has clear similarities with the grand mean t values (which involve $\sum_x (\bar{d}_{x.} - \bar{m}_{x.})$ in the numerator) but, by squaring the difference in means, avoids the problem of possible compensating negative and positive difference regions. Comparison with the second of the grand mean t-tests suggests that standardization using estimates of the variances of $\bar{d}_{x.}$ and $\bar{m}_{x.}$ might be more appropriate than standardization using the spatial average of the grid point time variances. This gives an alternative to SITES, viz.,

$$\mathrm{SITES1} = \sum_x (\bar{d}_{x.} - \bar{m}_{x.})^2 / [n_x (n_x - 1) S_2^2] \tag{6}$$

where S_2^2 is given by (3). P+B give no justification for using $\sigma_D \sigma_M$ as a divisor, but their primary motive is clearly the obvious one of producing a nondimensional statistic.

Apart from argument by analogy with the conventional t-test, however, SITES1 has no clear advantage over SITES. Both statistics have unknown sampling distributions, and Monte Carlo techniques must be employed to assess significance. Spatial autocorrelation must affect the sampling distributions, since the statistics both involve a spatial average. In this case, however, this would automatically be accounted for in a Monte Carlo simulated sampling distribution.

Grid point by grid point analysis. Here, the time means are compared at each grid point using a local test for the difference in means. The local test may be either one- or two-tailed; in the examples given later, we consider only the second case. If the d and m variances do not differ significantly, then the test statistic is

$$t = (\bar{d}_{x.} - \bar{m}_{x.})/S_x \tag{7}$$

where

$$S_x^2 = (s_{d,x}^2 + s_{m,x}^2)/(n_t - 1) \tag{8}$$

Significance of t is conventionally (but not necessarily correctly) assessed using Student's distribution with $2n_t - 2$ degrees of freedom (assuming Normally distributed data, no temporal autocorrelation, and equal d and m variances). For assessing field significance, an appropriate test statistic is the fractional number of successes (n_s/n_x), where "success"

refers to a locally significant result at some prescribed local significance level.

While local significance is relatively easily judged for a single grid point in isolation, the issue of field significance is more difficult to settle for two reasons. First, with a multiplicity of tests (one at each of many grid points), the probability of a significant result occurring by chance increases. (The expected number of successes is $E\{n_s\} = \alpha n_x$, where α is the prescribed local significance level.) Second, spatial autocorrelation will mean that not all tests are independent. The effective number of independent tests will be less than n_x, and the effective number of successes will be similarly (but not necessarily proportionally) reduced (this is explained in detail by *Wigley and Santer* [1988]). Multiplicity may be accounted for using the binomial distribution [e.g., *Livezey and Chen*, 1983], but assessing the effects of spatial autocorrelation requires a Monte Carlo analysis.

MRPP. For the special case considered here, comparison of only two spatial fields using the same number (n_t) of realizations of each, the test statistic is based on the generalized distance measure

$$\Delta_{tu} = \left[\sum_x (Y_{xt} - Y_{xu})^2 \right]^{\nu/2} \tag{9}$$

where Y may be either d or m and ν is an arbitrary exponent, usually taken to be 1 or 2. (In principle, Y may be any d or m variable.) The Δ_{tu} are first averaged over all (t, u) permutations ($t \neq u$) for both **D** and **M**, and then these two values are averaged to give

$$\delta_0 = [n_t(n_t - 1)]^{-1} \sum_{u<t} \left\{ \left[\sum_x (d_{xt} - d_{xu})^2 \right]^{\nu/2} + \left[\sum_x (m_{xt} - m_{xu})^2 \right]^{\nu/2} \right\} \tag{10}$$

To assess the significance of δ_0, the **D** and **M** arrays need to be combined and then repartitioned into two new arrays. The number of distinct repartitionings (given that the order is unimportant) is

$$N = (2n_t)!/[2(n_t!)^2]$$

For each repartitioning, a new δ value can be calculated using (10), and the significance of δ_0 can be assessed against the sampling distribution constructed from the N new δ values. If $\langle d \rangle$ and $\langle m \rangle$ differ noticeably, then δ_0 will have a small value relative to most of the sampling distribution values. (This can be seen clearly from (11) below. If $\langle d \rangle$ and $\langle m \rangle$ differ noticeably, combining, shuffling, and repartitioning **D** and **M** will generally increase both σ_D^2 and σ_M^2, so that δ_0 should be small relative to the sampling distribution of δ.) If N_0 is the number of δ values less than δ_0, then the observed significance level is

$$p = N_0/N$$

In practice it may be both too time consuming and computationally inefficient for one to calculate all the distinct δ values. Instead, random repartitionings can be made and the significance level assessed using the approximate cumulative sampling distribution so generated. From experience, 500–1000 random permutations are usually sufficient. This is equivalent to P+B's pool-permutation procedure. Alternatively, the random permutations may be used to estimate the mean, variance, and skewness of the sampling distribution, and these may then be used as parameters in a Pearson type III distribution, which the δ distribution approaches asymptotically [*Mielke*, 1985].

Since Δ_{ut} involves $(Y_{xt} - Y_{xu})^2$, MRPP is not affected by the possibility of compensating areas of negative and positive differences between d and m.

For the special case of $\nu = 2$, Mielke's δ_0 statistic can be expressed in terms of the sums of squares defined previously. Equation (10) becomes

$$\delta_0 = (\text{GSSD} + \text{GSSM} - \text{SSXD} - \text{SSXM})/(n_t - 1) = (\sigma_D^2 + \sigma_M^2)/(n_t - 1) \tag{11}$$

In the case where Y_{xt} is replaced by the spatial mean values, $\bar{Y}_{.t}$, (10) becomes

$$\delta_0 = (\text{SSTD} + \text{SSTM})/[n_x(n_t - 1)] \tag{12}$$

In spite of the simplified form for the case $\nu = 2$ given by (11), $\nu = 2$ is not necessarily the optimum choice. *Mielke* [1985] states that $\nu = 1$ is substantially superior in many cases (especially for non-Normal data). Clearly, $\nu = 1$ is less sensitive to distortion by isolated and possibly unrepresentative outliers.

In the present application, the δ statistic has some obvious drawbacks. In its most general form it is computationally demanding to calculate, and its interpretation is not straightforward. We will show below that it often gives information which overlaps with other, simpler, statistics.

Summary. For the last three tests, significance must be assessed using a permutation or Monte Carlo procedure. This has the advantage that some of the restrictions attendant on tests which use a theoretical null distribution are avoided. Note that in the present context this does not remove problems associated with temporal autocorrelation, but at least on a monthly or longer time scale, temporal autocorrelation is small for many meteorological variables.

Although all tests deal with mean values, the relationships between the tests are not immediately clear. Consider the grid point by grid point test compared with SITES. If there are a large number of t-test successes, then many of the terms $(\bar{d}_{x.} - \bar{m}_{x.})^2$ that are summed in SITES will be large, ensuring a large value for SITES. But how many successes are required to give a significant SITES value? Which of these two tests is the more sensitive?

One important difference is that the grid point t-tests require a local significance level to be specified. This could be viewed as being unnecessarily restrictive, since it provides only a discrete yes/no answer at each grid point rather than a continuous assessment as in SITES or δ_0. However, since the local significance level may be varied, the tests can provide additional (spatial) information that SITES and δ_0 cannot provide. For example, by using a set of different local significance levels, α = 1%, 5%, and 10%, say, one might distinguish between field significance due to an area of highly significant local differences (α = 1%) and field significance due to an area of weakly significant local differences (α = 10%). Equivalent (and additional) information could be obtained by mapping and examining the pattern of local observed significance levels.

Relationships between the tests will be discussed further in a later section.

3.2. *Comparison of Temporal Variances*

Variances are usually compared by testing the variance ratio using an F-distribution. In comparing **D** and **M**, this can be done grid point by grid point, paralleling the method for comparing means. P+B have introduced an alternative measure, the SPRED statistic. Generalizations of SPRED have been defined by *Preisendorfer and Mobley* [1982]. Here we will define other variance comparison statistics.

The SPRED statistic. P+B define SPRED by

$$\text{SPRED} + 2 = n_t \sum_x (s^2_{d,x} + s^2_{m,x})/\sigma_D\sigma_M \qquad (13)$$

The logic behind this statistic becomes clear when it is rewritten as

$$\text{SPRED} = (\sigma_D - \sigma_M)^2/\sigma_D\sigma_M$$

$$= \sqrt{\sigma^2_M/\sigma^2_D} + \sqrt{\sigma^2_D/\sigma^2_M} - 2 \qquad (14)$$

Thus if the ratio of the spatial means of the time variances is high (either $\sigma^2_D > \sigma^2_M$ or $\sigma^2_M > \sigma^2_D$), then SPRED will be noticeably greater than zero (SPRED = 0 when $\sigma^2_D = \sigma^2_M$). The SPRED value does not, however, indicate which of these mean variances is the higher. Since the sampling distribution of SPRED is unknown, its significance can only be assessed by Monte Carlo or permutation methods.

SPRET1. Since SPRED is effectively a variance ratio statistic, it seems logical to consider this variance ratio directly. We therefore define SPRET1 as the ratio of the spatial means of the time variances,

$$\text{SPRET1} = \overline{s^2_{d,x}}/\overline{s^2_{m,x}} = \sigma^2_D/\sigma^2_M \qquad (15)$$

This can be written in terms of sums of squares as

$$\text{SPRET1} = (\text{GSSD} - \text{SSXD})/(\text{GSSM} - \text{SSXM}) \qquad (16)$$

Either large or small values of SPRET1 will be significant, so SPRET1 can be tested using a two-tail test if one is only searching for a difference in variances, or using a one-tail test if one is searching for a directional difference. In this regard, SPRET1 is more useful than SPRED, but otherwise, since SPRED + 2 tends to either SPRET1 or 1/SPRET1 for high or low variance ratios, the two statistics provide equivalent significance information when SPRET1 is tested using a two-tailed test.

SPRET2. SPRET1 involves the ratio of spatial means of the time variances. A logical complement to this would be to consider the ratio of the time variances of the spatial means, SPRET2. Biased estimates of these variances are given by $V(\tilde{d}_{.t}) = \text{SSTD}/n_x n_t$ and $V(\bar{m}_{.t}) = \text{SSTM}/n_x n_t$ (Table 1). Instead of taking the ratio of these, we could choose to use the unbiased variance estimates:

$$\hat{V}(\tilde{d}_{.t}) = [\text{SSTD} - \text{SSED}/(n_x - 1)]/n_x(n_t - 1)$$

and

$$\hat{V}(\bar{m}_{.t}) = [\text{SSTM} - \text{SSEM}/(n_x - 1)]/n_x(n_t - 1)$$

However, since the significance of this statistic must generally be assessed using Monte Carlo methods, there is little to be gained by using the unbiased variances, and we therefore define SPRET2 as

$$\text{SPRET2} = V(\tilde{d}_{.t})/V(\bar{m}_{.t}) = \text{SSTD}/\text{SSTM} \qquad (17)$$

Grid point by grid point analysis. Here the time variances $s^2_{d,x}$ and $s^2_{m,x}$ are compared grid point by grid point using the test statistic

$$F = s^2_{d,x}/s^2_{m,x} \qquad (18)$$

and judging local significance using an F-distribution with $n_t - 1$, $n_t - 1$ degrees of freedom. Either a two- or one-tail test may be appropriate, depending on the hypothesis being tested (i.e., that the **D** and **M** variances differ, or that one is significantly greater than the other). In the examples given later we use a two-tail local test. Field significance can be assessed using the fractional number of successes (locally significant results). As for the grid point comparison of means, due account must be taken of multiplicity and of spatial autocorrelation in assessing field significance. Until a reliable method for estimating effective sample size is developed, the latter requires the use of Monte Carlo methods.

This type of analysis gives similar information to that obtained from SPRED or (more obviously) SPRET1. As in the case of local t-tests, grid point by grid point F-tests are both more restrictive (in requiring a local significance level to be prescribed and thus only a yes/no result at each grid point) and more flexible (in that the spatial character of the differences in variance can be identified by varying the prescribed local significance level, or by examining the spatial pattern of the observed significance levels).

3.3. *Comparison of Spatial Variances*

The two obvious measures here are the ratio of the time means of the spatial variances (SPREX1, analogous to SPRET1) and the ratio of the spatial variances of the time means (SPREX2, analogous to SPRET2).

SPREX1. The time mean of the **D** spatial variances, $\overline{s^2_{d,t}}$ is defined in Table 1. SPREX1 is defined as the ratio of the time-mean spatial variances, i.e.,

$$\text{SPREX1} = \overline{s^2_{d,t}}/\overline{s^2_{m,t}}$$

$$= (\text{GSSD} - \text{SSTD})/(\text{GSSM} - \text{SSTM}) \qquad (19)$$

SPREX2. SPREX2 is similar to SPREX1 except that the order of the variance and average operations is reversed; i.e., SPREX2 is the ratio of the spatial variances of the time-mean fields of **D** and **M**. A straightforward measure of this spatial variance is, for **D**, $V(\tilde{d}_{x.}) = \text{SSXD}/n_x n_t$ (see Table 1). An unbiased estimator of the corresponding population parameter is $\hat{V}(\tilde{d}_{x.}) = [\text{SSXD} - \text{SSED}/(n_t - 1)]/[n_t(n_x - 1)]$. For consistency with SPRET2, however, we define SPREX2 using the biased estimators,

$$\text{SPREX2} = V(\tilde{d}_{x.})/V(\bar{m}_{x.}) = \text{SSXD}/\text{SSXM} \qquad (20)$$

3.4. *Comparison of Spatial Patterns*

An obvious method here is to calculate the correlation coefficient between the two time-mean fields. P+B's SHAPE statistic is related to this measure, but it is determined not only by spatial similarities between the time-mean

fields, but also by similarities in the way the **D** and **M** spatial patterns evolve through time. The mean anomaly correlation provides an easily interpretable alternative to SHAPE.

Correlating the time-mean fields. The correlation coefficient between the time-mean fields is defined by

$$r=\left[\sum_x (\bar{d}_{x.}-\langle d\rangle)(\bar{m}_x-\langle m\rangle)\right]\Big/\left[n_x\sqrt{V(\bar{d}_x)V(\bar{m}_{x.})}\right] \quad (21)$$

where $V(\bar{d}_{x.})$ and $V(\bar{m}_{x.})$ are spatial variances of the time-mean fields (see Table 1). Hence

$$r = n_t\left[\sum_x \bar{d}_{x.}\bar{m}_{x.}-n_x\langle d\rangle\langle m\rangle\right]\Big/\sqrt{\text{SSXD SSXM}} \quad (22)$$

As noted above, the sample spatial variances are biased estimators of the corresponding population parameters. Thus an "unbiased" analog to r ($\hat{r}$) could be defined using $\hat{V}(\bar{d}_{x.})$ and $\hat{V}(\bar{m}_{x.})$ in (21). If n_t is small, r and $\hat{r}$ may differ noticeably, with $\hat{r}$ necessarily greater than r (see *Mitchell et al.* [1987] for some numerical examples). Whether or not one uses r or $\hat{r}$, however, is somewhat arbitrary. The conventional test for a correlation coefficient would be distorted by the presence of spatial autocorrelation, so the significance of both statistics must be assessed using Monte Carlo methods. Both statistics should yield similar observed significance levels.

There is an important distinction between a Monte Carlo test of r and the usual tests for correlation coefficients, which assume a specific sampling distribution. In the latter, the test is to determine whether the correlation coefficient differs significantly from zero or some a priori determined value. In the present situation, however, we are concerned to find out whether the observed correlation differs significantly from the unknown value near 1 which would arise if **D** and **M** were drawn from the same population. With a Monte Carlo significance assessment, this presents no problem.

The SHAPE statistic. P+B's SHAPE statistic is defined by

$$\text{SHAPE} + \text{SITES} + \text{SPRED} = L^2/\sigma_D\sigma_M \quad (23)$$

where

$$L^2 = \sum_x \sum_t (d_{xt} - m_{xt})^2 \quad (24)$$

is a gross measure of the separation between **D** and **M**. SHAPE can be expressed in the following form:

$$\text{SHAPE} = 2 - \left(2n_x \sum_t C_t\right)\Big/\sigma_D\sigma_M \quad (25)$$

where C_t is the covariance between the two anomaly fields at time t defined by

$$C_t = \sum_x (d_{xt} - \bar{d}_{x.})(m_{xt} - \bar{m}_{x.})/n_x \quad (26)$$

SHAPE and r differ in a very important way. The statistic r does not depend on the temporal evolution of the **D** and **M** fields, whereas, since SHAPE involves a spatial covariance at each time slice t, it must depend critically on the details of the temporal evolutions of both **D** and **M**. In many of the situations in which spatial fields are compared (e.g., equilibrium GCM validations or perturbation experiments), the precise temporal evolution is of no consequence, so SHAPE would be an inappropriate statistic for the comparison of spatial patterns in these cases.

SHAPE has a minor illogicality in its formulation when compared with a conventional correlation coefficient. The covariance in SHAPE is standardized using the spatial means of the time variances. However, in producing the equivalent correlation coefficient, r_t, from the covariances C_t, standardization is achieved using the spatial variances (at time t). It would therefore be more logical to replace $\sigma_D\sigma_M$ in the definition of SHAPE by the corresponding term using the time means of the spatial variances to give

$$\text{SHAPE1} = 2 - \left(2n_x \sum_t C_t\right)\Big/(\text{GSSD}-\text{SSTD})(\text{GSSM}-\text{SSTM}) \quad (27)$$

(recall that $\sigma_D\sigma_M$ = (GSSD − SSXD)(GSSM − SSXM)). Note that this same minor inconsistency occurs in P+B's SITES statistic; it led to our defining the alternative SITES1 (equation (6)). In both cases it arises because of the constraint of relating SITES, SPRED, and SHAPE directly to L^2, a constraint that is mathematically elegant, but unnecessary.

The mean anomaly correlation. If one were to define a statistic which was an indicator of spatio-temporal similarities in **D** and **M** independent of the "trinity" constraint of P+B's formalism, then the simplest indicator would be the time mean of the anomaly correlations

$$\bar{r} = \sum_t r_t/n_t \quad (28)$$

where

$$r_t = \sum_x [(d_{xt} - \bar{d}_{x.}) - (\bar{d}_{.t} - \langle d\rangle)] \cdot [(m_{xt} - \bar{m}_{x.}) - (\bar{m}_{.t} - \langle m\rangle)]/(n_x\bar{s}_{d,t}\bar{s}_{m,t}) \quad (29)$$

and

$$\bar{s}_{d,t} = \sum_x [(d_{xt} - \bar{d}_{x.}) - (\bar{d}_{.t} - \langle d\rangle)]^2/n_x \quad (30)$$

($\bar{s}^2_{m,t}$ similarly). Analysis of individual r_t values can provide useful insights into the **D** − **M** differences (see, for example, *Briffa et al.* [1986], where the **M** fields correspond to pressure pattern reconstructions based on a spatial array of tree-ring data).

As with r, since the sampling distributions of SHAPE (or SHAPE1) and $\bar{r}$ are unknown, their statistical significance can only be judged using permutation or Monte Carlo methods. There is an important distinction, however. For r the null distribution corresponds to identity of the populations from which **D** and **M** are drawn, whereas for SHAPE and $\bar{r}$, the null distribution corresponds to total dissimilarity in the temporal evolutions of the spatial anomaly patterns. With r therefore the test is for significant differences between the time-mean fields, whereas for SHAPE and $\bar{r}$ the test is for a significant similarity in the spatiotemporal characters of **D** and **M**.

TABLE 2. Summary of Test Statistics

Statistics	Designed to Test . . .	Defining Equations	H_0 Value*
T1	Differences between grand means	(1), (2)	0
T2	Differences between grand means	(1), (3)	0
T3	Differences between grand means	(1), (4)	0
SITES	Overall difference in means	(5)	†
SITES1	Overall difference in means	(6), (3)	†
NT1 and NT5	Differences in time means, grid point by grid point. NT1 and NT5 are the fraction of grid points with significant differences at the 1% and 5% level, respectively.	(7), (8)	1[5]
DELTA1 and DELTA2	Overall difference in means based on spatial mean values	(10) using $\tilde{Y}_{.t}$ and $\nu = 1$; (12) for $\nu = 2$	†
DELTA3	Overall difference in means using original data	(10) using Y_{xt} and $\nu = 2$ (equivalent to (11))	†
SPRED	Overall difference in temporal variances	(13), (14)	0
SPRET1	Overall difference in temporal variances	(15), (16)	1
SPRET2	Overall difference in temporal variances	(17)	1
NF1 and NF5	Differences in temporal variances, grid point by grid point (cf. NT1 and NT5 above)	(18)	1[5]
SPREX1	Overall difference in spatial variances	(19)	1
SPREX2	Overall difference in spatial variances	(20)	1
r	Differences in spatial patterns of time-mean fields	(21), (22)	0.9–0.99
SHAPE	Similarities in spatiotemporal evolution	(25), (26)	2
$\bar{r}$	Similarities in spatiotemporal evolution	(28)–(30)	0

*Value under the null hypothesis.
†Value is unknown.

The various statistics described above are summarized in Table 2.

4. EXAMPLES

In order to illustrate the use of this set of statistics, and to compare the types of information provided by the different statistics within each group, we will use observed mean sea level pressure (MSLP) data over the North America/North Atlantic/European region (20°–85°N, 180°W—60°E). The data are from the United Kingdom Meteorological Office gridded data set, which are on a 5° latitude by 10° longitude grid (20° longitude at 70°N or above). For further details, see *Williams and Van Loon* [1978] and *Jones* [1987]. We will examine two decades, 1951–1960 and 1971–1980 and test for possibly significant changes in climate. This particular region has been chosen for two reasons: its agricultural importance and relevance to a significant fraction of the world's population, and because it encompasses two of the main "centers of action" of the northern hemisphere circulation, the Azores High and the Iceland Low.

We have chosen to compare two periods of observed data rather than, for example, compare a GCM simulation with observed data, in order to obtain results where the differences are close to the border between overall significance and nonsignificance. For a number of published GCM control runs, the model simulations differ so obviously from observations that all tests of the mean give qualitatively identical results, namely, significant differences even at levels of 0.1% or less [*Santer*, 1988*a*; *Santer and Wigley*, 1989]. In order to usefully compare the significance levels yielded by different tests, one needs to compare spatial fields that are more nearly similar. If differences exist that are real, but not visually obvious, this will provide conditions in which different tests of the same characteristic might give noticeably different results. In such situations, the use of an approximate assessment of significance might lead to a spurious conclusion. The chosen decades were selected a posteriori on this basis; decades prior to 1931–1940 were excluded a priori because of data quality problems in high latitudes [*Jones*, 1987] and missing data in the Pacific.

4.1. *January and July Comparisons*

We begin by comparing the means, variances, and spatial patterns for January and July for the two decades. The decadal-mean MSLP patterns are shown in Figure 1, with the difference fields given in Figure 2. It is clear from Figure 1 that the MSLP patterns for the two decades are quite similar for both months. It is also clear from Figure 2 that there are noticeable differences in the time means in some parts of the study area. Are these differences statistically significant? In Figure 3 we show the ratio of the temporal variances grid point by grid point. There are noticeable differences in these variances between the two decades, but are they statistically significant?

To assess statistical significance, we have calculated test statistic values for all the statistics described above. The corresponding observed significance levels (p values) were estimated using the PPP of P+B. All p values reported here are one-tail values, i.e., they represent the fraction of permuted results which, depending on which tail is appro-

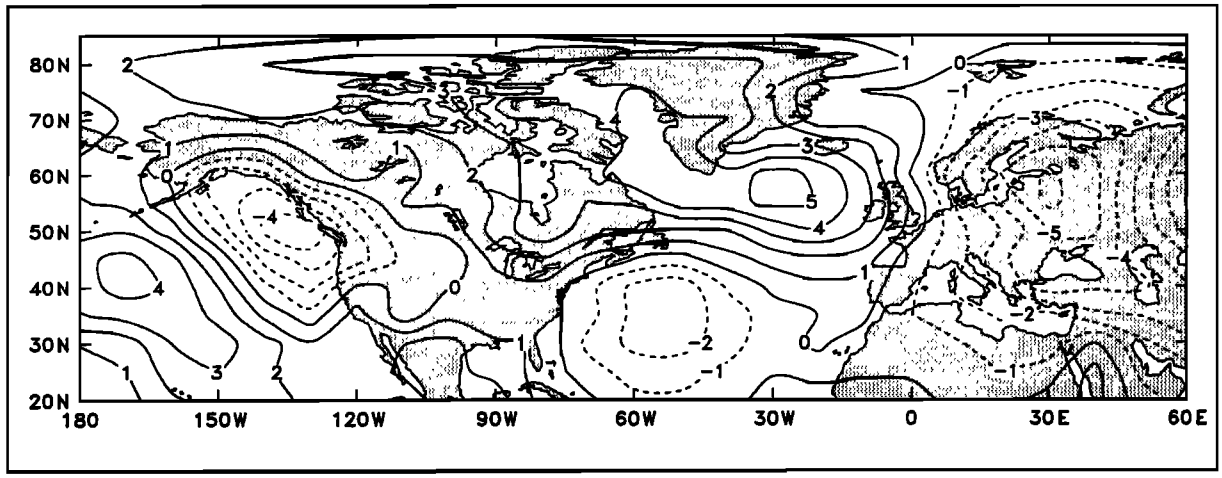

Fig. 2*a*

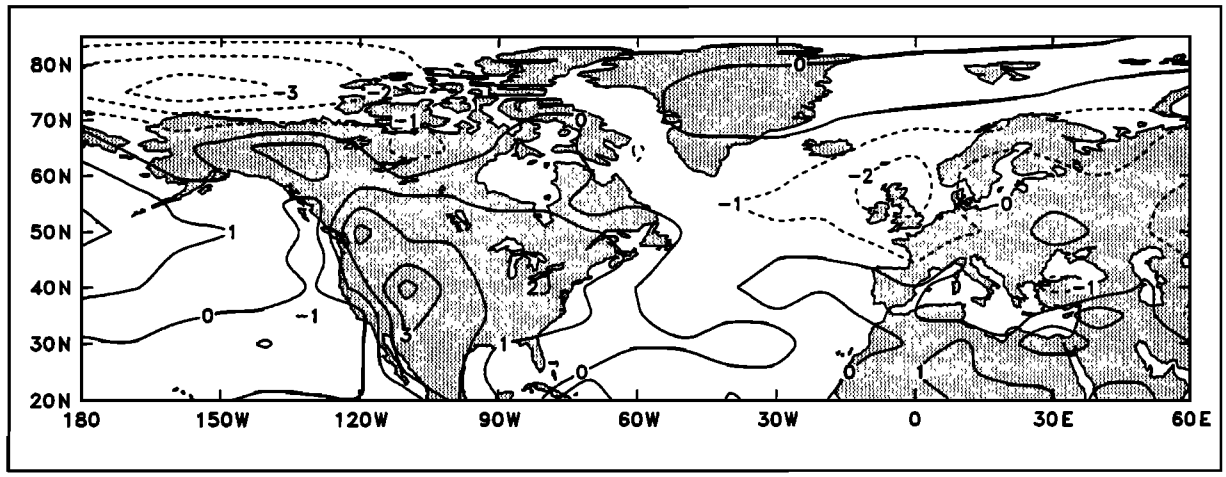

Fig. 2*b*

Fig. 2. Mean sea level pressure differences (millibars) for (*a*) January 1951–1960 minus January 1971–1980 and (*b*) July 1951–1960 minus July 1971–1980.

priate, either exceed or are less than the actual test statistic value. A p value close to zero or close to 1 would therefore represent a significant result.

In PPP the two samples are first combined to give $2n_t$ years (n_t = 10 here), each containing n_x (= 292) data points. (Occasional missing data points meant that n_x varied from 288 to 292.) The $2n_t$ years are then split randomly into two new samples of size n_t (in which spatial ordering is the same as in the original data), and the test statistic value recalculated. By performing a large number of randomizations, one can generate a null sampling distribution against which the test statistic for the original data can be compared. In the present case the total number of distinct permutations in which temporal ordering is unimportant is $20!/[2(10!)^2]$ = 92,378. We have found that at least 500 permutations are required to give stable results (we have used 1000). The number suggested by P+B, 50–100, is much less than this and is almost certainly too small [see *Efron*, 1987].

Traditional methods for assessing statistical significance in which the sampling distribution for the test statistic is known a priori, can give distorted results if the assumptions on which the theoretical sampling distribution is based are not satisfied. One of the most important of these assumptions is that individual data points are independent. Most meteorological data violate this assumption due to the existence of temporal and spatial autocorrelation. Permutation procedures have some clear advantages over traditional methods. First, in many cases they are immune to problems related to temporal and spatial autocorrelation, provided the autocorrelation structure is preserved by the chosen permutation process. Second, they can be applied even if the test statistic has an unknown sampling distribution. In PPP as applied here, spatial autocorrelation is preserved, but temporal autocorrelation is not. However, the variables considered, year-by-year values of gridded monthly mean MSLP, show no significant temporal autocorrelations.

January and July results for the 20 test statistics are given in Table 3. We consider the results pertaining to differences in means first.

The three grand-mean tests (T1, T2, T3) give virtually identical results in terms of significance levels (as one would expect). They show that there are no significant differences in the grand means at the 5% level. SITES and SITES1, also as expected, give virtually identical significance levels, and both indicate that there are no overall differences in means. DELTA1 and DELTA2 (based on Mielke's MRPP and using spatial means as the test variable), which also test overall differences in means, give the same result, i.e., no significant

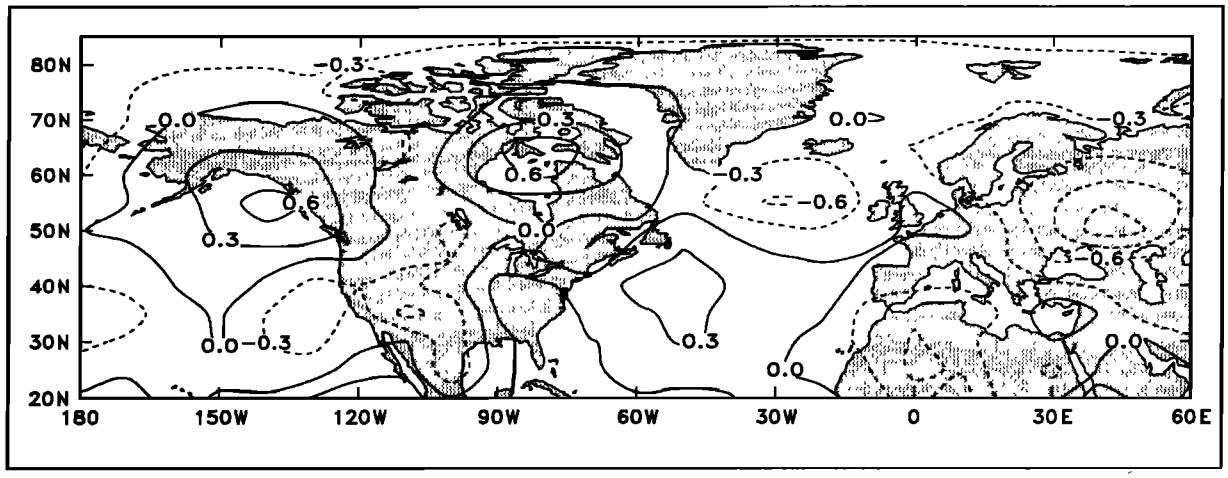

Fig. 3*a*

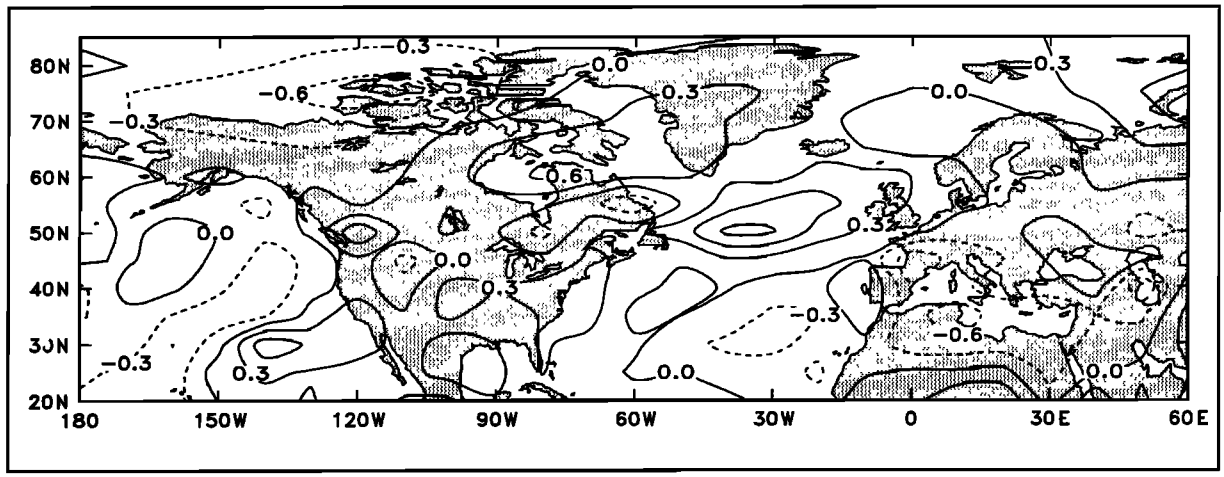

Fig. 3*b*

Fig. 3. Mean sea level pressure variance ratios for (*a*) January 1951–1960 divided by January 1971–1980 and (*b*) July 1951–1960 divided by July 1971–1980. The isopleths show the logarithm of the variance ratio in order to facilitate the contouring of the occasional grid points that have unusually high or low ratios.

difference. This is true also of DELTA3 (which uses the full original data set).

DELTA1 and DELTA2 give information which is equivalent to, but more restricted than that given by T1, T2, and T3. As presented here, the p values for T1, T2, and T3 are directional; i.e., a very low value of p would indicate a significant result with $\langle d \rangle \gg \langle m \rangle$, while a value of p close to 1 would indicate that $\langle d \rangle$ was significantly less than $\langle m \rangle$. The DELTA statistics, however, cannot be directional. For compatibility, significance for T1, T2, and T3 would have to be assessed using a two-tailed test (i.e., $\langle d \rangle \neq \langle m \rangle$ as the alternate hypothesis). If this is done, then the p values are approximately double those given in Table 1 (only approximately because the sampling distribution is not perfectly symmetrical when generated using only 1000 permutations). Significance levels for DELTA2 are then always identical to those for T3 (which considers the grand means as time averages of the spatial means). DELTA3, the more general form of the δ statistic, gives p values which are virtually identical to SITES/SITES1 (note that these three statistics are all nondirectional). The reason for these correspondences is not immediately obvious.

NT1 and NT5 give quite different results from the other

TABLE 3. Test Statistic Values and Observed Significance Levels (p Values) for MSLP Comparisons, 1951–1960 Versus 1971–1980

	January		July	
Statistic	Value	p	Value	p
T1	1.728	0.219	0.281	0.459
T2	0.722	0.219	0.095	0.459
T3	0.848	0.219	0.145	0.459
SITES	0.269	0.244	0.317	0.174
SITES1	0.091	0.246	0.022	0.173
NT1	9.030	0.014	10.270	0.001
NT5	15.630	0.047	14.730	0.018
DELTA1	0.204	0.360	0.137	0.586
DELTA2	1.932	0.414	0.886	0.889
DELTA3	52.410	0.244	8.901	0.171
SPRED	0.018	0.185	0.002	0.555
SPRET1	0.763	0.896	1.105	0.262
SPRET2	1.964	0.087	0.405	0.899
NF1	2.080	0.067	1.710	0.267
NF5	7.290	0.097	4.450	0.673
SPREX1	0.758	0.948	1.031	0.324
SPREX2	0.771	0.905	1.009	0.456
r	0.925	0.271	0.979	0.085
SHAPE	2.055	0.601	1.382	0.000
$\bar{r}$	0.020	0.417	0.262	0.003

statistics in this group. For both months, both NT1 and NT5 give significant results at the 5% level, showing that the number of grid points with significantly different time means is greater than one would expect to occur by chance. These results are consistent with the systematic deviations that are visually apparent in Figure 2.

The five temporal variance statistics form three clear groups. If SPRET1 is tested using a two-tailed test, then SPRED and SPRET1 give identical significance levels. However, SPRET1 can be used directionally, and it is these (one-tailed) results that are shown in Table 3. SPRET2 (like SPRET1, a directional statistic) gives results which differ markedly from SPRET1. These differences are illustrated further below. None of these statistics indicates any significant overall difference at the 5% level between the time variances of the two decades. NF1 and NF5 results lead to the same conclusion, although one value (NF1 in January) approaches significance at the 5% level. Note, however, that variance tests invariably have low power, so that with such small samples as these, any real differences would have to be quite large before they would lead to statistically significant results.

SPREX1 and SPREX2 show that there are no significant differences in the spatial variances (SPREX2 is almost significant at the 5% level in January). Although these two statistics give different observed significance levels, we will show below that they overlap considerably in information content.

The spatial pattern comparisons divide into two types. The correlation statistic r measures the degree of similarity of the time-mean patterns, with a significant value pointing to significant differences. As pointed out earlier, this differs from the conventional interpretation of a correlation coefficient, where a significant result would indicate the presence of common information. Both r values show that there are no strongly significant differences in the time-mean spatial patterns, although the July result is significant at the 10% level. With a test statistic value of 0.979 in July, it might appear strange that such a high value could indicate a marginally significant pattern difference. In this case, the mean correlation in the PPP sampling distribution is 0.986 with a standard deviation (sample size 1000) of 0.005 and skewness of -1.12, so 0.979 clearly lies in the tail of the sampling distribution.

The statistics SHAPE and $\bar{r}$ are indicators of similarities in the spatio-temporal evolution of the two fields. Significant results indicate that the time evolutions of the spatial anomaly patterns have similarities which cannot be attributed to chance. We will show below that they generally give equivalent information. The July results obtained here are most surprising, since they suggest that the year-to-year variations over 1951–1960 were similar to those over 1971–1980 (p values of 0.000 and 0.003). However, the degree of similarity is small, amounting to only about 7% of variance in common (based on the $\bar{r}$ value; the SHAPE statistic is singularly uninformative in this regard).

4.2. *Seasonal Cycle Comparisons*

Analysis of January and July data has shown that certain test statistics give similar information. To gain further insight into these similarities, we consider results over the entire seasonal cycle.

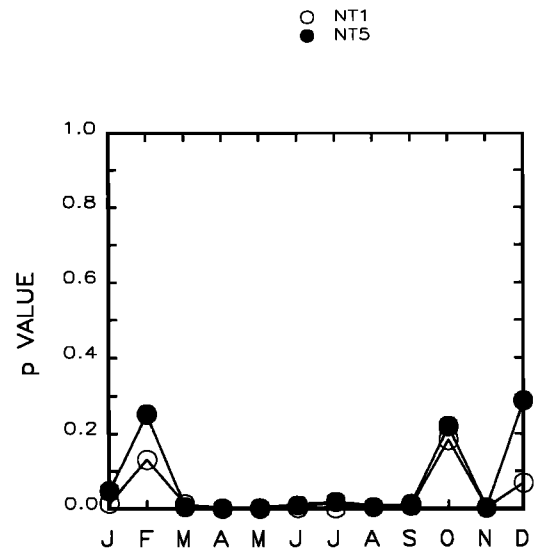

Fig. 4. Comparison of one-tailed p values for NT1 and NT5. NT1 and NT5 are the fraction of grid points with significant differences in the mean at the 1% and 5% levels (two-tailed local tests). Field significant differences occur in all months except February, October, and November.

Figure 4 compares NT1 and NT5. Both statistics show similar month-to-month variations in their significance levels, with NT1 values generally more significant (except in March and November).

Figure 5 compares NT1, SITES, and T1. (Note that the equivalence of SITES and SITES1 has already been demonstrated.) SITES, T1, and NT1 show seasonal variations which are similar, but the SITES and T1 values are invariably much less significant than NT1, and T1 is less significant than SITES for all months except January, February, and March. In general, of course, the sum of a set of univariate

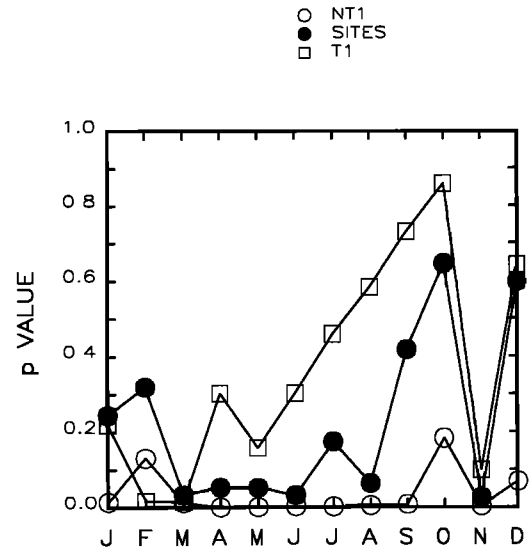

Fig. 5. Comparison of one-tailed p values for NT1, SITES, and T1 (note that, as used here, only T1 is a directional statistic). All three statistics show similar seasonal variations, but NT1 is generally much more sensitive. The exception is February. In this month a significant difference exists in the overall mean, but this is only weakly reflected in NT1.

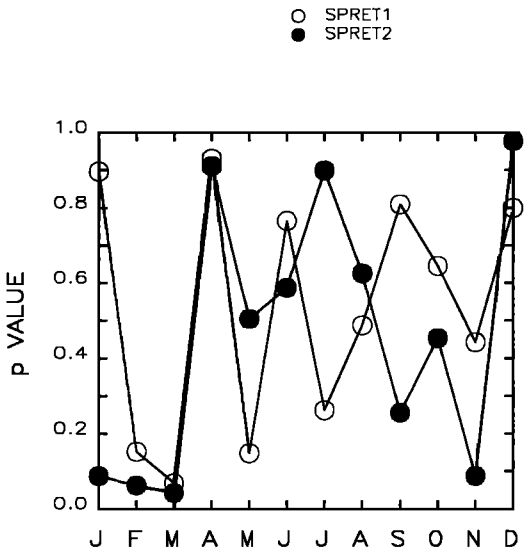

Fig. 6. Comparison of SPRET1 (ratio of the spatial means of the local time variances) and SPRET2 (ratio of the time variances of the spatial means); one-tailed p values.

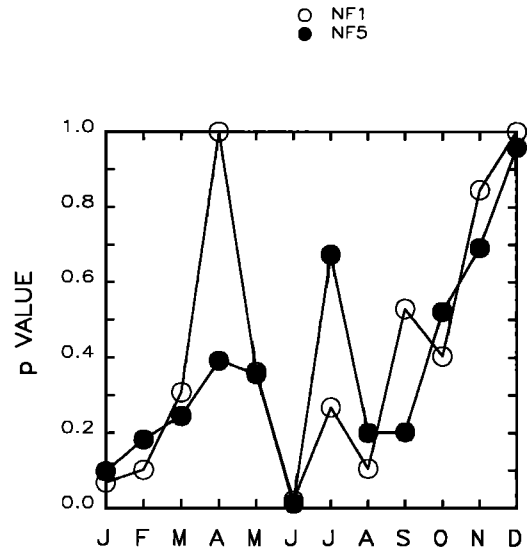

Fig. 7. Comparison of one-tailed p values for NF1 and NF5. NF1 and NF5 are the fraction of grid points with significant differences in variance at the 1% and 5% levels (two-tailed local tests). Field significant differences only occur in June. In April (NF1 only) and December the number of locally significant results is significantly less than one would expect to occur by chance.

tests need not lead to the same conclusion as a single multivariate test [*von Storch*, 1982], so the similarities here need not apply in other circumstances.

For SITES, only March, June, and November show significant differences at the 5% level, while only February, October, and December fail to reach significance at this level for NT1. The NT statistics therefore appear to be similar to, but more sensitive than SITES.

T1 results for February and March are significant at the 5% level. Significant results for T1 (or T2 or T3) can arise in two ways; either there is a small area of highly significant local differences in means, or there is a large area of smaller (possibly nonsignificant) differences in means. These two possibilities should be distinguished by the NT results. In February the latter possibility holds because the NT statistics are not significant in this month. In March, however, all mean statistics give significant results. This does not rule out the possibility of a large-scale bias, but it clearly shows that the character of the differences in March differs noticeably from that in February.

In Figure 6 we compare SPRET1 (which, in its two-tailed form, is equivalent to SPRED) and SPRET2. The seasonal variations are quite different, indicating that these two statistics give distinct and complementary information about overall changes in temporal variability. Indeed, they can lead to conclusions that are, at least superficially, in conflict (witness the results for January). In this case, SPRET1 indicates that 1951–1960 was less variable when the average local variability is considered, but more variable than 1971–1980 when the temporal variability of the spatial mean is considered. None of these results, however (with the exception of SPRET2 in March), is significant at the 5% level.

NF1 and NF5 are compared in Figure 7. In general, these two statistics show similar month-to-month variations. Significant differences between the decades occur in June (p = 0.020 for NF1, p = 0.011 for NF5), indicating that the temporal variability in 1951–1960 differed from that in 1971–1980 at a significant number of grid points (roughly twice as many points as would be expected by chance). Since the results given here employed a two-tail test at the grid point level, they do not show which decade (if either) was the more variable. (When one-tailed local tests were used, these showed that June 1971–1980 was consistently more variable than June 1951–1960.) A different type of significant result occurs in April (NF1 only, $1 - p = 0.000$) and December ($1 - p = 0.000$ for NF1, $1 - p = 0.042$ for NF5). This implies differences in variance which are less than one would expect to occur by chance if the two decades were drawn from the same population, a result which is not easy to interpret.

The spatial analogs of SPRET1 and SPRET2, viz.,

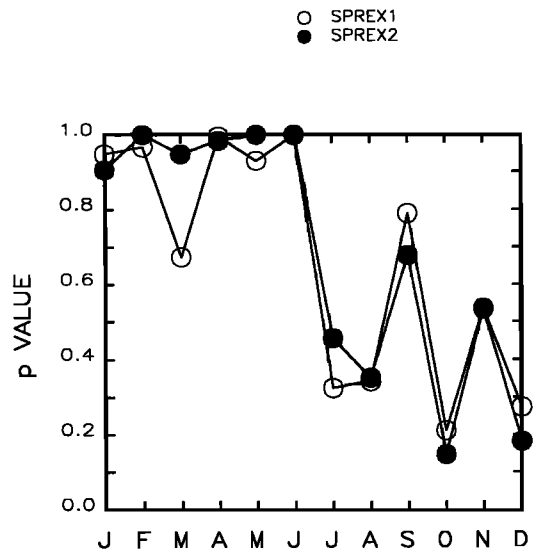

Fig. 8. Comparison of SPREX1 (ratio of the time-mean spatial variances) and SPREX2 (ratio of the spatial variances of the time means); one-tailed p values.

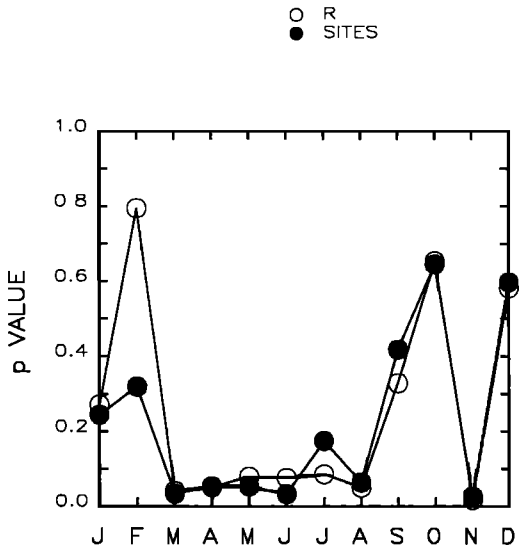

Fig. 9. Comparison of r (the spatial correlation coefficient between the time-mean fields) and Preisendorfer and Barnett's SITES statistic; one-tailed p values.

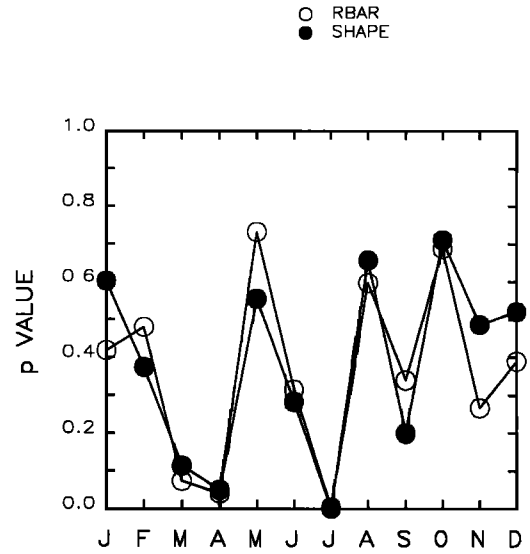

Fig. 10. Comparison of $\bar{r}$ (time mean of the spatial correlations between the anomaly fields) and Preisendorfer and Barnett's SHAPE statistic; one-tailed p values.

SPREX1 and SPREX2, are compared in Figure 8. Except for March, these two statistics tend to give similar information. They show significant differences in overall spatial variability in February, April, May, and June (1951–1960 less variable, $1 - p < 0.05$ for SPREX1, SPREX2, or both).

Figure 9 compares r with SITES. These two statistics show somewhat surprising similarities in the month-to-month variations of their observed significance levels. (Similarities between SITES1 and r are equally pronounced.) One might, however, expect that these statistics would give overlapping information, since they have an important term in common in their definitions, namely, $\sum_x \bar{d}_{x.}\bar{m}_{x.}$. This term is clearly crucial in the definition of r, since it encapsulates the covariance between the two time-mean fields. It is less obviously important in SITES and SITES1, but the empirical evidence presented here attests to its significant role. One can also argue for a similarity between r and SITES on the basis of NT1 (or NT5); NT1-SITES parallels are discussed above. If NT1 is significant, then this points to a spatially specific difference in the time-mean fields. If such a difference exists, then one might also expect r to point to a significant difference in the spatial patterns (although this is not always the case [see *Santer*, 1988*b*]). For SITES (and SITES1) therefore, although they give information about differences in means, their overall interpretation is quite complex.

Finally, $\bar{r}$ and SHAPE are compared in Figure 10. The parallel between their month-to-month variations is striking. As already noted, $\bar{r}$ is superior to SHAPE in that it can be interpreted directly in terms of common variance.

5. Conclusions

The statistical analysis of model validation, perturbation, and predictability experiments has been divided into tests which compare time-mean fields, temporal variances, spatial variances, and spatial patterns. In each group, we have identified a number of test statistics; some of these have been used previously, and others are new. Within any one group the various statistics must give overlapping information, so some tests may be superfluous. To assess the degree of overlap between different statistics, and to illustrate their use and interpretation, we have compared MSLP data over the North America/North Atlantic/Europe region for two decades, 1951–1960 and 1971–1980. This comparison has revealed some significant changes in climate between the two decades and has provided the following insights into the various test statistics.

For tests involving the time-means, grid point by grid point tests (NT1 and/or NT5) appear to be the most informative and the most sensitive. A grand-mean test (e.g., T1) should be performed as well in order to assess the significance of large-scale changes. SITES1 has no advantage over SITES. These statistics rarely add any unique and easily interpretable information. SITES overlaps with the spatial correlation statistic r, which is clearly preferable in examining pattern changes. As used here, the MRPP statistics DELTA1, DELTA2, and DELTA3 add no new information. DELTA2 is equivalent to the two-tailed version of T3, while DELTA3 is equivalent to SITES.

For tests of temporal variance differences, NF1 (and/or NF5) gives different information from the other tests. SPRET1, basically a clone of SPRED, is preferred because it gives directional information. SPRET2 apparently gives information that is different from that given by SPRET1. The two tests to detect differences in spatial variability, using SPREX1 and SPREX2, give information that overlaps considerably in most instances.

Only one test has been given which can be used to identify differences in spatial patterns. This uses the statistic r, the correlation coefficient between the time-mean spatial fields. The PPP method for significance testing allows one to identify whether or not significant pattern differences exist (under the null hypothesis that the two fields are drawn from the same population). In many cases, however, one may be interested in knowing whether the two fields have any significant common features (for which the null hypothesis

will have $r = 0$). The standard test cannot be used in this case because most meteorological fields have strong spatial autocorrelation, and this would invalidate the test assumptions. If spatial autocorrelation exists, quite large correlations between spatial fields can occur by chance, and further work is required to devise an appropriate permutation procedure for significance testing under the $r = 0$ null hypothesis.

The statistics SHAPE and $\bar{r}$ can be used to identify significant parallels between the combined spatial and temporal features of variables. They might be applied, for example, in testing models of the ENSO phenomenon, where both a characteristic response pattern and its temporal evolution must be simulated. The time-mean spatial anomaly correlation, $\bar{r}$, gives information which is equivalent to that given by SHAPE. However, $\bar{r}$ is preferred because it is easier to interpret physically and because its value is directly related to the overall common variance.

To summarize, we recommend that the statistics NT1, NT5, T1, SITES, NF1, NF5. SPRET1, SPREX1, and r be used routinely in quantitative evaluation of spatial field similarities and/or differences. In cases where temporal evolution is important, similarities can be best quantified and assessed using $\bar{r}$. In addition to tests using these easily interpreted statistics, other tests are possible based on Hotelling's T^2 statistic or the Mahalanobis D^2 statistic. These give information which, under some circumstances, may overlap with NT1, NT5, T1, and r, and further work is required to see what additional insights they can provide. Since $n_x \gg n_t$, in general, use of T^2 requires some form of spatial compression. The results will likely depend on the method of compression used (zonal means versus spectral decomposition versus principal components, etc.).

In a few of the cases considered here, the test statistics have mathematically tractable expressions for their theoretical null distributions under certain simple assumptions. However, these assumptions tend to be frequently violated by meteorological data. We have found that significance levels estimated using theoretical null distributions invariably differ markedly from those obtained using the PPP technique.

In a companion paper [*Santer and Wigley*, 1989] we will consider the application of the statistics discussed above to the validation of a number of general circulation models.

Acknowledgments. This work was funded by the Carbon Dioxide Research Division of the U.S. Department of Energy under grant DE-FG02-86-ER60397 and under contract W-7405-ENG-48 with the Lawrence Livermore National Laboratory. Help and advice from K. J. Keen and R. W. Katz are gratefully acknowledged.

REFERENCES

Barnett, T. P., and M. E. Schlesinger, Detecting changes in global climate induced by greenhouse gases, *J. Geophys. Res.*, *92*, 14,772–14,780, 1987.

Briffa, K. R., P. D. Jones, and T. M. L. Wigley, Climate reconstruction from tree rings, 2, Spatial reconstruction of summer mean sea-level pressure patterns over Great Britain, *J. Climatol.*, *6*, 1–15, 1986.

Chervin, R. M., Interannual variability and seasonal climate predictability, *J. Atmos. Sci.*, *43*, 233–251, 1986.

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

Edgington, E. S., *Randomization Tests*, Marcel Dekker, New York, 1987.

Efron, B., Better bootstrap confidence intervals, *J. Am. Stat. Assoc.*, *82*, 171–185, 1987.

Hasselmann, K., On the signal-to-noise problem in atmospheric response studies, *Meteorology of Tropical Oceans*, edited by D. B. Shaw, pp. 251–259, Royal Meteorological Society, London, 1979.

Jones, P. D., The early twentieth century Arctic high—Fact or fiction?, *Clim. Dyn.*, *1*, 63–75, 1987.

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

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

Mielke, P. W., Design and evaluation of weather modification experiments, in *Probability, Statistics and Decision Making in the Atmospheric Sciences*, edited by A. H. Murphy and R. W. Katz, pp. 439–459, Westview Press, Boulder, Colo., 1985.

Mitchell, J. F. B., C. A. Wilson, and W. M. Cunningham, On CO_2 climate sensitivity and model dependence of results, *Q. J. R. Meteorol. Soc.*, *113*, 293–322, 1987.

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

Preisendorfer, R. W., and C. D. Mobley, Data intercomparison theory, II, Trinity statistics for location, spread, and pattern differences, *Tech. Memo. ERL PMEL-39*, 91 pp., Pac. Mar. Environ. Lab., NOAA, Seattle, Wash., 1982.

Santer, B. D., Regional validation of general circulation models, *Clim. Res. Unit Res. Publ. 9*, 375 pp., Univ. of East Anglia, Norwich, England, 1988*a*.

Santer, B. D., Validation of sea-level pressure simulated by the ECMWF T21 model for the northern hemisphere, Climate simulations with the ECMWF T21 model in Hamburg, I, Climatology and sensitivity experiments, *Rep. 4*, pp. 65–98, Meteorol. Inst. der Univ. Hamburg, Federal Republic of Germany, 1988*b*.

Santer, B. D., and T. M. L. Wigley, Regional validation of means, variances, and spatial patterns in general circulation model control runs, *J. Geophys. Res.*, this issue.

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

Wigley, T. M. L., and B. D. Santer, Validation of general circulation climate models, in *Physically-Based Modelling and Simulation of Climate and Climatic Change, Part 2*, edited by M. E. Schlesinger, pp. 841–879, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1988.

Wigley, T. M. L., G. J. Kukla, P. M. Kelly, and M. C. MacCracken, Recommendations for monitoring and analysis to detect climate change induced by increasing carbon dioxide, in *Detecting the Climatic Effects of Increasing Carbon Dioxide*, edited by M. C. MacCracken and F. M. Luther, pp. 177–185, Carbon Dioxide Research Division, U.S. Department of Energy, Washington, D. C., 1985.

Williams, J., and H. H. van Loon, An examination of the northern hemisphere sea-level pressure data set, *Mon. Weather Rev.*, *104*, 1354–1361, 1978.

B. D. Santer, Max-Planck-Institut für Meteorologie, Bundesstrasse 55, 2 Hamburg 13, Federal Republic of Germany.

T. M. L. Wigley, Climatic Research Unit, University of East Anglia, Norwich, NR4 7TJ, England.

(Received August 1, 1988;
revised March 31, 1989;
accepted April 28, 1989.)