JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 93, NO. D9, PAGES 10,815–10,830, SEPTEMBER 20, 1988

The Principal Discriminant Method of Prediction: Theory and Evaluation

RUDOLPH W. PREISENDORFER[1]

Pacific Marine Environmental Laboratory, National Oceanic and Atmospheric Administration, Seattle, Washington

CURTIS D. MOBLEY

Joint Institute for the Study of the Atmosphere and Ocean, University of Washington, Seattle

TIM P. BARNETT

Climate Research Group, Scripps Institute of Oceanography, La Jolla, California

The Principal Discriminant Method (PDM) of prediction employs a novel combination of principal component analysis and statistical discriminant analysis. Discriminant analysis is based on the construction of discrete category subsets of predictor values in a multidimensional predictor space. A category subset contains those predictor values which give rise to a predictand (or observation) in that particular category. A new predictor value is then assigned to a particular category (i.e., a forecast is made) through the use of probability distribution functions which have been fitted to the category subsets. The PDM uses principal component analysis to define the multidimensional probability distribution functions associated with the category subsets. Because of its underlying discriminant nature the PDM is also applicable to problems in data classification. The PDM is applied to prediction problems using both artificial and actual data sets. When applied to artificial data the PDM shows forecast skills which are comparable to those of standard forecast techniques, such as linear regression and classical discriminant analysis. When applied to actual data in a forecast of the 1982–1983 El Niño, the PDM performed poorly. However, in forecasting winter air temperatures over North America, the PDM proved superior to other forecast techniques, after suitably filtering or smoothing the raw data in order to improve the signal-to-noise ratio. It is expected that the PDM will show its greatest advantage over other forecast techniques when the relation between predictors and predictand is nonlinear.

1. INTRODUCTION

Discriminant methods in general, and the Principal Discriminant Method (PDM) in particular, can be applied to forecasting problems in which it is desired to forecast a discrete state of the atmosphere or ocean. An example is the forecasting of seasonal temperatures as one of the three discrete states "above average," "average," or "below average." Because of its underlying discriminant nature the PDM can also be used in data classification. An example is the assignment of the observed state of the atmosphere to one of several discrete "climate types." A further application of the PDM is the linking of the output of a general circulation model (GCM) of the atmosphere with observed fields in order to produce model-output statistic (MOS) schemes of prediction. Our description of the PDM shows its essential form, so as to facilitate applications to any of the problems just mentioned.

The successful construction of category subsets in a multidimensional predictor space is a sine qua non of any discriminant method, along with the fitting of versatile probability density function (pdf's) to these subsets. The modifier "principal" in the name of the present method derives from the fact that for multiple predictors, essential use is made of principal component analysis (PCA) in order to determine appropriate probability density functions for the category subsets. This is the major difference between the PDM and standard discriminant analyses. In effect, it allows for irregular distribution of prediction data that consequently do not fit well-known pdf's, these pdf's being the heart of any discriminant method.

Another unique feature of the PDM is that of self-evaluation of predictive skill. This is supplied by three indices of skill: the potential predictability, the potential 0-class error and the potential 1-class error in the predictand categories. These indices along with their critical values, supplied by a Monte Carlo technique, help the user to decide how much confidence to place on a given prediction made by the PDM. Also, during the construction of the PDM's working parts, provision is made to test the method on an independent data set. This testing gives another indication of how well a data set is constituted to allow predictions of its variables' future states.

The exposition of the PDM, which is the main goal of this paper, will be made in two parts. The first part (section 2) treats the case of a single predictor, in which case the PDM reduces to a classical discriminant method. In real applications the single-predictor mode can yield much information about the potential predictability of a predictand by a given predictor, along with some information about the skill of the predictions. The single-predictor mode of the PDM can therefore stand as an independent, preliminary prediction method. The second part (section 3) treats the case of multiple predictors. It is expected that the predictability will increase when a single predictor is joined by several more predictors and when

[1]Deceased September 16, 1986.

Paper number 8D0169.
0148-0227/88/008D-0169$05.00

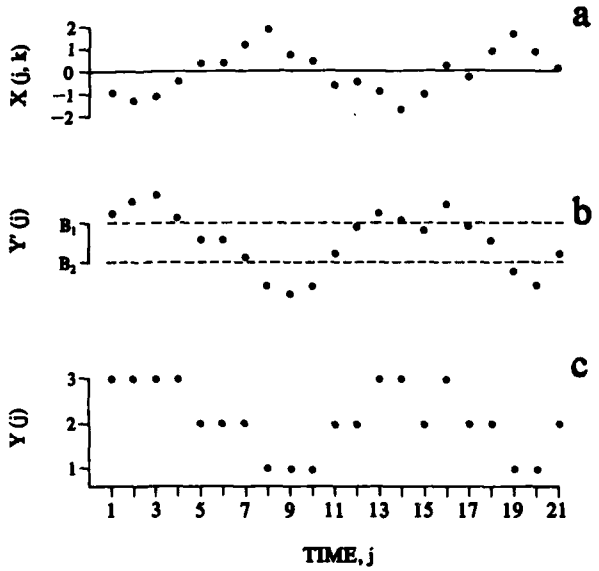

Fig. 1. Illustration of a predictor-predictand pair and a tercile categorization. (*a*) A standardized predictor time series $X(j, k)$, $j = 1, \cdots, N = 21$, where k is fixed. (*b*) The corresponding time series of the predictand values, $Y'(j)$; boundary values B_1 and B_2 are indicated. (*c*) The terciled values of the predictand, $Y(j)$.

the category subsets in the resultant multidimensional predictor space can be carved out of the swarm of data points there. It is in this mode that the PDM realizes its full power, via its application of principal component analysis to the multidimensional swarm of data points.

Section 4 discusses the results of using the PDM in various forecast situations. This rather brief discussion is intended to highlight some of the strengths and weaknesses of the PDM, a goal in concert with the theoretical nature of the rest of the paper.

This paper is a condensation of a technical memorandum [*Preisendorfer et al.* 1987], which can be consulted for a more detailed presentation, especially of the results discussed in section 4. The reader desiring an elementary discussion of discriminant analysis in its conventional statistical formulation is referred to *Lachenbruch* [1975]. Applications of the discriminant method in climate forecasting are given by *Harnack et al.* [1985]. For a discussion of principal component analysis, see Preisendorfer [1988].

2. THE SINGLE-PREDICTOR STAGE

It is assumed that we have available a data set consisting of simultaneous observations of both predictors and predictands. Such a data set is required in order to construct the PDM model. After the model has been constructed, it is capable of making forecasts when given new predictor values.

2.1. *The Predictor-Predictand Pair*

Let $X(j, k)$ denote the value of the kth predictor X at time j. It is convenient to standardize the predictor in time, so that the time series $X(j, k)$, $j = 1, 2, \cdots, N_0$, has zero mean and unit variance for each k, $k = 1, 2, \cdots, K$. Let $Y'(j)$ denote the value of the predictand Y' at the same time j. For example, in a model-output statistic setting, the various predictors $X(j, k)$ might be the sea surface temperature ($k = 1$), the sea level pressure ($k = 2$), the relative humidity ($k = 3$), etc., all at the same spatial location, and a particular predictand $Y'(j)$ might be the horizontal visibility at the same time j and at the same or a different location.

In order to use the predictive capabilities of the PDM, we introduce a time lag τ into $Y'(j)$, so as to pair $Y'(j + \tau)$ with $X(j, k)$, $\tau \geq 0$. For simplicity it will be assumed that τ has been introduced into $Y'(j)$, and we will retain the notation $X(j, k)$ and $Y'(j)$ for the lagged predictor-predictand pair, where now $j = 1, 2, \cdots, N$ labels the common ranges of times of the lagged pair. Hereafter it will be assumed that each predictor-predictand datum pair is statistically independent from other members of the data set. This condition can be tested and the original data suitably redefined to ensure independence if necessary. Several of the methods to be discussed later require this property of the data.

2.2. *Q-tiling the Predictand*

Divide the range of predictand values $\{Y'(j): j = 1, \cdots, N\}$ into Q intervals. By judicious choice of the boundary values $B_1, B_2, \cdots, B_{Q-1}$ between these intervals, we can "Q-tile" the predictand $Y'(j)$ into Q discrete categories. Let $Y(j)$ denote the value of the discrete category to which $Y'(j)$ belongs; thus $Y(j) = q$ if $Y'(j)$ falls into category q, $1 \leq q \leq Q$. Figure 1 illustrates these ideas for the case of $Q = 3$, called a "tercile categorization." In Figure 1 we define $Y(j)$ as follows:

$$Y(j) \equiv 1 \quad \text{if } Y'(j) < B_1$$

$$Y(j) \equiv 2 \quad \text{if } B_1 \leq Y'(j) < B_2$$

$$Y(j) \equiv 3 \quad \text{if } B_2 \leq Y'(j)$$

for $j = 1, \cdots, N$. There is no requirement that the boundary values be equally spaced or that the Q categories be equally populated after the Q-tiling of the predictand.

2.3. *The Discriminant Set*

The time series for the kth predictor $X(j, k)$ (Figure 1*a*) and the Q-tiled predictand $Y(j)$ (Figure 1*c*) can be combined to form a single diagram, called the discriminant diagram. Figure 2 shows the discriminant diagram corresponding to Figure 1. In this example one sees at a glance that large, positive predictor values tend to be associated with terciled predictand values in category 1, predictor values near zero are associated with category 2 predictand values, and large, negative predictor values tend to correspond to predictand values in category 3. The discriminant set consists of the N points $[X(j, k), Y(j)]$, $j = 1, 2, \cdots, N$.

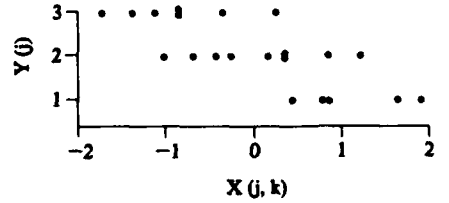

Fig. 2. The discriminant diagram corresponding to Figure 1*a* and Figure 1*c*, where k is held fixed as j runs from 1 to $N = 21$.

2.4. *Training and Testing Sets*

The discriminant set of N points is randomly split into two subsets of predetermined sizes, N_{tr} and N_{ts}. The subset containing N_{tr} points is the training set, and the subset containing N_{ts} points is the testing set. Typically, we choose $N_{tr} = 2N_{ts}$, so that two thirds of the N available data points can be used to "train," or to construct, the PDM; one third of the points can be used to "test," or to score, the PDM. Figure 3 shows a possible partition of the points of Figure 2 into training and testing sets. Let $X_{tr}(i, k)$, $i = 1, 2, \cdots, N_{tr}$, denote those values of $X(j, k)$ which fall into the training set. Likewise, let $Y_{tr}(i)$, $i = 1, \cdots, N_{tr}$, denote the corresponding values of $Y(j)$. Those points of the discriminant set which have been randomly assigned to the testing set are denoted by $[X_{ts}(i, k), Y_{ts}(i)]$, $i = 1, 2, \cdots, N_{ts}$. In order to fully utilize the training–testing set partition philosophy, it is necessary that the "training" data be statistically independent from the "testing" data. This is a critical factor in our procedure, and henceforth we assume that independence has been established (compare section 2.1).

2.5. *Category Subsets of Predictor Space*

The subset of predictor points in the training set which is associated with category q of predictand values is termed the qth category subset of the predictor space, denoted by C_q, $q = 1, 2, \cdots, Q$, whose elements are $C_q(i)$, $i = 1, 2, \cdots, M_q$. Figure 3 shows the three category subsets for the illustrated training set: C_1 with $M_1 = 3$, C_2 with $M_2 = 6$, and C_3 with $M_3 = 5$. The category subsets form the heart of the discriminant structure of the PDM.

2.6. *Fitting the Probability Density Functions*

Once the category subsets of predictor points have been obtained, any discriminant method, including the PDM, requires the fitting of probability density functions to these category subsets. A decisive point in the discriminant method can arise when choosing the specific form of the probability density function to be fitted to the category subsets. We choose the Gaussian distribution for this exposition, although it may be worthwhile in other applications to use a pdf specifically tailored to a given data set. The form of the Gaussian pdf for

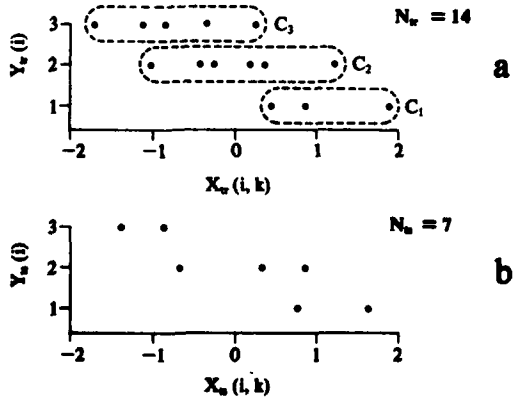

Fig. 3. A partitioning of the discriminant set shown in Figure 2 into (*a*) a training set and (*b*) a testing set. The category subsets of the training set are indicated.

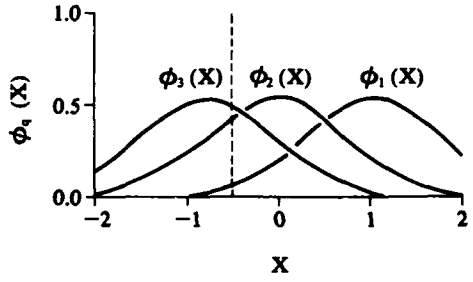

Fig. 4. The pdf's $\phi_1(X)$, $\phi_2(X)$, and $\phi_3(X)$ for the category subsets of Figure 3*a*.

category q is

$$\phi_q(X) = (2\pi\sigma_q^2)^{-1/2} \exp\left[-\frac{(X - \bar{X}_q)^2}{2\sigma_q^2}\right]$$

where $\bar{X}_q$ is the average over i of the qth category $\{C_q(i): i = 1, \cdots, M_q\}$ and σ_q^2 is the variance of this set of points.

Note that although the original data set $X(j, k)$, $j = 1, \cdots, N$, was standardized to zero mean and unit variance, the category subsets C_q in general have nonzero means and nonunit variances. Figure 4 shows the fitted Gaussian pdf's, $\phi_1(X)$, $\phi_2(X)$, and $\phi_3(X)$, for the category subsets of Figure 3*a*. Once the $\phi_q(X)$, $q = 1, \cdots, Q$, have been determined, the construction (or training) of the single-predictor PDM model is complete. Observe that implicit in the $\phi_q(X)$ is the fact that they were constructed for a particular realization of the training set. A different partition of the discriminant set into training and testing sets would yield somewhat different $\phi_q(X)$ functions.

2.7. *Making a Prediction*

Suppose a new predictor realization X' occurs for predictor k; i.e., $X' = X(j, k)$ for some time j. We wish to use the PDM model constructed earlier in order to make a predictand forecast for the new predictor value X'. Various strategies can be adopted regarding the manner in which the pdf's $\phi_q(X)$ are employed in making a forecast. Two of the more obvious are discussed in the following subsections.

2.7.1. *Maximum probability strategy.* Given a predictor value X', we compute $\phi_q(X')$ for each category $q = 1, \cdots, Q$ and note which q value, call it q', has the maximum pdf value. The prediction is then that $Y(j) = q'$. In Figure 4 we see, for example, that $X' = -0.5$ would yield a prediction of Y in category 3, $X' = 0.0$ would predict $Y = 2$, and so on.

2.7.2. *Bayesian strategy.* The maximum probability strategy is easily interpreted and computationally simple; however, it may not make the best use of the available information. The method of Bayesian inference is perhaps better suited to the problem at hand.

Strictly speaking, the $\phi_q(X)$ pdf's relate to conditional probabilities: namely, $\phi_q(X)$ gives the pdf of X, given that category q is observed. To fix this idea, let us write $\phi(X \mid q) = \phi_q(X)$. What we really need in order to make a forecast is the probability that category q occurs, given that a specific value of X occurs; let us denote this by $P(q \mid X)$. The category, call it q', with the greatest probability $P(q \mid X)$ for the given value of $X = X'$ is then the category forecasted by the PDM when X' is observed. Since the Q predictand categories are mutually exclusive and exhaustive, Bayes' theorem (see, for example,

Box and Tiao [1972, p. 10])

$$P(q \mid X) = P(X \mid q)P(q)\left[\sum_{q=1}^{Q} P(X \mid q)P(q)\right]^{-1}$$

$$= \phi(X \mid q)P(q)\left[\sum_{q=1}^{Q} \phi(X \mid q)P(q)\right]^{-1}$$

can be used to obtain the desired $P(q \mid X)$ values. Here $P(X \mid q)$ is the probability of X given q, which is just $\phi(X \mid q)$. $P(q)$, known as the a priori probability of category q occurring, lies at the heart of Bayesian inference. $P(q)$ is a measure of our knowledge about what forecast category will occur before the predictor value X is obtained. The selection of appropriate $P(q)$ values is a task which falls on the user of the Bayesian strategy and is an extra computation above those required for the maximum probability strategy.

If we were making a random forecast of $Y(j)$, it would be reasonable (but not necessary) to make the probability of randomly choosing category q proportional to the number of points of the training set which fall in category q. So a reasonable choice of $P(q)$ is

$$P(q) \equiv \frac{M_q}{N_{tr}} \qquad q = 1, \cdots, Q$$

It should be understood that in making this choice of $P(q)$ we are allowing information about the relative distribution of points in the category subsets to influence the PDM's forecast of the predictand when given a new predictor value X'. This is the whole point of the Bayesian strategy. Another choice of $P(q)$ could lead to an entirely different forecast being made for the same X' value. If we wish to make no use of our knowledge about the distribution of points in the category subsets, we can pick $P(q) = 1/Q$ for all q. This is the case of equally likely a priori distributions, for which the Bayesian strategy reduces to the maximum probability strategy.

2.8. *Potential Predictability*

The PDM as it now stands is ready to make predictions by whichever strategy is chosen in the previous paragraph. However, it is of great interest also to compute some measure of confidence in these predictions, i.e., to ascertain the expected forecast skill of the PDM. When the pdf's $\phi_q(X)$ are not well separated, then the predictions have low skill, no matter what prediction strategy we choose. Note, for example, in Figure 4 that for predictor values X' near 0.5 it is nearly equally probable that the predictand is in category 1 or 2, if we use the maximum probability strategy. Conversely, if the $\phi_q(X)$ are well separated, then the PDM has no difficulty in determining which pdf has the maximum value for a given X', and we have greater confidence that the predictions will be correct. Therefore a measure of our confidence in the predictions can be obtained via a measure of how well separated are the pdf's. One measure of this separation is given by the potential predictability index (PP). Note that this index is distinctly different from prior uses of "potential predictability" in the literature, for example, *Madden and Shea* [1978].

First define

$$P'(i, q) \equiv \phi_q[X_{tr}(i, k)]\left\{\sum_{q=1}^{Q} \phi_q[X_{tr}(i, k)]\right\}^{-1}$$

for $i = 1, \cdots, N_{tr}$, where $q = 1, \cdots, Q$, and k is held fixed. Note that

$$\sum_{q=1}^{Q} P'(i, q) = 1$$

If the pdf's are identical, $P'(i, q) = 1/Q$. Thus a measure of how far the pdf's are from being identical is

$$\sum_{q=1}^{Q}\left[P'(i, q) - \frac{1}{Q}\right]^2$$

Moreover, if the pdf's are perfectly separated, then

$$\sum_{q=1}^{Q}\left[P'(i, q) - \frac{1}{Q}\right]^2$$

$$= \left(1 - \frac{1}{Q}\right)^2 + \left(0 - \frac{1}{Q}\right)^2 + \cdots + \left(0 - \frac{1}{Q}\right)^2$$

where the first term on the right-hand side of the equation results from the single occurrence of $P'(i, q) = 1$, in the sum, and the remaining terms on the right-hand side, $Q - 1$ in number, result from $P'(i, q) = 0$. Therefore

$$\sum_{q=1}^{Q}\left[P'(i, q) - \frac{1}{Q}\right]^2 = \frac{Q - 1}{Q}$$

Thus we are led to define

$$\mathrm{PP}(i) \equiv \frac{Q}{Q - 1}\sum_{q=1}^{Q}\left[P'(i, q) - \frac{1}{Q}\right]^2$$

Clearly, $\mathrm{PP}(i) = 1$ if the pdf's are perfectly separated and $\mathrm{PP}(i) = 0$ if the pdf's are identical. Finally, we define the potential predictability, PP, as

$$\mathrm{PP} \equiv \frac{1}{N_{tr}}\sum_{i=1}^{N_{tr}} \mathrm{PP}(i)$$

Thus PP has the property $0 \le \mathrm{PP} \le 1$ and is a measure of how distinct the pdf's are: PP approaches zero as the pdf's become identical (and our confidence in a prediction decreases), and PP approaches 1 as the pdf's become widely separated (and our confidence in a prediction increases). This definition for PP is consistent with the choice of the maximum probability strategy for making a forecast, as discussed in 2.7.1. If the Bayesian strategy of section 2.7.2 is chosen, the definition must be modified slightly by using

$$P'(i, q) \equiv P[q \mid X = X_{tr}(i, k)]$$

$$= \phi_q[X_{tr}(i, k)]P(q)\left\{\sum_{q=1}^{Q} \phi_q[X_{tr}(i, k)]P(q)\right\}^{-1}$$

which reduces to the previous definition of $P'(i, q)$ if the a priori distributions $P(q)$ are chosen to be equally likely, i.e., $P(q) = 1/Q$.

PP is implicitly indexed by k for the particular predictor $X(j, k)$ in question. Moreover, PP depends on the particular partition of the discriminant set into training and testing sets. Thus one should make several (say Ω) random partitions of the discriminant set and compute PP for each. Then, in the final tally the average PP (AVGPP) over all partitions should be taken:

$$\mathrm{AVGPP}(k) = \frac{1}{\Omega}\sum_{\omega=1}^{\Omega} \mathrm{PP}(k, \omega)$$

where we now explicitly show the predictor (k) and partition

1-class error score. Then define

$$a_0 \equiv \frac{1}{N_{ts}} \quad \text{[number of 0-class errors]}$$

$$a_1 \equiv \frac{1}{N_{ts}} \quad \text{[number of 1-class errors]}$$

clearly, a_0 and a_1 satisfy

$$0 \le a_0 \le 1$$

$$0 \le a_1 \le 1$$

The larger a_0 is, the better the PDM has forecasted the testing set values, and the smaller a_1 is, the better the PDM has performed. Unlike PP, $\hat{a}_0$, and $\hat{a}_1$, which are based on the fitted pdf's defining the PDM model, a_0 and a_1 are actual forecast scores made by the PDM when applied to an independent testing set. Our studies of the PDM in section 4 will make use of the training and testing sets in the manner just discussed: the PDM will be defined using the training set, and its performance will then be evaluated using the testing set. The a_0 and a_1 scores are a convenient means of presenting forecast skill when discrete forecast categories are used. (See, for example, *Preisendorfer and Mobley* [1984] for the use of a_0 and a_1 in scoring seasonal climate forecasts.)

2.11. *Significance Tests for Class Errors*

The Monte Carlo procedure, used in section 2.9 to determine the 5% critical value for potential predictability, is equally applicable to the determination of critical values for $\hat{a}_0$, $\hat{a}_1$, a_0, and a_1. For each of the 100 realizations of the random data set R, we can compute $\hat{a}_0$ and $\hat{a}_1$ from the associated training set, and we can compute a_0 and a_1 scores from the associated testing set. We then determine the 5% upper critical levels, $\hat{a}_0(96)$ and $a_0(96)$, and the 5% lower critical values, $\hat{a}_1(05)$ and $a_1(05)$. Significantly good predictions will have $\hat{a}_0$ and a_0 scores that equal or exceed $\hat{a}_0(96)$ and $a_0(96)$, respectively. Significantly good predictions will have $\hat{a}_1$ and a_1 scores that equal or are less than $\hat{a}_1(05)$ and $a_1(05)$, respectively. Note that when more than one predictor is considered (section 3.1), estimation of significance level becomes more complicated.

2.12. *Ranking and Screening Single Predictors*

The net result of this section is the ability to individually rank (for a given predictand $Y(j)$) the predictors $X(j, k)$, $k = 1, \cdots, K$, in terms of their PP, $\hat{a}_0$, $\hat{a}_1$, a_0, and a_1 scores. Those predictors that have significant potential predictability and class-error scores become candidates for further consideration in the multiple-predictor stage. Predictors that have nonsignificant scores as single-predictors of a predictand are unlikely to add useful information if they are combined with other predictors in the multiple-predictor stage, and they therefore can be dropped from further consideration. It is important to remember here that as one considers more and more predictors, the probability of finding an apparently "good one" by chance increases. In fact, if one considers K different predictors, then an appropriate 5% critical level for any single predictor is $(0.05)^K$. Parsimony is obviously called for in the original definition of the predictor pool.

There are obviously other methods of ranking the predictors than those outlined here. Multiple-correlation analysis could be used in place of the simple correlation described earlier to affect the ranking. Similarly, a redefinition of the predictors in terms of their principal components and subsequent ranking by eigenvalue size represents a very different approach to predictor ordering (compare section 4.2). Whatever method one uses, it is necessary to avoid a large predictor pool or risk the chance of obtaining false results.

3. THE MULTIPLE-PREDICTOR STAGE

After performing the single-predictor, ordinary discriminant analyses of section 2 on each predictor $X(j, k)$, $k = 1, \cdots, K$, we have, for a fixed predictand $Y(j)$, a set of predictors ordered by their potential predictability scores. We drop from further consideration any predictors which did not have statistically significant PP scores in the single-predictor stage, so that $K_s \le K$ predictors remain. We now turn our attention to the task of constructing the PDM model in its multivariate setting.

We choose the predictor with the highest potential predictability score as the first predictor to be included in the multiple-predictor PDM model. We then must screen the remaining $K_s - 1$ predictors in order to select those which, when combined with the first predictor, yield a multiple-predictor model which is, in some sense, optimum.

3.1. *Correlational Screening of Predictors*

Suppose we have already selected $L - 1$ predictors, $L = 2, \cdots, K_s - 1$. Let these selected predictors be $X(j, l)$, $l = 1, \cdots, L - 1$. Let the remaining set of unselected predictors be denoted by $W(j, u)$, $u = 1, \cdots, U$; $U + L - 1 = K_s$. Let $\rho[u, l]$ denote the correlation between the indicated predictors. The number

$$\rho_{\max}(u) \equiv \text{Max } \{|\rho[u, l]|\} \qquad l = 1, \cdots, L - 1$$

is a measure of the distance between the uth unselected predictor $W(j, u)$ and the set of $L - 1$ previously selected predictors $X(j, l)$. The larger $\rho_{\max}(u)$ is, the closer $W(j, u)$ is to $\{X(j, l), l = 1, \cdots, L - 1\}$ as a whole.

When choosing a new candidate predictor for addition to the previously selected predictors, we choose that predictor $W(j, u)$ which has the minimum correlation magnitude, $\rho_{\max}(u)$. In so doing, we are selecting that predictor which is least correlated with the existing predictors and therefore most likely to add new information to the model. If u' is the value of u giving the minimum $\rho_{\max}(u)$, then we set $X(j, L) = W(j, u')$, $j = 1, \cdots, N$. This correlational screening is the first step in choosing the Lth predictor. Whether or not this candidate predictor is retained in the PDM model will depend on its effect on the PP, $\hat{a}_0$, and $\hat{a}_1$ scores, to be discussed in section 3.9.

3.2. *The L-Dimensional Discriminant Set and Related Subsets*

Having added a candidate Lth predictor, we now have a set of L predictors which at each time j form a vector $\mathbf{X}(j) \equiv [X(j, 1), X(j, 2), \cdots, X(j, L)]$ in euclidean L-space E_L. As the time index j varies, $\mathbf{X}(j)$ moves about in E_L. The category-valued predictand $Y(j)$ concurrently changes with j. The set of all ordered pairs $[\mathbf{X}(j), Y(j)]$, $j = 1, \cdots, N$, constitutes the L-dimensional discriminant set.

The L-dimensional discriminant set is randomly split into two parts, exactly as in section 2.4. The result is a set of

(ω) indices. When comparing two possible predictors for a given predictand, the one with the higher AVGPP will represent the higher predictability, on average.

2.9. *Monte Carlo Significance Test for PP*

While one predictor may have a higher potential predictability than another, for a given predictand, it is possible that neither is significant in the statistical sense. This possibility can be tested via a Monte Carlo approach. Let a random number generator choose a class q at each time j; define a new array $R(j) = q$, $j = 1, \cdots, N$, and replace Y by R (a random version of Y). The probability of randomly assigning a particular q value to $R(j)$ should be made proportional to the relative frequency of occurrence of the qth category in the Q-tiling of the original data set, so that the Monte Carlo test will simulate as closely as possible the real experiment.

We can now use the given predictor set $X(j, k)$ and the newly defined random predictand $R(j)$ to produce training and testing sets, as in section 2.4, and to carry through all the subsequent steps to obtain a value of PP. This entire process can then be repeated, after generating a new realization of the random predictand R, to obtain another value of PP for a random relation between predictor and predictand. This process can be repeated to generate, say, 100 values of PP for random predictor-predictand connections. These 100 values can be ordered from smallest to largest; call them PP(1) for the smallest to PP(100) for the largest. The 5% critical value for PP is then determined from the ninety-sixth smallest PP value, PP(96). Thus the probability that a randomly produced PP value will equal or exceed PP(96) is approximately 0.05. Therefore if the PP value determined for the actual predictor-predictand pair satisfies PP $\geq$ PP(96), we will say that PP is significant at the 5% level.

If one wants to establish a critical value for AVGPP(k), then the Monte Carlo simulation is conducted so as to mimic the generation of AVGPP(k), as described in section 2.8. Thus one randomly produces Ω realizations of PP(k, ω), finds their average, and goes through this average-finding procedure 100 times in all. The ninety-sixth smallest randomly generated AVGPP value then gives the 5% critical value for AVGPP.

We note also that there are other measures of separation of the category swarms. For example, Hotelling's T^2 test (the multivariate generalization of Student's t test) can be used to test the significant separation of a pair of category means $\bar{X}_q$. However, such tests often depend on assumptions of normality or independence of events. The potential predictability measure of separation was developed in an attempt to have a nonparametric test.

2.10. *Class Errors*

The potential predictability gives us one measure of how well a particular predictor can be expected to forecast predictand values. Another straightforward indicator of how well a prediction method is doing, when predicting categories, is to count the number of predictions that are correct (0-class errors) and the number of predictions that are off by one category (1-class errors). In the PDM we shall do this two ways: we will determine the potential 0- and 1-class errors, $\hat{a}_0$ and $\hat{a}_1$ respectively, using the training set, and we will determine the actual 0- and 1-class errors, a_0 and a_1, using the testing set.

2.10.1. *Potential errors*: $\hat{a}_0$ and $\hat{a}_1$. Recall the probabilities $P'(i, q)$, which were defined when developing the PP index (using either the maximum probability or Bayesian strategies). For each i value, find the maximum of the Q probabilities, $\{P'(i, q): q = 1, \cdots, Q\}$ and let $q'(i)$ be the q value for which $P'(i, q)$ is a maximum. We now define the potential 0-class error as

$$\hat{a}_0 \equiv \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} P'[i, q'(i)]$$

Note that as the pdf's $\phi_q(X)$ become well separated, $P'[i, q'(i)]$, and consequently $\hat{a}_0$, approach 1. As the pdf's become identical, $P'[i, q'(i)]$ and $\hat{a}_0$ approach $P(q')$, which for the Bayesian case is $1/Q$. Therefore $\hat{a}_0$ is another measure, based on the pdf's $\phi_q(X)$, of how confidently we can expect the PDM to make a correct category forecast.

But even if the PDM makes an incorrect forecast, it is clearly better to have a forecast that misses by only one category than to have a forecast that misses by two or more categories. For example, if category 1 is observed, a forecast of category 2 is closer to the truth than is a forecast of category 3. Thus it is useful to have a measure of how likely it is that the PDM will err by only one category, if it indeed makes an incorrect forecast. Toward this end, we define

$$\hat{P}(i, 1) \equiv 0$$

$$\hat{P}(i, 2) \equiv P'(i, 1)$$

$$\hat{P}(i, 3) \equiv P'(i, 2)$$

$$\vdots$$

$$\hat{P}(i, Q + 1) \equiv P'(i, Q)$$

$$\hat{P}(i, Q + 2) \equiv 0$$

The idea here is to have $P'[i, q'(i) - 1] = 0$ if $q'(i) = 1$ and $P'[i, q'(i) + 1] = 0$ if $q'(i) = Q$. Then define

$$\hat{a}_1 \equiv \frac{1}{2} \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \{\hat{P}[i, q'(i)] + \hat{P}[i, q'(i) + 2]\}$$

A moment's reflection shows that $\hat{a}_1$ is a measure of the probability that a category one less or one greater than the correct forecast category will be selected, if indeed the $q'(i)$ value gives a false forecast. As the pdf's $\phi_q(X)$ become well separated, $\hat{a}_1$ approaches 0; as the pdf's become identical, $\hat{a}_1$ approaches $1/Q$. Thus we have

$$0 \leq \hat{a}_1 \leq \frac{1}{Q} \leq \hat{a}_0 \leq 1$$

The larger $\hat{a}_0$ is, the better $X(j, k)$ may predict $Y(j)$, and the smaller $\hat{a}_1$ is, the better $X(j, k)$ may predict $Y(j)$.

2.10.2. *Actual errors: a_0 and a_1.* After the PDM has been constructed, or trained, using the training set $[X_{tr}(i, k), Y_{tr}(i)]$, we can apply the PDM to the testing set predictors, $X_{ts}(i, k)$, and can verify the predictions it makes against the actual observations for the testing set, $Y_{ts}(i)$. It is again crucial that the members of the testing set be statistically independent from the training set. Each time the PDM makes a correct forecast, we tally one to the 0-class error score, and each time the PDM forecast errs by one category, we tally one to the

L-component vectors $\mathbf{X}_{tr}(i)$, $i = 1, \cdots, N_{tr}$, containing those elements of $\mathbf{X}(j)$ randomly falling into the training set, and another set of vectors $\mathbf{X}_{ts}(i)$, $i = 1, \cdots, N_{ts}$, containing the remaining elements of $\mathbf{X}(j)$. The associated sets of predictands $Y_{tr}(j)$ and $Y_{ts}(j)$ are defined just as before.

We can now define subsets of E_L, the setting of the predictor space, that are associated with each of the Q predictand categories. The logic of this definition is the same as that of section 2.5. Thus we set $\mathbf{C}_q(i) = \mathbf{X}_{tr}(i)$ if $Y_{tr}(i) = q$; the number of points tallied to $\mathbf{C}_q(i)$ is M_q.

It is to the subsets $\mathbf{C}_q$, $q = 1, \cdots, Q$, of E_L that we will eventually fit L-dimensional probability density functions. However, before fitting the pdf's, we perform a preliminary analysis of the L-dimensional category subsets using principal component analysis (PCA). It is in this application of PCA that the PDM parts company with classical discriminant analysis.

3.3. *Binary PCA Decomposition of Category Subsets*

Let us consider, for didactic purposes, the case of two predictors ($L = 2$) and a terciled predictand ($Q = 3$). Figure 5 shows three swarms of (artificially generated) points in E_2, representing the three category subsets. In classical discriminant analysis, each category subset would be fitted with a bivariate normal pdf. For a point swarm shaped like that of category 2, the bivariate normal pdf would probably be quite satisfactory; Figure 6 shows the category 2 swarm and the best fit binormal pdf. However, for an irregularly shaped swarm, such as category 1 of Figure 5, the bivariate normal pdf is clearly a poor representation of the actual shape of the category subsets. Figure 7 shows the best fit bivariate normal pdf for category 1. Since discriminant methods depend upon having pdf's which accurately delineate the category subsets, we could not expect accurate forecasts from a model based on fits as poor as that of Figure 7, and standard discriminant analysis will fail.

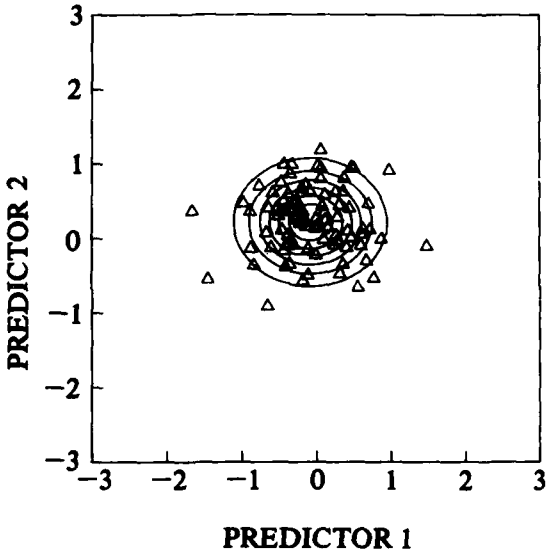

Fig. 6. The category 2 point swarm of Figure 5 and the probability contours of the best fit bivariate normal pdf.

Principal component analysis enables us to systematically and objectively subdivide an arbitrarily shaped category swarm into a number of smaller point swarms in E_L. If each of the smaller swarms is then roughly elliptical in shape (in terms of hyperellipses in E_L), then a multinormal pdf can be well fitted to each smaller swarm. The critical need for parsimony in this subdivision process should be kept in mind as the reader proceeds through the next several sections. The pdf describing the original, irregularly shaped category swarm can

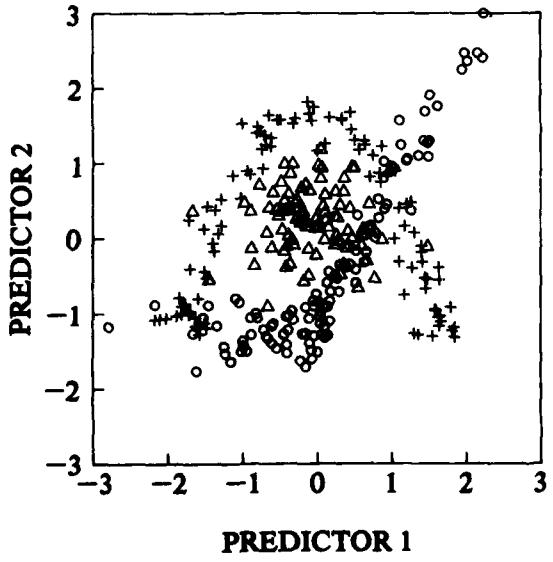

Fig. 5. An illustration of three category swarms $\mathbf{C}_1$, (pluses; $M_1 = 99$ points), $\mathbf{C}_2$ (triangles; $M_2 = 89$), and $\mathbf{C}_3$ (circles; $M_3 = 112$) in E_2.

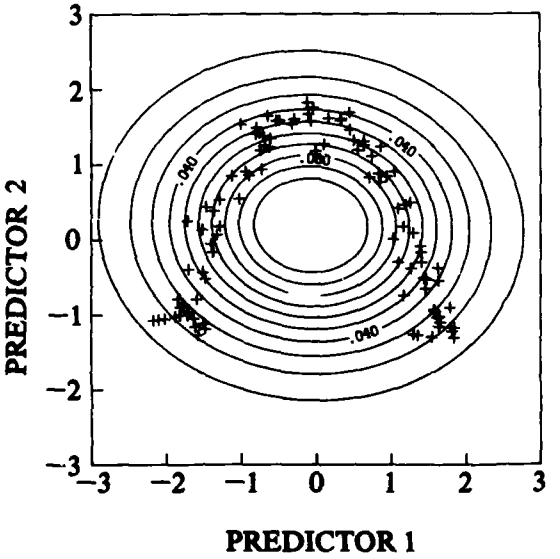

Fig. 7. The category 1 point swarm of Figure 5 and the probability contours of the best fit bivariate normal pdf.

both computationally expensive and overly strenuous, in the sense that category swarms are subdivided just because they are nonspherical. Swarms can deviate greatly from a spherical shape and can still be adequately fit by a multivariate normal pdf; it is the sinuous shapes (compare Figure 8) and multimodal (or clustered) point swarms that need to be decomposed.

3.4.2. *Strategy 2.* One simple way to terminate the PCA subdivision process is to simply force all initial category swarms $\mathbf{X}_q$ to undergo a fixed number of subdivisions, say, to level 2, as shown in Figure 9. This procedure seems to work fairly well in practice, although it should not be applied blindly. For instance, the category 2 swarm of Figure 5, which was nearly spherical to begin with, seems little distorted by decomposing it into, say, the four subswarms of a level 2 decomposition. If $\mathbf{X}_q$ is sinuous, as are categories 1 and 3 of Figure 5, then a level 2 decomposition goes a long way toward generating a reasonable resolution of the original swarm, but without getting too near the noise level.

3.4.3. *Strategy 3.* It is the sinuous shape of the data distribution that causes the poor definition of pdf's and hence the need for PDM subdivision. Thus we can envision measures of the skewness and kurtosis of the data swarms that could be used to decide if partitioning is required. It is clear that such higher-moment measures of the data swarm will have to be able to discern category 1 (Figure 5) distributions from elliptical distributions, since the latter are well represented by multidimensional Gaussians. We will not pursue such measures here.

3.5. *Fitting pdf's to the Terminal Nodes*

Let us suppose that the qth category subset $\mathbf{X}_q$ has been decomposed into a number of terminal nodes $\mathbf{X}_q(\alpha_1, \cdots, \alpha_K)$. Let $\mathbf{T}_q(t)$ denote the tth terminal node $\mathbf{X}_q(\alpha_1, \cdots, \alpha_K)$ of $\mathbf{X}_q$, and let NT_q be the number of terminal nodes of $\mathbf{X}_q$; $t = 1, 2, \cdots, NT_q$. Thus $NT_q = 1$ for the case of no decomposition of the original category subset, $NT_q = 4$ for a level 2 decomposition like that of Figures 8 and 9, and so on. Let $N_q(t)$ denote the number of points $N_q(\alpha_1, \cdots, \alpha_K)$ in the tth terminal node;

$$\sum_{t=1}^{NT_q} N_q(t) = M_q$$

The centroid of $\mathbf{T}_q(t)$ is located at $\bar{\mathbf{T}}_q(t)$. Finally, let $\mathbf{S}_q(t)$ be the $L \times L$ covariance matrix of $\mathbf{T}_q(t)$, with determinant $\| \mathbf{S}_q(t) \|$ and inverse $\mathbf{S}_q^{-1}(t)$.

The best fit multivariate normal pdf for the tth terminal node $\mathbf{T}_q(t)$ is then

$$\phi_q(t, \mathbf{X}) = (2\pi)^{-L/2}(\| \mathbf{S}_q(t) \|)^{-1/2} \cdot \exp\{-0.5[\mathbf{X} - \bar{\mathbf{T}}_q(t)]^T \mathbf{S}_q^{-1}(t)[\mathbf{X} - \bar{\mathbf{T}}_q(t)]\}$$

(It is assumed that $\| \mathbf{S}_q(t) \| \neq 0$, so that $\mathbf{S}_q^{-1}(t)$ exists; if this is not the case, the PCA decomposition leading to this terminal node is not made, and the parent swarm is declared terminal.) $\mathbf{X}$ is an arbitrary point in E_L. $\mathbf{S}_q^{-1}(t)$ is readily obtained from the eigenvalues and eigenvectors obtained in the PCA of $\mathbf{T}_q(t)$, namely,

$$\mathbf{S}_q^{-1}(t) = (M_q - 1) \sum_{j=1}^{L} \frac{1}{l_j} \mathbf{e}_j \mathbf{e}_j^T$$

3.6. *Assembling the pdf's*

A multivariate normal pdf is fitted to each terminal node $\mathbf{T}_q(t)$ of $N_q(t)$ points, $t = 1, \cdots, NT_q$. We define a weighting function $W_q(t) = N_q(t)/M_q$, so that

$$\sum_{t=1}^{NT_q} W_q(t) = 1$$

The probability distribution function for the qth category subset is then taken to be

$$\Phi_q(\mathbf{X}) = \sum_{t=1}^{NT_q} W_q(t)\phi_q(t, \mathbf{X})$$

for $q = 1, \cdots, Q$ and $\mathbf{X}$ in E_L. These pdf's $\Phi_q(\mathbf{X})$ define the desired PDM model.

Figure 7 showed the binormal pdf for the category 1 point swarm of Figure 5; this is the case of $NT_q = 1$, or no PCA decomposition of the category set. Figure 10 shows the contours of $\Phi_1(\mathbf{X})$ when determined by a level 2 decomposition, as illustrated in Figures 8 and 9 and discussed in section 3.4.2. This pdf is clearly a much more realistic description of the category 1 swarm than is the pdf of Figure 7. If the PCA decomposition is allowed to proceed until just before the minimum point requirement $N_q(t) > L$ is violated, the category 1 point swarm of Figure 5 is reduced to 23 terminal nodes. Figure 11 shows the tree diagram of this maximum possible decomposition. Figure 12 shows the $\Phi_1(\mathbf{X})$ contours determined from the terminal nodes of Figure 11. This pdf gives a very sharp delineation of the category subset, but the fine structure of the probability contours is clearly being determined by the individual points of the category subset, which may be undesirable, as discussed in section 3.4.

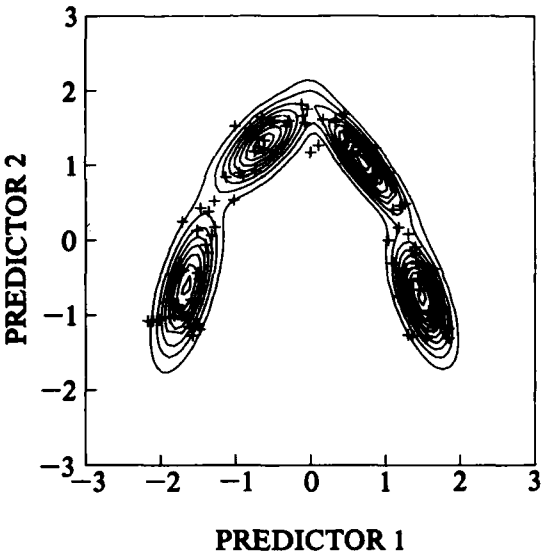

Fig. 10. The category 1 point swarm of Figure 5 and the probability contours of $\Phi_1(\mathbf{X})$, as determined by a level 2 PCA decomposition.

3.7. *Making a Prediction*

Just as in the single-predictor case, we must choose a prediction strategy (maximum probability, Bayesian, or another) for using the pdf's $\Phi_q(\mathbf{X})$ to make a prediction. If the maximum probability strategy is chosen, then, given a new predictor realization $\mathbf{X}'$ (now an L-dimensional vector), we evaluate $\Phi_q(\mathbf{X}')$, $q = 1, \cdots, Q$. The prediction is then that the predictand falls into category q', where q' is the q value corresponding to the maximum $\Phi_q(\mathbf{X}')$, $q = 1, \cdots, Q$. If the Bayesian

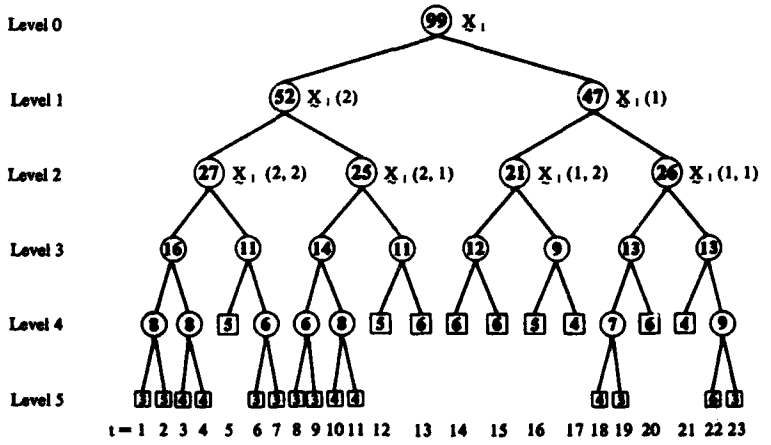

Fig. 11. The tree diagram showing the maximum possible decomposition of the category 1 subset of Figure 5. The circles represent the $\mathbf{X}_1(\alpha_1, \cdots, \alpha_n)$ subsets, and the numbers within the circles give the number of points in the subswarm, $N_1(\alpha_1, \cdots, \alpha_n)$. Terminal nodes $\mathbf{T}_1(t)$ are represented by boxes; the enclosed numbers give $N_1(t)$.

strategy is chosen, the a priori probabilities can be set to $P(q) = M_q/N_{tr}$, as in the single-predictor case, and the pdf's $\Phi_q(\mathbf{X}) \equiv \Phi(\mathbf{X} \mid q)$ are used in Bayes' formula.

3.8. *Potential Predictability, Class Errors, and Significance Tests*

These matters all proceed in exact analogy to the single-predictor case. Thus in computing the potential predictability index for the maximum probability strategy, we first compute

$$P'(i, q) = \Phi_q[\mathbf{X}_{tr}(i)]\left\{\sum_{q=1}^{Q} \Phi_q[\mathbf{X}_{tr}(i)]\right\}^{-1}$$

for $q = 1, \cdots, Q$ and $i = 1, \cdots, N_{tr}$. The only difference from the single-predictor case is that we are now using the L-dimensional training set values $\mathbf{X}_{tr}(i)$ in the multivariate pdf's $\Phi_q(\mathbf{X})$. Subsequent formulas leading to PP or AVGPP are unchanged. Likewise, the modifications required for the Bayesian strategy are trivial.

The potential predictability is now measuring the separation of pdf's in an L-dimensional space. Figures 13–15 show three sets of pdf's, as determined for the example point swarms of Figure 5, where $L = 2$. Figure 13 (reproducing parts of Figures 6 and 7) shows in superposition the contours of equal probability of the three best fit binormal pdf's, $\phi_q(\mathbf{X})$, as would be obtained in classical discriminant theory. The potential predictability for these pdf's is PP = 0.39, when using the maximum probability forecast strategy. Figure 14 shows the pdf's, $\Phi_q(\mathbf{X})$, as obtained by level 2 PCA, as illustrated in Figures 8–10. The eye can now easily distinguish the three pdf's determined from the three point swarms of Figure 5, and the potential predictability has risen to PP = 0.77. Figure 15 shows the pdf's as determined from the maximum possible PCA decomposition of the category swarms, as shown in Figures 11 and 12. These pdf's show even better separation; as

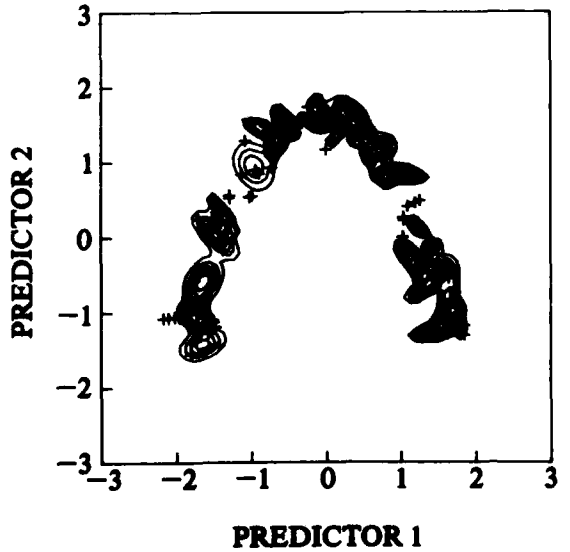

Fig. 12. The category 1 points of Figure 5 and the $\Phi_1(\mathbf{X})$ probability contours, as constructed from the 23 terminal nodes of Figure 11.

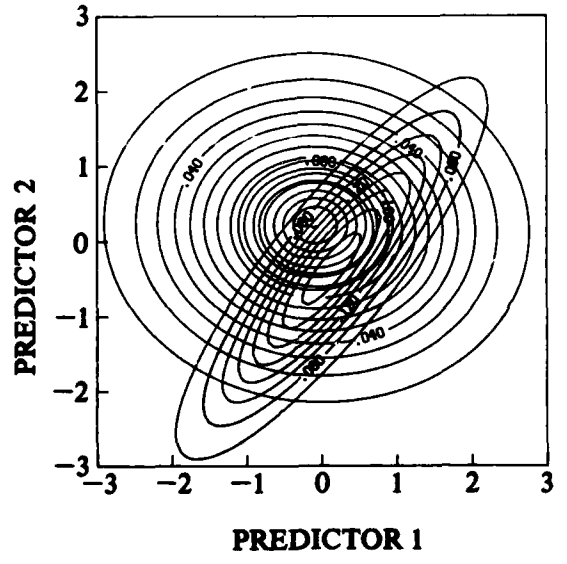

Fig. 13. Contours of equal probability of the three binormal pdf's $\Phi_q(\mathbf{X})$, $q = 1, 2, 3$, fitting the three category subsets of Figure 5. The contour interval is different for each of the three pdf's.

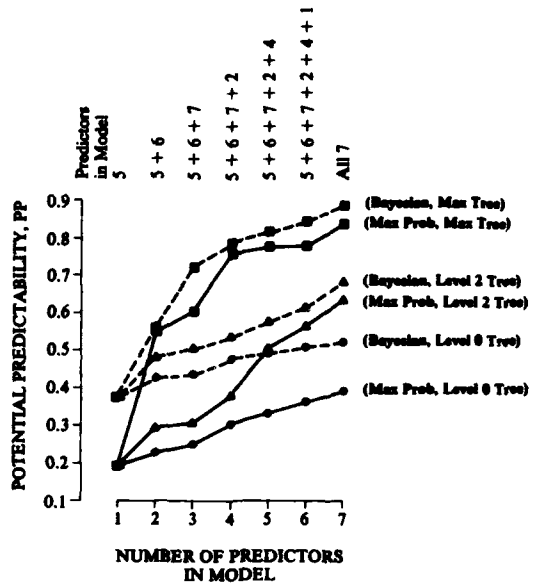

Fig. 17. Potential predictability values for various PDM models (time lag $\tau = 0$). The solid curves are for the maximum probability forecast strategy, and the dashed curves are for the Bayesian strategy. Dots are for no PCA decomposition of the category swarms (a level 0 decomposition, equivalent to classical discriminant analysis), triangles are for a level 2 PCA decomposition, and squares identify the curves for which the maximum possible number of PCA decompositions was performed.

3. All else being equal, PP increases as the number of PCA decompositions of the category swarms increases.

4. All else being equal, PP increases as more predictors are added to the model.

Similar results were found for $\hat{a}_0$ and $\hat{a}_1$, e.g., $\hat{a}_1$ decreases (the model becomes better) as predictors are added, all else being equal, and so on. This behavior is consistent with our expectations and with the high likelihood that much of the apparent skill is artificial.

Figure 18 shows the dependence of PP on the time lag τ between predictor and predictand, for the case of a Bayesian forecast strategy and a level 2 PCA decomposition of category swarms. We note that the PP scores decrease somewhat as τ increases from 0 to 4 months for the two predictor model, but that the PP scores are relatively independent of τ for the five-predictor model.

Figure 18 was generated for two-predictor and five-predictor models in which the particular predictors in the model were held fixed (i.e., predictors 5 and 6 in the first case and predictors 5, 6, 7, 2, and 4 in the second case). In general, we would expect that the best predictors for one time lag might not be the best for another time lag. Indeed, for $\tau = 0$ or 1, predictor 5 has the highest PP of any single predictor, whereas for $\tau = 2$, 3, or 4, predictor 2 (Barnett's $U2$) has the highest PP. However, for the present data set this dependence is weak: for $\tau = 4$, predictor 2 has PP = 0.336 and predictor 5 has PP = 0.316.

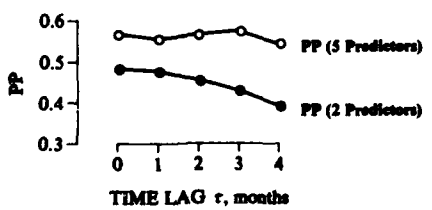

Fig. 18. PP scores for the two-predictor (solid circles) and five-predictor (open circles) PDM, as a function of time lag τ.

However, when the various PDM models of Figures 17 and 18 were applied to the testing set, the performance of the PDM was quite disappointing. Indeed, the PDM's tercile category forecasts failed even to show the presence of the 1982-1983 El Niño, let alone accurately predict its onset. Careful investigation into the cause of the PDM's failure showed that the raw data are so noisy that the category swarms cannot be adequately distinguished: the points for the extreme categories 1 and 3 are nearly lost in the swarm of points for category 2. The associated pdf's $\Phi_q(\mathbf{X})$ are correspondingly overlapping; a result anticipated in point 1, cited earlier. Given such data, neither the PDM nor any similar technique can be expected to show any useable degree of forecast skill.

4.2. *Using Filtered Predictors*

If the poor performance of the PDM in the El Niño forecast is indeed due to noise in the data, then perhaps filtering or smoothing the raw predictor values will increase the signal-to-noise ratio and thereby allow the PDM to extract the information needed to make its forecast. To investigate this possibility, a series of forecasts was made using two types of filters:

1. A seven-point running mean was applied to each predictor time series. Thus each predictor value $X(j, k)$, $k = 1, \cdots, K$, was replaced by a smoothed value, $X_s(j, k)$, given by

$$X_s(j, k) \equiv \frac{1}{7} \sum_{j'=j-3}^{j+3} X(j', k)$$

The 3 months at the beginning and end of the 476-month time series were left unsmoothed. The PDM analysis then proceeded as before, but now using the $X_s(j, k)$ as predictors.

2. As before, the training set $\mathbf{X}_{tr}$ was selected to be the first $N_{tr} = 396$ months of each of the $K = 7$ predictors. A PCA was then performed on the training set to get

$$\mathbf{A} = \mathbf{X}_{tr} \cdot \mathbf{E}$$

where $\mathbf{E} \equiv [\mathbf{e}_1, \cdots, \mathbf{e}_7]$ is the 7×7 matrix of empirical orthogonal functions (EOFs) and $\mathbf{A} = [\mathbf{a}_1, \cdots, \mathbf{a}_7]$ is the 396×7 matrix of principal components. The principal component time series $\mathbf{a}_j = [a_j(1), \cdots, a_j(N_{tr})]^T$, $j = 1, \cdots, K$, were ordered by the size of their associated eigenvalue and were used as the predictors in training the PDM, rather than using the original $X(j, k)$ as predictors (compare section 2.12). Since the $\mathbf{a}_j$ are orthogonal, we can do no further predictor ranking using correlations between predictors (compare section 3.1).

The testing set $\mathbf{X}_{ts}$ was defined as before to be the predictors from 1980 to 1986. However, before making a forecast using the testing set, we replaced $\mathbf{X}_{ts}$ by amplitudes $\mathbf{A}_{ts}$, defined by

$$\mathbf{A}_{ts} \equiv \mathbf{X}_{ts} \cdot \mathbf{E}$$

where $\mathbf{E}$ is the EOF matrix of the training set. We thus performed the same transformation on the training and testing sets, so that the $\mathbf{A}_{ts}$ values can be used in the probability distribution functions Φ_q of the PDM.

A series of experiments was made to compare the forecasts made using the filtered predictors with the forecasts mode using the raw data. The Bayesian forecast strategy and a level

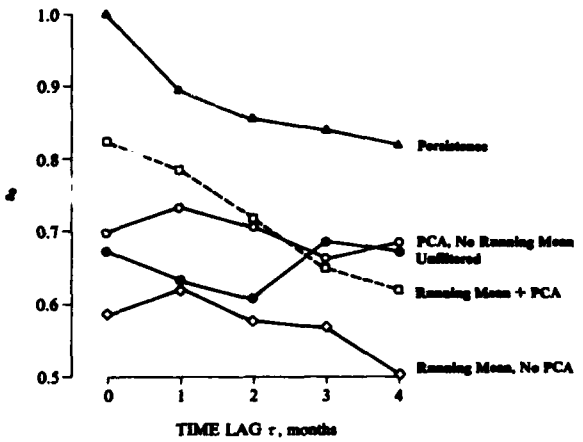

Fig. 19. The a_0 scores for various two-predictor PDM models: solid circles, unfiltered predictors 5 and 6; diamonds, predictors 5 and 6 with a seven-point running mean; open circles, principal components 1 and 2; squares, seven-point running mean, then PCA and using principal components 1 and 2; triangles, persistence of the predictand values.

2 decomposition of the category swarms were chosen. Figure 19 compares the a_0 scores of the various two-predictor models. We note first that the a_0 scores obtained after applying the seven-point running mean filter to predictors 5 and 6 are in fact lower than the scores obtained using unfiltered predictors 5 and 6. However, if we perform a PCA and then use principal components 1 and 2, the a_0 scores are generally higher than the scores of the unfiltered two-predictor model out to $\tau = 2$. These results can be interpreted as follows. The running mean is a low-pass temporal filter which leaves a low-frequency, but possibly still random, time series. The spatial correlations between the predictor time series are relatively unchanged by the temporal smoothing. The PCA operation, on the other hand, is a spatial filter, and the resulting time series $\mathbf{a}_1$ contains spatially coherent information from all of the original predictor regions. Time series $\mathbf{a}_2$ also contains spatially coherent information from all of the original predictor regions, though of a spatial pattern which is distinct from that of $\mathbf{a}_1$. Thus merely filtering high-frequency noise from the predictor time series does not improve the a_0 scores, whereas using the spatially coherent signal from all of the original predictor regions does lead to a better set of predictors $\mathbf{a}_1$ and $\mathbf{a}_2$. Figure 19 also shows that if we first apply the seven-point running mean to each of the original seven predictors and then perform a PCA on the smoothed time series, we get greatly improved a_0 scores for short time lags, although the a_0 scores are degraded for longer time lags. This latter effect may simply be due to statistical uncertainty in the estimation of the a_0.

For reference purposes, Figure 19 also shows the a_0 scores obtained by persistence; that is, the observed category at time j is used as the forecast for time $j + \tau$. (For $\tau = 0$, then, persistence uses the observed category to forecast itself and obtains a perfect score of $a_0 = 1$.) Since the SST anomaly categories, as terciled, are quite persistent, persistence attains a high a_0 score. In a similar fashion, climatology, which always forecasts tercile category 2, attains a score of $a_0 = 0.725$ owing to the chosen terciling scheme. Neither persistence nor climatology can forecast the onset of an El Niño, however, and thus are not valid competitors in actual forecast situations.

We also note that scores like a_0 are overall measures of a forecaster's performance over the time span of the testing set. If we are interested only in forecasting the onset of an El Niño, then a low a_0 score does not necessarily imply poor model performance, nor does a high a_0 score imply success in the forecast.

Figure 20 shows the actual category forecasts made by the smoothed two-predictor PDM model using $\mathbf{a}_1$ and $\mathbf{a}_2$ as predictors. We see that the PDM forecasts are similar to that of the Barnett model for small τ: a rise to the above-normal category, followed by a fall to the below-normal category. But the PDM's longer-lead forecast missed the peak of the event and also failed to predict the longevity of the warming. Thus even though the preliminary PCA spatial filtering of the noisy wind fields helped the model, it has not been able to extract the same information from the original data set as did the linear prediction model. In essence, the PP scores suggest that there is so much variability between El Niño events that the requisite pdf's are poorly defined, and so the PDM should fail. Further, the 1982–1983 event was quite unusual for a variety of reasons and so may not fit well into the statistical structure determined from the training set. These problems withstanding, we still should expect the PDM to fail in 1983 for the same reason the linear prediction model failed [cf. *Barnett*, 1984].

In summary, the PDM did not perform particularly well on

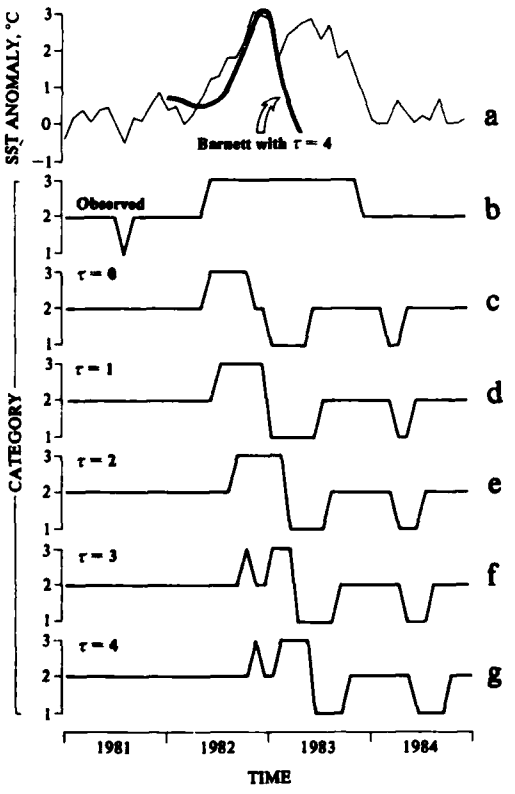

Fig. 20. Category forecasts made using principal components 1 and 2 (circles of Figure 19). (*a*) The actual SST anomalies and the forecast made by the Barnett model, with $\tau = 4$ months. (*b*) The observed tercile categories (a perfect forecast). Figures 20c–20g forecasts for time lags τ, as shown.

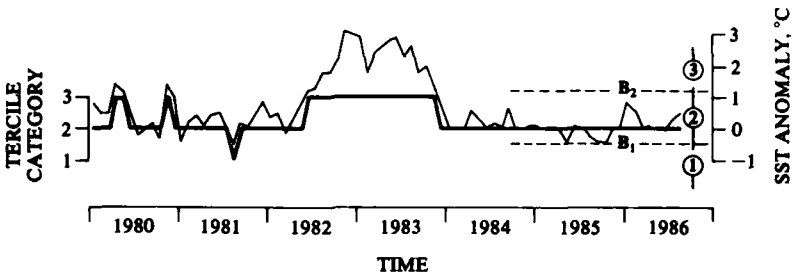

Fig. 16. The testing set for prediction of the 1983 El Niño. The light curve and the scale at the right show the actual SST anomalies. The heavy curve and the scale at the left show the corresponding tercile category values.

4. APPLICATION OF THE PDM

This section is intended to show some of the strengths and weaknesses of the PDM. Thus we show a forecast scenario in which the method does not do particularly well and one where it apparently does better than other conventional forecast schemes. The first example is particularly illuminating, for there we intercompare results obtained using some of the different strategies discussed earlier, thereby giving the reader a feeling for the sensitivity of the PDM to the details of its construction. Since this is essentially a theoretical paper, the discussion of the applications is brief. Additional examples of the PDM in operation are given by *Preisendorfer et al.* [1987].

4.1. *Forecasting the El Niño of 1982–1983*

Barnett [1984] addressed the problem of statistically forecasting sea surface temperature (SST) anomalies in the equatorial Pacific using wind anomalies as predictors, during the 1982–1983 El Niño. That study used an advanced regression model which related the SST anomalies in the predictand regions to the prior wind anomalies in the predictor regions. The study showed, among other things, that it was possible to forecast the onset of El Niño, as measured by SST anomalies in a region off the coast of Peru, using wind anomalies from various regions in the central Pacific. These forecasts were successful at lead times of up to 4 months. Although the model did an acceptable job of forecasting the onset of the 1982–1983 El Niño, it failed to accurately predict the decline of the El Niño, for reasons discussed in the 1984 paper. It was felt that a repetition of this study would be another means of evaluating the PDM's forecast ability.

The data set consists of monthly wind and temperature anomalies for the 476 months from January 1947 to August 1986. There are four regions of the equatorial Pacific for which u-component (east–west) wind anomalies are available, and three regions for which there are v-component (north–south) wind anomalies. Thus there are seven possible predictors (labeled $1, \cdots, 7$ and corresponding to Barnett's $U1$, $U2$, $U3$, $U4$, $V1$, $V2$, and $V3$, respectively). The predictand SST anomalies were terciled, so that only the extreme events would fall outside the "normal" category. Inspection of the SST record shows that if boundaries $B_1 = -0.5°C$ and $B_2 = 1.2°C$ are selected (see section 2.2), then slightly less than one sixth of the anomalies fall into category 1 (below normal SST), somewhat more than two thirds fall into category 2 (normal SST), and slightly less than one sixth fall into category 3 (above normal SST). The above-normal category, so defined, contains only anomalies which are greater than two standard deviations from the mean, which is a reasonable definition of El Niño. The 396 months from January 1947 to December 1979 were taken to be the training set, and the 80 months from January 1980 to August 1986 were taken to be the testing set. The training set contains several El Niños, so we thought that the PDM should have a good opportunity to define the category pdf's. The 1982–1983 event stands out prominently in the testing set, as is seen in Figure 16. Furthermore, the testing set is largely independent of the training set, although there is substantial autocorrelation within each set (compare section 2.1 and section 2.4).

The PDM was applied in various configurations:

1. Both maximum probability and Bayesian strategies were used. In the Bayesian case the priors were made proportional to the number of points in the category (compare section 2.7).

2. Category swarms were forced to undergo a predetermined number of PCA subdivisions, either zero (as seen in Figure 13), 2 (as seen in Figure 14), or the maximum possible number (as seen in Figure 15), as discussed in section 3.4.

3. The potential predictability was used to measure the separation of the category pdf's, although the 5% significance levels were computed only in the single-predictor cases (owing to computational expense).

4. The individual predictors were rated by their potential predictability scores in order to select the first predictor. Subsequent predictors were added to the model in the order given by the correlations, as described in section 3.1. Models containing 1–7 predictors were compared.

For a time lag of $\tau = 0$, predictor 5 (wind in region $V1$) has the highest potential predictability score of any individual predictor. If the maximum probability strategy is chosen, this value is $PP = 0.196$; the 5% significance level is $PP(96) = 0.019$, so that PP is significant. For the Bayesian strategy, $PP = 0.377$ and $PP(96) = 0.316$, so that PP is once again significant. Predictor 5 thus becomes the first predictor of the PDM model. Predictor 6 (wind in region $V2$) is least correlated with predictor 5, and therefore becomes the next predictor added to the model. With two or more predictors in the model, we also have the possibility of forecast skills depending on the number of PCA decompositions of the category sets. Figure 17 shows the dependence of the potential predictability on the form of the PDM model. In Figure 17 we note the following behavior of the potential predictability:

1. The relatively low initial PP values, while significant, indicate that the category pdf's are not very distinct. We immediately expect that the PDM, as constituted for this problem, will not perform well.

2. All else being equal, PP is greater for the Bayesian forecast strategy than for the maximum probability strategy.

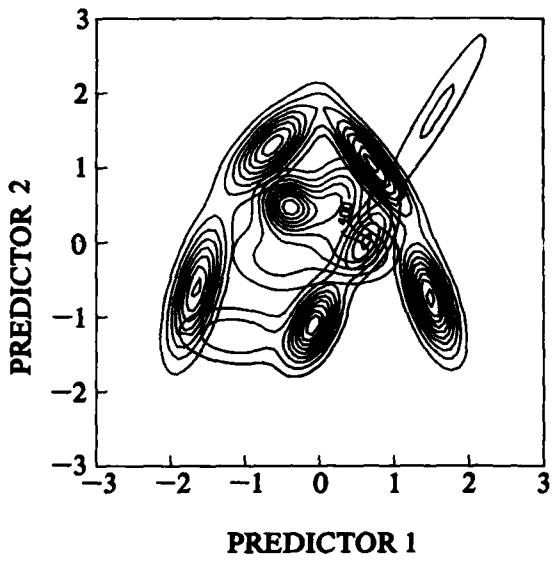

Fig. 14. Contours of equal probability of the three pdf's $\Phi_q(\mathbf{X})$, as determined from a level 2 PCA decomposition of each category subset of Figure 5. Contour intervals vary.

verified by their PP value of PP = 0.87, but the noise in the data (i.e., the positions of the individual points) has clearly affected the pdf's themselves. If the pdf's of Figure 15 were used for actual forecasting, it might often occur that predictor values $\mathbf{X}'$ would "fall into the gaps" of these irregularly shaped pdf's in such a manner as to cause the point to be ascribed to the wrong pdf, thus giving an incorrect forecast.

Given the probabilities $P'(i, q)$, the potential class errors $\hat{a}_0$ and $\hat{a}_1$ are immediately available. The actual class errors a_0 and a_1 are now computed from the multipredictor testing set $\mathbf{X}_{ts}(i)$, $i = 1, \cdots, N_{ts}$.

Monte Carlo experiments for determining 5% significance

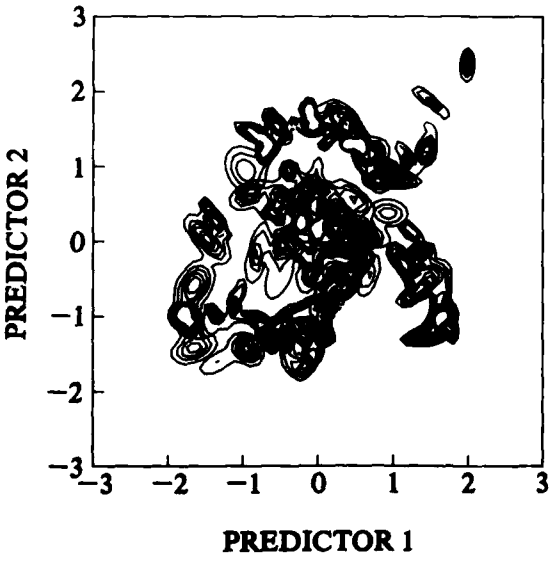

Fig. 15. Contours of equal probability of the three pdf's $\Phi_q(\mathbf{X})$, as determined from the maximum possible PCA decomposition of the category subsets of Figure 5. Contour intervals vary.

levels on PP, $\hat{a}_0$, $\hat{a}_1$, a_0, and a_1 proceed, in principle, as before. Now, however, when the randomly generated predictand $R(i)$ is analyzed using the multivariate predictors, it is necessary to perform a full PCA decomposition in order to get the needed pdf's (as described in sections 3.3 to 3.6). This PCA analysis becomes prohibitively expensive when it must be repeated 100 times in a Monte Carlo experiment. Thus in practice, the 5% significance levels may not be available.

3.9. *Final Screening of the Candidate Predictor*

We recall from section 3.1 that we have admitted a candidate Lth predictor to the PDM model, based upon the correlation screening described there. We now may use the information gathered in the previous paragraph to decide whether or not to keep the candidate predictor in the model. Let $\hat{a}_0(L-1)$ and $\hat{a}_1(L-1)$ denote the $\hat{a}_0$ and $\hat{a}_1$ scores obtained from the PDM model before the candidate Lth predictor was admitted (if $L = 2$, we have the single predictor potential class errors available). Let $\mathrm{PP}(L)$, $\hat{a}_0(L)$, and $\hat{a}_1(L)$ be the scores obtained after the candidate Lth predictor was admitted. Moreover, let $\mathrm{PP}(96; L)$, $\hat{a}_0(96; L)$, and $\hat{a}_1(05; L)$ be the appropriate 5% critical values, as determined by Monte Carlo simulations. We then accept the candidate Lth predictor, $X(j, L)$, into the PDM model, if the following conditions hold:

Condition 1

$$\mathrm{PP}(L) \geq \mathrm{PP}(96; L)$$

Condition 2

$$\hat{a}_0(L) > \hat{a}_0(L-1) \qquad \hat{a}_1(L) \leq \hat{a}_1(L-1)$$

Condition 3

$$\hat{a}_0(L) \geq \hat{a}_0(96; L) \qquad \hat{a}_1(L) \leq \hat{a}_1(05; L)$$

If these three conditions are not satisfied, we delete the candidate predictor from the model and return to section 3.1 to select the next candidate predictor. We continue in this manner until all possible predictors have been examined, at which time the PDM model is complete.

Condition 1 is simply the requirement that the model have a statistically significant potential predictability. Condition 2 is the requirement that the addition of the Lth predictor improve the potential class error scores, and condition 3 expresses the requirement that the model's potential class error scores be statistically significant. Conditions 1 and 3 can be relaxed by using, say, a 10% significance level instead of the 5% level shown. Condition 2 cannot be relaxed. For complete rigor the critical level should decrease as the number of possible predictors is increased. This allows for the probability that one of the predictors will, by shear chance, appear useful (compare section 2.12).

3.10. *Scoring the PDM Model*

Once the PDM model is complete, we can compute the actual class errors a_0 and a_1, using the testing set $\mathbf{X}_{ts}(i)$, $i = 1, \cdots, N_{ts}$, generated during the examination of the final predictor which was admitted to the model. These a_0 and a_1 scores, together with the information shown in conditions 1, 2, and 3 in section 3.9, are the data by which we measure the PDM model's actual and potential skills.

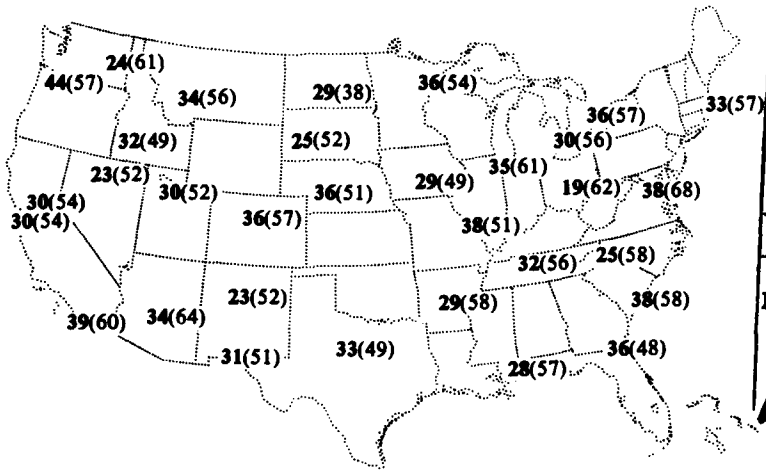

Fig. 23. Percentage a_0 scores obtained from Monte Carlo experiments in which the category forecast was made at random. The expected value is 33. The numbers in parentheses show the a_0 scores of the two-predictor PDM, from Figure 22.

4. Forecasts were made for winter at a lead time of one season, e.g., fall SLP predicting winter temperature. The scores are shown as "percent correct category forecasts" and thus a randomly made forecast has an expected value of 33%. Note that these are actual forecast skills, since the testing sets in no way entered the predictor screening or PDM pdf construction.

The results of the single-predictor experiments are shown in Figure 21. Monte Carlo simulations showing a_0 forecast skill values averaged over the entire United States in excess of 50% are significant at the 95% level. The highest forecast skills are in the eastern and western thirds of the country; there is only a modest drop-off in skill in the central region. The former result is in accord with earlier studies [cf. *Barnett*, 1981; *Barnett and Preisendorfer*, 1987]. However, the levels of skill for the PDM are higher than in the earlier studies. More importantly, the PDM appears to exhibit significant skill for the central United States, a region where the linear prediction models failed.

Adding an additional predictor and allowing the possibility of a single PCA subdivision of the category subset gives the results shown in Figure 22. Skill scores are increased typically by 15–25%, particularly in the northeast part of the country. Notice that the central region, where skills were lowest, exhibits only a small change in skill with the increased model complexity.

These results can be contrasted with the average forecast skills obtained from the Monte Carlo experiments (Figure 23). The expected value of 33% is indeed realized on average over the entire United States. The forecasts made by the PDM are clearly much better than random chance: the PDM's a_0 score, averaged over the entire United States; is 53%.

We conclude that the PDM performance in forecasting winter air temperature over the United States is highly statistically significant. Further, the skill levels are substantially higher than those obtained by earlier methods. If it is eventually found that the predictability of the climate system is associated with short-lived climatic "regimes" in nature, as suggested by the results of *Barnett and Preisendorfer* [1987], then the PDM offers one of the few statistical techniques for long-range forecasting and diagnostics.

Acknowledgments. This work was supported in part by the U.S. Tropical Ocean-Global Atmosphere (TOGA) Program (NOAA/NA85AA-D-AC132), in part by the U.S. Climate Program Office via the Experimental Climate Forecast Center at the Scripps Institution of Oceanography (NOAA/NA86-AA-D-CP104), in part by the National Science Foundation (ATM85-13713), and in part by the Office of Naval Research Oceanic Biology Program (N00014-87-K-0525). This paper is contribution 39 from the Joint Institute for the Study of the Atmosphere and Ocean and contribution 978 from the NOAA Pacific Marine Environmental Laboratory. Word processing was performed by Ryan Whitney and the figures were drawn by Gini Curl.

REFERENCES

Barnett, T. P., Statistical prediction of North American air temperatures from Pacific predictors, *Mon. Weather Rev., 109*, 1021–1041, 1981.

Barnett, T. P., Prediction of the El Niño of 1982–83, *Mon. Weather Rev., 112*, 1403–1407, 1984.

Barnett, T. P., and R. W. Preisendorfer, Origins and levels of monthly and seasonal forecast skill for North American surface air temperatures determined by canonical correlation analysis, *Mon. Weather Rev., 115*, 1825–1850, 1987.

Box, G. E. P., and G. C. Tiao, *Bayesian Inference in Statistical Analysis*, 588 pp., Addison-Wesley, Reading, Mass., 1972.

Harnack, R., M. Cammarata, K. Dixon, J. Lanzante, and J. Harnack, Summary of U.S. Seasonal Temperature Forecast Experiments, in *Proceedings, 9th Conference on Probability and Statistics in Atmospheric Sciences*, pp. 175–179, American Meteorological Society, Boston, Mass., 1985.

Lachenbruch, P. A., *Discriminant Analysis*, 128 pp., Hafner Press, New York, 1975.

Madden, R. A., and D. Shea, Estimates of the natural variability of time averaged temperatures over the United States, *Mon. Weather Rev., 106*, 1695–1703, 1978.

Preisendorfer, R. W., *Principal Component Analysis in Meteorology and Oceanography*, edited by C. D. Mobley, 425 pp., Elsevier Science, New York, 1988.

Preisendorfer, R. W., and C. D. Mobley, Climate forecast verifications, United States mainland, 1974–83, *Mon. Weather Rev., 112*, 809–825, 1984.

Preisendorfer, R. W., C. D. Mobley, and T. P. Barnett, The principal discriminant method of prediction: Theory and evaluation, *NOAA Tech. Memo. ERL PMEL-71*, 76 pp., Pac. Mar. Environ. Lab., Seattle, Wash., 1987. (Available as NTIS PB87-209276 from Nat. Tech. Inf. Serv., Springfield, Va.)

T. P. Barnett, Climate Research Group, Scripps Institute of Oceanography, La Jolla, CA 92093.

C. D. Mobley, Pacific Marine Environmental Laboratories, NOAA/ERL, NOAA Building 3, 7600 Sand Point Way, NE, Seattle, WA 98115.

(Received August 5, 1987;
revised February 15, 1988;
accepted February 23, 1988.)