Multivariate Recurrence Analysis

FRANCIS W. ZWIERS

Canadian Climate Centre, Downsview, Ontario, Canada

HANS VON STORCH

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

(Manuscript received 20 June 1988, in final form 30 June 1989)

ABSTRACT

Recurrence analysis was introduced to infer the degree of separation between a "control" and an "anomaly" ensemble of, say, seasonal means simulated in general circulation model (GCM) experiments. The concept of recurrence analysis is described as a particular application of a statistical technique called multiple discriminant analysis (MDA). Using MDA, univariate recurrence is easily generalized to multicomponent problems. Algorithms that can be used to estimate the level of recurrence and tests that can be used to assess the confidence in a priori specified levels of recurrence are presented.

Several of the techniques are used to reanalyze a series of El Niño sensitivity experiments conducted with the Canadian Climate Centre GCM. The simulated El Niño response in DJF mean 500 mb height are all estimated to be more than 94% recurrent in the tropics and are estimated to be between 90% and 95% recurrent in the Northern Hemisphere between 20° and 60°N latitude.

Discrimination rules that can be used to classify individual realizations of climate as members of the control or "experimental" ensemble are obtained as a by-product of the multiple recurrence analysis. We show that it is possible to make reasonable inferences about the state of the eastern Pacific sea surface temperature by classifying observed DJF 500 mb height fields with discrimination rules derived from the GCM experiments.

1. Introduction

The concept of recurrence analysis was recently introduced to the meteorological literature by Storch and Zwiers (1988). In that paper, which we will subsequently refer to as SZ, the concept was dealt with in a univariate manner. In the present paper the concept is extended to the multivariate domain.

Recurrence analysis is a tool that can be employed in the analysis of climate experiments with general circulation models (GCMs), and in model–model and model–reality comparisons. The aim is to characterize aspects of the two climates that are significantly different and recurrent. For example, in an El Niño sensitivity experiment such as the one discussed in SZ, the aim is to discover aspects of the GCM's response to a given sea surface temperature (SST) anomaly that are likely to recur each time a new, independent realization of the perturbed climate is simulated.

Geometrically the concept is very simple: the difference between two climates is "recurrent" if very little overlap between the probability distributions describes the realizations of climate states in the two climates. A schematic diagram of the concept is given in Fig. 1. We suppose for the moment that climate states can be represented by bivariate Gaussian distributions. Figure 1 shows two such distributions corresponding to two different climates. The centroids of the two populations are sufficiently separated that in any statistical test with reasonably sized samples we would come to the conclusion that the means of the two climates are well separated. On the other hand, there is overlap between the two distributions, and hence the parent population of a realization that is chosen randomly from one of the two climates cannot be determined with complete reliability. Thus, the difference between the two climates is not 100% recurrent. A precise definition of the concept in the case of a univariate state "vector" is given in SZ. An appropriate definition for the multivariate case will be given below.

In section 2 we discuss the connection between recurrence analysis and multiple discriminant analysis and review some of the rich statistical literature on the latter subject. Several multivariate techniques for estimating the degree of recurrence are described in section 3. Some multivariate tests for a priori specified levels of recurrence are described in section 4. A reanalysis of the El Niño sensitivity experiments described by SZ, using multivariate techniques rather than univariate techniques, appears in section 5. The paper concludes with a brief discussion in section 6.

Corresponding author address: Dr. Francis W. Zwiers, Canadian Climate Centre, 4905 Dufferin Street, Downsview, Ontario, Canada M3H 5T4.

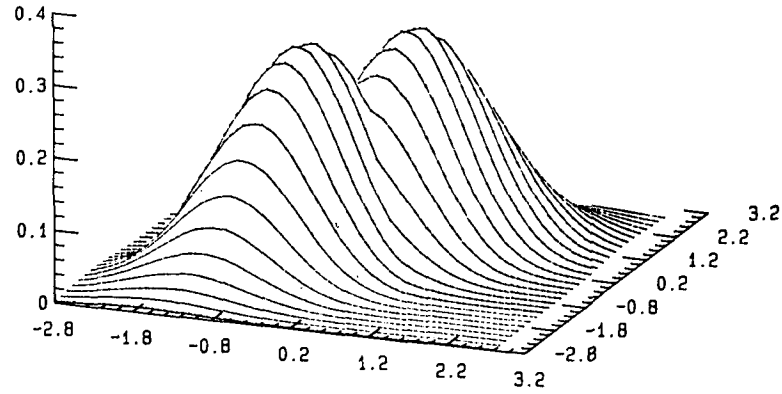

FIG. 1. A schematic illustration of the probability density functions of two Gaussian populations defined on a two-dimensional state space. The means of the two distributions are separated by the vector $(1, 1)^t$ and they have common variance–covariance matrix $\Sigma = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}$.

2. Multiple discriminant analysis

The concept of recurrence analysis described by SZ is very similar to that used by statisticians in multiple discriminant analysis (MDA). In MDA it is presumed that observations come from one of two (or several) populations. Samples taken from the two (or more) populations in question are used to estimate a rule with which to classify future observations of unknown origin as belonging to one of the parent populations. Having estimated such a discrimination rule, probabilities of misclassification can be estimated, and the "significance" of the rule can be tested. Anderson (1984) contains a good description of MDA in a rather general setting. Penfield and Koffler (1985) give a short overview of some of the statistical literature on the subject.

The probabilities of misclassification are directly related to the degree of recurrence of the climate response: If the degree of recurrence is small the probability of misclassification will be large and vice versa. Alternately, a probability of misclassification (or level of recurrence) can be specified a priori and a test can be conducted to determine whether it is possible to discriminate between the two climates to that degree. Thus the tools of MDA provide an immediate generalization of recurrence analysis to multivariate problems.

a. A brief introduction

To begin the discussion we describe briefly the derivation of Fisher's linear discriminant function (LDF), which is often used in MDA. First suppose that we have an l-dimensional state vector $\mathbf{Z}$ which was observed in one of two climates (the "experimental" Π_e and the "control" Π_c). The problem is to determine which of the two climates generated the observed vector. If $f_e(\mathbf{z})$ and $f_c(\mathbf{z})$ are the probability density functions of state vectors observed in the experimental and control climates respectively, then a reasonable approach is to form a likelihood ratio and to classify $\mathbf{Z}$ according to the rule

$$\mathbf{Z} \in \begin{cases} \Pi_e, & \text{if} \quad f_e(z)/f_c(z) \geq 1 \\ \Pi_c, & \text{otherwise.} \end{cases} \tag{1}$$

If the relative costs of misclassification, say c_e (the cost of classifying $\mathbf{Z} \in \Pi_e$ when $\mathbf{Z}$ actually comes from Π_c) and c_c, are known, and if the prior probabilities q_e and q_c that $\mathbf{Z}$ comes from the experimental or control climates are known, then an "optimal" rule that minimizes expected cost is given by

$$Z \in \begin{cases} \Pi_e, & \text{if} \quad f_e(z)/f_c(z) \geq k \\ \Pi_c, & \text{otherwise} \end{cases} \tag{2}$$

where $k = c_e q_c / c_c q_e$.

Under the Gaussian assumption and after taking logs (1) reduces to

$$Z \in \begin{cases} \Pi_e, & \text{if} \quad (z-\mu_c)^t \Sigma_c^{-1}(z-\mu_c) - (z-\mu_e)^t \\ & \quad \times \Sigma_e^{-1}(z-\mu_e) \geq \dfrac{l}{2}(\ln|\Sigma_e| - \ln|\Sigma_c|) \\ \Pi_c, & \text{otherwise.} \end{cases} \tag{3}$$

The quantity $(\mathbf{z}-\mu_c)^t \Sigma_c^{-1}(\mathbf{z}-\mu_c) - (\mathbf{z}-\mu_e)^t \Sigma_e^{-1}(\mathbf{z}-\mu_e)$ is referred to as the "quadratic discrimination function" (QDF). Fisher's linear discrimination function (LDF) is obtained by making the further assumption that control and experimental variance–covariance matrices, Σ_c and Σ_e, are equal. In this case we have

$$Z \in \begin{cases} \Pi_e, & \text{if} \quad z^t \Sigma^{-1}(\mu_e - \mu_c) \\ & \quad - (\mu_e + \mu_c)^t \Sigma^{-1}(\mu_e - \mu_c)/2 \geq 0, \\ \Pi_c, & \text{otherwise} \end{cases} \tag{4}$$

where Σ is the common variance–covariance matrix.

It is easily shown that the discrimination statistic given by

$$W(\mathbf{z}) = \mathbf{z}^t\boldsymbol{\Sigma}^{-1}(\mu_e - \mu_c) - (\mu_e + \mu_c)^t\boldsymbol{\Sigma}^{-1}(\mu_e - \mu_c)/2 \quad (5)$$

has a Gaussian distribution with mean $\pm\ \nabla^2/2$ and variance ∇^2 where ∇^2 is the Mahalanobis distance:

$$\nabla^2 = (\mu_e - \mu_c)^t\boldsymbol{\Sigma}^{-1}(\mu_e - \mu_c). \quad (6)$$

The sign of the mean of $W(\mathbf{z})$ depends upon whether $\mathbf{Z}$ comes from the experimental or control climate. Thus the probabilities of misclassification are

$$\Pr\{e|c\} = \Pr\{c|e\} = \Phi(-\nabla/2) \quad (7)$$

where $\Phi(\cdot)$ denotes the distribution function of a standard Gaussian random variable.

In Fig. 1 the difference of means vector $\mu_e - \mu_c$ is given by $(1, 1)^t$ and the common variance covariance matrix is given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}$$

so that the Mahalanobis distance $\nabla^2 = 4/3$ and the probabilities of misclassification are 12.4%. Because the means of the experimental and control populations are given by $(0.5, 0.5)^t$ and $(-0.5, -0.5)^t$, respectively, the LDF is given by $W(\mathbf{z}) = \frac{2}{3}z_1 + \frac{2}{3}z_2$, where z_1 and z_2 are the two components of the bivariate state vector. A contour diagram of the two distributions is shown in Fig. 2 as is the line $W(\mathbf{z}) = 0$. The decision $\mathbf{Z} \in \Pi_e$ is made if the observed $\mathbf{Z}$ lies on or above the line; the decision $\mathbf{Z} \in \Pi_c$ is made if the observed $\mathbf{Z}$ lies below the line.

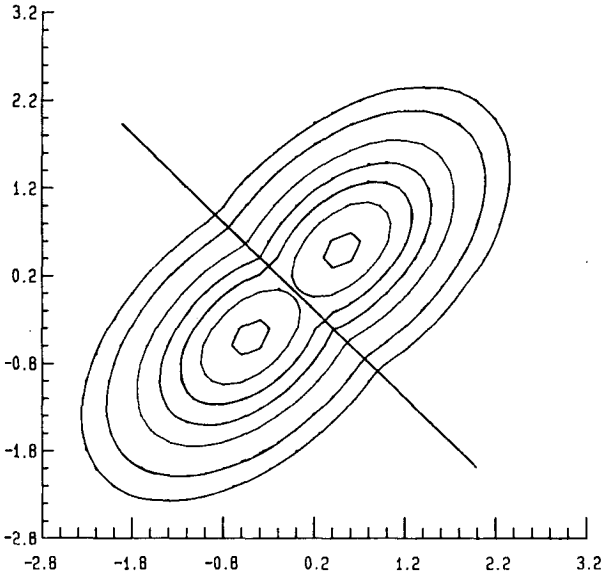

FIG. 2. As in Fig. 1, except a contour plot is shown. The discrimination function $W(z) = 0$ is also plotted.

Although there are many ways in which to approach MDA problems (see Penfield and Koffler 1985; Das Gupta 1973) the most frequently taken approach is to assume that $\mathbf{Z}$ has a Gaussian distribution with the same variance–covariance matrix in both climates and to replace the unknown parameters in (4) with estimates constructed from random samples of state vectors obtained from both climates. These samples are frequently referred to as "training samples" because they are used in the estimation of the discrimination rule. Under these circumstances the discrimination statistic W is given by

$$W(\mathbf{z}) = \mathbf{z}^tS^{-1}(\bar{x}_e - \bar{x}_c) - (\bar{x}_e + \bar{x}_c)^tS^{-1}(\bar{x}_e - \bar{x}_c)/2 \quad (8)$$

where S is the pooled sample variance–covariance matrix which is given by

$$S = [\sum_{i=1}^{n_e}\sum_{j=1}^{n_e}(x_{e_i} - \bar{x}_e)(x_{e_j} - \bar{x}_e)^t + \sum_{i=1}^{n_c}\sum_{j=1}^{n_c}(x_{c_i} - \bar{x}_c)(x_{c_j} - \bar{x}_c)^t]/[n_e + n_c - 2]. \quad (9)$$

Discrimination rule (4) is replaced with

$$\mathbf{Z} \in \begin{cases} \Pi_e, & \text{if } W(z) \geq 0. \\ \Pi_c, & \text{otherwise.} \end{cases} \quad (10)$$

This approach is not particularly robust when data are non-Gaussian. In this case an approximate transformation to the Gaussian distribution is often made by using a rank replacement technique such as that described by Conover and Iman (1980). Specifically, the observations in both training samples are combined into a single large pool of observations. Each entry in each climate state vector contained in the pool is ranked among the corresponding entries in every other state vector in the pool. The resulting vectors of ranks are resegregated according to climate, and the resulting samples of vectors of ranks are used to conduct a multiple discrimination analysis. Monte Carlo experiments conducted by Conover and Iman (1980) indicate that little is lost by using this rank replacement when data are Gaussian and that there is considerable improvement in misclassification errors when data are non-Gaussian.

b. Misclassification errors and recurrence analysis

MDA has two roles in recurrence analysis: to provide a diagnostic description of the degree of recurrence in a simulated climate's response to experimental conditions and to provide a foundation for statistical tests of a priori chosen levels of recurrence.

One useful diagnostic of the degree of recurrence of the response that MDA can provide is an estimate of the probability of misclassification, or the *misclassifi-*

cation error rate. A small misclassification error rate indicates very little overlap of the two ensembles of climate state vectors, and hence a recurrent response. It should be noted that when we speak of the misclassification error rate, we follow the conventions of the statistical literature on MDA and refer to $\Pr\{e|c\}$, the probability of incorrectly classifying $\mathbf{Z} \in \Pi_e$ when in fact $Z \in \Pi_c$. While it is true that $\Pr\{e|c\} = \Pr\{c|e\}$ in the Gaussian setup, this is not always the case when data are non-Gaussian.

In fact, three misclassification error rates can be considered in discrimination problems. These are the *optimal error rate,* $\mathcal{E}_{\text{opt}}$, obtained by classification procedure (1) in which the distributions and all their parameters are known; the *conditional error rate,* $\mathcal{E}_{\text{cond}}$, obtained by a discrimination rule that has been estimated from training samples; and the *expected error rate,* $\mathcal{E}_{\text{exp}}$, the expectation of the conditional error rate taken over all possible training samples.

In climate problems both the optimal and conditional error rates are of interest. The optimal error rate is directly related to the idea of recurrence as defined by SZ and is in fact a measure of the separation between the experimental and control climates in an ensemble sense. An appropriate general multivariate definition for p-recurrence consistent with the univariate definition given by SZ is

General definition. The difference between two climates is said to be *p-recurrent* if the optimal error rate for discrimination procedure (1) is $1 - p$.

A special parametric form of this definition is obtained if the Gaussian assumption is appropriate. In this case we have the

Gaussian definition. Suppose that two climates can be characterized by Gaussian climate state vectors $\mathbf{Z}_e$ and $\mathbf{Z}_c$ that have a common variance–covariance Σ. Then the difference between the two climates is said to be *p-recurrent* if $\Phi(\nabla/2) = p$ where ∇^2 is given by (6).

That is, with the Gaussian assumption and common variance–covariance matrices, the difference between two climates is p-recurrent if $(\mu_e - \mu_c)'\Sigma^{-1}(\mu_e - \mu_c) \geq Z_p$ where the latter is the pth quantile of the standard normal distribution. Note that a response that is 50% recurrent has corresponding Mahalanobis distance $\nabla^2 = 0$ and a discrimination procedure (1) which has no skill in distinguishing between control and experimental climates. Also note that, when the dimension of the observed "state vector" is one, the difference between two climates is p-recurrent according to the Gaussian definition if $|\mu_e - \mu_c| \geq \sigma Z_p$ where σ is the common standard deviation. In the example of Figs. 1 and 2 the difference between the illustrated populations is 87.6% recurrent.

The conditional error rate is of interest because it indicates how well we can distinguish between the two climates in question on the basis of the available data. This is a diagnostic quantity that tells us something about the utility of studying the samples in hand in more detail. If the conditional error rate is large, we have an indication that the two samples are not well separated and that they may not contain a great deal of information about differences between the control and experimental climates. The conditional error rate is perhaps the more relevant of the two.

We use both the concepts of optimal and conditional error rate in this paper. An estimator of recurrence will always have the form $1 - \hat{p}$, where $\hat{p}$ is an estimator of the conditional error rate. On the other hand, tests of hypothesis are concerned with ensemble properties of the climates in question. Hence a level of recurrence specified a priori in a test of hypothesis will always have the form $1 - p$, where p is an optimal error rate.

3. Estimating the degree of recurrence

By virtue of the above definition of recurrence, any estimator of the conditional classification error rate may also be regarded as an estimator of the degree of recurrence of the difference between the control and experimental climates. That is, one minus the estimated classification error rate is an estimate of the degree of recurrence. The literature on MDA contains descriptions and simulation studies of many classification error rate estimators. We will briefly describe a few estimators in this section and will also briefly touch on the question of their uncertainty.

a. Error rate or recurrence estimation

The statistical literature contains many studies on error rate estimators including Lachenbruch and Mickey (1968), Efron (1983), Snappin and Knoke (1984, 1985), Chernick et al. (1985) and Page (1985). Most studies assess error rate estimators only as estimators of the conditional error rate using criteria such as mean absolute error and mean squared error. Among the studies cited above, only Page (1985) assesses error rate estimators relative to the optimal error rate as well as the conditional error rate.

Error rate estimators fall into two groups: namely, parametric and nonparametric. Parametric estimators are constructed assuming that the climates are Gaussian with common variance–covariance matrices and that discrimination rule (4) is used. The simplest of these estimators, referred to as the "D-method" in the literature, is computed by replacing the Mahalanobis distance ∇^2 in (7) with the sample equivalent D^2. The latter is given by

$$D^2 = (\bar{x}_e - \bar{x}_c)'S^{-1}(\bar{x}_e - \bar{x}_c). \qquad (11)$$

Page (1985) shows that this estimator is optimistically biased as an estimator of both the optimal and conditional error rates. Considerable improvement is ob-

tained if S^{-1} in (11) is replaced with an unbiased estimator of Σ^{-1}. The result is a "shrunken" estimator of the Mahalanobis distance, which is given by $DS^2 = (n_e + n_c - l - 3)D^2/(n_e + n_c - 2)$ and is subsequently substituted into (7). The resulting error rate estimator is referred to as the "DS-method" in the literature.

Page (1985) reviews several other error rate estimators as well. In agreement with an earlier simulation study reported by Lachenbruch and Mickey (1968), the estimator with the best overall performance among those tested by Page was the "OS-estimator." This estimator is based on an asymptotic approximation for the conditional error rate (Okamoto 1963, 1968) which is expressed in terms of n_e, n_c, l and ∇^2. When DS^2 is used as an estimator of ∇^2 the result is an error rate estimator (referred to as the "OS-method") that has low mean absolute error both as an optimal and a conditional error rate estimator. The OS-estimator is described in appendix A. Another estimator that is also based on an asymptotic expansion and performed well as a condition error rate estimator (but not as an optimal error rate estimator) in Page's Monte Carlo experiments is the "M-method" estimator described by McLachlan (1975). It is described in appendix B.

Nonparametric error rate estimators use only the information contained in the training samples. As opposed to asymptotic expansions, they do not rely on assumed parametric models such as the Gaussian distribution. The simplest of these, the "R-method," applies the estimated discrimination rule to every observation in the control climate training sample, and counts the number of times an observation is misclassified. This estimate of the conditional error rate is often referred to as the "apparent error rate." As with the D-method, it is optimistically biased (Lachenbruch and Mickey 1968; McLachan 1976). Lachenbruch (1967) proposed the "leave-one-out" method (referred to as the "U-method") as a nearly unbiased estimator of conditional error rate. An observation is selected from the control climate training sample, a discrimination rule is estimated from the remaining observations, and the selected observation is classified using this rule. This process is repeated for every observation in this training sample, and the number misclassified is used to estimate the conditional error rate.

Several authors have suggested and studied many other nonparametric conditional error rate estimators (Lachenbruch and Mickey 1968; Efron 1983; Chernick et al. 1985; Snappin and Knoke 1985) as well as the R and U estimators. Nonparametric error rate estimators are generally robust relative to parametric estimators but as a rule do not perform as well as parametric estimators when the data actually come from Gaussian distributions. It is very difficult to recommend a particular error rate estimator that works well in all situations. The results of the various simulation studies that have been reported suggest that the choice of estimator depends somewhat upon the dimension of the observed state vectors and the "distance" between the control and experimental climate ensembles. As a general rule, however, it appears that Efron's 0.632 estimator may be one of the best nonparametric error rate estimators (Efron 1983; Chernick et al. 1985).

The 0.632 estimator is a weighted average of two estimators, R and e_0, defined by

$$\hat{\mathcal{E}}_{\text{cond}} = 0.368R + 0.632e_0 \tag{12}$$

where R is the apparent error rate and e_0 is a bootstrap estimate of the error rate. Specifically, bootstrap samples are taken from the training samples; these samples are used to estimate a discrimination rule; and the observations in the control climate training sample that were excluded from the corresponding bootstrap sample are classified with the estimated rule. The average misclassification rate over many such bootstrap samples is reported as e_0. This is somewhat different than the standard bootstrap estimator of $\hat{\mathcal{E}}_{\text{cond}}$ described by Efron because e_0 estimates the error rate directly while the standard bootstrap estimates the bias of the apparent error rate. Efron's choice of 0.632 as the weight which is placed on e_0 was determined more or less heuristically: The asymptotic proportion of a training sample contained in each bootstrap sample is 0.632.

b. *Uncertainty of estimates*

An estimate of the degree of recurrence is not very useful without some indication of its uncertainty. Unfortunately, this aspect of error rate estimation has not been discussed a great deal in the MDA literature. As we have seen, most error rate estimators are quite complex and thus it is not usually possible to derive estimates of their standard errors. There are, however, a couple of exceptions to this statement.

McLachlan (1975) was able to show that the M-method estimate of the conditional error rate is asymptotically a Gaussian random variable with variance $\sigma_M{}^2$ (given in appendix B), which is a function of the Mahalanobis distance ∇^2. Thus an asymptotic $(1 - \alpha) \times 100\%$ confidence interval for this parametric error rate estimator is given by $M \pm \sigma_M Z_{1-\alpha/2}$ where $\sigma_M{}^2$ is evaluated using the shrunken estimator of ∇^2 and $Z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

Each classification in the leave-one-out procedure may be thought of as an almost independent binomial trial with probability $1 - p$ of success where p is the degree of recurrence (Lachenbruch 1967). Thus the standard error of U (and $1 - U$) is approximately

$$\sigma_U = \{p(1 - p)/n_c\}^{1/2}. \tag{13}$$

A corresponding estimate is given by $\hat{\sigma}_U = \{U(1 - U)/n_c\}^{1/2}$ and an approximate $(1 - \alpha) \times 100\%$ confidence interval is given by $U \pm \hat{\sigma}_U Z_{1-\alpha/2}$.

4. Multivariate tests for recurrence

We have already described several diagnostic measures of the degree of recurrence of the difference between two climates in the form of both parametric and nonparametric error rate estimators. In this section we will describe some procedures that can be used to look for evidence of a priori specified levels of recurrence. The appropriate null and alternate hypotheses for all tests that we will discuss are

$\mathbf{H}_0$: The experimental climate response

is less than p-recurrent (14a)

$\mathbf{H}_a$: The experimental climate response

is at least p-recurrent (14b)

where p has a specific value which is chosen a priori.

Storch and Zwiers (1988) describe several univariate statistical tests that can be used to look for specific levels of recurrence. These tests can be characterized as one of two types. In the first approach the a priori specified minimum degree of recurrence is used to develop a parametric noncentral difference of means test. We will describe two such tests for the multivariate case. These tests will be developed directly without using any of the tools of MDA. In the second approach the degree of recurrence is estimated (either parametrically or nonparametrically) and a test is conducted to see if this estimate is consistent with the a priori *specification*. We will use the tools of MDA to describe one test of this kind. Neither approach is perfect. The former assumes that the Gaussian model with common variance–covariance matrices and independent climate realizations is correct. The latter is not constrained as strongly by such assumptions, but does not take into account the fact that a conditional error rate estimator is used to estimate the optimal error rate. For all tests it is implicit that the training samples and individual realizations of the climate state vectors are independent of each other.

a. Parametric tests

The parametric tests that we consider are both based on the Hotelling T^2 statistic, which is the multivariate analogue of the Student's t-statistic. The Hotelling's T^2 statistic (Morrison 1976) is given by

$$T^2 = [n_e n_c/(n_e + n_c)]D^2. \tag{15}$$

Under the null hypothesis

$$F = ((n_e + n_c - l - 1)/[l(n_e + n_c - 2)])T^2 \tag{16}$$

is distributed as a noncentral F random variable with l and $n_e + n_c - l - 1$ degrees of freedom (df) and noncentrality parameter

$$\delta^2 = [n_e n_c/(n_e + n_c)]\nabla^2. \tag{17}$$

According to the Gaussian definition of p-recurrence the relationship between p and the Mahalanobis distance ∇^2 is given by

$$p = \Phi(\nabla/2). \tag{18}$$

Thus the noncentrality parameter of the F distribution may be specified a priori as

$$\delta^2 = 4[n_e n_c/(n_e + n_c)]Z_p^2 \tag{19}$$

where Z_p is the pth quantile of the standard Gaussian distribution. Having derived the noncentrality parameter for the test of recurrence from null hypothesis (14a) it is necessary to determine a "p-value" for the computed T^2 by integrating the density function of the noncentral F distribution. The null hypothesis is then rejected if the computed p-value (which is the significance of the observed T^2) is less than α. Determining the p-value directly from the density of the noncentral F distribution is a nontrivial problem because this distribution does not have a closed analytic form. Fortunately, there is a good approximation to this distribution that only requires the numerical integration of the incomplete Beta function. It is described in appendix C.

An equivalent approach to the testing problem using the same parametric setting is to ask, "At what level of recurrence is the observed T^2 just large enough to reject (14a) at the α significance level?" If this level is greater than p (14a) can be rejected at the α level. In order to apply this approach it is necessary to estimate the value of the noncentrality parameter for which (15) is just barely sufficient evidence to reject (14a). That means we must find δ^2 such that T^2 is equal to the critical value of the α level test of significance. In other words, we solve the following integral equation for $\hat{\delta}^2(\alpha, T^2)$:

$$\alpha = \int_F^\infty g(f; l, n_e + n_c - l - 1, \hat{\delta}^2(\alpha, T^2))\mathrm{d}f \tag{20}$$

where $g(\cdot; l, n_e + n_c - l - 1, \hat{\delta}^2(\alpha, T^2))$ is the probability density function for the noncentral F distribution with l and $n_e + n_c - l - 1$ df and noncentrality parameter $\hat{\delta}^2(\alpha, T^2)$, and F is given by (16). The solution is then substituted into (19) to obtain a Gaussian quantile, say $Z_{\hat{p}(\alpha,T^2)}$, and finally the Gaussian distribution function is evaluated at $Z_{\hat{p}(\alpha,T^2)}$ to obtain an estimate of p. This estimate is given by

$$\hat{p}(\alpha, T^2) = \Phi([n_e n_c/(n_e + n_c)]^{1/2}\hat{\delta}(\alpha, T^2)/2). \tag{21}$$

If $\hat{p}(\alpha, T^2)$ is greater than the value specified in (14a), the hypothesis is rejected. Note that throughout this development we have written $\hat{\delta}$ and $\hat{p}$ as explicit functions of α and T to emphasize this dependence.

This test is in fact equivalent to the noncentral T^2 test described above, but it has the added advantage that it is possible to obtain additional diagnostic information by evaluating and plotting $\hat{p}(\alpha, T^2)$ for var-

ious values of α. Note that $\hat{p}(\alpha, T^2)$ is a monotonically increasing function of α which takes values in the interval [0.5, 1.0]. Also note that the function is only defined for those significance levels α at which T^2 is sufficient evidence to reject the null hypotheses of equality of means. If ${T_1}^2 > {T_2}^2$ then $\hat{p}(\alpha, {T_1}^2) > \hat{p}(\alpha, {T_2}^2)$ for all α in the interval [0, 1] for which both functions are defined.

b. *A nonparametric test*

A nonparametric test of (14) can be constructed easily and directly when the U-method is used to estimate the misclassification error rate. With the U-method (and to a lesser extent, the R-method) we can think of the classifications of the withheld observations as n_c very nearly independent trials (see Lachenbruch 1968). Therefore, under the null hypothesis that the response is p-recurrent, the total number of observations that are misclassified by the cross-validation procedure has approximately a binomial distribution with n_c trials and probability of "success" $1 - p$. It is therefore possible to determine a p-value for the observed U and hence conduct a test of (14a) based on the U-method. When $n_c(1 - p) > 5$ and $n_c p > 5$ the Gaussian approximation to the binomial distribution may be used. In this case $1 - U$ can be standardized as

$$Z = (1 - U - p)/[p(1 - p)/n_c]^{1/2} \qquad (22)$$

and a p-value can be determined by determining the area under the standard Gaussian curve to the right of Z.

5. An application

As an example of a multivariate recurrence analysis, we reanalyze the El Niño sensitivity GCM experiments that have already been considered in SZ. The experiments were conducted with the Canadian Climate Centre (CCC) GCM (Boer et al. 1984a,b). The control ensemble consists of 76 December–January–February (DJF) means that were extracted from three extended control simulations using climatologically varying sea surface temperature (SST). (The analysis reported in SZ used only 30 control DJF means.) The experimental ensemble consists of 5 DJF means extracted from simulations with anomalous SST in the equatorial Pacific. The SST anomalies used were twice the Rasmusson and Carpenter (1982) anomaly (2RC), minus twice this anomaly (−2RC) and the observed DJF anomaly during the 1982/83 El Niño. Results of the 1982/83 simulation have previously been described by Boer (1985).

In SZ we considered 500 mb height and temperature and found almost identical results for these two parameters. For the sake of brevity, we limit ourselves to 500 mb height in this paper. This is a less than ideal choice in tropical regions because of the baroclinic nature of the tropical response to anomalous SSTs. However, we made our less than optimal choice so that we could use a locally available dataset of twice-daily 1000 and 500 mb height analyses in our analysis of the sensitivity experiments.

Using univariate recurrence analysis we showed in SZ that there was a locally significant highly recurrent response in large portions of the tropics (Figs. 4–8 in SZ) and in a few extratropical locations. In that paper we also conducted a multivariate significance analysis on the equality of means: In all three experiments the mean DJF 500 mb heights in the control and the disturbed climates were inferred to be significantly different.

a. *Data compression*

To estimate the degree of recurrence it is necessary to use discrimination rule (10) or to calculate statistic D^2 (11). Thus the inverse of the estimated covariance matrix S (9) must exist, implying that the number of samples must be at least as large as the number of degrees of freedom.

In our example the data are given on a 64 by 32 Gaussian grid and hence represent points in a 2048-dimensional vector space. On the other hand, only 81 control and experimental realizations of the DJF mean are available in any one climate comparison. It is thus necessary to reduce the degrees of freedom drastically prior to the multivariate recurrence analysis. A good way to do this is to project the raw 64 by 32 gridpoint fields onto a few a priori chosen "guess patterns" (cf. Storch 1987). A recurrence analysis is subsequently performed in the low-dimensional space spanned by the guess patterns.

The best guess patterns are those which incorporate some prior knowledge about the expected signal. Such patterns should be chosen in such a way that they span a (preferably low dimension) subspace of the observing space which is likely to contain most of the signal variance. A choice of patterns that does not satisfy this basic requirement would result in a situation in which statistical inferences are technically correct but irrelevant because they concern small, obscure components of the signal. Knowledge about appropriate guess patterns might come from similar previous experiments, observations, or simplified theory. Empirical orthogonal functions (EOFs) may be used as guess patterns in the absence of appropriate prior knowledge because they are known to approximate observed variability in an optimal way.

We will conduct analyses with the first five and first ten EOFs of simulated DJF mean 500 mb height in two equatorially centered latitude bands. For the present example we also have available a priori "observational knowledge" in the form of Northern Hemisphere mean 500 mb height fields for winters (DJF) 1955 to 1984. This time interval includes a series of El Niño events. The range of variation in the 500 mb height anomaly

patterns which corresponds to these El Niño events (specifically 1957, 1963, 1965, 1969, 1972 and 1982; Wright 1984) are taken to represent variation that is likely to occur in similar future events. They are therefore used as guess patterns in our study (section 5c). The 500 mb height analyses are only available at latitudes north of approximately 20°N and thus the corresponding anomaly patterns are useful only for examining the extratropical response of the simulated climate to anomalous SSTs in the Northern Hemisphere.

b. Using EOFs as guesses

We conducted our initial analysis using EOFs as guess patterns in two latitudinal bands; the "tropical" band which extends from 30°S–30°N and the "global" band which extends from 60°S–60°N. The data for our analyses, which were provided on a Gaussian grid, were first interpolated to a 5° by 5° latitude–longitude grid and then averaged over 10° latitude by 20° longitude grid boxes. EOFs were computed from the variance–covariance matrix of the 76 control samples for each latitude band. We conducted analyses using the first five and first ten EOFs as guess patterns in our analysis. The first five (ten) EOFs of the tropical band explain 67% (83%) of the variance of 500 mb height simulated in that band by the control simulations. The corresponding figure for the global band is 56% (76%). Because the variance of 500 mb height is small in the tropics, the EOFs of the global band represent spatial structures that are primarily extratropical. We therefore expect a strong signal in the tropical band and a weaker signal in the global band because most of the signal is confined to the equatorial Pacific.

1) ESTIMATED LEVEL OF RECURRENCE

The level of multiple recurrence was estimated using the D, DS, M, OS, U and 0.632 methods. If possible, the standard error was also estimated. The results obtained from the raw data and from the rank-transformed data (Conover and Iman 1980) are very consistent indicating the appropriateness of the Gaussian assumption. Therefore only results obtained using the untransformed reduced data are displayed in Table 1.

Table 1 reveals some interesting characteristics of the various error rate estimators. The optimistic bias of the D method is clearly evident when estimates of recurrence made with this method are compared with those made by other parametric methods. Differences between DS, M and OS are always smaller than the corresponding estimated standard error of M. Seemingly, the only real distinction among parametric estimators lies between D and the rest, and we conclude that DS is an adequate parametric estimator of multiple recurrence.

The nonparametric estimates of recurrence made by the U and 0.632 methods are consistent with the parametric estimates but are clearly more variable. This consistency is another indication of the validity of the Gaussian assumption in the present problem.

There are some interesting differences between the results obtained using only five EOFs as guess patterns and those obtained using ten EOFs (Table 2). In the tropical band, the estimated level of recurrence is at least 96% (99.9%) when five (ten) EOFs are used as guess patterns. When five EOFs are used the 1982/83 signal is strongest with an estimated level of recurrence of at least 99.9%; the 2RC signal is weakest with estimated levels of recurrence of approximately 98%. When ten EOFs are used very strong recurrent signals are detected in all three experiments. The ranking of the signals is different (−2RC is the strongest; 2RC is the weakest) but less relevant because of the strength of the signal in all cases. The differences between the two analyses show, particularly in the 2RC and −2RC experiments, that the simulated atmosphere's response to the prescribed anomalous boundary conditions has a structure which is not well represented by only a few EOFs of the control climate. Apparently these responses are not easily described in terms of the predominant modes of variation which are simulated in the control climate in the tropical region. This is reasonable because El Niño excited tropical disturbances are known not to be part of the internal atmospheric variability but rather, are due to external forcing.

The estimated levels of multiple recurrence are weaker if global data are considered. When five EOFs are used as guess patterns there are only small differences between experiments in the strength of the measured response. The strongest (weakest) signal is observed in the −2RC (2RC) simulated mean DJF height which is estimated to be approximately 88% (84%) recurrent with a standard error of approximately 3%. When ten EOFs are used there are somewhat larger differences in the strength of the measured response. The strongest (weakest) signal is observed in the 2RC (−2RC) experiment with approximately 98% (93%) recurrence with a standard error of approximately 1.2% (2.3%). As anticipated, features that contribute strongly to the signal in the tropical band are not well represented by the global region EOFs of the control climate because these EOFs represent primarily extratropical variation. Even so, the use of ten EOFs does enable us to capture considerably more of the signal than the use of only five EOFs.

2) MINIMUM SIGNIFICANT LEVEL OF RECURRENCE

The minimum level of significant recurrence, $\hat{p}(\alpha, T^2)$, is the least number p so that null hypothesis H_0^p may be rejected with risk α given the observed T^2. Curves of $\hat{p}(\alpha, T^2)$ for tropical and global 500 mb height using the first five and ten EOFs as guess patterns are shown in Fig. 3.

Figure 3 shows that in the case of the tropical band, all responses are consistent with at least 99% recurrence at the 5% significance level with the exception of the

TABLE 1. Results of a multiple recurrence analysis of the El Niño sensitivity experiments when five EOFs are used in the EOF truncation of the data. The row labeled T^2 contains values of the T^2 statistic for the comparison of control and experimental means indicated in the column headings. Rows labeled *D*, DS, *M*, OS, *U* and 0.632 contain corresponding estimates of the degree of recurrence. Entries in parentheses are corresponding estimates of standard error of the recurrence estimates. Standard error estimates for *M* and *U* are evaluated using expressions (B2) and (13) respectively. Rows labeled H_0 contain *p*-values for tests of the hypotheses indicated using the test statistics indicated in parentheses. All probabilities are given in percent.

	30°S–30°N			60°S–60°N		
Experiment	2RC	−2RC	1982/83	2RC	−2RC	1982/83
T^2	98.3	197.4	232.5	20.2	28.9	21.9
D	99.9	>99.9	>99.9	85.0	89.3	86.0
DS	98.6	>99.9	>99.9	84.1	88.4	85.0
M	98.6	>99.9	>99.9	84.1	88.3	85.0
	(0.9)			(3.4)	(2.9)	(3.3)
OS	98.3	99.9	>99.9	84.7	88.5	85.5
U	96.1	100.0	100.0	82.9	85.5	84.2
	(2.2)			(4.3)	(4.0)	(4.2)
0.632	96.3	>99.9	100.0	82.7	88.4	84.0
$H_0^{50\%}(T^2)$	0.0	0.0	0.0	0.4	0.0	0.2
$H_0^{84\%}(T^2)$	0.0	0.0	0.0	65.1	32.8	58.2
$H_0^{84\%}(U)$	0.2	0.0	0.0	60.6	36.3	48.0

response to the 2RC anomaly at the five EOF data truncation. This response is consistent with only 95% recurrence at the 5% significance level. Apparently a considerable portion of the signal is projected onto tropical band EOFs 6–10 in the 2RC case.

We see somewhat similar behavior in the global band in the sense that a considerable portion of the signal in the 2RC case is projected onto (global band) EOFs 6–10. The response to the 2RC anomaly in the global band is the largest of all measured global responses when projected onto ten EOFs and the least of all measured global responses when projected onto five EOFs. A distinguishing feature of the global responses is that they are much more clearly delineated in Fig. 3 than the tropical responses. They are also much less interesting, however, because they are not nearly as strong. The measured responses range from being consistent with approximately 92% recurrence at the 5% level (2RC, 10 EOFs) to 67% recurrence at the 5% level (2RC, 5 EOFs). In the latter case the estimated level of multiple recurrence is approximately 84% with a standard error of 3.4%, suggesting that a realization of DJF mean 500 mb height from the 2RC experiment would be misclassified as belonging to the control climate approximately 9%–23% percent of the time. This is evidence for considerable overlap of the populations of seasonal means of these climates.

The use of ten EOFs in the data truncation uncovers stronger evidence for recurrence both in the tropical and global bands. In the case of the tropical band, the observed response is consistent with a minimum of 99% recurrence at the 5% significance level in all three

TABLE 2. As Table 1, except ten EOFs are used in the EOF truncation.

	30°S–30°N			60°S–60°N		
Experiment	2RC	−2RC	1982/83	2RC	−2RC	1982/83
T^2	222.1	389.2	294.5	88.7	48.0	51.0
D	>99.9	>99.9	>99.9	98.5	94.5	95.0
DS	>99.9	>99.9	>99.9	97.8	93.1	93.7
M	>99.9	>99.9	>99.9	97.6	92.8	93.4
				(1.2)	(2.3)	(2.2)
OS	99.9	>99.9	>99.9	97.4	93.0	93.5
U	100.0	100.0	100.0	98.7	90.8	89.5
				(1.3)	(3.3)	(3.5)
0.632	100.0	100.0	100.0	98.7	93.4	93.0
$H_0^{50\%}(T^2)$	0.0	0.0	0.0	0.0	0.0	0.0
$H_0^{84\%}(T^2)$	0.0	0.0	0.0	0.1	12.4	9.1
$H_0^{84\%}(U)$	0.0	0.0	0.0	0.0	5.4	9.5

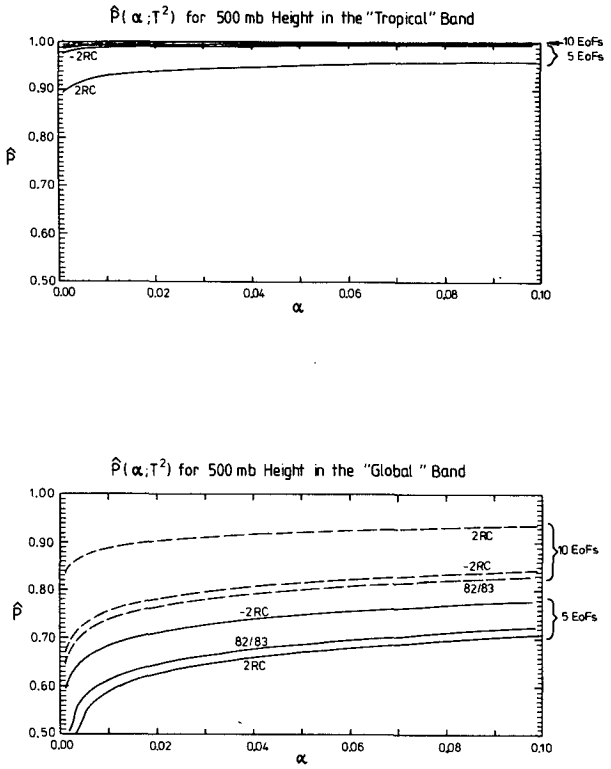

FIG. 3. Function $\hat{p}(\alpha, T^2)$ evaluated using values of the T^2 statistic representing differences between control and experimental DJF 500 mb height in the 30°N–30°S and 60°N–60°S latitude bands. Seasonal means were projected onto the first five (ten) EOFs of the control sample in the same latitude band.

experiments. In the case of the global band the 2RC experiments give rise to a response that is at least 92% recurrent at the 5% level. The magnitudes of the −2RC and 1982/83 responses are similar and are consistent with at least 80% recurrence at the 5% level. In the case of the −2RC experiment the estimated level of recurrence is approximately 93% with a standard error of 2.3% suggesting that realizations of DJF mean 500 mb height from the −2RC climate would not be correctly identified only 2%–12% of the time. In other words, there is a lot of information by which to characterize the results of the experiments even in the case of the weakest response.

3) T^2 AND U SIGNIFICANCE TESTS

Tests of hypotheses that the response is 50% recurrent and no more than 84% recurrent have been conducted using the parametric T^2 approach. Rejecting the null hypothesis $H_0^{50\%}$ is equivalent to rejecting the null hypothesis of equal means. The 84% level was chosen because in the univariate Gaussian case it is equivalent to a separation of means of at least two standard deviations (see SZ). The results are included in Tables 1 and 2.

The tropical height signal in all three experiments is significantly recurrent at the 84% level using both data truncations. Because $H_a^{84\%} \Rightarrow H_a^{50\%}$ the null hypothesis of equal means must also be rejected.

A global nonzero 84%-recurrent signal is identified at the 10% significance level in the height fields of the warm SST anomaly experiments but not in the cold SST anomaly experiment −2RC (10 EOF truncation). The $H_0^{84\%}$ could not be rejected for any experiment when only five EOFs were used for data truncation. The null hypothesis of equal means, however, is easily rejected at the 5% significance level in all experiments using both data truncations. This is not a great surprise; the relatively large control sample that is available implies that the T^2-test for equal means will be quite powerful and thus quite sensitive to small signals.

The results obtained with the U-method testing $H_0^{84\%}$ are quite similar to those obtained with the T^2 test with one exception.

4) CROSS-CLASSIFICATION OF RESPONSES

A by-product of the multiple recurrence analysis is discrimination rule (10), which can be used to make a decision about whether a particular mean DJF 500 mb height anomaly stems from a simulation forced with positive (negative) SST anomalies or a control run. To test the power of the discrimination rules found in the three SST anomaly experiments we used the discrimination rule based on one experiment to classify the individual outcomes of the other two experiments. The results for the "tropical" and "global" latitudinal belts are given in Table 3.

In the tropical area, the discrimination rules based on EOF truncation of the data appear to be quite powerful. Using the empirical evidence (based on the first five EOFs) of the 2RC (1982/83) experiment, five (three) DJF means of the other warm SST anomaly experiment, 1982/83 (2RC), are classified as being affected by a warm equatorial SST anomaly. When the response is projected onto ten EOFs the result is improved somewhat with four 2RC DJF means classified as belonging to the 1982/83 ensemble. Correctly, none of the −2RC DJF means is identified as being influenced by anomalous positive El Niño type SSTAs when either EOF truncation is used.

The results for the "global" data, 60°S–60°N, indicate that the EOF-based "global field" discrimination rules are unreliable, even when the higher EOF truncation is used. Only one of the outcomes of the warm SST anomaly experiments is classified as being forced by a warm equatorial SST anomaly. One of the outcomes of the cold anomaly experiment is also classified in this way.

c. Using observed anomalies as guesses

We found above that the DS estimator is adequate in the present circumstances and that the T^2- and the

TABLE 3. Classification of five individual response patterns simulated in the three GCM experiments (row headings) using the discrimination rule (10) based on the outcome of the other experiments (column headings). The data compression is based on EOF truncation using the first five (ten) EOFs. The entry "5" in the "2RC" column and "1982/83(30°S–30°N)" row (5 EOFs) indicates that all five tropical anomalies simulated in the five "1982/83" experiments are classified as belonging to the "2RC" ensemble and not to the control ensemble, if the empirical knowledge of the "2RC" experiment is used to formulate the classification rule.

		Five EOFs			Ten EOFs		
Region	Experiment	2RC	−2RC	1982/83	2RC	−2RC	1982/83
30°S–30°N	2RC	—	0	3	—	0	4
	−2RC	0	—	0	0	—	0
	1982/83	5	0	—	5	0	—
60°S–60°N	2RC	—	0	0	—	0	1
	−2RC	1	—	1	1	—	1
	1982/83	0	0	—	0	0	—

U significance tests yield equivalent results. Therefore, we limit ourselves only to the DS estimator and the T^2-test in this subsection. Anticipating that the bulk of the El Niño related information is located in the tropics and midlatitudes, we disregard all data north of 60°N. Analyses of the six observed positive SST anomaly guess patterns are available only in the Northern extratropics at latitudes greater than 20°N. Therefore, the tropical band considered in this subsection covers only 20°–30°N and the global band is taken to be 20°–60°N.

1) ESTIMATED LEVEL OF RECURRENCE

In the tropical band the estimated level of recurrence ranges between 94% (−2RC) and 99% (2RC) as is shown in Table 4. In the global band the estimates range between 90% (1982/83) and 95% (−2RC). For the tropical band the estimated level of recurrence is comparable to (less than) that obtained with five EOFs for 2RC (−2RC, 1982/83). This weak result using the observed guess patterns is not surprising because information about the expected form of the response is missing from most of the tropical band that was used in conjunction with the EOFs. In the case of the global band, the estimates of recurrence are comparable to those obtained with the ten EOF data truncation even though there are fewer guess patterns which are known over less than half of the tropical band used with the EOFs. This clearly shows that the Northern Hemisphere extratropical observed guess patterns are very efficient tools for extracting information about the nature of the model's response to anomalous equatorial forcing.

TABLE 4. As in Table 2 except results are for a multiple recurrence analysis of the El Niño sensitivity experiments when observed El Niño anomalies are used as guess patterns. Only results for the DS estimator and the T^2-test are shown. All probabilities are given in %.

	20°–30°N			20°–60°N		
Experiment	2RC	−2RC	1982/83	2RC	−2RC	1982/83
T^2	166.4	51.0	56.6	50.0	56.4	34.2
DS	99.9	94.2	95.1	94.1	95.1	90.1
$H_0^{50\%}(T^2)$	0.0	0.0	0.0	0.0	0.0	0.0
$H_0^{84\%}(T^2)$	0.0	3.3	1.6	3.5	1.7	22.9

2) T^2 SIGNIFICANCE TESTS

Curves of $\hat{p}(\alpha, T^2)$ for near tropical and extratropical Northern Hemisphere 500 mb height projected onto the six observed El Niño 500 mb height anomalies are shown in Fig. 4. In the tropical band the responses are found to be at least 84% recurrent at a significance level no greater than 3.3%. In the global band the responses to the warm and cold doubled Rasmussen and Carpenter anomaly are found to be at least 84% recurrent at a significance level no greater than 3.5%. It is rather puzzling that the response to the observed 1982/83 SST anomaly, which is estimated to be 90% recurrent, is the weakest response of the three experiments. While the response in this experiment was found to be significantly greater than 50% recurrent, its level of recurrence was not found to be significantly greater than 84%. This weak result shows that the response of the CCC GCM to the observed 1982/83 SST anomaly in the NH extratropical region is considerably different from most of the six El Niño DJF 500 mb height anomalies that are used as guesses.

3) CLASSIFICATION OF THE EXPERIMENTS

Having found that the recurrence analysis using observed El Niño guesses is competitive in the global band with the analysis using ten EOFs, we repeat the cross-classification done in section 5b (4) using the discrimination rules based on the observed El Niño guess patterns. The result is summarized in Table 5.

The result for the near-tropical latitude belt, 20°–30°N, lies between that obtained with the two EOF approaches for the tropical band, 30°S–30°N. The results for the midlatitude band, 20°–60°N, is much improved compared with Table 3. None of the warm

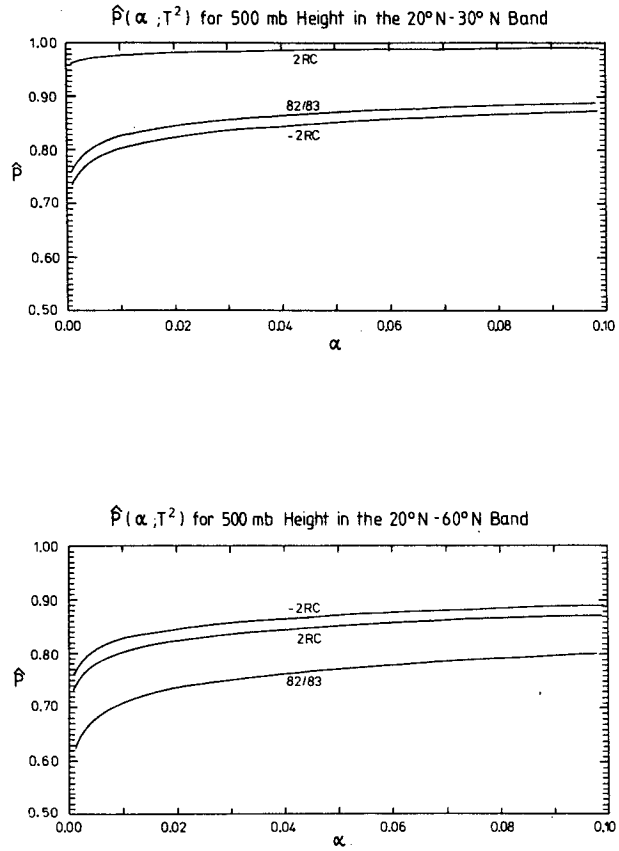

FIG. 4. As in Fig. 3, except for differences between control and experimental DJF 500 mb in the 20°–30°N and 20°–60°N latitude bands. Seasonal means were projected onto the six observed El-Niño guess patterns.

event anomalies is categorized as being a cold event anomaly and vice versa. Three of the 2RC simulations and one of the 1982/83 simulations are correctly classified as being modified by the presence of warm equatorial SST anomalies.

TABLE 5. Classification of five individual response patterns simulated in the three GCM experiments (row headings) using the discrimination rule (10) based on the outcome of the other experiments (column headings). The data compression is based on six observed El Niño guess patterns. The entry "4" in the "1982/83" column and "2RC/20°N–30°N" row indicates that four tropical anomalies simulated in the five "2RC" experiments are classified as belonging to the "1982/83" ensemble and not the control ensemble, if the empirical knowledge of the "1982/83" experiment is used.

Region	Experiment	2RC	−2RC	1982/83
20°N–30°N	2RC	—	0	4
	−2RC	0	—	0
	1982/83	0	0	—
20°N–60°N	2RC	—	0	1
	−2RC	0	—	0
	1982/83	2	0	—

d. *Classification of observed El Niño year anomalies*

Recurrence analysis classification rules obtained using ten EOFs and the six observed guess patterns were used to classify observed DJF 500 mb height for 1955 to 1984 inclusive. For the purposes of this classification, rules based on an EOF truncation of the data were rederived using only data in the near-tropical and extratropical latitude bands. EOFs of the control climate appropriate to the regions under consideration were employed. The results of these area-restricted recurrence analyses are summarized in Table 6. With the exception of the response to the 1982/83 experiment these results are comparable to those obtained with the observed guesses. In the case of the 1982/83 experiment the EOFs clearly do a better job of identifying the signal.

The classifications which were made as a byproduct of the area restricted analysis are summarized in Fig. 5. The diagrams display pairs of box plots indication the distribution of Wright's (1984) eastern tropical Pacific SST index for the nonwarm (noncold) and warm (cold) classifications. The box plots schematically depict the median values (the central horizontal bar), the 25th and 75th percentiles (the lower and upper horizontal bars, respectively) and the extremes (the endpoints of the vertical line segments) of the SST index values observed in classified years for each rule and classification.

As can be seen from Fig. 5, the combined use of observed guesses in the near-tropical band and empirical knowledge from the −2RC experiments (Fig. 5a) resulted in the best classification of observed DJF 500 mb height means into cold and noncold events. In this case all four generally recognized cold events (1955, 1970, 1973 and 1975) were correctly classified as cold events. Five of the remaining six realizations of DJF mean 500 mb height that were classified as cold events (1961, 1962, 1964, 1967, and 1984) are years with below average values of the SST index. The last year classified (1983) has an average SST index. The combined use of observed guesses in the near-tropical band and empirical knowledge from the 2RC experiments

TABLE 6. As in Table 2 except results are for a multiple recurrence analysis of the El Niño sensitivity experiments when the first ten EOFs of the control climate DJF 500 mb means are used as guess patterns. EOFs were computed for the 20°–30°N "near-tropical" band and the 20°–60°N "extratropical" band. Only results for the DS estimator and the T^2-test are shown. All probabilities are given in %.

	20°–30°N			20°–60°N		
Experiment	2RC	−2RC	1982/83	2RC	−2RC	1982/83
T^2	190.8	56.5	102.7	89.0	43.2	62.5
DS	>99.9	94.6	98.5	97.8	92.0	95.5
$H_0^{50\%}(T^2)$	0.0	0.0	0.0	0.0	0.0	0.0
$H_0^{84\%}(T^2)$	0.0	5.1	0.0	0.0	19.5	2.6

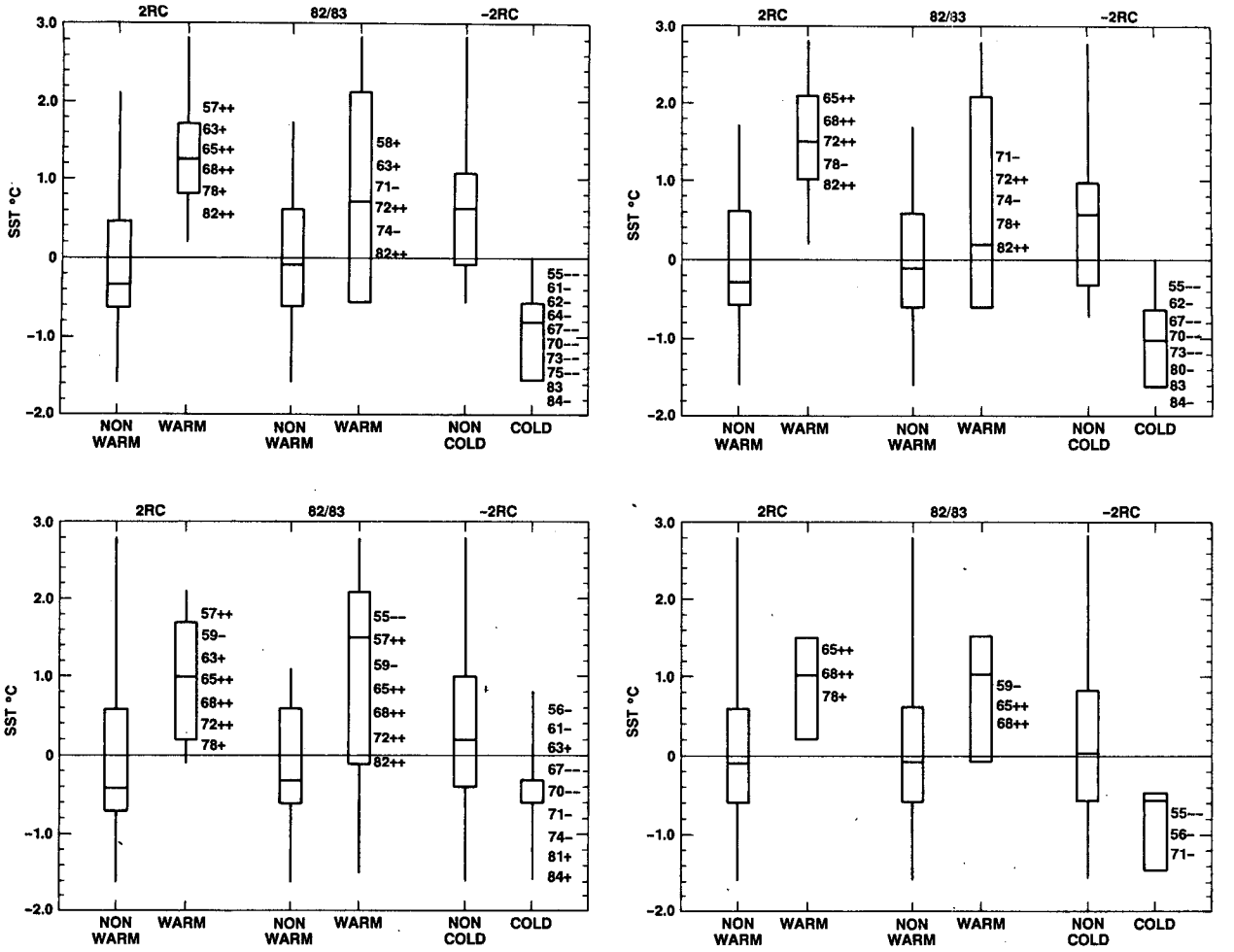

FIG. 5. Distribution of Wright's SST index by classification rule and classification. The horizontal line in the interior of each box indicates the median of the distribution. The lower and upper edges of each box indicates, respectively, the 25th and 75th percentiles of the distribution. The ends of the vertical line segments indicate the extremes of the distributions. Plus (minus) signs indicate years in which Wright's index is greater (less) than the mean. One sign indicates that the index has a value which differs from the mean by less than one standard deviation; two signs indicate that its value is different from the mean by at least one standard deviation. (a) Distributions for rules using the observed guess patterns in the 20°–30°N latitude band. (b) As in (a), but for rules using ten EOFs. (c) Distributions for rules using the observed guess patterns in the 20°–60°N latitude band. (d) As in (c), but for rules using ten EOFs.

resulted in the best classification of observed DJF 500 mb height means into warm and nonwarm events. In this case four of the six El-Niños are correctly classified as warm events. The two additional years that were classified as warm years (1968, 1978) are years with above normal values of the SST index. Note that the entire distribution of SST index values for years classified as warm (cold) years lies above (below) the mean index value in the case of the rule based on the 2RC (−2RC) experiment. Also note that the interquartile ranges of the nonwarm (noncold) and warm (cold) distributions are clearly separated. The rule based on the 1982/83 experiment was less successful, but the warm classification still corresponds to above average values of the SST index.

Classifications made using the first ten EOFs in the near-tropical band as guess patterns (Fig. 5b) were very similar to those made with the observed anomaly patterns. In the case of rules based on the 2RC and −2RC experiments the classifications resulted in the same pleasing separation of distributions of SST index values.

Classifications made in the extratropical region using the observed guess patterns (Fig. 5c) were surprisingly successful. The 2RC rule misclassified only one year with a below normal SST index (1959) and correctly identified four warm events. The 1982/83 rule identified five of the six warm events, but made a gross error by also classifying the 1955 cold event as a warm event. The −2RC rule identified only two of four cold events and it misclassified three years with above normal SST index. These rules likely have a considerable amount of artificial skill because the dataset of anomaly

patterns that is classified includes patterns used in their derivation. This is underscored by the performance of the extratropical rules based on EOFs (Fig. 5d), which were much more conservative than those using the observed patterns.

6. Discussion

In this paper we have successfully extended the concept of recurrence analysis to multivariate climate comparison problems by using the tools of multiple discriminant analysis. We have described several estimators and tests of recurrence. The latter address hypotheses about differences in the *ensemble* means of two climates while the former gives an indication of how much may be learned about differences between the two climates from the available samples.

We have also described the assumptions that are implicit in the various procedures. The Gaussian assumption can be relaxed by employing nonparametric procedures or by applying a rank transformation prior to an analysis. On the other hand the independence assumption (independent samples and independent realizations within samples) is implicit in both parametric and nonparametric procedures and application of a rank transformation will not appreciably ameliorate problems due to lack of independence. The effect of the latter is to optimistically bias estimates and tests of recurrence.

The techniques were successfully applied to a set of previously analyzed El Niño sensitivity experiments conducted with the CCC GCM. The results of this analysis are consistent with but more easily interpreted than those obtained in the univariate recurrence analysis described by SZ. We have seen that it is possible to derive discrimination rules that can differentiate reliably between realizations from the CCC GCM control climate and from CCC GCM climates with anomalous boundary forcing, even when the response is considered in large tropical and extratropical regions.

It was possible to derive discrimination rules that could discriminate between warm (cold) and nonwarm (noncold) years with a considerable degree of reliability by using rather limited empirical knowledge from sensitivity experiments conducted with the CCC GCM. This satisfying result indicates that the GCM not only responds significantly to anomalous SSTs, but that it does so in a way that the observed extratropical Northern Hemisphere atmosphere apparently recognizes. Thus, we conclude that the sensitivity experiments simulate, in a recurrent way, at least some of the characteristics of a true El Niño.

The example illustrates that the simple DS estimator of recurrence will be adequate for many climate comparison problems provided that the dimensionality of the data has been suitably reduced; however, there are good reasons to use several parametric and nonparametric procedures and to analyze both the reduced data and its rank transformation. Inconsistency among a group of estimators or tests indicates that the assumptions required for some of these procedures are not being satisfied. When this is the case the estimator or test with the least restrictive set of assumptions should be given the greatest weight.

Acknowledgments. The authors gratefully acknowledge the support for this work provided by the Max-Planck-Institut für Meteorologie to Francis Zwiers for the purpose of facilitating this collaboration.

APPENDIX A

Description of the OS Error Rate Estimator

Okamoto (1963, 1968) described an asymptotic expansion for the probability of misclassifying an observation from the control population as an observation from the experimental population when both populations are Gaussian with common variance–covariance matrix, and when classification decisions are made with Fisher's linear discrimination statistic (8). This expansion is given by

$$\Pr\{e|c\} = \Phi(u) + \frac{a_1}{n_c} + \frac{a_2}{n_e} + \frac{a_3}{n} + \frac{b_{11}}{n_c^2} + \frac{b_{22}}{n_e^2} + \frac{b_{12}}{n_c n_e} + \frac{b_{13}}{n_c n_e} + \frac{b_{23}}{n_c n_e} + \frac{b_{33}}{n^2} + O(n^{-3}) \quad \text{(A1)}$$

where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function and

$$n = n_c + n_e - 2$$

$$a_1 = (d_0^4 + 3ld_0^2)/(2\nabla^2)$$

$$a_2 = [d_0^4 - (l-4)d_0^2]/(2\nabla^2)$$

$$a_3 = (l-1)d_0^2/2$$

$$b_{11} = [d_0^8 + 6(l+2)d_0^6 + (l+2)(9l+16)d_0^4 + 20l(l+2)d_0^2]/(8\nabla^4)$$

$$b_{22} = [d_0^8 - 2(l-10)d_0^6 + (l-6)(l-16)d_0^4 + 4(l-4)(l-6)d_0^2]/(8\nabla^4)$$

$$b_{12} = [d_0^8 + 2(l+8)d_0^6 - 3(l^2-10l-16)d_0^4 - 12l(l-6)d_0^2]/(4\nabla^4)$$

$$b_{13} = (l-1)[d_0^6 + 3(l+4)d_0^4 + 6(l+4)d_0^2]/(4\nabla^2)$$

$$b_{23} = (l-1)[d_0^6 - (l-8)d_0^4 - 2(l-4)d_0^2]/(4\nabla^2)$$

$$b_{33} = [l-1)((l+1)d_0^4 + 4ld_0^2]/8$$

and

$$d_0^2 = -u\phi(u)$$

$$d_0^4 = (3u - u^3)\phi(u)$$

$$d_0^6 = -u(15 - 10u^2 + u^4)\phi(u)$$

$$d_0^8 = (105u - 105u^3 + 21u^5 - u^7)\phi(u)$$

$$u = -\nabla/2$$

where $\phi(\cdot)$ is the standard Gaussian probability density function.

The OS-estimator of the misclassification error rate (Page 1985) is obtained by substituting the shrunken estimator

$$\mathrm{DS} = [[(n_e + n_c - l - 1)/(n_e + n_c - 2)] \times (\bar{x}_e - \bar{x}_c)^t S^{-1}(\bar{x}_e - \bar{x}_c)]^{1/2} \quad \text{(A2)}$$

for ∇ in the expressions above.

APPENDIX B

Description of the *M* Error Rate Estimator

McLachlan (1975) described the *M*-method estimator of the misclassification error which, like the OS estimator, is based on an asymptotic expansion for the probability of misclassifying. The *M* estimator is given by

$$\begin{aligned} M = {} & \Phi(-\mathrm{DS}/2) + \phi(-\mathrm{DS}/2)\{\{(l-1)/\mathrm{DS}\}/n_c \\ & + (\mathrm{DS}/32)\{4(4l-1) - \mathrm{DS}^2\}/n \\ & + (l-1)(l-2)/(4\mathrm{DS}n_c^2) \\ & + \{(l-1)/64\}\{-\mathrm{DS}^3 + 8(2l+1)\mathrm{DS} \\ & + (16/\mathrm{DS})\}/(n_c n) \\ & + (\mathrm{DS}/12288)\{3\mathrm{DS}^6 - 4(24l+7)\mathrm{DS}^4 \\ & + 16(48l^2 - 48l - 53)\mathrm{DS}^2 \\ & + 192(-8l+15)\}\}/n^2 \end{aligned} \quad \text{(B1)}$$

where Φ, ϕ, u, n and DS are as defined in appendix A. McLachlan also demonstrated that M is asymptotically approximately a Gaussian random variable with mean $\Pr\{e|c\}$ and variance $\sigma^2(\nabla^2)$. The latter is given by

$$\begin{aligned} \sigma^2(\nabla^2) = {} & \{\phi(-\nabla/2)^2\}\left\{\frac{1}{n_c} + (\nabla^2/8)/n \right. \\ & + \{\nabla^2 + 4(3l-4) + (l^2 - 4l + 5)(16/\nabla^2)\}/ \\ & \qquad (4n_c)^2 + \{(\nabla^2 - 2l)/8\}/(n_c n_e) \\ & + \{\nabla^4 + 2(11l - 16)\nabla^2 + 8(5l-4)\}/(64n_c n) \\ & + \{\nabla^4 - 2(l+4)\nabla^2 + 8l\}/(64n_e n) \\ & + \{2\nabla^6 + 16(2l-5)\nabla^4 - 32(4l-13)\nabla^2\}/ \\ & \qquad \left.(32n)^2\right\}. \end{aligned} \quad \text{(B2)}$$

APPENDIX C

Tiku's Approximation of the Noncentral *F*-Distribution

Tiku (1965, 1966) and Pearson and Hartley (1976) describe an accurate approximation for the distribution of a noncentral F random variable F' with noncentrality parameter δ^2 and with ν_1 and ν_2 degrees of freedom. The approximation, which requires the numerical evaluation of the incomplete beta function, is given by

$$\Pr\{F' > F\} \approx \frac{1}{B(\alpha, \beta)} \int_0^{y_0} x^{\alpha-1}(1-x)^{\beta-1} dx \quad \text{(C1)}$$

where $B(\alpha, \beta)$ is the beta function and

$$\alpha = \nu_2/2, \quad \beta = \nu_1'/2$$

$$y_0 = \left\{1 + \frac{\nu_1'}{\nu_2}\frac{(F+c)}{h}\right\}^{-1}$$

$$c = \frac{\nu_2}{(\nu_2 - 2)}\left\{h - \frac{(\nu_1 + \delta^2)}{\nu_1}\right\}$$

$$h = \frac{\nu_1'}{\nu_1}\frac{1}{(2\nu_1' + \nu_2 - 2)}\frac{H}{K}$$

$$\nu_1' = \frac{(\nu_2 - 2)}{2}\left\{\left(\frac{E}{E-4}\right)^{1/2} - 1\right\}$$

$$E = H^2/K^3$$

$$K = (\nu_1 + \delta^2)^2 + (\nu_2 - 2)(\nu_1 + 2\delta^2)$$

$$H = 2(\nu_1 + \delta^2)^3 + 3(\nu_1 + \delta^2)(\nu_1 + 2\delta^2)(\nu_2 - 2) + (\nu_1 + 3\delta^2)(\nu_2 - 2)^2.$$

REFERENCES

Anderson, T. W., 1984: *An Introduction to Multivariate Statistical Analysis.* Second ed. Wiley and Sons, 675 pp.

Boer, G. J., 1985: Modeling the Atmospheric Response to the 1982/83 El Niño. *Coupled Oceanic–Atmosphere Models,* J. C. J. Nihoul, Ed., Elsevier, 7–17.

——, N. A. McFarlane, R. Laprise, J. D. Henderson and J.-P. Blanchet, 1984a: The Canadian Climate Centre spectral atmospheric general circulation model. *Atmos. Ocean,* **22,** 397–429.

——, —— and ——, 1984b: The climatology of the Canadian Climate Centre general circulation model as obtained from a five-year simulation. *Atmosphere: Atmos. Ocean,* **22,** 430–475.

Chernick, M. R., V. K. Murthy and C. D. Nealy, 1985: Application of bootstrap and other resampling techniques: Evaluation of classifier performance. *Pattern Recognition Letters,* **3,** 167–178.

Conover, W. J., and R. L. Iman, 1980: The rank transformation as a method of discrimination with some examples. *Commun. Statist.-Theor. Meth.,* **A9,** 465–487.

Das Gupta, S., 1973: Theories and methods in classification: A review. *Discriminant Analysis and Applications,* T. Cacoullos, Ed., Academic Press, 77–138.

Efron, B., 1983: Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.,* **78,** 316–331.

Lachenbruch, P. A., 1967: An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics,* **23,** 639–645.

——, and M. R. Mickey, 1968: Estimation of error rates in discriminant analysis. *Technometrics,* **10,** 1–11.

MacLachan, G. J., 1975: Confidence intervals for the conditional probability of misclassification in discriminant analysis. *Biometrics,* **31,** 161–167.

——, 1976: The bias of the apparent error rate in discriminant analysis. *Biometrika,* **63,** 239–244.

Morrison, D. F., 1976: *Multivariate Statistical Methods,* second edition. McGraw-Hill, 415 pp.

Okamoto, M., 1963: An asymptotic expansion for the distribution of the linear discriminant function. *Annals of Math. Statist.,* **34,** 1286–1301.

——, 1968: Correction to "An asymptotic expansion for the distribution of the linear discriminant function." *Annals of Math. Statist.,* **39,** 1358–1359.

Page, J. T., 1985: Error-rate estimation in discriminant analysis. *Technometrics,* **27,** 189–198.

Pearson, E. S., and H. O. Hartley, 1976: *Biometrika Tables for Statisticians, Vol. II.* Biometrika Trust, 385 pp.

Penfield, D. A., and S. L. Koffler, 1985: Nonparametric discrimination. *Encyclopedia of Statistical Sciences, Vol. 6,* S. Kotz, N. L. Johnson and C. B. Read, Eds., Wiley and Sons, 324–328.

Rasmussen, E. M., and T. H. Carpenter, 1982: Variations in tropical sea surface temperature and surface wind fields associated with the Southern Oscillation/El Niño. *Mon. Wea. Rev.,* **110,** 354–378.

Snappin, S. M., and J. D. Knoke, 1984: Classification error rate estimators evaluated by unconditional mean squared error. *Technometrics,* **26,** 371–378.

——, and ——, 1985: An evaluation of smoothed classification error-rate estimators. *Technometrics,* **27,** 199–206.

Storch, H. von, 1987: A statistical comparison with observations of control and El Niño simulations using the NCAR CCM. *Beitr. Phys. Atmos.,* **60,** 464–471.

——, and F. W. Zwiers, 1988: Recurrence analysis of climate sensitivity experiments. *J. Climate,* **1,** 157–171.

Tiku, M. L., 1965: Laguerre series forms of non-central χ^2 and F-distributions. *Biometrika,* **52,** 415–427.

——, 1966: A note on approximating to the non-central F-distribution. *Biometrika,* **53,** 606–610.

Wright, P. B., 1984: Relationships between indices of the Southern Oscillation. *Mon. Wea. Rev.,* **112,** 1913–1919.