Numerical Model–Reality Intercomparison Tests Using Small-Sample Statistics

RUDOLPH W. PREISENDORFER

Pacific Marine Environmental Laboratory, NOAA, Seattle, WA 98115

TIM P. BARNETT

Climate Research Group, Scripps Institution of Oceanography, La Jolla, CA 92039

(Manuscript received 21 December 1982, in final form 12 April 1983)

ABSTRACT

When a numerical model's representation of a physical field is to be compared with a corresponding real observed field, it is usually the case that the numbers of realizations of model and observed field are relatively small, so that the natural procedure of producing histograms of pertinent statistics of the fields (e.g., means, variances) from the data sets themselves cannot be usefully carried out. Also, it is not always safe to adopt assumptions of normality and independence of the data values. This presents the confident use of classical statistical methods to make significance statements about the success or failure of the model's replication of the data. Here we suggest two techniques, of determinable statistical power, in which small samples of spatially extensive physical fields can be made to blossom into workably large samples on which significance decisions can be based. We also introduce some new measures of location, spread and shape of multivariate data sets which may be used in conjunction with the two techniques. The result is a pair of new data intercomparison procedures which we illustrate using GCM simulations of the January sea-level pressure field and regional ocean model simulations of the near-shore velocity field off South America. We include with these procedures a method of determining the spatial and temporal locations of non-random errors between the model and data fields so that models can be improved accordingly.

1. Introduction

In the construction of numerical models, say, general circulation models (GCM's) of the earth's atmospheric and oceanic regimes, there is a stage wherein model data fields are to be compared with observed data fields. The models of interest here are those which produce data fields over large numbers of spatial grid points for limited periods of time. These two conditions on the data field (many space points, few time samples) produce two basic problems when one attempts to assess how well the model's data field fits the corresponding observed data fields. The first problem is how to measure objectively the goodness of fit between the fields of model and reality. The second problem is how to assign statistical significance to that measure of fit. Ideally, one would like to test both the first and second moments of model-produced fields for significance. Here we introduce some procedures designed to solve these problems under the many-point, small-sample constraints naturally arising in GCM developments. The new methods differ from those suggested previously (e.g., Chervin and Schneider, 1976; Katz, 1982a,b) and do not appear to have the problems associated with those earlier techniques (cf. e.g., Hasselmann, 1979; Storch, 1982; Katz, 1982b).

In the solution of the first problem, one must choose a measure of fit that is objective, as distinguished from subjective. In the latter case, one casts an eye over (say) a computer-graphic output of a GCM's sea-level atmospheric pressure field and gains an impression of how realistic the output is. Obvious mismatches between reality and the model's output can initially help decide the model's fidelity in reproducing natural events. As the model's fidelity increases, there will occur a point in the model-reality comparisons when an intuitively-based decision is no longer readily forthcoming; the visual differences between the model and real fields are apparently still there, but they are no longer flagrant. At this point it also becomes difficult to determine if changes in the model physics change the model's performance. Thus there comes a stage when there are many spatial points over which subtle differences must be objectively noted, decided upon, and somehow summarized. There is also the desirability of placing these summary decisions into some quantifiable form. The resultant decisions should be communicable using simple numerical indexes of the discerned goodness of fit. Finally, the measures of fit should be adjustable to allow both local and global intercomparisons.

In the solution of the second problem, namely, that of statistical significance of the fit, one faces the fact
that the sample size is small. This means that, in constructing the sample space associated with the statistical background of the problem, one is limited to discrete sample spaces—and thus one faces the prospect that the analytic methods of continuous sample spaces are not available. Moreover, because the sample is small and its parts may not always be statistically independent, one cannot confidently rely on any one of the classical continuous probability density functions such as the normal, Student-\(t\), chi-square and \(F\)-distributions. The limited populations of real fields and associated computer-generated fields are too sparse and statistically interdependent to allow one confidently to assign the normal distribution or its relatives to stand in for the probability distributions actually encountered.

Despite these limitations it is still possible to proceed using accepted principles of statistical inference. In particular, one can draw upon some basic notions in nonparametric small-sample theory. In this theory, one takes what sparse data samples one has and generates by various permutation techniques the associated sample space for one's statistical problem. For some examples of the permutation technique in simple contexts, see Hollander and Wolfe (1973). Along these general lines, see also Mielke et al. (1981). In our permutation procedures, we shall demonstrate that decisions can be reached as to the statistical significance of the fit of a model to its associated portion of reality. Objective judgments can be made as to the reliability of the decision (via prior studies of power tests). Moreover, the objective judgments can be quantitatively communicated, and thus one can take a step closer to a scientific evaluation of the performance of a GCM.

2. Overview of tests

The tests we develop here are based on a set of three statistics carved out of a single measure of separation between two data sets (\(D\) and \(M\)) given to us in the form of \(n \times p\) matrices. This measure of separation of \(D\) and \(M\) is the square of the Euclidean distance (the norm) between the two sets when they are represented as two points in a Euclidean space \((E_{np})\) of dimension \(np\). By an algebraic rearrangement, the expression for the norm can be represented as a sum of three terms which depict the differences between three principal geometric properties of the sets \(D\) and \(M\), each set now considered as an \(n\)-point swarm in a common \(p\)-dimensional Euclidean space, \(E_p\). These differences, illustrated in Fig. 1, are the separation of the centroids of the two swarms, the difference in radial scale of the two swarms, and the difference in spatial-temporal evolution of the two swarms in \(E_p\). These three differences are respectively designated by the mnemonic names: SITES, SPRED and SHAPE. These difference-measures serve, within this study, to solve the first of the basic problems of data intercomparison, namely, that of defining natural measures of separation between two data sets.

The second problem, that of statistical significance of the measures of separation, is solved by constructing an appropriate reference distribution for each of the same three statistics: SITES, SPRED and SHAPE. We shall consider two general procedures for such a construction.

The first of these reference-distribution constructions (which we shall denote by PPP) starts with the union (\(U\)) of the two data swarms \(D\) and \(M\) in \(E_p\), thereby forming a single swarm of \(2n\) points. The swarm \(U\) is then repeatedly partitioned randomly into two \(n\)-point swarms \(U_1\) and \(U_2\). The statistics \(S\) (SITES, SPRED and SHAPE) of the pair \(U_1, U_2\) can then be computed for each realization (different partition). The set of numerical values of the statistics formed by the class of all generated partitions can be used to form a reference distribution of values for the respective types of separations. Experience shows that about 100 partitions are needed to form a usable reference distribution of the values of the three statistics. Once such a distribution is at hand, we can compute
the separation \( S_0 \) (SITES, SPRED or SHAPE) of the two original sets \( D \) and \( M \) and determine its location along the associated reference distribution of \( S \)-values. From this determination we can decide if \( S_0 \) is unusually large, unusually small, or just run-of-the-mill relative to its reference distribution of \( S \)-values. In other words, we can determine the significance level of \( S_0 \).

The second method of reference-distribution construction (which we shall denote by APP) works in a different way: each data set \( D \) and \( M \) is partitioned randomly into pairs \( D_1, D_2 \) and \( M_1, M_2 \). The separations \( S \) between \( D_1 \) and \( D_2 \) and between \( M_1 \) and \( M_2 \) are determined. This is done many (around 100) times. The \( S \) values from each set of pairs are collected into their reference distributions, and these are denoted by DD and MM, respectively. Concurrently using these same partitions, we apply the separation measure \( S \) to the pair of sets \( D_1, M_2 \) and also to \( M_1, D_2 \), to form two more reference distributions denoted by DM and MD, respectively. Comparing the auto reference distribution DD with the cross reference distribution DM (or alternately, MM with MD), we can determine the significance of the separations (in the sense of SITES, SPRED or SHAPE) of the two given data sets \( D \) and \( M \). Clearly, if \( D \) and \( M \) are drawn from the same population then the auto and cross reference distributions DD and DM will on average be equal. Otherwise, we can determine the disparity of DD and DM by seeing where on the domain of DD the median of DM falls. This fixes a level of significance of the measure \( S \) for \( D \) and \( M \). Now to the details.

3. Trinity statistics

Suppose we have two space–time data sets for some physical field, such as sea surface temperature or sea level pressure. To fix ideas, suppose that one set has been observed and the other determined by a model intended to reproduce the observed phenomenon.\(^1\) These sets are represented as a collection of field values:

\[
D: \{d(t, x); \ t = 1, \ldots, n; \ x = 1, \ldots, p\}, \quad (3.1)
\]

\[
M: \{m(t, x); \ t = 1, \ldots, n; \ x = 1, \ldots, p\}. \quad (3.2)
\]

Here \( t, x \) are integers which serve as time and location indexes. We can think of \( D \) and \( M \) in (3.1), (3.2) either as \( n \times p \) matrices, as two points in a Euclidean space \( (E_{np}) \) of dimension \( np \), or as two swarms of \( n \) points each in a Euclidean space \( (E_p) \) of dimension \( p \). Each of these representations will be used in the developments below.

We are interested in obtaining various measures of separation of \( D \) and \( M \). A fundamental measure of separation is given by

\[
L^2(D, M) = \sum_{i=1}^{n} \sum_{x=1}^{p} [d(t_i, x) - m(t, x)]^2, \quad (3.3)
\]

which considers \( D, M \) as two points in \( E_{np} \). By imagining \( D, M \) as \( n \)-point swarms in \( E_p \), i.e.,

\[
D: \{d(t); \ t = 1, \ldots, n\}, \quad (3.4)
\]

\[
M: \{m(t); \ t = 1, \ldots, n\}, \quad (3.5)
\]

where (with \( T \) denoting transpose)

\[
d(t) = [d(t, 1), \ldots, d(t, p)]^T, \quad (3.6)
\]

\[
m(t) = [m(t, 1), \ldots, m(t, p)]^T, \quad (3.7)
\]

we can write (3.3) as

\[
L^2(D, M) = \sum_{i=1}^{n} \|d(t_i) - m(t_i)\|^2. \quad (3.8)
\]

Here, in general, for any vector \( \nu = [\nu(1), \ldots, \nu(p)]^T \) in \( E_p \),

\[
\|\nu\|^2 = \nu^T \nu = \sum_{x=1}^{p} \nu^2(x). \quad (3.9)
\]

Natural indicators of location of the \( n \)-point swarms \( D \) and \( M \) are their centroids:

\[
d = \frac{1}{n} \sum_{i=1}^{n} d(t_i), \quad m = \frac{1}{n} \sum_{i=1}^{n} m(t_i). \quad (3.10)
\]

Moreover, natural indicators of the radial scale of \( D \) and \( M \) are the scatters of \( D, M \) about their centroids:

\[
\sigma_D^2 = \frac{1}{n} \sum_{i=1}^{n} \|d(t_i) - d\|^2 \quad \text{(3.11)}
\]

\[
\sigma_M^2 = \frac{1}{n} \sum_{i=1}^{n} \|m(t_i) - m\|^2 \quad \text{(3.11)}
\]

With these definitions, we may arrange (3.8) to read

\[
L^2(D, M) = n\|d - m\|^2 + \left(\sigma_D - \sigma_M\right)^2
\]

\[
+ 2[\sigma_D\sigma_M - \sum_{i=1}^{n} (d(t_i) - d)^T (m(t_i) - m)]. \quad (3.12)
\]

Inspection of (3.12) suggests that we normalize each term using the product \( \sigma_D\sigma_M \) occurring in the third term, thereby introducing dimensionless measures of separation. The resultant form of (3.12) is expressible as:

\[
\text{DIST2} = \text{SITES} + \text{SPRED} + \text{SHAPE}, \quad (3.13)
\]

\(^1\) The methods apply equally well to Data–Data or Model–Model comparisons.
where we have written

\begin{align}
\text{DIST2} & \quad L^2(D, M)/\sigma_D\sigma_M \\
\text{SITES} & \quad n||d - m||^2/\sigma_D\sigma_M \\
\text{SPRED} & \quad (\sigma_D - \sigma_M)^2/\sigma_D\sigma_M \\
\text{SHAPE} & \quad 2\left\{1 - \sum_{t=1}^n \left[ \frac{(d(t) - d)^T}{\sigma_D} \right]^2 \left[ \frac{(m(t) - m)}{\sigma_M} \right]^T \right\}
\end{align}

(3.14) (3.15) (3.16) (3.17)

The trinity of statistics, SITES, SPRED and SHAPE, carved out of DIST2, solve the first of the two main problems of this study. They are invariant under general affine transformations of \( E_p \), so that, in particular, they will give the same values independent of temperature or pressure units of \( D, M \).

An examination of these definitions shows that SITES is the desired measure of separation of the centroids of the swarms \( D, M \), i.e., it is a measure of the differences of the mean fields of \( D \) and \( M \) (cf. Fig. 1). SPRED measures the difference in the radial scatter of \( D \) and \( M \), and so compares the levels of variability, or second moments, of \( D \) and \( M \). The third measure, SHAPE, is seen to involve spatial correlations of \( n \) \( D \)- and \( M \)-derived vectors in \( E_p \). Observe that \( u(t) = (d(t) - d)/\sigma_D \) and \( v(t) = (m(t) - m)/\sigma_M \) are, for each \( t = 1, \ldots, n \), vectors in \( E_p \). If \( u(t) \) and \( v(t) \) point in the same direction for every \( t \), then \((d(t) - d) = (\sigma_D/\sigma_M)(m(t) - m)\), and so SHAPE = 0, meaning these vectors in \( E_p \) move in exact unison, spinning out two \( n \)-point swarms in \( E_p \) which have the same spatial configuration about their centroids. Thus, the locations \( m, d \) and the radial scales \( \sigma_M, \sigma_D \) of \( M \) and \( D \) may differ, but the relative evolution of \( D \) and \( M \) through space/time will be the same whenever SHAPE = 0.

In the present introductory study of these concepts, we shall concentrate only on SITES and SPRED, being the simpler members of the trio within DIST2. The statistic SHAPE, as it turns out, because of its more complex internal spatial- and time-sensitive structure, is less powerful than SITES and SPRED for discerning differences in \( D, M \) using the present permutation procedures; its detailed study and its logical descendants are given elsewhere (Preisendorfer and Mobley, 1982b).

4. Developing the statistical background: Ideal world

We begin our exposition of the reference-distribution-construction procedures with a relatively ideal case for data intercomparisons, namely, the case where we have a workably large supply of data sets. In geophysics such data abundance is rarely found. However, if the data are sufficiently numerous, then they can be used to form a reference distribution which will allow a decision as to whether or not two data matrices \( M, D \) are significantly separated in the sense of SITES or SPRED.

The inclusion of this example will serve several purposes in the present study. First of all, it will alert some readers to the possibility of forming their own probability distributions (histograms) when they fortuitously encounter relatively data-rich settings. Second, it will encourage certain modelers and data collectors to think of deliberately increasing their own data records (each in his own way) so that, in time, they can plot their own reference distributions and thereby become increasingly independent of procrustean-type statistical intercomparison analyses. Finally, and perhaps most importantly for the present study, the inclusion of this example will throw into relief the relatively data-sparse settings we will have to deal with in the remaining two operational procedures of Section 5. For examples of the use of the present ideal settings, see Preisendorfer and Mobley (1982a). The present procedure, which we term Experimental-Observational Procedure (EOP), is as follows.

A. Given: \( n \times p \) data sets \( D, M \) (over region \( R_0 \), time interval \( I_0 \) where \( p \) is the number of points in \( R_0 \); \( n \) the number of temporal samples at each point).

B. Question: Is \( \text{STSTC}(D, M) \) significantly large to permit one to say \( D \) and \( M \) are statistically different? (STSTC = SITES or SPRED).

C. Data space: Available collection \( D_1, \ldots, D_\Omega \), \( \Omega \geq 15 \), of \( n \times p \) data sets over region \( R \) and time interval \( I \), which are relevant to \( \Omega \) A, B above.

D. Sample space: All ordered pairs \( (D_i, D_j) \), \( i, j = 1, \ldots, \Omega \geq 15, i \neq j \), are in the sample space.

E. Statistics: Form the statistic values \( S_k = \text{STSTC}(D_i, D_j), k = 1, \ldots, r \leq \omega = \Omega(\Omega - 1)/2 \).

F. Reference distribution: Order the \( S_k \) in \( \Omega \) E according to increasing size; and relabel: \( S_1 \leq S_2 \leq \cdots \leq S_{\text{crit}} \leq \cdots \leq S_{\Omega} \), \( \text{crit} = [(1 - \alpha)\rho] \) where \( \alpha \) is the significance level of the test, \( 0 < \alpha < 1 \).

G. Null hypothesis \( H_0 \): \( \text{STSTC}(D, M) \) is in the sample space.

H. Accept \( H_0 \): If \( \text{STSTC}(D, M) \leq S_{\text{crit}} \),

Reject \( H_0 \): If \( \text{STSTC}(D, M) > S_{\text{crit}} \).

In the latter case, \( \text{STSTC}(D, M) \) is significantly large; in the former case, it is not.
COMMENTS

In § A, the set \( D \) could be the values of sea level pressure (SLP) at \( p = 300 \) points on a certain region \( R_0 \) of the ocean averaged over January for \( n = 5 \) different Januarys; \( M \) could be a GCM simulation of SLP over the same points of \( R_0 \) for five separate realizations of January. The question in § B is then, e.g., "Is SITES (\( D, M \)) significantly large?" In the present version of EOP, by definition, there is available a set of SLP averages over at least 15 sets of five Januarys in the region \( R_0 \). These could be drawn, e.g., from a hundred-year data set. The sample space in § D consists of at least \( \Omega(\Omega - 1)/2 \) = 105 points, so that in § E one can compute, say, \( r = 100 \) values of \( S_k \). These values are ordered as in § F and if \( \alpha = 0.05 \), then crit = [0.95\%] = 95. The null hypothesis in § G states that the given \( D \) and \( M \) are a pair (\( D, M \)) that forms a run-of-the-mill falling controllably within the sequence of reference values in § F. This notion is made precise in § H. If, then, SITES(\( D, M \)) is significantly large, say its value falls beyond \( S_{\text{crit}} \), where crit = 95 is defined above, then we would reject \( H_0 \) and declare with confidence 95% that (\( D, M \)) does not belong to the sample space. Physically this implies that the mean fields (first moments) of \( D \) and \( M \) are different from each other with confidence 95%.

The central feature of EOP, the one that distinguishes it from PPP and APP discussed below, is the step in § C wherein it is explicitly required that there are available at least 15 \( n \times p \) sets \( D_i \), which are physically relevant to the given data set. This means, e.g., that, if \( D \) is an array of SLP measurements over some region \( R_0 \) of the Pacific and consists of monthly averages then so, too, should the \( D_i \) be SLP monthly averages over \( R_0 \). Naturally \( M \) should then also be SLP monthly average values. But \( M \) values need not always be over the same spatial or temporal domain as \( D \). There is, accordingly, in the gathering of the Data Space \( D_1, \ldots, D_9 \) considerable freedom as to choice of region \( R_0 \) and time location of the samples—as long as the choice leads to a mathematically well-defined and physically reasonable sample space.

Observe that the sample space constructed in § D is open-ended, in that more elements can be added beyond those generated from the present selection of the data space. Thus the "sample space" referred to in § D is some larger set containing the explicitly constructed set of ordered pairs. We observe, finally, that the sample space is built on the assumption that the temporal order of the \( D_i \) is immaterial to the question at hand. Thus we are not interested in the fact that \( D_i \) may precede \( D_j \) in time (say) as we are simply in the separation STSTC(\( D_i, D_j \)) of some attribute (SITES or SPRED) of the two sets. If in some investigation time order of \( D_i, D_j \) is important, this should be stated in the question and the appropriate sample space constructed.

5. Developing the statistical background: Real world

In real world situations, there is generally no sufficiently large data space available to use the EOP above. There are two procedures that still may be of help in resolving the question of the statistically significant separation of \( D, M \) in either the SITES or SPRED sense. Note again that both the methods are nonparametric and require no assumption about the statistical structure of the reference distributions built up from \( D, M \). The procedures are discussed below.

a. Pool-Permutation Procedure (PPP)

A. Given: \( n \times p \) data sets \( D, M \) (over region \( R_0 \), time interval \( I_0 \), where \( p \) is the number of points in \( R_0 \), \( n \) the number of time samples at each point).

B. Question: Is STSTC(\( D, M \)) significantly large? (STSTC = SITES or SPRED).

C. Data space: \( D \) only.

D. Sample space: The present elements of this are constructed from \( D, M \) as follows:

(i) Represent \( D \) as \{\( d(1), \ldots, d(n) \}\) and \( M \) as \{\( m(1), \ldots, m(n) \}\), where \( d(t), m(t), t = 1, \ldots, n \), are vectors in \( E_0 \).

(ii) Form the union of \( D, M \), i.e., pool the \( m(t) \) and \( d(t) \) vectors to find \( U = \{u(1), \ldots, u(n), u(n + 1), \ldots, u(2n)\} = \{d(1), \ldots, d(n), m(1), \ldots, m(n)\} \).

(iii) Construct a random permutation \( \phi \) of the set of integers \( 1, 2, \ldots, n, n + 1, \ldots, 2n \).

(iv) Partition \( U \) via \( \phi \): \( U_1, U_2 = [\{u(\phi(1)), \ldots, u(\phi(n))\}, \{u(\phi(n + 1)), \ldots, u(\phi(2n))\}] \).

(v) Repeat (iii), (iv) \( r \) times, \( r \geq 100 \), thereby forming \( r \) ordered pairs \( (U_1, U_2), \ldots, (U_{r1}, U_{r2}) \). These constitute elements of the sample space.

E. Statistics: Form the statistic values \( S_k = STSTC(U_1, U_2)_k, k = 1, \ldots, r \).

F. Reference distribution: Order the \( S_k \) in § E according to increasing size; and relabel: \( S_{(1)} \leq S_{(2)} \leq \cdots \leq S_{(r)} \), crit = [(1 - \( \alpha \))\%], where \( \alpha \) is the significance level of the test.

G. Null hypothesis \( H_0 \): (\( D, M \)) is in the sample space.

H. Accept \( H_0 \): If STSTC(\( D, M \)) \( \leq S_{(1)} \).

Reject \( H_0 \): If STSTC(\( D, M \)) > \( S_{(1)} \).
In the latter case, STSTC(D, M) is significantly large; in the former case it is not.

The present procedure has the following rationale. We envision the two data sets D, M as two n-point swarms of points in \( E_p \). The null hypothesis in effect says that the labels "\( d(t) \)" and "\( m(t) \)" used to identify the points of each swarm, are interchangeable in arbitrary ways. The null hypothesis \( H_0 \) above is therefore shorthand for the assumption that D and M are randomly drawn from the same parent population—the one whose members are generated (say) by some common, statistically steady physical process. Thus D, M, under this assumption, are just as likely to have evolved by this process as any of the other members of the pairs \( U_1, U_2 \) of n-point subsets of the union U of M and D. Given this assumption, the remaining \( \S \) E–H follow just as in the EOP of Section 4.

It should be noted that the number of elements of the sample space formed in \( \S \) D can be quite large even for modest sample sizes n in D and M. For example, if \( n = 10 \), then there are at least \( (2n)!/(n!)^2 \) = 184756 elements in the sample space. Our experience shows that 100 random reorderings of the data/model sets \( r = 100 \) are adequate for most studies. The sample space defined as in (v) of \( \S \) D, although potentially quite large, nevertheless is a finite sample space which we visualize as a subset of some larger sample space that could be generated had we more to work with than just D. Therefore the null hypothesis \( H_0 \) in step G implies that there is room for (D, M) in this larger sample space along with all the \( (2n)!/(n!)^2 \) other pairs \( (U_1, U_2) \). Whether or not (D, M) is in the sample space, is decided in \( \S \) H. Observe that the temporal order of evolution of the rows of (D, M), and hence that of the rows of \( (U_1, U_2) \) is presumed immaterial to the question at hand (cf. comments on EOP).

b. Auto-cross Permutation Procedure (APP)

Another procedure that may be followed in the presence of small samples so as to lead to significance decisions concerning separation of data sets D, M is the following.

A. Given: \( n \times p \) data sets D, M (over region \( R_0 \), time interval \( I_0 \), where \( p \) is the number of points in \( R_0 \), \( n \) the number of time samples at each point).

B. Question: Is STSTC(D, M) significantly large? (STSTC = SITES or SPRED).

C. Data space: D only.

D. Sample space: The present elements of this are constructed from D, M as follows:

(i) Represent D as \( \{d(1), \ldots, d(n)\} \) and M as \( \{m(1), \ldots, m(n)\} \), where \( d(t), m(t), t = 1, \ldots, n \), are vectors in \( E_p \).

(ii) Construct a random permutation \( \phi \) of the set of integers \( 1, \ldots, n \). Let \( n = 2m \).

(iii) Partition D and M via \( \phi \), and form \( m \times p \) data sets, as follows:

\[
\begin{align*}
D_1, D_2 & = \{d(\phi(1)), \ldots, d(\phi(m))\}, \{d(\phi(m+1)), \ldots, d(\phi(n))\} \\
D_1, M_2 & = \{d(\phi(1)), \ldots, d(\phi(m))\}, \{m(\phi(m+1)), \ldots, m(\phi(n))\} \\
M_1, M_2 & = \{m(\phi(1)), \ldots, m(\phi(m))\}, \{m(\phi(m+1)), \ldots, m(\phi(n))\} \\
M_1, D_2 & = \{m(\phi(1)), \ldots, m(\phi(m))\}, \{d(\phi(m+1)), \ldots, d(\phi(n))\}
\end{align*}
\]

(D-based partitions)

(M-based partitions)

(For the case of \( n = 2m + 1 \), partition according to \( \{\phi(1), \ldots, \phi(m+1)\}, \{\phi(m+2), \ldots, \phi(n)\}\).)

(iv) Repeat (ii), (iii) \( r \) times, \( r \gg 1 \), thereby forming \( r \) ordered pairs \( (D_1, D_2), \ldots, (D_1, D_2)_r \), along with \( (D_1, M_2), \ldots, (D_1, M_2)_r \), from the D-based sets. These two sets form the sample spaces of current interest and are, respectively, the auto and cross sample spaces.

E. Statistics: Form the statistics \( A_k = STSTC(D_1, D_2)_k \), \( C_l = STSTC(D_1, M_2)_l \), \( k, l = 1, \ldots, r \).

F. Reference distribution: Order the auto statistics \( A_k \) in \( \S \) E according to increasing size; and relabel. Do likewise with the cross statistics \( C_l \):

\[
A_{(1)} \leq A_{(2)} \leq \cdots \leq A_{(r)} \leq \cdots \leq A_{(r/2)}
\]

\[
C_{(1)} \leq C_{(2)} \leq \cdots \leq C_{(r/2)} \leq \cdots \leq C_{(r)}
\]

where \( \alpha \) is the significance level of the test. Moreover:

\[
C_{(1)} \leq C_{(2)} \leq \cdots \leq C_{(r/2)} \leq \cdots \leq C_{(r)}
\]

where \( C_{(r/2)} \) = median of cross statistics \( C_{(r)} \), \( (r/2) \) taken to nearest larger integer.

G. Null hypothesis \( H_0 \): The pair \( (D_1, M_2) \), where \( l = (r/2) \), is in the auto sample space.

H. Accept \( H_0 \): If \( C_{(r/2)} \leq A_{(r)} \), where \( l = (r/2) \), in the auto sample space.

Reject \( H_0 \): If \( C_{(r/2)} > A_{(r)} \).
In the latter case, a randomly selected element \((D_1, M_2)_k\) of the cross sample space will lie beyond the critical level of the auto sample space at least 50% of the time. In such a case we would conclude that, with at least probability 0.50, that STSTC(D, M) is significantly large with confidence \((1 - \alpha) \times 100\%\). In this way we infer something about the separation of \(D\), \(M\) by a detailed examination of a large number of separations of paired subsets of \(D\) and \(M\).

The rationale for this procedure is basically that of PPP. The null hypothesis \(H_0\) is shorthand for the statement that \(D\) and \(M\) arise from the same population. If both \(D\) and \(M\) are drawn from the same population, then, as before, the identity-tag switching is permissible, and a sample space is readily forthcoming. Also, as a consequence of \(H_0\), the \(D\)-based auto and cross reference distributions are equal on average, and so the median \(C_{(1/2)}\) will rarely fall beyond \(A_{\text{crit}}\) (when \(\alpha = 0.05\), say), and \(H_0\) would be accepted at least \((1 - \alpha) \times 100\%\) of the time.

It should be noted that the conclusion following from a rejection of \(H_0\), namely, that, with at least probability 0.50, STSTC(D, M) is significantly large, is a much stronger type of conclusion than that which occurs in EOP or PPP: We make \(r\) trials at sampling where STSTC(D1, M2)k falls in the auto reference distribution of D. In the usual tests for significance, only one trial is made. This multiple-trial procedure arises in the present APP procedure by constructing the auto and cross reference distributions and seeing where the median of the latter falls on the range of the auto reference distribution. Thus we make a decision about STSTC(D, M), using \((D_1, M_2)_k\) values from \(C_{(1/2)}\).

We have not used the \(M\) based partition constructed above since we know from experiments and logical arguments that, for given parent populations for \(D\) and for \(M\), on the average, the cross distribution for \(M_1, \ M_2\) equals that for \(D_1, \ M_2\), and hence that the reference distributions of the \(M\) based partitions will, on the average, produce the same results (acceptance or rejection of \(H_0\)) as the \(D\) based partitions. Thus either set of partitions may be used in practice.

The relatively strong conclusions of the present procedure come from the possibility of making an auto reference from \(D\) (and hence automatically also from \(M\)). In order for this to be a usable procedure in practice, \(n\) must be somewhat larger than that in PPP, all other things being equal. The power tests below, however, indicate that \(n\) need not be much larger. Yet, it is clear that one cannot get more information out of \(D\) using APP, than out of \(D\) using PPP, unless the \(m \times p\) data sets of the above partition procedure in \(D\) (iii) have \(m\) approaching \(n\) in the PPP setting. We note finally that the temporal order of the rows of the \(D\), \(M\) pair and that of \(D_1, \ \ D_2, \ M_1, \ M_2\), is, by assumption, immaterial to the question at hand (cf. comments in EOP).

6. Tests of PPP and APP

a. Test data sets

To obtain a first impression of the operating properties of the two more frequently required procedures (PPP and APP), we conducted a set of controlled tests using artificial data sets for \(D\) and \(M\). Examinations of the EOP method are reported elsewhere (Preisendorfer and Mobley, 1982a).

The general test forms of \(D\), \(M\) are given by

\[
M = F + R_M, \quad (6.1)
\]

\[
D(\mu, \sigma) = [\mu E + \sigma F] + R_D, \quad (6.2)
\]

where \(R_D, R_M\) are \(n \times p\) matrices whose respective row vectors are drawn independently and randomly from a \(p\)-variate Gaussian population of zero mean and unit variance; \(F\) is an \(n \times p\) matrix whose entry in the \(r\)th row and \(x\)th column is given by

\[
f(t, x) = a \sum_{j=1}^{p} v_j(t) u_j(x), \quad (6.3)
\]

where for \(j = 1, \ldots, p; x = 1, \ldots, p; t = 1, \ldots, n\)

\[
u_j(x) = \left[2/(p + 1)\right]^{1/2} \sin[j \pi x/(p + 1)]
\]

\[v_j(t) = \cos(2\pi r_j t/M)
\]

\[r_j = \sin[j \pi t/(2(p + 1))]/\sin[p \pi t/(2(p + 1))]
\]

Thus \(F\) represents the deterministic part of both \(M\) and \(D(\mu, \sigma)\), and is derived from a solution of a set of differential equations governing a set of coupled harmonic oscillators of adjustable frequencies. The matrix \(D(\mu, \sigma)\) is an \(n \times p\) matrix where \(\mu\) and \(\sigma\) serve as location and scale parameters of the \(n\)-point swarm associated with \(D(\mu, \sigma)\) in \(E_p\). Moreover, \(E\) is an \(n \times p\) matrix all of whose entries are 1. The amplitude parameter \(a\) in \(F\) allows adjustment of the basic signal-to-noise ratio. Unless otherwise stated, \(a = 1\) throughout the discussion below. Finally, we set \(M = 12\) for all tests.

b. Power curves for SITES, SPRED via PPP, APP

The power curves for the statistics and their reference–distribution procedures arise as follows. In each step \(H\) of the PPP and APP, a decision process is provided for the acceptance or rejection of the null hypothesis \(H_0\). A good test procedure will generally often decide to reject the null hypothesis when in fact the null hypothesis does not hold. The higher the percentage of such rejections of \(H_0\) when \(H_0\) is false, the more powerful the test procedure.\(^2\)

In Fig. 2 we display the power of the SITES test under the PPP. We have chosen, for illustrative pur-

\(^2\) In general, the power curves will depend on the assumed form of \(D\) and \(M\) (see Section 6c).
poses, \( n = 36, p = 24, r = 100 \), and have let the significance level \( \alpha \) of the test be 0.10. This \( \alpha \) defines the probability of rejecting the null hypothesis \( H_0 \) when \( H_0 \) is true. Consider the curve labeled \( \sigma = 1 \). This was produced first of all by setting \( \sigma = 1 \) in (6.1) and (6.2). The \( \mu \) was set equal to 0.20 and ten realizations of the matrices \( M(\omega), D(0.20, 1; \omega), \omega = 1, \ldots, 10 \), were constructed. These realizations were produced by ten independent realizations each of \( R_D \) and \( R_M \). The SITES test was then applied to these 10 pairs. Only once out of these 10 trials did the SITES test arrive at the decision to reject \( H_0 \). The constructed form of \( H_0 \) is clearly false in the present set of trials, since \( \mu \neq 0 \). [The populations from which \( D(\mu, \sigma) \) and \( M \) are drawn have different locations in \( E_p \).] Hence for \( \mu = 0.20 \) the SITES test, using PPP, has low power. However, when \( \mu = 0.22 \), and \( \sigma = 1 \), it turned out that \( H_0 \) was rejected 9 out of 10 times. Fig. 2 shows that, as \( \sigma \) becomes larger, the power of SITES rises to 1.0 for ever smaller \( \mu \) values. In general, each point on the curves is the percentage of successes (\( H_0 \) rejections) resulting from 10 realizations of \( M \) and \( D(\mu, \sigma) \) for the given \( \mu, \sigma, n \) and \( p \) values.

Fig. 3 shows the power of the SITES test under APP and for the same parameters at the PPP test: \( n = 36, p = 24, r = 100 \) and \( \alpha = 0.10 \). The first thing to notice is the uniformly lower power of the APP tests relative to the PPP tests. Thus the \( \sigma = 2 \) curve, e.g., in PPP occurs at generally lower \( \mu \) values than that of the APP. This is a mixed blessing for PPP. Of course, one wants a powerful test; however, notice that the PPP curves, for all their higher power, are generally steeper than their APP counterparts. The latter, therefore, are “more forgiving” in a data intercomparison by both waiting a little longer before declaring a significant separation and by doing this over a larger \( \mu \)-interval as the curve rises from 0.00 to 1.00. For example, too abrupt an increase or decrease of power with small \( \mu \) changes would make tuning a GCM difficult: for example, as the modeler changes his parameters slightly, by increasing or decreasing them, the model may alternately be declared significant or not significant.

Examine next, the \( n \) behavior of the power curves (Figs. 2, 3, lower). Here we set \( \sigma = 1, p = 24 \) and \( \alpha = 0.10 \), and determine the \( n \) dependence of the power curves of SITES for PPP and APP, respectively. The power curves for both procedures indicate increasing power with sample size \( n \), as expected.

The SPRED power curves for both procedures are shown in Figs. 4 and 5 for \( \mu = 0, p = 24 \) and \( \alpha = 0.10 \). The PPP curves exhibit considerably higher power than the APP curves. Once again, as explained above, this is a mixed blessing for PPP. This disparity in power of the two procedures could be used to advantage in different situations. Thus in tuning a GCM or in generally less rigorous intercomparisons, APP could be used, while a stringent pair of SITES and SPRED tests can be made via PPP.

A detailed study has been made of the forms of the reference distributions, comparing those produced by
APP and PPP with "benchmark" distributions using Monte Carlo procedures (MCP). Moreover, comparisons were made between SITES and SPRED and classical $T^2$ and $\lambda$ (respectively, SITES type and SPRED type) measures of separation of $n$-point data swarms in $E_p$. The results are given, along with other intercomparison test studies, in Preisendorfer and Mobley (1982a). Also given in that reference are detailed descriptions and discussions of the power tests. The comparisons of the APP and PPP reference distributions with the MCP distributions under controlled artificial data conditions show that the APP curves tend to be closer to the "true" (benchmark) distributions under more general conditions than those of PPP. While this in itself does not place a final endorsement on the APP tests, it does indicate that, of PPP and APP, the sampling structure of APP is perhaps closer in form to that of the classical procedures.

c. Threshold detection with real data

A moment's reflection will show that the foregoing procedures can be used in real life situations to determine the threshold for Data-Model differences that the statistical tests can detect. This can be done simply as follows: Replace $D$ in (6.2) by the observed data set and $M$ in (6.1) by perturbations of $D$. These perturbations are constructed by changing the mean/variance of $D$ in predetermined ways. Application of the power tests to $D$ and its perturbed approximant allows one to determine the magnitude of the difference that the tests will detect with, say, 90% confidence, given the basic data set which with one has to work. The result of sensitivity studies, such as discussed here, offers a dimension not found in most statistical tests—but one which seems important in the model/data intercomparison.

7. Diagnosing model problems

In this section we offer an approach for diagnosing problem areas in models which fail the significance tests of Section 5. Specifically, we will develop a method for isolating regions of the space–time continuum where the differences between model and data are largest. The analysis will concentrate on possible difficulties associated with second moments (SPRED, SHAPE), since they represent a more difficult diagnostic problem than those associated with the first moment (SITES). In the latter case, one needs simply to subtract the mean fields of $D$ and $M$ to isolate, in space, the areas of largest discrepancy.

Using the notation of Section 3 we define the model/data difference field, $\Delta(x, t)$, as

$$\Delta(x, t) = [d(x, t) - d(x)] - [m(x, t) - m(x)],$$

(7.1)

where, e.g., $d(x)$ is the $x$th component of $d$. We next represent $\Delta(x, t)$ by means of a truncated sum of empirical orthogonal functions (EOFs):

$$\Delta(x, t) = \sum_{n=1}^{p'} \lambda_n a_n(x) e_n(x),$$

(7.2)

where the $n$th eigenvector is $e_n(x)$, the $n$th principal component is $a_n(x)$ with associated eigenvalue $\lambda_n$. Properties of such decompositions are available in the literature (e.g., Preisendorfer et al., 1981). Notice that we have trimmed the representation of $\Delta(x, t)$ down to $p'$ terms from the full $p$ terms. Sets of selection rules are available to estimate $p'$ and hence the part of $\Delta$ that is different from uncorrelated noise (cf. Preisendorfer et al., 1981).

The application of these selection rules to the difference field $\Delta(x, t)$ constitutes a form of significance test on the second moments $D$ and $M$. If the $D$ and $M$ differ in a way indistinguishable from white noise (or other preselected noise distributions), then we would find that $p' = 0$. In this case, both SPRED and SHAPE would tend to show $D = M$ to within the statistical confidence bounds of the analysis. If $p' > 0$ then it is a simple matter to synthesize the difference field between data and the model as shown in (7.2). The average spatial properties of the difference field are carried by the $e_n(x)$ and hence they show the spatial regions where, on average, the model and data differ the most and in a non-random way. Similarly the $a_n(x)$ show the time structure of the difference field and hence give information on the time scales of errors and their possible relations to, say, time-dependent forcing functions. The $e_n$ and $a_n$, to-

![Fig. 4. Power curves for SPRED statistic as generated using PPP. Parameters are as defined in Fig. 2.](image-url)
gether, can be considered as crude spatial and temporal spectra of the difference field, allowing the modeler to define the scales where the model performs both well and poorly.

Armed with the space–time details of the difference field, a knowledgeable modeler should be able to “fix” the problems. On the other hand, the error may not be pertinent to the problem under study. For instance, a climate model designed to reproduce the seasonal cycle may do quite well, but it may perform poorly at daily–weekly time scales and small space scales. This latter failing could result in failing the tests of Section 5. However, as long as the larger time and space scales are faithfully represented, the investigator may put up with inaccuracy which lies within the “noise level” of the climate study. Thus, cases are imaginable where, even though the model failed the tests of Section 5, it would still be useful. In such cases the model outputs could be suitably smoothed to eliminate the high-frequency (erroneous) non-information. The structure of the $e$, and $a$, would then suggest the construction of the necessary smoothing operator.

8. Applications

We now use the techniques developed in Sections 5 and 7 to explore the behavior of selected numerical models. Note that our applications are for demonstrative purposes only. They are not intended as full tests of the models discussed. We are grateful to R. Chervin (NCAR) and G. Heburn (NORDA) for making their model results available to us for these example calculations.

a. General circulation model (GCM) example: Model-data intercomparison

R. Chervin provided us with five realizations of “typical” January sea level pressure (SLP) fields for the Northern Hemisphere from an early version of the NCAR GCM, each realization differing from the others only by small, random perturbations in the initial conditions. Each realization is the result of averaging the last 30 days of a 60-day integration. The energetic region between 20°–80°N, 10°W–180° was singled out for comparison with observation. The model output was projected onto a 5° latitude by 10° longitude grid. Therefore, the $M$ matrix was of dimensions $5 \times 234$.

The main problem in carrying out the model–data intercomparison was to define in the observations a set of “typical” Januarys. Since the GCM studies don’t define “typical,” we have taken it to mean “average.” Thus, for comparison, we selected the five Januarys from the data that were closest to the long-term average January. The method was as follows: let $p(x, t) =$ SLP field with $x =$ (latitude, longitude) coordinate and $t =$ time. Twenty-four years of January data were used so $t = 1, 2, \ldots, 24$. We define

$$p(x) = \frac{1}{24} \sum_{t=1}^{24} p(x, t),$$

$$\sigma(x) = \left( \frac{1}{23} \sum_{t=1}^{24} (p(x, t) - p(x))^2 \right)^{1/2}.$$

We next defined a “separation index” $\lambda(t)$ such that

$$\lambda(t) = \sum_{x} \left[ \frac{p(x, t) - p(x)}{\sigma(x)} \right]^2.$$

Note that all (latitude, longitude) coordinates have equal weight in $\lambda$. Finally, the five Januarys with the smallest values of $\lambda$ were selected as most “typical” and hence most appropriate for comparison with their GCM counterparts.

The question posed was, “Did the GCM reproduce the observed mean January SLP field in the domain in interest?” This question is answered via the SITES statistic. The reference distribution (the permutation CDF) was constructed according to Section 5 using PPP. The actual model/data value of SITES was then entered on the CDF (Fig. 6). The actual SITES value is far above the 95% critical value of the distribution. We conclude with confidence exceeding 99% that in the region of interest the GCM’s “typical” January SLP field is different, in the sense of SITES, from a “typical” observed SLP.

b. Regional ocean circulation example: Model-data intercomparison

Heburn (1980) constructed a three-dimensional model to account for the ocean circulation off the coast of South America. The model is started from

![GCM vs OBSERVATIONS](SLP FIELD)

**Fig. 6.** Intercomparison of the location (SITES) of sea-level pressure fields produced by NCAR GCM vs. observations for typical January. The solid line is the cumulative (reference) distribution function (CDF) obtained by pool, permutation procedure (PPP). The actual value of the SITES statistic between the GCM runs and the data realizations is shown on the SITES axis.
rest and forced with the observed local wind and long-shore pressure gradient to simulate a propagating Kelvin wave. Open ocean boundary conditions are based also on observations. The model results are to be compared with month-long current meter observations from 30 positions within the model’s domain. The data were obtained during the spring of 1977 as part of the CUEA program. A typical example of the time history of longshelf current, from a current meter and the comparable model’s output at the current meter location, is shown on Fig. 7.

The question we posed was, “Does the model produce the degree of variability observed in the current meter data?” It would thus be difficult to digest, simultaneously, 30 pictures, such as Fig. 7, and provide any type of quantitative answer to this question. Both the PPP and APP SPRED test (Section 5), were applied to the 30-day time series (four observations per day) for the 30 current meters and the model-simulated currents for the current meter locations. In this case \( D \) and \( M \) had dimensions \( 120 \times 30 \).

The tests led to an ambiguous result (Fig. 8). The actual model-data value of SPRED was larger than any comparable value obtained via the PPP reference distribution. It may thus be concluded, with confidence greater than 99%, that the magnitude of the variability in the ocean model was distinctly different than that occurring in the data. However, in the lower panel of Fig. 8, the median SPRED value from the APP cross-statistics (Section 5b, \( F \) and \( F \) F) was nearly in the center of the auto-statistics values. In this test we conclude that the model and data have similar levels of variability. Such apparent dichotomies might be expected based on the power tests (Section 6) and vividly illustrate the more forgiving nature of the APP procedure. In view of the comments at the end of Section 6b, one might choose to conclude that the model has reasonably reproduced the observed long-shore current.

c. Model-model intercomparison

The method of Section 5 may also be used to test differences associated with the incorporation of different physical processes in model hierarchies. The regional ocean model of Heburn (1980) readily allowed such an example calculation. We use Model I to denote the regional ocean model described in Section 8c. This model does not include the process of air-sea heat exchange. Heburn next constructed Model II so that it was identical to Model I in every way (boundary and initial conditions, wind forcing, etc.) except that air-sea heat exchange was included as an additional forcing term to better represent the sea surface temperature field. The result for each model was a set of simulated, month-long current fields at 30 common spatial points. Thus \( D = M \), now represents Model I and \( M = M_II \) represents Model II, and \( n = 120 \), \( p = 30 \).

The question posed was “Does the inclusion of air-sea heat exchange in Model II lead to an appreciably different degree of variability in the simulated current data?” Application of the PPP and APP SPRED test to the two sets of model output provides an answer as shown in Fig. 9. The PPP test shows the magnitudes of the variability in Models I and II match with confidence 20% or equivalently, are disparate, with confidence 80%. Hence the inclusion of air-sea heat exchange has modified the variability of Model II relative to Model I, but not enough to give a clear statistical difference. The APP result (Fig. 9 lower) suggests also that Models I and II differ little in their

![Fig. 7. Time histories of long-shelf current velocity component (cm s\(^{-1}\)) from current meter (data) and as simulated by the regional ocean model of Heburn (1980).](image)
Fig. 8. Intercomparison of the variability (SPRED) between regional ocean model's simulated current field and the observed current fields. The cumulative (reference) distribution functions are obtained from pool, permutation procedure (PPP) [upper] and auto-cross procedure (APP) [lower].

Fig. 9. Comparison of the magnitude of variability between two different ocean models. Model I contains no air-sea heat exchange. Model II is identical to I except that it contains air-sea heat exchange as an additional forcing function. The cumulative (reference) distribution functions are obtained from PPP (upper) and APP (lower).

Fig. 10. Mode 1 eigenvector (upper) and principal component (lower) for the normalized difference field between Model I and the observations. The mode captured 49% of the variance in the difference field.

variability. These results together suggest that the process of air-sea heat exchange did not significantly affect the degree of variability of the current field in Model II relative to Model I.

d. Diagnosing model problems

We return to the intercomparison of Section 8b and try to determine the space-time structure of the model field that led to its failure of the SPRED test according to the PPP method. The difference field was formed according to Eq. (7.1) for the longshore component of the flow. Each time series in $\Delta$ was next normalized to have unit variance so as to obtain a picture of relative error. The EOF's of this normalized field were computed and the subsequent eigenvalues examined. A value of $p' = 2$, determined by Rule N of Preisendorfer et al. (1981) showed that 62% of the variance in the $\Delta$ field was noise, suggesting substantial room for model improvement.

The first eigenvector, accounting for 49% of the variance in $\Delta$, is shown on the upper panel of Fig. 10. The display has been converted to longitude-depth coordinates in order to display the model-data discrepancies in a physical space. Several features are immediately obvious:
1) All components of the first eigenvector are negative values, showing the errors are coherent over the entire measurement domain.

2) The discrepancies, while somewhat larger at mid-depth, are of the same relative magnitude over the sample water column. This suggests potential problems with large-scale forcing functions or perhaps in reproducing scales of motion larger than the study area.

3) The distribution of observations in physical space was not ideal for testing the model; e.g., what was the character of the near-surface error field between 75°34' and 75°40'?

The principal component that goes with EOF1 is shown in the lower panel of Fig. 10. This panel suggests the model errors are associated with relatively long time scales of the order of 100 hours. This is a characteristic time interval between synoptic events and so the error may be related to inexact specification of the forcing functions.

The discussion above was only for demonstrative purposes. With additional information (e.g., the model forcing functions), a more substantive discussion based on an EOF decomposition of the vector field or a complex vector EOF decomposition designed to look at the propagation of errors within the Δ field (e.g., Barnett, 1983) would be possible.

9. Conclusions

The amount of current effort in modeling various aspects of atmospheric and oceanic dynamics requires some objective means to compare the models with reality. It would be fair to say that, at this stage of modeling efforts, there are no adequate all-purpose procedures by which to judge model performance. There is no need for such objective judgments when a model is obviously not performing well. However, as a model’s fidelity to reality increases, there is a point where a quantitative, communicable verification of the model’s performance would be desirable.

In the present study, we have defined and tested two statistical procedures for deciding on the relative location (mean value in time or space) and spread (variability in time or space) of two data sets.

Our confidence in these tests has been established by repeated applications, to controlled artificial data sets, over a great variety of location and spread conditions. Moreover, our applications to real data sets in this study, and other applications recorded elsewhere, have added to our confidence in the general reasonableness and utility of the procedures. Our present conclusion, therefore, is that we have some tentatively workable measures of distance between the various attributes (location, spread, shape) of extensive multivariate small-sample data sets and also procedures by which to decide on the statistical significance of those measures. Finally, we have suggested a method by which the major discrepancies between model fields and the observed fields they seek to represent can be objectively defined in the space–time continuum.

Acknowledgments. The authors are greatly indebted to Bob Chervin (NCAR) and George Heburn (NORDA) for providing the model fields necessary to demonstrate the methods discussed in this paper. The generous contributions of both individuals to this paper are appreciated. The work was performed under support from National Science Foundation Grant ATM-7918206 (TPB) and support from the Pacific Marine Environmental Laboratory of NOAA (RWP).

REFERENCES


