WEATHER ELEMENT PREDICTION BY DISCRIMINANT ANALYSIS

L.J. WILSON
ECMWF

ABSTRACT

Multiple Discriminant Analysis (MDA) is a statistical procedure where linear combinations of a set of variables are sought to maximise the separation between categories of the dependent variable. MDA has been used as a basis for operational forecasts of probability of precipitation amount at 74 Canadian stations for one and a half years with considerable success. MDA forecasts were selected for use in operations following the results of a series of comparative tests between the MDA-based equations and equations based on the Regression Estimation of Event Probability (REEP) method (Miller, 1964). It was also necessary to demonstrate the superiority of the MDA-based forecasts to the Analog-based forecasts that were operational at the time.

In this presentation the MDA method is outlined, and comparison tests with REEP using Canadian data are described. Examples of the forecasts and results of the tests are presented to demonstrate that MDA is a useful alternative to regression for some meteorological applications.

1. INTRODUCTION

In Canada, statistical methods have been used to generate objective forecasts using the output of the Canadian NWP model (now a spectral model). Although our experience with statistical interpretation is not as extensive as in the United States, we have in fact produced routine maximum and minimum temperature forecasts out to the third day for 10 years (Wilson and Yacowar, 1980). These forecasts, using the "perfect prog" formulation, have shown a gradual improvement as the driving model has improved. More recently, we have experimented with different statistical methods for the preparation of probability of precipitation amount forecasts. First, an analog-base system was implemented, but further tests showed that, at least in the short range, the Multiple Discriminant Analysis and REEP methods could produce superior forecasts, and a combined system was implemented which used MDA for forecasts to 36 hours and REEP for 48 and 60 hour forecasts. The reasons for opting for a combined system are discussed below.

As the MDA method is perhaps less well-known than the regression technique, a brief description of MDA is included below. Most of the mathematical details have been omitted and can be found in Miller, 1962. Following the MDA description, brief descriptions of REEP and some standard scoring rules are given in preparation for the discussion of the tests on Canadian data. Finally the test results are shown.

2. MULTIPLE DISCRIMINANT ANALYSIS

2.1 Definition and Use

If one were asked to define MDA in a few words, it would be necessary to include two important points in the words chosen. Firstly, MDA is a linear procedure. It shares that characteristic with most other statistical methods used in meteorology - multiple linear regression, canonical correlation, principle components analysis etc. MDA is therefore equivalent to regression in that sense. Secondly, MDA is designed to aid in the separation of predictand variables into one of several categories. MDA is therefore intended for use in predicting to which predictand category an event belongs, and its output is always in probabilistic terms.

In meteorology, MDA is best suited to those applications where the predictand is, or can be categorised for some reason. Types of situations where categorised predictands and probability forecasts may be most suitable include:

(a) Non-numerical predictands, such as precipitation type.

(b) Predictands which are effectively categorised due to a highly non-normal or multimodel distribution. For example, total cloud amount usually has a U-shaped distribution with a minimum near the mean value. Cloud amount could be easily divided into two or three categories using as thresholds points where the distribution density is a relative minimum.

(c) Predictands for which categorical or probability forecasts are desired for operational reasons. For example, ceiling base height is a quasi-continuous variable, but only specific thresholds are of interest in aviation forecasting. Precipitation amount is another example. There may be considerable interest in forecasting the probability of an extreme rainfall amount in a given period. The use of categorised predictands provides a means of tuning the forecast product to meet the needs of specific users. In practice, the choice of categories must always be made considering both the user's needs and the limitations of the available database. It is not possible to run an MDA procedure to obtain a forecast for a category for which there are no events in the dependent dataset. If there are a few events in the dependent dataset, it is possible to obtain a forecast, but that forecast will not be very reliable. When choosing categories, it is often necessary to compromise between users needs and reliable (accurate) forecasts.

MDA, as a statistical technique, may be applied to meteorological problems using any or all of the three application formulations, classical, perfect prog and model output statistics. The three application methods dictate only the way in which the data is used in development and testing of equations; they are independent of the statistical technique used. Furthermore, the same sets of advantages and disadvantages of the three application methods apply whether MDA or regression is used.

MDA is a two-step process. The first step is to find linear combinations of predictors ("discriminant functions") which provide a basis for determining which category an event belongs to. At the end of this step, one only has a set of equations. The second step is to use the equations to determine probabilities of category membership based on input predictor values. Technically, the MDA is complete after the first step, but the second step is needed to produce a forecast probability.

2.2 Graphical Analogy

The easiest way to understand the first step of the MDA procedure is by looking at a graphical analogy. For this purpose, consider a simple situation with two predictand categories and two predictors. Suppose a sample were available containing model forecasts of relative humidity, vertical velocity and verifying observations of precipitation. If all events where precipitation did not occur are grouped into category 1 and all precipitation events are grouped into category 2, the predictor values may be plotted as dots for category 1 and crosses for category 2. The resulting scatter plot may, ideally, look like figure 1.

Ellipses can be drawn on figure 1 to enclose a specific percentage of the events. Such centile contours will only be perfect ellipses if the data is approximately normally distributed within categories. If the data points are visually projected onto either one of the axes it is evident that the overlap area would be relatively large, as indicated by the ticks on the axes, and the within-group variance (the "spread" of the data points) would also be relatively large. If a line is drawn parallel to the line joining the two category means, it is evident that projection of the data points onto this line would produce sharper within-group distributions and minimise the overlap.

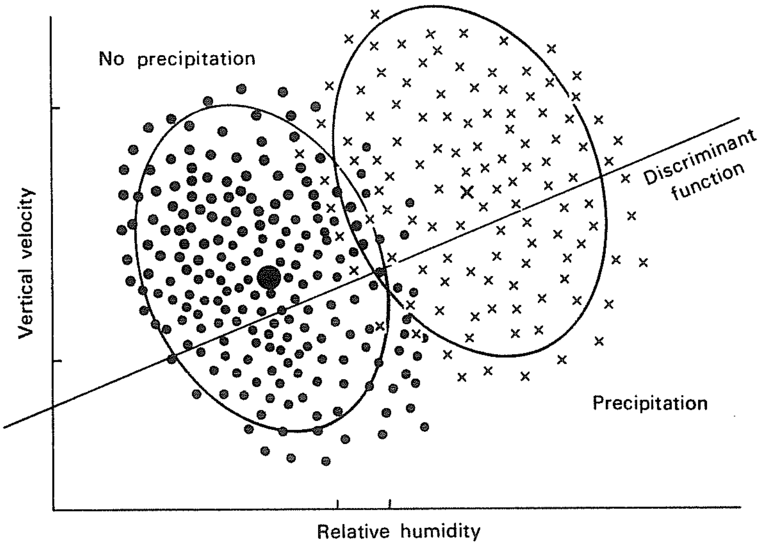

Fig. 1 Schematic representation of scatter plot of relative humidity versus vertical velocity for no rain cases (dots) and rain cases (crosses). Ellipses enclose a specific percentage of events in each category. The large dot and large cross indicate the category average of the two variables.

The visual analogy can be extended only so far. If there were three categories it would be possible to find two lines which would meet the criteria of maximising separation of groups while minimising within group dispersion. However, as figure 2 shows, one line separates the groups much more successfully than the other.

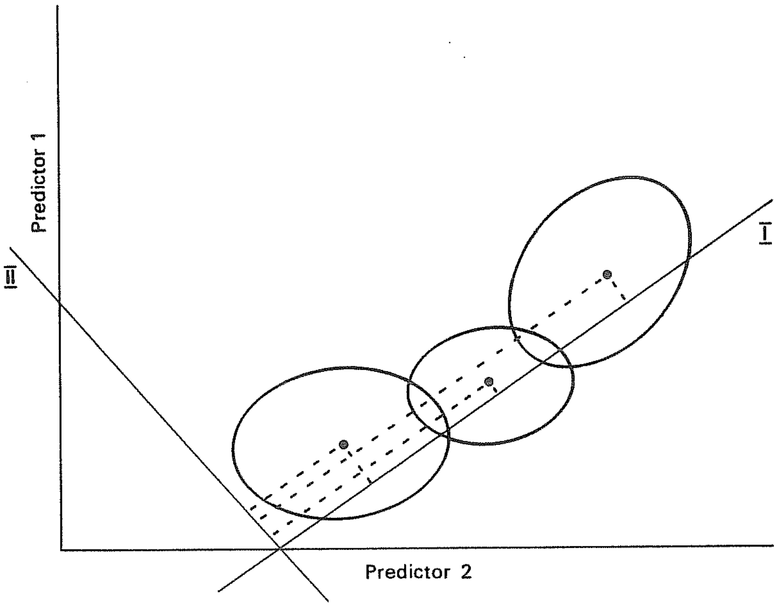

Fig. 2 Schematic representation of discriminant function orientation for three group-two predictor case.

The graphical analogy suggests the characteristics of the discriminant functions quite clearly. Discriminant functions are linear combinations of the predictors which maximise between-group separation and minimise within-group dispersion. In fact, it is the between-groups/within-groups dispersion ratio that is maximised. Most of the discriminant power available in the predictors is concentrated in the first one or two discriminant functions, as indicated by figure 2. It is possible to find in general (G-1) discriminant functions where G is the number of groups. Should the number of predictors be very small, the number of discriminant functions will be limited to the number of predictors if it is less than (G-1). Discriminant functions are independent in a statistical sense, but not orthogonal.

Calculation of discriminant functions is in some ways analagous to the performance of principle components analysis prior to regression analysis. The aim is similar: to determine uncorrelated linear combinations of the predictors which maximise the predictive power for the intended purpose. In both cases, the intent is to replace the original set of predictors with the new smaller set of predictors for which predictive power is maximised. The number of variables in the analysis is thereby reduced. For MDA, the "predictive power" is given by the between-groups/within-groups variance ratio.

Figure 3 shows a "real" example of a scatter plot of data as a function of two predictors. In this application, there were 6 predictors used in the MDA and the first two, total precipitation from the ECMWF model, and 1000 mb geopotential height, were plotted. The predictand was precipitation amount in 4 categories, less than 1.0 mm, 1 to 10 mm, 10 to 25 mm and more than 25 mm accumulation in 24 hours. The three lines labelled I, II and III are the discriminant functions for this case. It can be seen from the category means that there is a reasonable separation of groups with lower precipitation amounts related to low values of the model precipitation and high values of the 1000 mb height, and higher precipitation amounts related to high values of the model precipitation and low 1000 mb heights. However, there is considerable scatter within categories. The first function is not quite parallel to a line through the group means because the within-group dispersions are such that the maximum in the dispersion ratio is achieved with a slightly steeper-sloped line. There are three functions in this case because there are 4 categories; the lines here represent the projection of those lines into 2-dimensional predictor space. There is not much difference between functions II and III in terms of the two predictors plotted here; there is more separation in terms of others of the 6 predictors in the analysis.

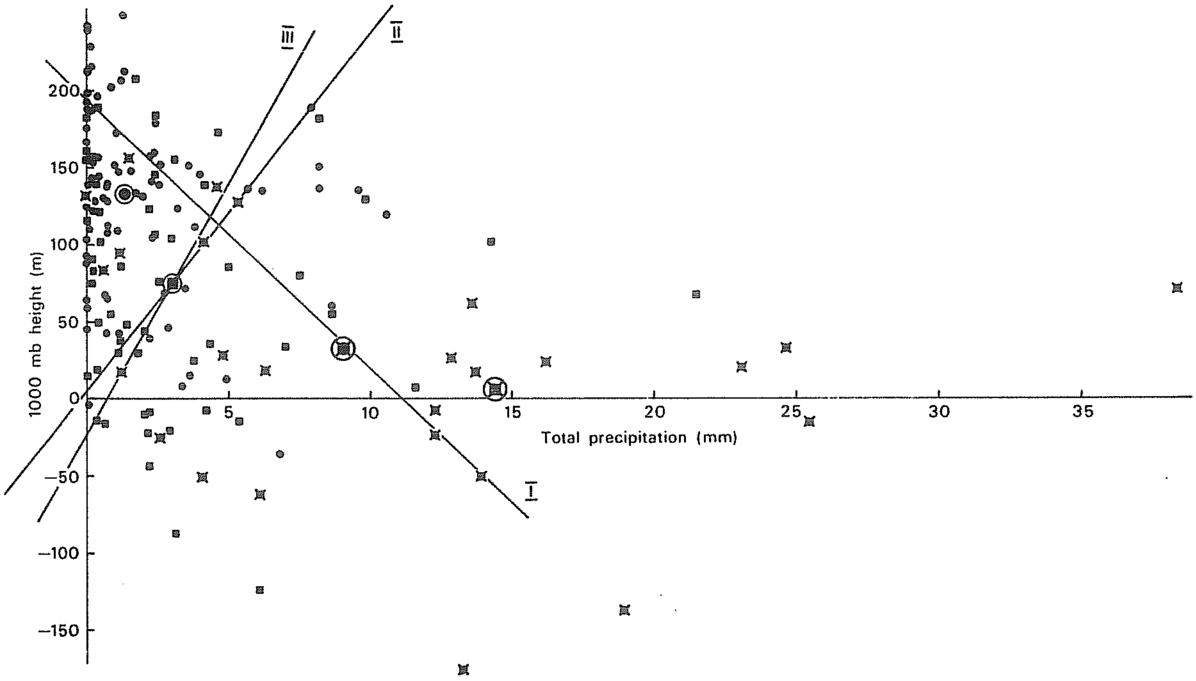

Fig. 3 A realistic example of a scatter plot of a dataset as a function of two predictors. Dots are category 1 events; squares are category 2 events, crosses category 3 events, and crossed squares are category 4 events. The oblique lines are discriminant functions. Encircled symbols are category means.

2.3 Mathematical Procedure - Step I

The mathematical procedure for MDA is described in Miller, 1962 and is covered by Glahn elsewhere in this volume. A brief summary follows. The dataset to be analysed is reorganised and partitioned according to predictand categories. From the reorganised dataset, two matrices are formed, a within-groups dispersion matrix and a between-groups dispersion matrix. The within-groups matrix has the variances of the predictors with respect to the group means on the diagonal, and covariances of predictor pairs, referenced to the group means, as the off-diagonal elements. The between-groups matrix has the variance of the group means for each predictor with respect to the overall sample ("grand") mean on the diagonal, and the covariances of group means for pairs of predictors off the diagonal. The two matrices are symmetric and have the property that their sum is the variance-covariance matrix used in regression analysis. The matrix product $W^{-1}B$ is formed from the within-(W) and between-(B) groups matrices. It is this matrix product that contains the discriminant information for the predictors in the analysis.

The next step is to find the characteristic roots of $W^{-1}B$ by solving the determinantal equation

$$\det(W^{-1}B-\lambda I) = 0 \qquad (1)$$

where I is the identity matrix. The associated eigenvectors are determined by solving the equations

$$(W^{-1}B-\lambda_j I)\bar{V}_j = 0 \qquad (2)$$

As mentioned above, there will be (G-1) eigenvectors where G is the number of groups, unless the number of predictors is smaller in which case the number of eigenvectors will equal the number of predictors. The eigenvectors are the coefficients of the predictors in the discriminant functions, and the associated eigenvalues λ_j are ordered such that the largest is associated with the first, most useful discriminant function, and their relative magnitudes indicate the amount of discriminating power associated with the functions.

2.4 Calculation of Probabilities - Step II

Once the discriminant functions have been obtained, the analysis may be considered to be complete. However, the functions must now be used to calculate probabilities of group membership. Two classes of methods are available, parameteric methods which involve assumptions about the distribution of data in the dependent sample, and non-parametric methods, which allow determination of probabilities directly from the discriminant function values in the dependent sample. Non-parametric methods are impractical for use in operations because the dependent sample, or at least the dependent data transformed to discriminant space must be retained in storage and called up each time a forecast is to be made. For large numbers of MDA equations, the storage requirements quickly become prohibitive. The parametric method such as that described by Miller, 1962, is much more suitable in operations because it requires storage of only a square matrix of within-group dispersions for the dependent sample in discriminant space, vectors of group means for each discriminant function, and a vector of group frequencies of occurrence in the dependent sample. For a typical 5 predictor, 4 group analysis, this amounts to 25 data values.

The two assumptions made in the parametric method are that the within-group dispersions are equal for all groups, and the data is normally distributed within groups. These assumptions are necessary to permit the use of the multivariate normal distribution in the probability computations with parameters estimated from the dependent sample. (Various tests are available for testing for normality. The easiest way might be to examine scatter plots of the data in discriminant space). Should the within-group distributions prove to be highly non-normal, another theoretical distribution could be used with parameters estimated from the dependent data or a transformation used to normalise the data. Non-normality of the data has not appeared to have a serious effect on the generated probabilities in meteorological applications to date.

Once an appropriate theoretical distribution has been established, and its parameters estimated from the dependent sample, the procedure for obtaining probabilities is straightforward: each event is transformed to discriminant space by multiplying predictor values by corresponding coefficients of the discriminant functions. Then individual values of the discriminant functions are converted to distances from the group means. The probabilities are given by Bayes' Rule

$$P(g|x_1----x_t) = \frac{P(x_1---x_t|g)\cdot q_g}{\sum\limits_{g=1}^{G} P(x_1---x_i|g)\cdot q_g} \qquad (g=1,---,G) \qquad (3)$$

where q_g is the a priori probability of the occurrence of category g estimated from the dependent data and $P(x_1---x_t|g)$ represents the probability of occurrence of the discriminant function values x_1---- x_t given that group g occurs. t is the number of discriminant functions used to obtain probabilities and the probability values are obtained from the multivariate normal distribution,

$$P(\bar{x} \ |g) = \frac{|\Sigma^{-1}|^{\frac{1}{2}}}{(2\pi)^{t}/2} \exp\left(-\tfrac{1}{2}(\bar{x}-\bar{\mu})'\Sigma^{-1}(\bar{x}-\bar{\mu})\right) \qquad (4)$$

where $\bar{x} = (x_1---x_t)$ is the vector of t discriminant function values and $\bar{\mu} = (\mu_1---\mu_t)$ is the vector of means of the discriminant functions for group g estimated from the dependent sample, and Σ is the within-group dispersion for the discriminant functions, again estimated from the dependent sample.

3. PREDICTOR SELECTION

3.1 Selection of predictors to offer to statistical procedures - prescreening

In meteorological applications it is normally necessary to select a few predictors from the vast numbers available. The ultimate aim is to include only those predictors that will produce useful forecasts. The procedures whereby a subset of available predictors is selected from the available predictors are generally called "screening". Screening is really a two-stage procedure. In the first stage, which might be called "pre-screening", the meteorologist makes use of his knowledge of the physical relationship of predictors and predictand to determine which predictors should be offered to the statistical program and how they should be modified. The second stage is the objective predictor screening that selects predictors on the basis of the increase in the value of a statistic that represents the quality of the equation for the intended purpose.

Pre-screening is to some extent an interactive procedure. Knowledge of physical relationships can be used at first, but it is always worthwhile to examine scatter plots of predictor vs predictand. These may reveal that a non-linear transformation of the original predictor may provide a better fit to the data. With MDA, where group separation is important, non-linear transformations that spread out group means are useful.

Other factors should be considered when pre-screening. For example model verification studies can provide a lot of information about reliability of a predictor. A predictor may be linearly related to the predictand on the basis of a given data set, but have such a large variance associated with it that it is unreliable for use with independent data. Other examples of unreliable predictors are those which have changed definition during the dependent sample period.

Another factor to consider is the availability of the predictors in operations. After the equations are developed and tested, they are generally implemented to run in real time using real time output of the driving model. It is entirely possible that the dataset available for development and testing may contain more predictors, or at least different predictors, than are usually available on a routine daily basis. Theoretically, any predictors used in development could be made available to the equations in operations, but it might be quite difficult, especially if the equations are to be implemented at a location different from the site where the driving model is run. It is therefore advisable to offer only predictors that are relatively easily obtained in operations.

In summary, there are three major factors to consider in pre-screening:

1. Physical relationships with predictand.
2. Reliability of predictor.
3. Availability of predictors in operations.

It is likely that more improvement in statistical equations can be obtained by careful pre-screening than by simply adding more predictors.

3.2 Objective Screening in MDA

The objective screening procedures available for use in MDA are the same as those available in regression analysis. All objective screening procedures operate by comparing the value of a dependent sample statistic for an equation containing the predictors under test with a standard value or a value obtained from a previous test. The test statistic is designed to measure the goodness of fit of the equation for the intended purpose. For MDA, an appropriate test statistic is the Mahalanobis D^2 value, which measures the overall discriminant power of the predictors under test. The Mahalanobis statistic takes the form

$$Dp^2 = ((n-1) - G.P) \text{ trace } (W^{-1}B) \qquad (5)$$

where W and B are the within and between groups dispersion matrices, n is the number of independent events in the dependent sample, G is the number of categories and P is the number of predictors currently in the equation. If the events in the dependent sample are not independent (e.g. they are serially correlated), the value of n used in the computation should be reduced accordingly.

The Mahalanobis statistic is postulated to have an χ^2 distribution for large n, and it can therefore be used to test the discriminant power of the P predictors by comparing calculated values with tabulated χ^2 values for a specific significance level and P(G-1) degrees of freedom. However, statistical significance does not guarantee meteorological significance, nor does meteorological significance guarantee statistical significance. It is possible, and perhaps even justifiable, therefore, to use simpler methods of determining whether each candidate predictor contributes significantly to the equation.

There are 4 screening procedures available for use in meteorological applications:

1. Optimal
2. Forward selection
3. Forward stepwise
4. Backward elimination

Of these, only the first guarantees that the best equation will be chosen, but does so at the expense of computing time. In the optimal method, all combinations of the predictors are tried, beginning with one predictor, then two predictors and so on. The number of combinations becomes very large very fast, and computer time normally limits the selection to the best 5 predictors. This method can be efficient if the number of candidate predictors is relatively small, but this is usually not the case in meteorological applications.

The backward elimination method is rarely used, because it does not guarantee a better equation than the two forward methods, and is much less efficient.

The two forward methods are popular because they provide a reasonable selection of predictors while operating very efficiently. Forward selection is the simpler of the two methods, but generally does not obtain quite as good an equation. This is the method used at present in the MDA program available at ECMWF. The procedure is as follows:

1. For all predictors offered, calculate the Mahalanobis statistic. Choose as the first predictor the one giving the highest value.

2. Recalculate the Mahalanobis statistic for all predictors in combination with the first. The second predictor is the one that increases the Mahalanobis value by the greatest amount.

3. Repeat step 2 for 3 predictors, 4 predictors etc. until no predictor increases the Mahalanobis value by a significant amount, say 5%.

The threshold percentage (the cutoff criterion) can be set at any suitable value. Higher values result in selection of fewer predictors, lower values result in selection of more predictors. The use of a percentage increase in the Mahalanobis value as a cutoff criterion is a simple way of discontinuing selection. However, it guarantees that selection will always occur, even if no predictors are significant. This may be a disadvantage and it is useful to do a more rigorous statistical test on the final equation. The number of predictors selected is affected not only by the cutoff criterion, but also by the sample size since that is included in the Mahalanobis statistic. If the sample is small, it is useful to raise the cutoff criterion to avoid selecting too many predictors. Our experience suggests that 150 independent events is sufficient to justify a 5% cutoff in the MDA procedure.

If resources are available to test different cutoff criteria, a plot of the Mahalanobis statistic as a function of the number of predictors in the equation, such as figure 4, is useful. Normally, the rate of increase of the Mahalanobis value decreases as more predictors are added, and the best stopping point is soon after the initial strong rise in the Mahalanobis statistic. In the case of figure 4, this would be about 5. Stopping too soon causes rejection of useful discriminant power, and stopping too late causes a better fit, but reduces the reliability of the equation on independent data. Called overfitting, the inclusion of too many predictors is an important problem with small datasets and is diagnosed only by poor performance on independent data.

Forward stepwise selection is similar to forward selection, except that after each step, all predictors are tested for significance as if each is entered last. This "backward look" ensures the removal of any predictor that has been rendered useless by the addition of a highly correlated predictor at a later stage. Forward stepwise selection is considered to be the best selection method (Draper and Smith, 1981), and it is not noticeably less efficient than forward selection. It would probably be worthwhile to convert the MDA program to use forward stepwise screening.

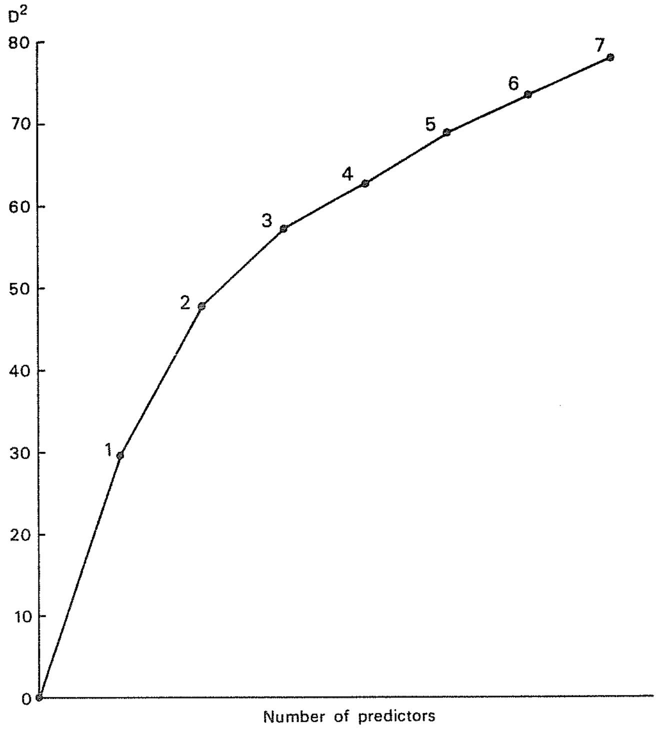

Fig. 4 Plot of the Mahalanobis statistic as a function of the number of predictors in the equation for a typical screening cycle.

4. ASSESSMENT OF PROBABILITY FORECASTS

Verification of probability forecasts is a complete subject by itself. Included here are only the aspects of verification pertaining to the results of the studies on Canadian data, and those aspects which relate to the handling of the output of MDA.

Verification of probability forecasts can be handled in two ways. The probability forecasts can be verified directly as probabilities or the probability output can be converted to categorical forecasts which can then be verified. Verification of probabilities directly is the true measure of the skill of the MDA forecasts (or any other probability forecasts) since conversion to categorical forecasts invariably involves altering the forecasts in some way. Normally, the probability forecasts would be left as probabilities, and the user would convert them to categorical forecasts according to his needs. Classification into categories requires the use of one of many objective "decision rules" or "strategies", which the user is in the best position to formulate according to his needs. However, verification of probability forecasts (example, probability of precipitation amount) for scientific purposes is likely to include a comparison with a direct model output of the same parameter. Comparison of a probability forecast and a continuous variable forecast is only possible if the two are made equivalent by converting both into categorical forecasts.

Conversion of probability forecasts to categorical forecasts for comparison purposes can be done by using simple strategies. One such strategy states "The forecast category is the one with the highest associated forecast probability". This strategy maximises percentage of correct categorical forecasts. Another strategy that can be used is "The forecast category is the one where the probability forecast shows the greatest positive difference from the climatological probability". This rule generally causes overforecasting of rare categories. In the results presented below, the former decision rule has been adopted.

There are very many verification measures available. Those that have been used in assessment of the results on the Canadian data are briefly described below.

A. For probability verification

1. Brier Score (PS)

$$PS = \frac{1}{2N} \sum_{i=1}^{G} \sum_{j=1}^{N} (P_{ij} - O_{ij})^2 \qquad (6)$$

where

P_{ij} is the forecast probability for the i^{th} category and j^{th} event

O_{ij} is the observation vector component for the i^{th} category and j^{th} event (0 or 1)

The Brier score is in fact the mean squared probability error, and is negatively oriented (0 is a perfect score). The Brier score should be used only to verify 2 category (dichotomous) variables.

2. Rank Probability Score (RPS)

$$RPS = 1 - \frac{1}{G-1} \sum_{i=1}^{G} \left(\sum_{j=1}^{i} P_j - \sum_{j=1}^{i} 0_j \right)^2 \qquad (7)$$

where i and j are both category indices.

The Rank probability score is equivalent to the Brier score, but for cumulative probabilities, and is identical to the Brier score for the two-category situation. The RPS is intended for use in multi-category situations where the categories have an implicit ranking, for example a categorised continuous variable such as precipitation amount. The RPS accounts for nearness in the sense that some credit is given for a high probability forecast in an adjacent category to the one which occurred. The RPS is positively oriented (in the form given above) and has a range of 0 to 1.

3. Skill Scores

Skill scores have the general form

$$\text{Skill} = \frac{S_F - S_{STD}}{S_p - S_{STD}} \qquad (8)$$

where S is any score, positively oriented, F refers to the forecast, STD to a standard such as climatology and P refers to a perfect score. A skill score is thus the fractional improvement of the forecast over the standard forecast, using the score S as a measure. Skill scores can be defined for negatively oriented scores as well, but they take a slightly different form. It turns out that a skill score defined in this way using the Brier score is the reduction of variance if the standard forecast is climatology. Both climatology and persistence are frequently used as standards. For medium range forecasting, climatology generally represents a stronger competitor than persistence.

A skill score based on the RPS is

$$\frac{RPS(F) - RPS(C)}{1-RPS(C)} \tag{9}$$

if climatology is used as the standard forecast. The RPS for climatology is the score achieved when climatological probabilities are inserted as the forecast. Skill scores have the range (- ∞ to 1) and tend to be highly variable when the standard forecast is nearly perfect. Positive values represent improvement over the standard, negative values indicate that the standard forecast is better.

It is worthwhile noting that there are two important characteristics that a probability forecast should possess: reliability and sharpness (resolution). The reliability is the accuracy of the forecast, the degree to which the forecast agrees with the actual relative frequency of the event. Reliability can be measured simply by plotting a reliability table, where the sample is divided into categories according to deciles of forecast probability and the actual relative frequency of occurrence of the event is plotted against the forecast probability represented by each category.

Sharpness is the degree to which the forecast probability provides a basis for decision-making, that is, it is the degree to which the probability forecast approaches a categorical forecast. Sharpness can be assessed by partitioning the sample according to the observed category, and calculating the average forecast probabilities for each partition. One should see a high average probability associated with the correct category and low average probabilities associated with the others.

Both the Brier score and RPS measure reliability and sharpness together, and it is possible to partition either score to obtain pure measures of reliability and sharpness.

B. For categorical verification

Contingency tables and associated scores are usually used for this purpose. For a standard 3 category table, the scores are listed below.

Observed

		1	2	3	
FCST	1	a	b	c	J
	2	d	e	f	K
	3	g	h	i	L
		M	N	O	T

J = a+b+c

T = J+K+L = M+N+O

1. Percent correct $= \dfrac{a+e+i}{T}$ (10)

2. Threat score $= \dfrac{a}{J+M-a}, \quad \dfrac{e}{K+N-e}, \quad \dfrac{i}{L+O-i}$ (11)

$$= \frac{\text{number correct}}{\text{number forecast + number observed - number correct}}$$

The threat score is lowered if one tries to over-forecast a rare category to "catch" the occurrences of this category.

3. Bias $= \dfrac{J}{M}, \dfrac{K}{N}, \dfrac{L}{O} = \dfrac{\text{no.forecast}}{\text{no.observed}}$

The Bias provides a check whether categories are forecast with the correct frequency. Perfect bias is 1.

4. Heidke skill score

$$HS = \frac{a+e+i - \frac{JM+KN+LO}{T}}{T - \frac{JM+KN+LO}{T}} \qquad (12)$$

$$= \frac{\text{no. correct - no. correct by chance}}{\text{total - no. correct by chance}}$$

The Heidke skill score is in the same format as the skill scores discussed above, where the standard is chance. The Heidke skill score can easily be reformulated using persistence or climatology as a standard, but it is the "chance" form that is most frequently used.

5. **REEP (Regression Estimation of Event Probabilities)**

REEP (Miller, 1964) is a regression procedure where the predictand is binary. In the usual application, a continuous predictand is categorised and each category is represented as a separate binary predictand, taking the value 1 if the event occurred in that category and 0 if it did not. Each binary predictand is submitted to a stepwise regression to select predictors and separate equations are obtained for all predictands. If the predictand is in 4 exhaustive categories, this means there will be 4 equations, only 3 of which are independent. Application of the equations to individual events produces values usually, but not always within the range 0 to 1. These are interpreted as the probability of occurrence of the category represented by the equation. If probabilities outside the range 0 to 1 are generated, they are renormalised.

For consistency purposes, it is advisable to ensure that all equations for a given predictand contain the same predictors. This implies that all the predictands must be screened at once, with each predictor chosen to be the one that is best for any one of the predictands, then forced into all equations. An alternative selection method is to use MDA to screen for predictors, then force all selected predictors into all of the REEP equations. This method seems to work because REEP and MDA are equivalent procedures. They can be proven to be equivalent for the two-category case, and are thought to be equivalent for multiple categories. Experiments at ECMWF have supported the equivalency in multiple category situations. Predictors selected by MDA were all selected as the first or second in a REEP screening on the same data.

6 CANADIAN EXPERIMENTS IN PROBABILITY OF PRECIPITATION AMOUNT (POPA) FORECASTING

6.1 Background of the project

The Canadian Weather Service includes as major components a central office in Montreal and 7 major regional weather offices. The central office is responsible for running the operational NWP model (a spectral model) and sending the output to all major regional offices. The central office, like ECMWF, prepares no products for end-users and issues no forecasts directly to end-users. The major regional offices have this responsibility, each forecasting for an area the size of France. They are large operations in themselves, employing 30 to 40 meteorologists each, and operating a mini-computer. Development of statistical products is done in Montreal and in Toronto at the headquarters. These products are generally intended for implementation centrally, that is, at Montreal, since the driving model is run there. However, the regional offices have sufficient computing power to do local studies of their own, some of which follow statistical procedures discussed in this volume.

The POPA project began with a request from a regional office to supply probability forecasts to assist in agriculture forecasting. As this was the first such request, it was treated as an opportunity to test and evaluate several forecast methods. An analog method was implemented first, because it was the simplest, and tests continued on MDA and REEP-based procedures. After tests demonstrated the superiority of REEP and MDA to the analog method for at least the first 48 hours, the operational method was changed to a combination of the MDA and REEP forecasts. The reasons for adopting a combination of the two methods are discussed below. The REEP and MDA procedure has been operational since December 22, 1980. Forecasts are issued from the central office twice a day for 74 Canadian stations.

6.2 Method

The predictand is probability of precipitation amount, in percent for 12 hour precipitation accumulation. Forecasts are for four categories:

1. No precip: a trace or less
2. Light : 0.2 to 2.0 mm
3. Moderate : 2.0 to 10.0 mm
4. Heavy : > 10 mm

The four categories are exhaustive and therefore each set of probabilities sums to 1. The predictand is valid for a single station only.

The development sample consisted of 10 years of upper air analysed data and surface reported data. For the POPA equations, upper air analysed and derived predictors were offered for screening. The 68 predictors (34 at the beginning of the 12 hour valid period and 34 at the end) are listed in table 1. Separate equations were derived for 00Z and 12Z runs and for each of 4 seasons, Dec. to Feb., March to May, June to August and Sept. to Nov. Total sample sizes for each equation averaged about 500 cases.

Both REEP and MDA methods followed the "perfect prog" approach, whereby concurrent observed (or analysed) predictor data is used for equation development, and predicted values of the predictors are obtained from NWP model output to produce forecasts of the predictand. With the POPA equations for example, forecasts of precipitation amount probability for the period 12 to 24 hours ahead are made using spectral model output parameters valid 12 and 24 hours ahead.

The development procedure is shown schematically in figure 5.

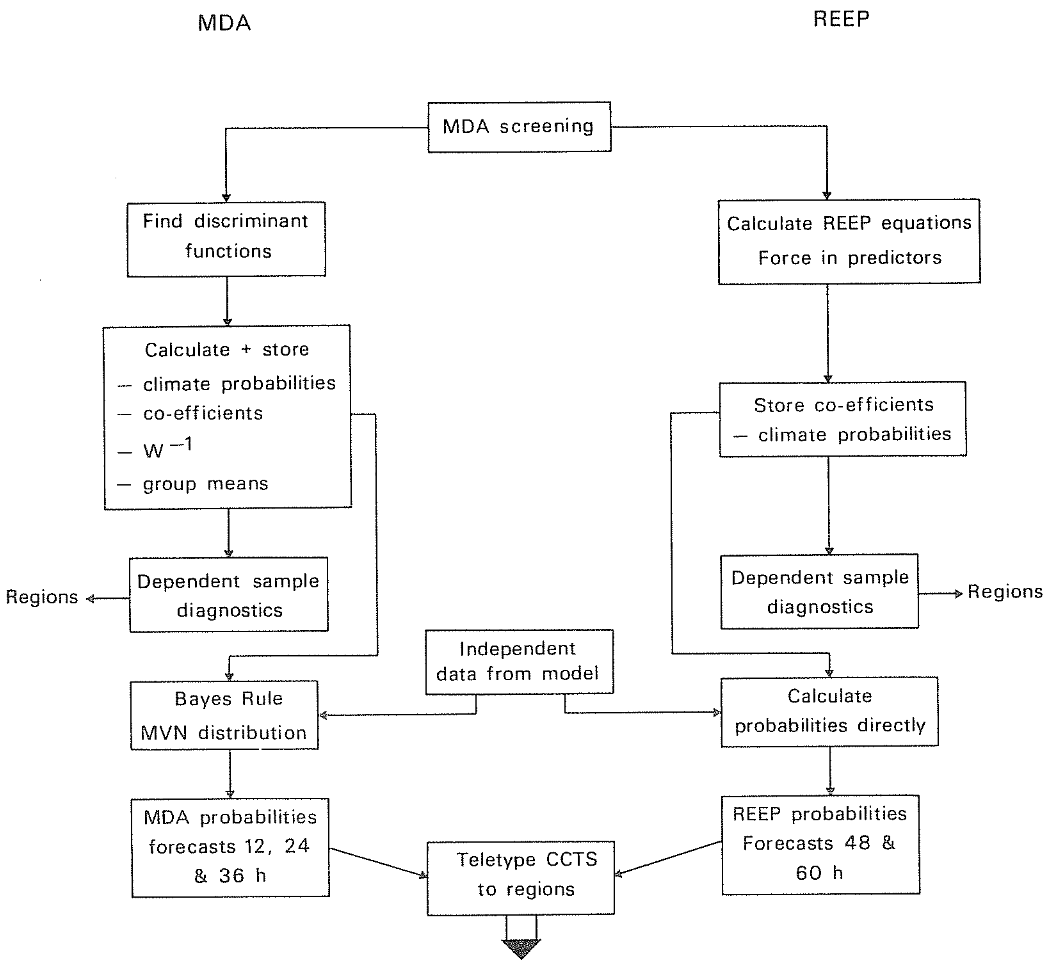

Fig. 5 Schematic representation of MDA and REEP equation development and implementation in Canada.

TABLE 1 POTENTIAL PREDICTORS

1.	Z1000
2.	Z850
3.	Z700
4.	Z500
5.	T850
6.	T700
7.	T500
8.	850(T-TD)
9.	700(T-TD)
10.	500(T-TD)
11.	1000 MB UG
12.	1000 MB VG
13.	850 MB UG
14.	850 MB VG
15.	700 MB UG
16.	700 MB VG
17.	500 MB UG
18.	500 MB VG
19.	500 MB ABS. VORT.
20.	1000-500 THICKN. ADV.
21.	500 MB ABS.VORT.ADV.
22.	850 MB TEMP.ADV.
23.	500 MB TEMP.ADV.
24.	1000 MB RELAT.VORT.
25.	GEORGE K INDEX
26.	TOTAL TOTALS INDEX
27.	SHOWALTER INDEX
28.	850 MB DELTA T E-W
29.	850 MB DELTA T N-S
30.	850 MB DELTA (T-TD) E-W
31.	850 MB DELTA (T-TD) N-S
32.	700 MB DELTA (T-TD) E-W
33.	700 MB DELTA (T-TD) N-S
34.	500 MB DIFFLUENCE

The data was first stratified into seasons, hours of the day (00 and 12Z) and by predictand category. The predictor set was then screened using the multiple discriminant analysis screening procedure. The predictor sets thus determined were forced into equations for both techniques, coefficients were calculated and stored. Both techniques were applied to the dependent sample to generate diagnostics which were then sent to users in regional offices on request. With the coefficients for all 592 equations for each technique stored, attention turned to tests on independent data. For this purpose, model output had to be treated exactly the same way as the development data for the purpose of deriving predictors. The largest test on independent data was conducted on Ontario stations data and is described below.

The operational form of the equations uses a set of derivation programs to calculate the necessary predictor values from the model output, then reads the necessary coefficients from a file. Forecasts are produced daily for both techniques and stored for later verification, but only one set of forecasts is transmitted to regions.

The predictors selected for the various stations showed considerable consistency in space even though each station was treated separately. East of the Ontario-Manitoba border, 98% of the equations contained the 1000 mb geostrophic relative vorticity at the end of the valid period as the first predictor. For this area, significant precipitation usually occurs ahead of a moving synoptic storm. For the Prairie provinces, over 50% of the equations contain a low level geostrophic U-velocity as a predictor. That represents the tendency of rain to occur with extended periods of upslope (easterly) winds. For British Columbia, a mountainous area, a stability index was frequently chosen, signalling the importance of orographically induced convective precipitation.

6.3 Verification Results

In this section, highlights of the comparative verification of the methods is shown. The techniques were run in an operationally realistic fashion, (i.e. using prognostic predictor data) and 5 months of forecasts for 12 Ontario stations were verified in several ways. Three aspects of the verification are shown: Average score values for the entire dependent data set, contingency tables for the entire sample, and two individual cases.

6.3.1 Scores

Figure 6 shows the skill score as a function of forecast projection time for all 3 methods. The skill score used is based on the Rank probability score, and is referenced to climatology (eq. 9). Climatology in this case is the estimated frequencies of occurrence for the 10 year dependent sample.

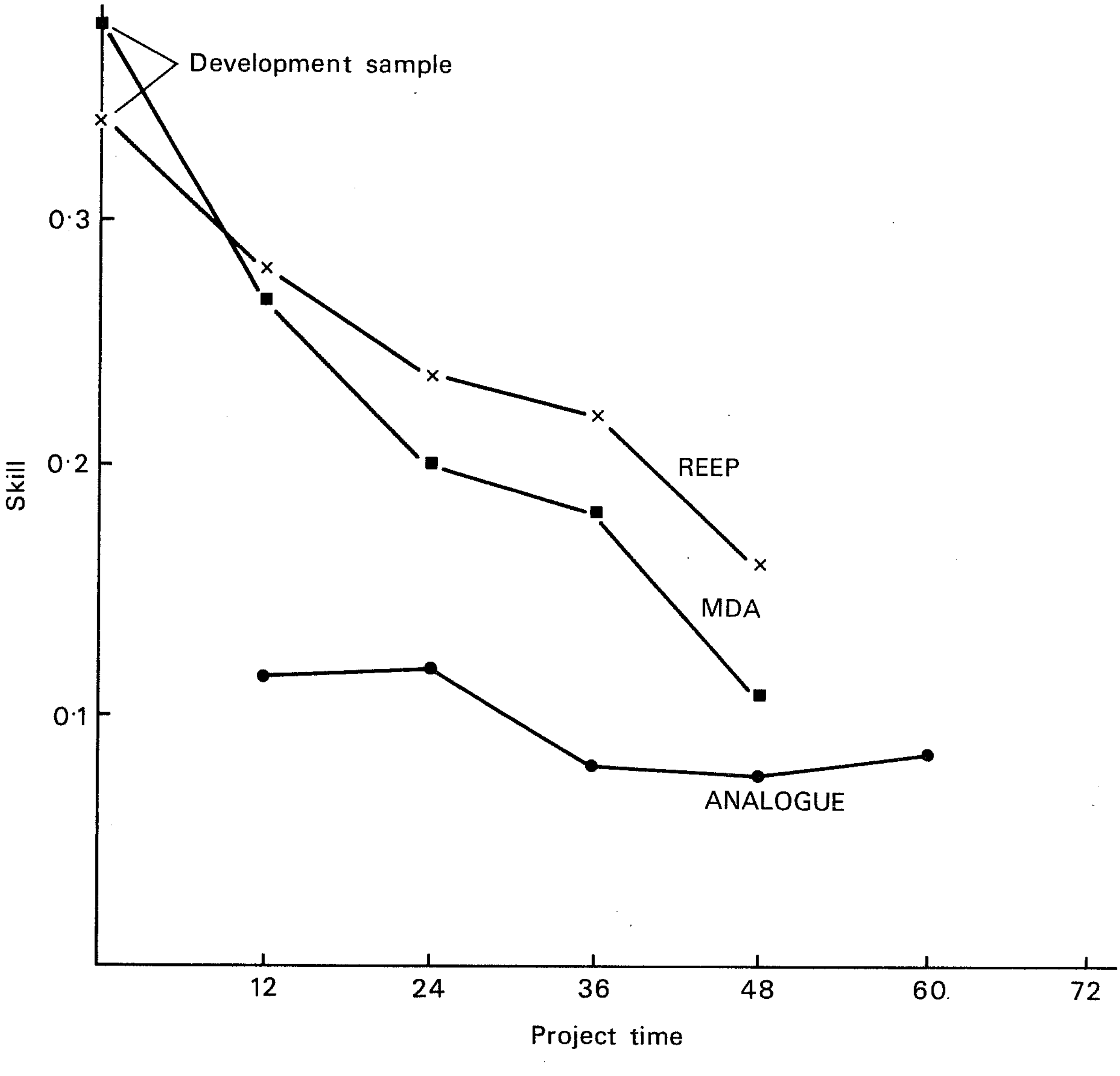

Fig. 6 For all stations skill versus projection time for winter and spring.

Figure 6 shows superiority for MDA and REEP over the analog technique out to at least 48 hours, and the expected loss of quality with increasing projection time shows up. The analog forecasts, however, do not degrade significantly with increasing projection time. We have attributed this effect to the analog method's inability to catch smaller synoptic scale storms - the analogs chosen are dominated by the long wave structure.

On average, REEP outperforms MDA slightly despite MDA's better fit to the dependent data. As shown below, REEP is a conservative method (forecasts towards climatology and predictions are not sharp)while MDA often produced sharp forecasts.
A conservative method generally shows a better score when a quadratic scoring rule is used because sharp forecasts are heavily penalised when wrong. For this reason, average score values do not necessarily give all the information required to choose the "best" procedure.

6.3.2 Contingency tables

Figure 7 shows contingency tables for the REEP and MDA methods for the entire 5 month independent sample of 3528 forecasts. The tables were generated by choosing as the predicted category the category with the highest forecast probability.
The following points are noteworthy:

(i) REEP is conservative - never forecasts a heavy precipitation case, despite 75 occurrences. MDA, however, forecasts 20 correctly and 37 close.

(ii) Both methods overforecast no precipitation cases (bias greater than 1) and underforecast precipitation cases. MDA forecasts have the smaller biasing problem. It should be noted that bias problems can be largely overcome by various other methods of choosing categorical forecasts, all of which add complexity to the procedure and make its interpretation more difficult. Perfect bias is bias=1.0, the event is forecast as often as it is observed.

		REEP Observed Group					MDA Observed Group				
		1	2	3	4		1	2	3	4	
FORECAST	1	2241	638	203	28	3110	2156	540	133	10	2839
	2	60	101	99	16	276	100	145	88	8	341
	3	15	25	71	31	142	54	65	123	37	279
	4	0	0	0	0	0	6	14	29	20	69
		2316	764	373	75	3528	2316	764	373	75	3528

REEP: Percent correct = .684

Biases 1.34 .361 .381 .00

MDA: Percent correct = .693

Biases 1.23 .446 .748 .92

Fig. 7 Contingency tables for a 12 hour forecast for all 12 stations for winter and spring.

(iii) REEP achieves a large number of correct no precipitation cases, but MDA achieves a high total number correct, 2444vs 2413, despite the slight superiority of REEP at 12 hours in the skill score. If it is considered that correct precipitation forecasts are more important than correct no precipitation forecasts (precipitation events are less common), the superiority of MDA becomes more significant.

6.3.3 Case examples

The case example shown in figure 8 shows that considerably different probability sets are produced by the two methods, based on the same set of spectral model forecast data. The case is for a 36 hour period, Ottawa equations, February 22, 12Z to February 24, 00Z. For each of the 3 valid periods and each technique, four sets of probabilities are listed, generated from 4 spectral prog. issues valid at that valid period. In this case, a storm was approaching Ontario from the southwest and the earlier progs. (36 and 48 hours) were slow, causing both MDA and REEP to predict significant precipitation in the 2nd and 3rd periods. As the verifying time approached, the progs indicated precipitation in the first and second periods mostly and both MDA and REEP responded. MDA responded more clearly and sharply to the change but REEP also showed the general trend.

MODEL	PROJECTION (hrs)	FORECAST PERIOD											
		FEB. 22nd 12-24 GMT				FEB. 23rd 00-12 GMT				FEB.23rd 12-24 GMT			
		CAT. 3 OBSERVED				CAT. 4 OBSERVED				CAT. 1 OBSERVED			
M	12	.22	.22	.40	.15	.04	.16	.65	.16	.88	.11	.01	.00
D	24	.35	.26	.34	.05	.01	.09	.52	.38	.81	.17	.02	.00
A	36	.74	.17	.09	.00	.00	.00	.08	.92	.71	.24	.05	.00
	48	.93	.04	.02	.00	.00	.00	.04	.96	.25	.38	.34	.03
	Category	1	2	3	4	1	2	3	4	1	2	3	4
R	12	.41	.18	.26	.15	.19	.29	.38	.14	.79	.15	.06	.00
E	24	.46	.18	.24	.13	.12	.31	.40	.17	.71	.20	.08	.01
E	36	.65	.12	.16	.08	.00	.33	.46	.21	.63	.22	.12	.03
P	48	.81	.00	.08	.11	.00	.27	.49	.24	.37	.27	.24	.11

Fig. 8 Case study period 12 GMT 22 February to 00 GMT 24 February at Ottawa.

Other notable points made by figure 8 are:

(i) The sharpness of the MDA output. For example, the 48 hour forecast for the second period indicates a 96% chance of heavy precipitation.

(ii) The tendency of REEP to forecast either no precipitation or category 3 (moderate). An explanation for this is that the physics separating rain events from no rain events is not necessarily the same as the physics which determines the amount of rain. Because the sample is dominated by no rain cases, both REEP and MDA will bias themselves in predictor selection toward the rain-no rain problem, and there will be a tendency to lump the rain predictions into the middle of the range. We tried a straight POP prediction followed by a conditional POPA prediction, but found the extra effort did not produce any better results. Again, techniques are available for choosing "best" categories, based on comparisons with climatology.

Another case example is shown in figures 9 to 11. Figure 9 is an analysis of the forecast probabilities for the MDA output, 24 hour forecast for April 14, 1981 at 12 GMT, figure 10 is the same for the REEP forecasts and figure 11 is an analysis of the verifying precipitation amounts. On the forecast charts, two sets of lines are shown: the solid lines give the forecast probability of precipitation calculated by adding the probabilities of the three precipitation categories. The dashed lines give the forecast probabilities of the extreme event, category 4, The contours on the verifying analysis are selected according to the three category thresholds, 0.2, 2.0 and 10.0 mm. It can be seen from this case that the probabilities are generally consistent in space, partly due to spatial consistency in the model forecasts, and partly due to the fact that similar predictors appear in the equations for nearby stations

The sharpness of the MDA forecasts is again evident, with over 90% probability of category 4 forecast at several stations in southern Ontario. It can be seen also that the general structure of the observed precipitation area has been caught, the only major error being the partial miss at Moosonee on James Bay. This case and others that have been plotted and analysed show that an important effect of the statistical forecasts is to superimpose local precipitation climatology on whatever synoptic scale precipitation is suggested by the model.

6.4 Comments on the results

When these results were presented to regional meteorologists, a preference was indicated for the sharper MDA forecasts. It is, however, evident from the results that the sharper MDA forecasts result in lower average skill scores, due to the existence of occasional large errors in the probability forecasts. The MDA forecasts therefore must be carefully assessed together with the model output which produced them. The MDA forecasts would be unsuitable for situations where assessments would

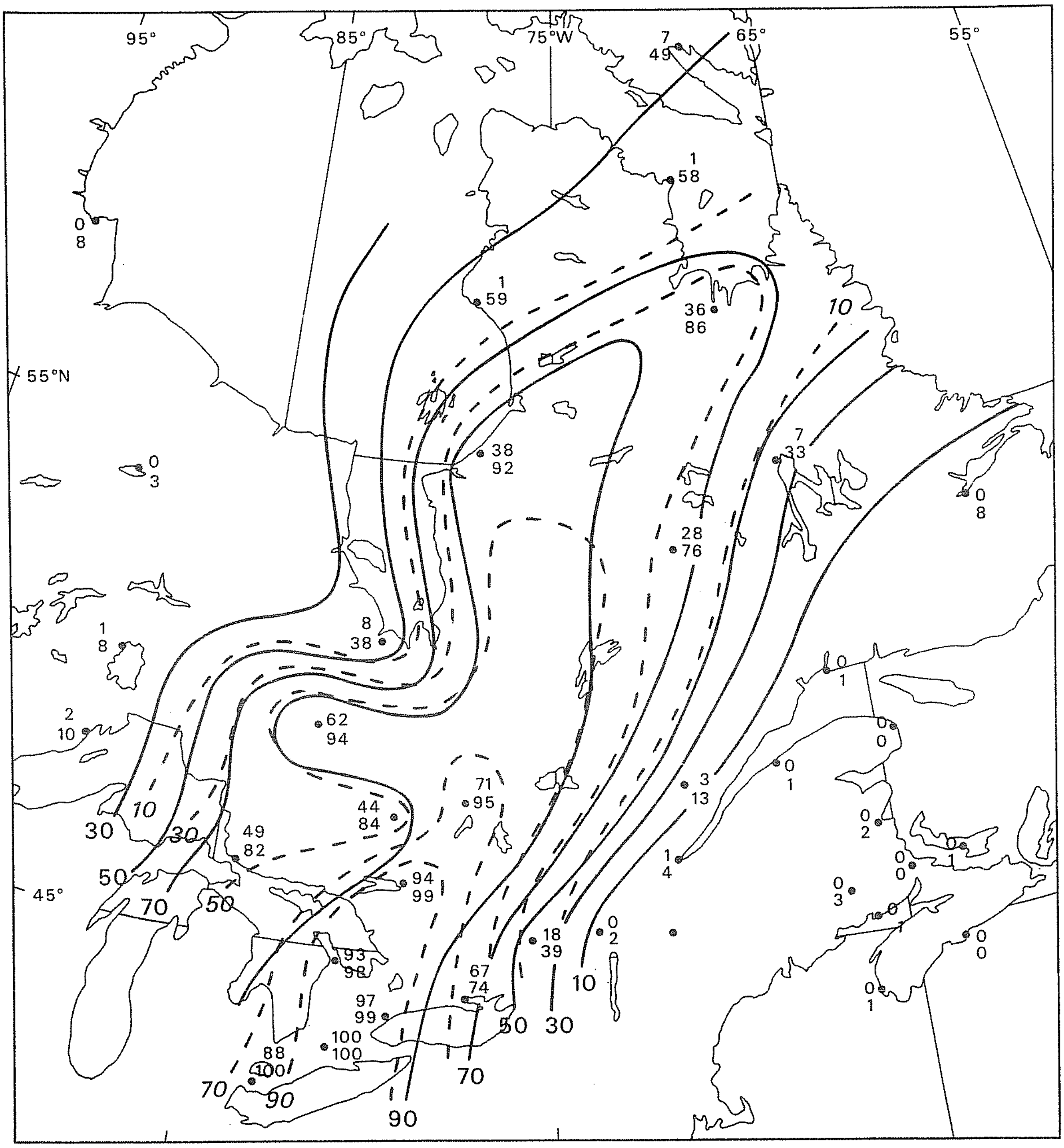

Fig. 9 Case example of MDA forecasts. Contour interval is 20%. Dashed lines are probabilities of category 4 occurrence and solid lines are probability of precipitation.

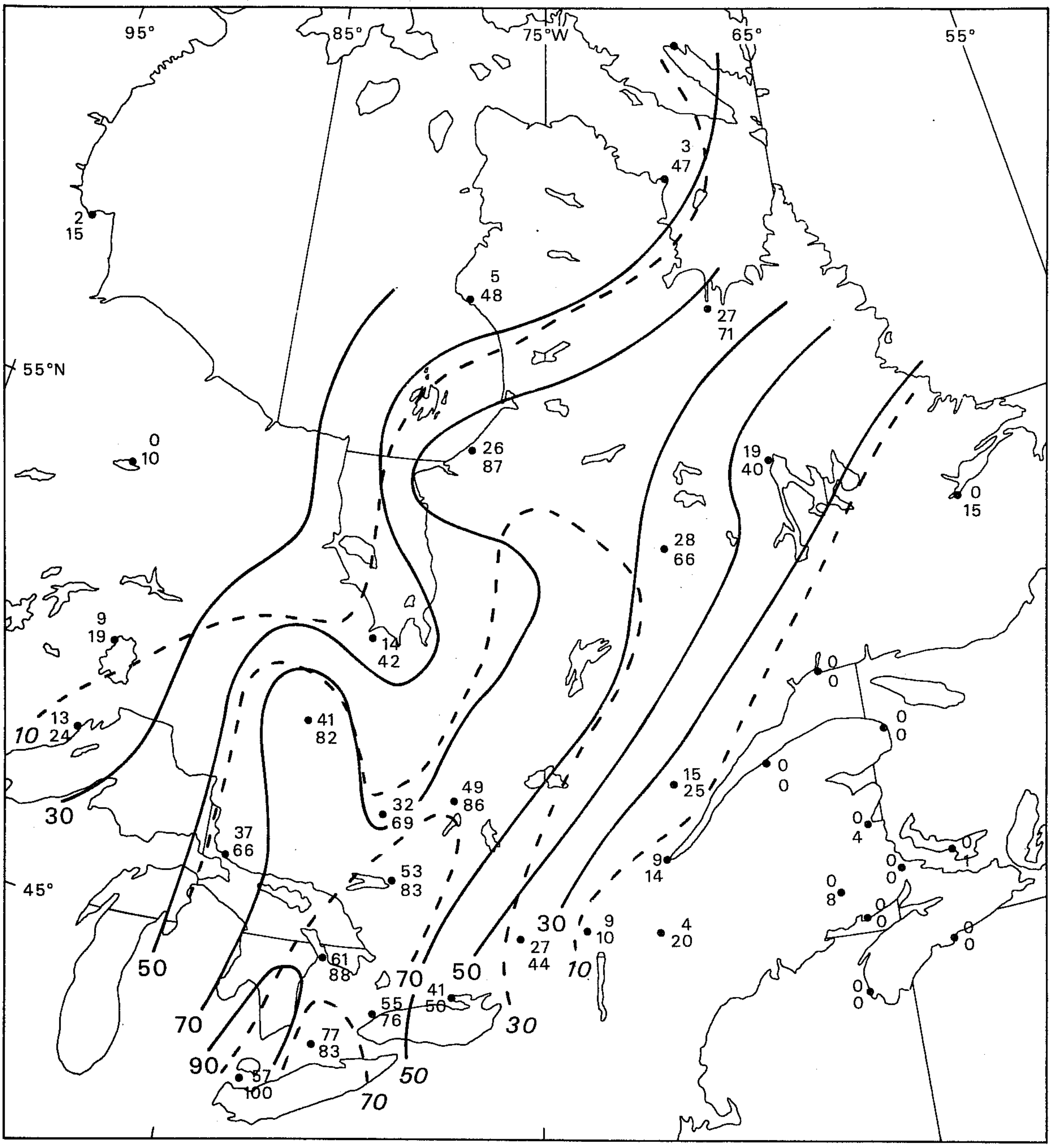

Fig. 10 As for Fig. 9 but for REEP forecasts.

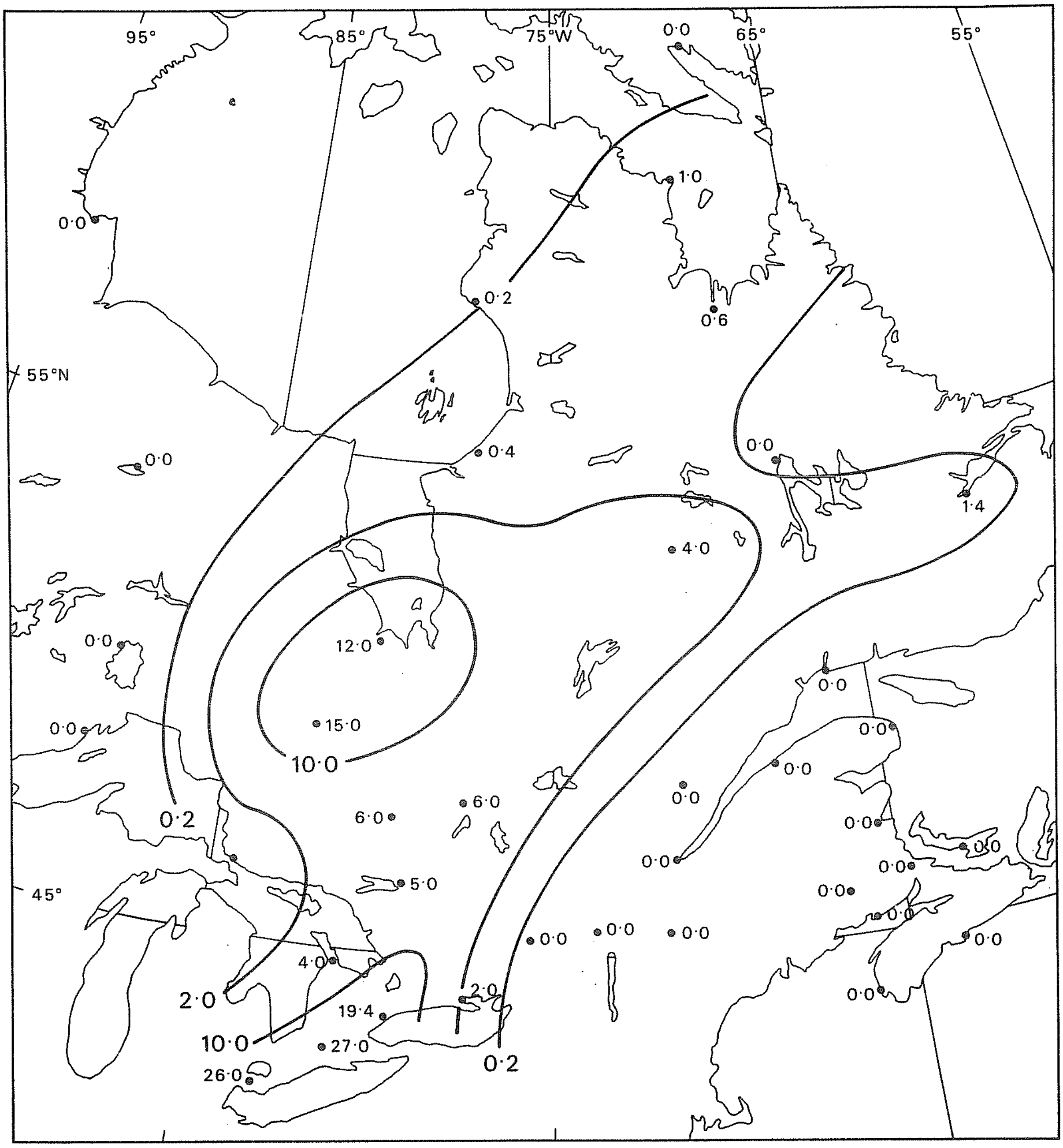

Fig. 11 Verifying precipitation amounts, 12 hours ending at 12 GMT 14 April 1981. Contours are drawn category thresholds.

not be done, for example, the longer range period. As our forecasters concentrate their effors on the short range period out to 36 hours, we elected to supply the MDA forecasts out to 36 hours, opting for the smoother REEP forecasts for 48 and 60 hours.

The general philosophy was to provide a "transparent" forecast - one which is a clear interpretation of the model output. As our forecasters also receive the model output used in the POPA forecasts, they would be in a position to assess both and at least subjectively modify the probabilities if desired. This philosophy also affected the pre-screening of predictors. The predictor sets were limited to familiar quantities such as vorticity advection, that the forecasters would be in a position to assess more confidently.

Use of the perfect prog method is to some extent hazardous for probability forecasting. The method always leads to forecasts over the complete range of probabilities regardless of the skill of the driving model. This means that forecasts are likely to become unreliable if the model skill decreases, unless the model is damped towards climatology as projection time increases. The unreliability shows up mainly as an overforecasting of high probabilities and an underforecasting of low probabilities. The MOS method on the other hand, produces reliable forecasts by restricting the range of forecast probabilities as the model skill decreases. This occurs because both MDA and REEP will predict more towards the mean (climatological) probability as the statistical relationship becomes weaker. MOS therefore obtains reliability at the expense of sharpness, and supplies an accurate (reliable) but nearly useless forecast (in comparison to climatology) at longer forecast ranges. Such a method will score well in the sense that any predictive information available in the model should show up in the statistical forecast as an improvement over climatology, however slight.

Perfect prog methods sacrifice reliability to obtain sharpness. Such a method will not score well if model accuracy is low, especially with a quadratic scoring rule, but may be useful in situations where pure model interpretation is desired. The choice of perfect prog or MOS, for probability forecasts, must be made in consideration of the user. If the product is not for assessment and modification by meteorologists, MOS is undoubtedly the preferred choice with an inaccurate model. If the product will be assessed and modified by meteorologists, perhaps the pure model interpretation provided by the perfect prog. method will be preferred.

7. REFERENCES

Draper, N.R. and H. Smith, 1981: Applied Regression Analysis, Second Edition, Wiley and Sons, Inc., New York, New York, p.308.

Miller, R.G., 1962: Statistical Prediction by Discriminant Analysis Met. Monograph Vol. 4 No. 25, American Met. Soc., Boston, Mass. 54pp.

Miller, R.G., 1964: Regression Estimation of Event Probabilities, Tec. Rep. 7411-121 Contract cwb-10704, Travellers Center, Inc. Hartford, Conn.

Wilson, L.J. and N. Yacowar, 1980: Statistical Weather Element Forecasting in the Canadian Weather Service. Proceedings of the WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting, Nice, France, September 8-12, 1980.