INTERNATIONAL ATOMIC ENERGY AGENCY
UNITED NATIONS EDUCATIONAL, SCIENTIFIC AND CULTURAL ORGANIZATION
INTERNATIONAL CENTRE FOR THEORETICAL PHYSICS
I.C.T.P., P.O. BOX 586, 34100 TRIESTE, ITALY, CABLE: CENTRATOM TRIESTE

SMR/534-8

ICTP/WMO WORKSHOP ON EXTRA-TROPICAL AND TROPICAL
LIMITED AREA MODELLING
22 October - 3 November1990

"Statistical Weather Forecasting"

by Harry R. Glahn

Presented by:
C. DE SIMONE
Servizio Meteorologico dell'Aeronautica
Rome, Italy

8
Statistical Weather Forecasting

Harry R. Glahn

1. INTRODUCTION

Statistical weather forecasting, in its broadest sense, has undoubtedly been practiced for thousands of years. All that is necessary is for someone to collect some data, someone to process it, and someone to use the results to make a forecast. Ancient man, seeing a dark cloud approaching and thinking that rain was likely, would be practicing statistical weather forecasting even if he had no knowledge of the physical processes involved. However, in this chapter we use the term statistical weather forecasting to mean forecasting through the use of a formal statistical analysis of the data, with the results of that analysis being clearly stated.

Statistical forecasting is a branch of objective weather forecasting, the other branch being numerical weather prediction. Allen and Vernon (1951) have defined an objective forecast as "... a forecast which does not depend for its accuracy upon the forecasting experience or the subjective judgment of the meteorologist using it. Strictly speaking, an objective system is one which can produce one and only one forecast from a specific set of data." (Subjective judgment is, of course, used in the development of the system.) Occasionally, these restrictions are relaxed slightly or some subjectivity may enter into the definition of "a specific set of data." For instance, an observation of temperature at a certain location may be needed as input to an objective scheme, and this observation may not be available for some reason. It may, then, have to be estimated from other data. Even though this estimate is made subjectively and requires skill on the part of the meteorologist, the forecast would probably still be called objective.

Some statistical techniques are very simple, whereas other procedures are more complicated. Various forms of scatter diagrams and histograms fall into the first category. Discriminant analysis and logit analysis are examples of the latter category.

In the early years of operational numerical weather prediction, competition rather than cooperation dominated the relationship between those individuals engaged in developing statistical models and those researchers concerned with developing numerical models. Each group thought that its approach was the best way to proceed and that the other branch of objective weather prediction was not necessary. Even though the barriers between the two groups have not yet vanished, each group has become much more tolerant of the other group's viewpoint. Statistical modelers now use the results from (rather than compete with) numerical models, and numerical modelers recognize the usefulness of properly applied statistical procedures.

In this chapter, we will review the three general methods of application of statistical models and describe the statistical techniques that have been applied to weather prediction. Emphasis will be placed on those techniques that have been used operationally. Other discussions of statistical models employed in objective weather forecasting can be found in Allen and Vernon (1951), Gringorten (1955), Panofsky and Brier (1958), U. S. Navy (1963), Glahn (1965), and Miller (1977).

2. METHODS OF APPLICATION

2.1 Classical Method

Before the days of numerical models, statistical techniques necessarily incorporated the time lag. That is, if one wanted to develop a scheme for forecasting the maximum (max) temperature for tomorrow, the input would consist only of observational data available at the time that the forecast was to be made. This situation can be expressed as

$$\hat{Y}_t = f_1(\underline{X}_0), \tag{1}$$

where $\hat{Y}_t$ is the estimate (forecast) of the predictand (dependent variable) Y at time t and $\underline{X}_0$ is a vector of observational data (independent variables) at time 0. (The observations are not necessarily all made at time 0 but must be available at that time.) This technique has become known as the "classical" approach for lack of a better name (Klein, 1969). In application, the input is the same as in development.

2.2 Perfect Prog Method

As numerical models were implemented and improved, it was recognized that their output must be exploited to the greatest possible extent. However, these models did not predict many of the weather variables with which users were concerned - for instance, max

temperature. This situation led to the development of the perfect prog (prog for prognostic) technique (Klein et al., 1959).

A concurrent relationship between the predictand variable and the predictor variables is developed, which can be expressed as

$$\hat{Y}_0 = f_2(\underline{X}_0), \tag{2}$$

where $\hat{Y}_0$ is the estimate of the predictand Y at time 0 and $\underline{X}_0$ is a vector of observations of variables that can be predicted by numerical models. The time relationship need not be exactly concurrent, but it is much more nearly so than in the classical technique. Even though $\hat{Y}_0$ is an estimate, it is not a "forecast" in the sense of "looking ahead"; it is more appropriately called a "specification."

In application, $\underline{\hat{X}}_t$ is inserted into Eq. (2) to provide a forecast $\hat{Y}_t$:

$$\hat{Y}_t = f_2(\underline{\hat{X}}_t). \tag{3}$$

The vector $\underline{\hat{X}}_t$ is obtained from numerical model output. This approach assumes that the model output is "perfect" (hence, the name "perfect prog").

2.3 Model Output Statistics Method

Although the perfect prog technique makes use of numerical model output, it is not necessarily true that the statistical relationship between Y and $\underline{X}$ at time 0 is the best relationship for time t when $\underline{X}_t$ is estimated by numerical models as in Eq. (3). In order to overcome this problem, the model output statistics (MOS) technique was developed (Glahn and Lowry, 1972). In this approach, a sample of model output is collected and a statistical relationship is developed, which can be expressed as

$$\hat{Y}_t = f_3(\underline{\hat{X}}_t), \tag{4}$$

where $\hat{Y}_t$ is the estimate of the predictand Y at time t and $\underline{\hat{X}}_t$ is a vector of forecasts from numerical models. The numerical model predictions $\underline{X}_t$ need not be limited to time t but could be valid either before or after time t; however, the projection times of the different variables will usually be grouped around t. In application, Eq. (4) is used as developed.

2.4 Comparison of Classical, Perfect Prog, and MOS Techniques

Table 1 summarizes the development and application aspects of the three techniques. Since the classical technique does not depend on numerical models, it is most useful for very short-range forecasting. The strength of most numerical models lies in predicting events several hours to a few days in advance. For predictions of up to 4 hours, say, simple statistical models and even persistence may be quite good in comparison to numerical models or statistical forecasts derived from them. The classical technique is relatively simple to use, observations are usually abundant for model development, and there is no dependence on a numerical model to complicate the application.

For many purposes, the perfect prog technique gives quite good results. Since Eq. (2) is based entirely on observations (or simple calculations made from them), a large data sample usually can be obtained to ensure a stable relationship. (A stable relationship is one that will give similar results on dependent and independent data.) The availability of observations also may allow useful stratifications of the data. That is, different relationships can be developed for different months of the year, hours of the day, etc. In addition, as numerical models become more accurate, forecasts based on Eq. (3) will improve even without redevelopment of the functional relationship f_2.

For medium range forecasting, MOS is the best technique if (a) a sufficient sample of model output can be obtained for development and (b) the model does not undergo major changes. Use of MOS usually requires more planning than the other techniques because the model output desired may not be saved without special arrangements. The major disadvantage is that a relationship f_3 developed for one model may not hold for another model. Therefore, if the operational model is changed substantially, a new relationship should be developed. This redevelopment can be done only after the new model has been used for a long enough period to obtain an adequate data sample. At the time that this chapter was written, changes in the National Weather Service (NWS) models being employed by the National Meteorological Center (NMC) have not presented serious problems in MOS applications.

Changes in numerical models that might materially affect MOS applications could be any of three types: (a) the model produces the same output variables, but the overall skill is higher; (b) the model produces the same output variables and the overall skill is about the same, but the error characteristics are different; or (c) the model produces different output variables with or without an increase in skill. (We assume that a new model would not be implemented if the skill level was below that of the old model.) In the first case, use of the new model would probably decrease the skill of MOS forecasts slightly (without redevelopment) unless the model skill was increased

Table 1. Development and application equations for the classical, perfect prog, and MOS techniques.

Technique	Development equation	Application equation
Classical	$\hat{Y}_t = f_1(\underline{X}_0)^a$	
Perfect prog	$\hat{Y}_0 = f_2(\underline{X}_0)$	$\hat{Y}_t = f_2(\underline{\hat{X}}_t)$
MOS	$\hat{Y}_t = f_3(\underline{\hat{X}}_t)^a$	

[a]Development and application equations are identical.

considerably. If the skill did increase markedly, then the MOS skill would probably also improve. For a new model with equal skill but different error characteristics, the MOS skill would undoubtedly decrease. If the new model didn't produce the same output variables, the old variables would have to be estimated (by interpolation or other computations from the new variables). A decrease in MOS skill would likely result unless the new model was considerably more skillful.

Even with these potential problems, it is likely that MOS will be used operationally more than perfect prog and will produce better forecasts for many years to come. At some point, some of the applications may shift to perfect prog. However, when the predictand is a dichotomous event (e.g., precipitation/no precipitation) and the statistical relationship estimates the probability of that event, MOS will always be superior to perfect prog. MOS incorporates the inaccuracies of the numerical model, and as the skill becomes small for large projections (i.e., long lead times) the estimated probability will approach climatology. Perfect prog will not give this result; that is, the possible range of predictions in the perfect prog approach is just as great for long-time projections (say 5 days) as for short-time projections (say 12 hours), unless the numerical model itself becomes much smoother with time and perhaps approaches climatology. Therefore, perfect prog probabilities will not be reliable (i.e., will not correspond to observed relative frequencies). Figure 1 shows schematically the relationship between MOS and perfect prog probabilities as a function of projection.

Although relatively little experience to date has been obtained in applying Eq. (4) to a model other than that on which it was developed, the evidence available suggests that the decrease in skill is minimal. Major operational models share many of the same characteristic errors - incorrect phase of systems at long projections, missed cyclogenesis, etc. The forecasts prepared by two models frequently look more like one another than either looks like reality. That is, major error characteristics for different models are similar. As long as this situation exists, relationships developed on one model can be applied to another model without major loss in skill.

All of the statistical models presented in the following sections, such as scatter diagrams or regression, can be used with any of the techniques discussed above. In the following sections, an estimate of a predictand may at times be called a "forecast" even though no time projection is actually involved.

3. HISTOGRAMS

Perhaps the simplest statistical model one might apply is the histogram. Figure 2 shows the relative frequency of frozen precipitation at Salt Lake City as a function of the 1000-500 mb thickness

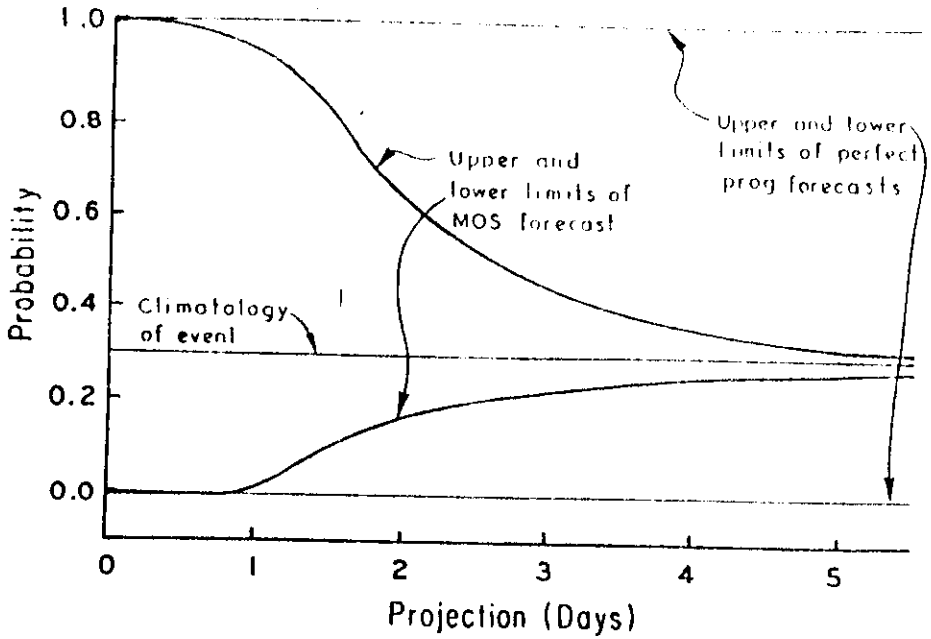

Figure 1. Schematic diagram of upper and lower limits of MOS and perfect prog forecasts as a function of projection.

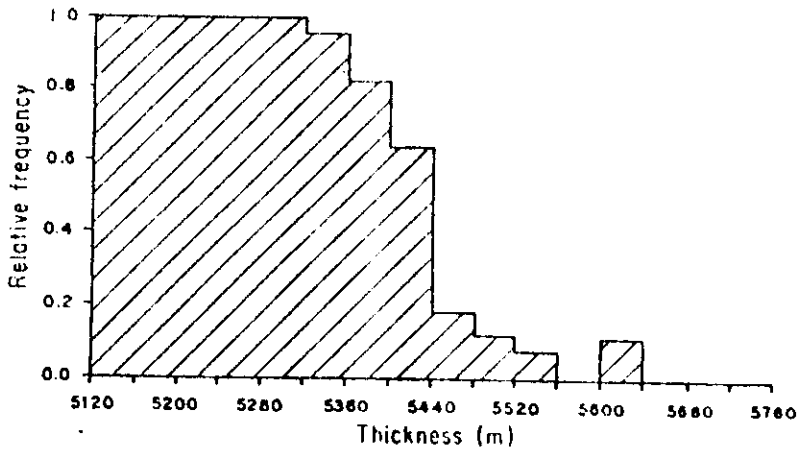

Figure 2. Relative frequency of frozen precipitation at Salt Lake City, Utah, as a function of forecast 1000-500 mb thickness (after Glahn and Bocchieri, 1975).

forecast by the primitive equation (PE) model (Shuman and Hovermale, 1968). The relative frequencies have been calculated for 40 m intervals. The interval must be wide enough to encompass several cases in the region of greatest concern and yet be small enough to give sufficient detail to be useful. The intervals need not be the same width.

The histogram can be applied directly. However, Figure 2 poses a question: "Should the relative frequency for each thickness band be used exactly as plotted?" This question really has two parts, one involving smoothing and one involving interpolation. Is there any reason to believe that the relative frequency should be higher for the 5600-5640 m band than for the bands on either side? If not, smoothing is suggested. Also, should one use 19% for 5441 m and jump to 64% for 5439 m? If not, interpolation is suggested. In any case, judicious use of histograms can produce a useful objective tool, and a computer is not required for its development or use.

4. SCATTER DIAGRAMS

Another model rivaling the histogram in its simplicity is the scatter diagram. It is primarily a noncomputer technique and was used as early as the beginning of this century by Besson (1905). The technique has also been called graphical regression and was studied in detail by Brier (1946). It has been used extensively by the U.S. Weather Bureau, now the National Weather Service, since the mid-1940's and several papers appeared in the Monthly Weather Review circa 1950 illustrating the use of this model. A typical paper from this period is that by Thompson (1950).

In its simplest form, coordinate axes are established on a diagram such that the ordinate represents the dependent variable or predictand and the abscissa represents a single independent variable or predictor. Points are then plotted on this diagram depicting the available data sample. Finally a line can be drawn by eye which seems to fit the data points. In application, a forecast of the predictand is found by reading the ordinate value of the line at the abscissa value of the predictor. Such a completed diagram is shown in Figure 3.

Usually, however, one wants to use two or more predictors. In this case, the coordinate axes should be the values of the pair of predictors, and the predictand values are plotted at the points on the diagram representing the data sample. An analysis is then made of the plotted data. The analysis is subjective and will depend on the skill of the analyst. The analyst must be careful not to "over analyze" the data, especially in regions where few data points exist. In general, the analysis should be rather smooth and, in case of doubt, known physical relationships may furnish a key to correct analysis.

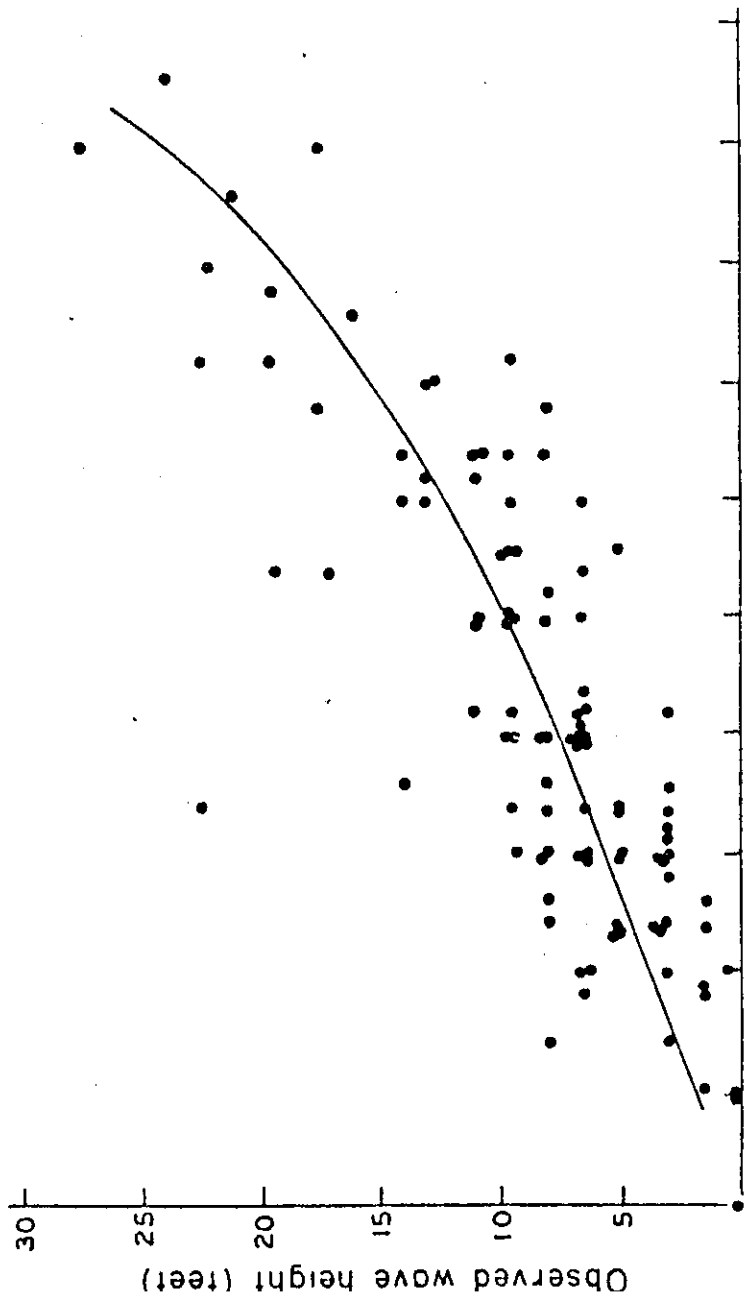

Figure 3. Example of one-predictor scatter diagram (after Pore and Richardson, 1969). Plotted data are observed wave height versus observed wind speed at ocean station vessels. A

Figure 4 shows an example of a two-predictor scatter diagram taken from Thompson (1950). Precipitation amount can be forecast with Figure 4 and the observed values of the two predictors, x_1 (700 mb height at Oakland) and x_2 (San Francisco minus Los Angeles sea level pressure difference). If one wishes to use more than two predictors, other diagrams can be plotted and analyzed. For instance, predictors 3 and 4 can be combined on a diagram similar to Figure 4. The predictand value "forecast" from Figure 4 can be called predictor x_5. Similarly, the predictand value estimated from x_3 and x_4 can be called x_6. Then x_5 and x_6 can be the coordinates on another scatter diagram in which the actual predictand values are plotted as a function of x_5 and x_6. (A variation of this procedure is to plot deviations between the preliminary estimates x_5 and x_6 and the actual values Y.) An analysis of these values will then define estimates of Y given values of x_5 and x_6. Thompson (1950) presents an example of a six-predictor scatter diagram procedure.

The scatter diagram model is very simple in principle, yet it allows for any degree of complexity that the data warrant. Its success will depend on the analyst's ability to choose meaningful predictors, as will the success of any technique. Thompson (1950) offered the following comments concerning the analysis: "While the meteorological relationships brought out by the primary graphical combination of each pair of variables may ... be discussed from a physical standpoint, and thereby the reasonableness of the isograms checked, very little can be said about the secondary combinations. Here the complexity of the joint relationships, as well as the probable effect of other variables not considered in the integration, defeats any attempt to supply a theoretical or physical justification for the distribution of the isograms. Consequently the construction of these charts must depend almost entirely upon an analysis of the data."

Scatter diagram analysis is very useful when resources are limited and only small amounts of data are available. It does not lend itself easily to processing by electronic computer, and the method itself implies hand analysis [although some individuals, including Freeman (1961), have attempted to automate the process]. For this reason, other techniques are usually to be preferred when samples consisting of several thousand cases and a computer are available. Also, no reliable significance test exists for the scatter diagram procedure to determine whether added predictors will lead to increases in forecast accuracy on new data. Therefore, a forecasting system based on scatter diagrams should always be tested on new data if at all possible.

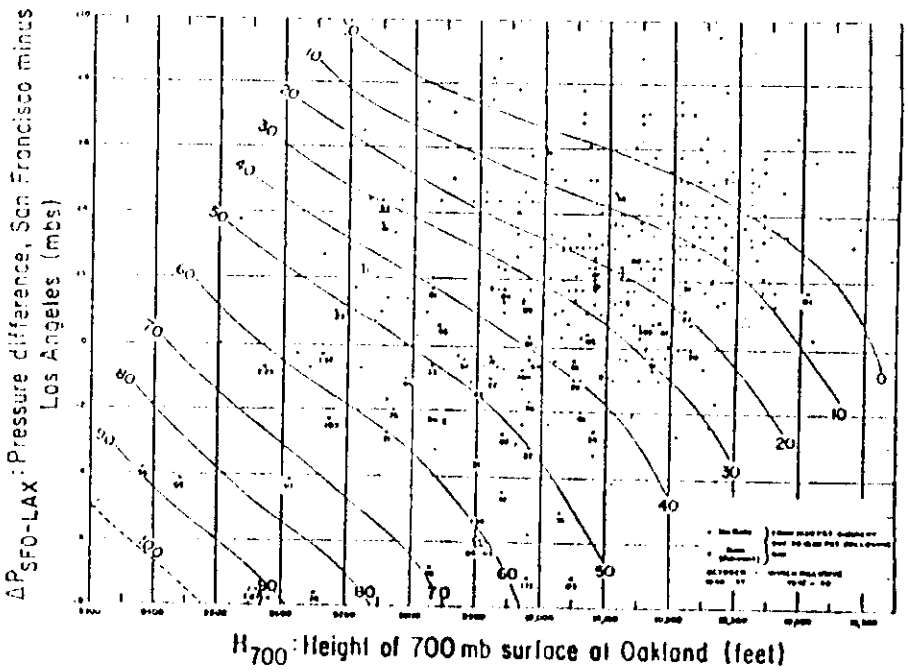

Figure 4. Precipitation amount plotted as a function of 700 mb height at Oakland versus the San Francisco minus Los Angeles sea level pressure difference. Isolines representing rainfall amount have been adjusted to a scale of 0 to 100 for input to another diagram (after Thompson, 1950).

5. REGRESSION

It has become increasingly clear during the last 20 years that the use of large samples is very desirable in the solution of meteorological prediction problems. Three reasons can be cited for this conclusion. First, the autocorrelation of many meteorological variables does not approach zero unless observations are taken quite some time apart. Therefore, the number of degrees of freedom for such data is much less than the sample size. Second, many variables usually can be found that possess a relationship to the predictand, and one frequently wants to include - or at least test the desirability of including - a large number of these potential predictors. In the application of most models, this process uses up many degrees of freedom. Third, the distribution of the predictand is frequently highly skewed, so that the very weather situations that are most important to predict occur very infrequently. A large sample is necessary to include a representative number of such situations.

The use of large samples and the inclusion and testing of many predictors necessitates the use of an electronic computer and a model that lends itself to computer application. Linear regression is such a model.

5.1 Simple Linear Regression

The simple linear regression model is of the form

$$Y = \alpha + \beta X + \epsilon,$$

where Y is the predictand, X is a predictor, α and β are parameters, and ϵ is the error term. The predicted value of Y is $\hat{Y}$, where

$$\hat{Y} = a + bX,$$

in which a and b are estimates of the parameters.

In this model, the sum of squares of the observed errors (i.e., the e_i's) is minimized over a dependent sample of size n:

$$\min \sum_{i=1}^{n} e_i^2 = \min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \min \sum_{i=1}^{n} (Y_i - a - bX_i)^2.$$

Taking partial derivatives of Σe_i^2 with respect to a and b and setting each derivative equal to zero yields the so-called normal equations:

$$an + b \sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i = 0,$$

$$a \sum_{i=1}^{n} X_i + b \sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i Y_i = 0.$$

Solving for a and b gives

$$b = \frac{n \sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

$$a = \frac{1}{n} \sum_{i=1}^{n} Y_i - \frac{b}{n} \sum_{i=1}^{n} X_i.$$

This one-predictor model is useful for illustrative purposes and can be applied to situations such as that shown in Figure 3. Usually, however, several predictors need to be considered. As a result, one is led to multiple linear regression.

5.2 Multiple Linear Regression

If we form matrices from data samples of size n for the predictand and p predictors, then

$$\underline{x} = \underline{X} - \overline{\underline{X}}$$

and

$$\underline{y} = \underline{Y} - \overline{\underline{Y}}$$

are nxp and nx1 matrices, respectively, in which each column is the deviation from the mean of the corresponding original variable. Variance-covariance matrices can be calculated as follows:

$$\underline{S}_{11} = \frac{1}{n} \underline{x}'\underline{x},$$

$$\underline{S}_{12} = \underline{S}'_{21} = \frac{1}{n} \underline{x}'\underline{y},$$

$$\underline{S}_{22} = \frac{1}{n} \underline{y}'\underline{y} = \hat{\sigma}_Y^2,$$

where a prime denotes a matrix transpose. The multiple regression equation, derived in a manner analagous to the one-predictor case in Section 5.1, is

$$\hat{Y} = \underline{X}\, \underline{S}_{11}^{-1}\, \underline{S}_{12} - \overline{\underline{X}}\, \underline{S}_{11}^{-1}\, \underline{S}_{12} + \overline{Y}. \tag{5}$$

Associated with Eq. (5) are a reduction of variance (RV) and a multiple correlation coefficient (R) that are defined as follows:

$$RV = R^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2},$$

$$= \frac{\text{variance of } Y - \text{error variance of } \hat{Y}}{\text{variance of } Y}.$$

These quantities are easily calculated from the variance-covariance matrices:

$$RV = R^2 = \frac{\underline{S}_{21}\, \underline{S}_{11}^{-1}\, \underline{S}_{12}}{\underline{S}_{22}} = \frac{\underline{S}_{21}\, \underline{S}_{11}^{-1}\, \underline{S}_{12}}{?}$$

[illegible] joint distribution of predictand and predictors, except that no predictor may be an exact linear function of one or more other predictors. In that case, the inverse of $\underline{S}_{11}$ could not be determined; that is, it would be singular. For real meteorological data and sample sizes much larger than the number of parameters (i.e., p + 1), this problem seldom arises.

Under certain conditions, analysis of variance can be used for testing the significance of the reduction of variance, R^2, and of individual terms in the equation. The conditions are that the sample be drawn randomly from a multivariate normal population and that no preselection of variables be made using the same sample on which the regression equation is developed. The F value corresponding to R^2 is calculated according to line 2 in Table 2. This calculated value can then be compared to the tabled F value for a desired α level (probability of Type I error) with p and n-p-1 degrees of freedom.

Table 2. Analysis of variance table for reduction of variance associated with multiple linear regression model.[a]

Source	Sum of squares[b]	Degrees of freedom	Mean square[b]	F value
Total	1	n-1		
Regression equation - p predictors	R_p^2	p	$\frac{R_p^2}{p}$	$\frac{R_p^2\,(n-p-1)}{(1-R_p^2)\,p}$
pth predictor in regression equation	$R_p^2-R_{p-1}^2$	1	$R_p^2-R_{p-1}^2$	$\frac{(R_p^2-R_{p-1}^2)(n-p-1)}{1-R_p^2}$
[illegible]	[illegible]	[illegible]	[illegible]	

[a]Patterned after Panofsky and Brier (1958).

[b]All entries in these columns should be multiplied by $n\hat{\sigma}_Y^2$. This factor cancels out in computing F.

Suppose R^2 is significant in a particular situation, and one wonders whether a particular predictor is really adding any predictive information over and above the other p-1 predictors. Consider that predictor as the last or pth predictor. Then the appropriate F value is shown on line 3 of Table 2 and has 1 and n-p-1 degrees of freedom. This test is valid, under the conditions stated above, provided that the choice of which predictor to test is not based on an analysis of the data sample. This topic will be discussed further in Section 5.3.

The model discussed in this section is a linear model, and other models may be more appropriate. However, if the population distribution is multivariate normal, then this model is the best model that can be found. If a researcher wants to "screen" out from a much larger set of possible predictors the p predictors to include in the equation, then the approach described in the next section can be used.

5.3 Screening Regression

Screening regression, as the term is usually defined in meteorology, combines multiple linear regression with an objective method of selecting a "good" set of predictors to use in the equation from a larger set of m potential predictors. Since regression finds the solution that minimizes the estimated error variance on the dependent sample, it is logical to choose a set of predictors that would be better than any other set for reducing this error variance. However, if p predictors were to be picked from a set of m predictors, then the number of combinations for even a moderate value of m is quite large (unless p is very small or approaches m). Specifically, the number of such combinations is

$$C_p^m = \frac{m!}{p!(m-p)!}.$$

It is usually not feasible to compute all these combinations, so some shortcut must be taken to find a "good" set that may not be the "best" set.

Screening can be done in one of several ways. The simplest is what may be called forward selection. This procedure consists of first selecting the one predictor from the total set of m predictors under consideration that reduces the variance of the predictand more than any other possible predictor, then choosing the predictor that together with the first one selected reduces the variance more than any other such combination of two predictors, and continuing the selection procedure on a "one at a time" basis until the additional reduction of variance afforded by any predictor is very small. This procedure insures that the first predictor selected is the best

single predictor, but it does not insure that the first two chosen are the best pair, etc. This stepwise selection was discussed as early as 1940 by Wherry (1940) and was introduced into the meteorological literature by Miller (1958) following some unpublished work by Bryan (1944).

The question arises as to how many predictors to choose. One might be tempted, after selecting p-1 predictors, to test the additional reduction of variance given by the pth predictor using the F value computed in the third line of Table 2. This procedure was suggested by Lubin and Summerfield (1951) and is sometimes done. However, it must be considered only as a stopping criterion. That is, one must not attach any particular significance level to it. The reason for this limitation is that the test is being performed on the next best predictor and not on a predictor that has been selected at random.

Miller (1958) has suggested a modification to the standard F test that compensates for the testing of the best of several remaining potential predictors. Instead of using the critical value $F_{(1-\alpha)}$ at each selection step, he suggests using the value $F^*_{(1-\alpha)} = F_{1-\alpha/(m-p+1)}$ at the pth selection step with some desired probability of Type I error α. This criterion is rather harsh, since it assumes (approximately) that the (m-p+1) tests that could be performed at the pth selection are independent and, in the absence of additional complications, tends to lead to the selection of too few predictors if some accepted value of α such as 0.05 is used (Zurndorfer and Glahn, 1977). Additional complications could include highly nonnormal distributions or non-zero autocorrelation for the predictors. The test proposed by Miller (op. cit.) will also tend to compensate for nonzero autocorrelations. However, it should be remembered that, because of all the complications, this test is primarily a stopping procedure and no exact level of significance should be attached to it.

The decision as to the exact number of predictors to select is many times overemphasized. The mean square error for independent data is usually not very sensitive to the number of predictors in the equation within rather broad limits. Figure 5 shows schematically the kind of results that have been obtained from experiments (e.g., see Bocchieri and Glahn, 1972). R^2 on dependent data always increases with the addition of another predictor but tends to "level out" so that little is to be gained in terms of the mean square error of the predictand by including more than, say, 12 terms. The mean square error on test data need not decrease monotonically. A small test sample will frequently cause the mean square error curve to be irregular. Also, for a large number of predictors, the R^2 test sample curve will usually turn downward. However, a broad, flat maximum will generally be found where, for practical purposes, the

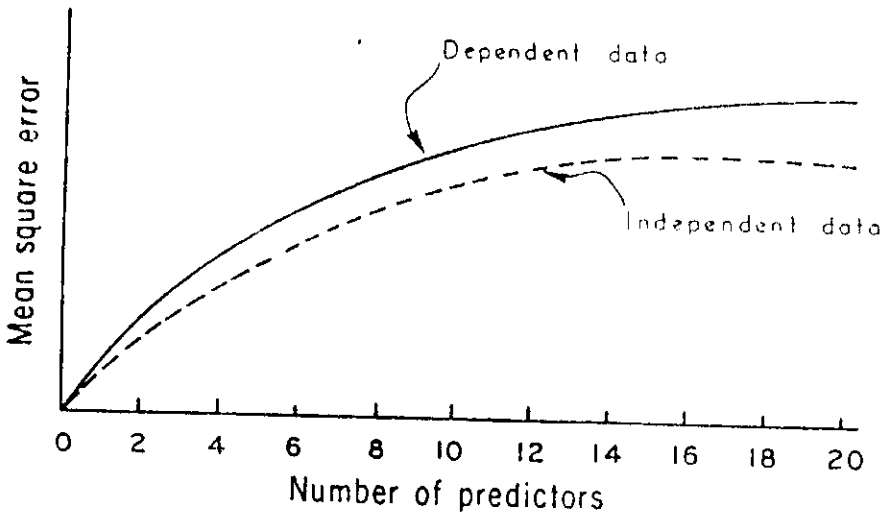

Figure 5. Schematic diagram of mean square error as a function of number of predictors.

predictions are of equal quality. For this reason, almost any "practical" stopping procedure is quite adequate, such as (a) when the added reduction of variance of the next predictor is less than 0.005, (b) when 12 predictors have been selected, (c) when the reduction in mean square error is less than, say, 0.05°F for temperature or 0.2 mph for wind speed, or (d) Miller's $F_{(1-\alpha)}$.

Another version of screening regression is to find the reduction of variance for all m predictors, and then start eliminating predictors one by one until some stopping criterion is met. This backward elimination procedure also does not yield the unique best set, and significance testing for it has been inadequately studied. However, some simple stopping procedure similar to those methods described for forward selection can be used. A complication could occur if one predictor were an exact linear function of a set of other predictors. Then S_{11} would be singular. This possibility is very unlikely with real meteorological data, unless one were to actually formulate a predictor from a linear relationship. For instance, one could include only two of the following three predictors: 500 mb height, 1000 mb height, and 1000-500 mb thickness.

Still another algorithm combines the above two procedures. Forward selection is done with an F test being performed at each step. When the added reduction of variance is insufficient to be judged significant, the procedure is stopped. However, between each selection step, all the variables selected up to that point (and not subsequently discarded) are tested for significance. The least significant is discarded if it does not meet the test, and again all those remaining are tested until none is discarded.

Forward selection screening regression has been used more than any other computer-oriented model for statistical weather prediction. Many studies were made at the Travelers Research Center, Inc. in the late 1950's and early 1960's using this procedure in combination with the classical approach. For instance, Veigas et al. (1958) produced an objective method for predicting the behavior of hurricanes in the western Atlantic and Gulf of Mexico that was subsequently used operationally by the National Weather Service. More recently, the Techniques Development Laboratory (TDL) of the National Weather Service has developed many operational products based on this model. Most of these products involve the MOS technique, but some forecasts are based on perfect prog and classical procedures. These present-day uses of regression analysis are discussed in Section 11.

5.4 Regression Estimation of Event Probabilities

The screening regression model can be used when the predictand is binary. For instance, the predictand might take on the value of 1 when an event occurred and 0 when it didn't. The regression

equation, then, can be thought of as yielding the probability (or relative frequency) of the event for realistic combinations of predictor values. This approach was used by Mook (1948) and Lund (1955), but no extensive application was made of it until Miller (1964) and others began using it at the Travelers Research Center in the 1960's. They dubbed it REEP for regression estimation of event probabilities. Miller (op. cit.) realized that the probabilistic model also held for multiple categories. For instance, if ceiling height is divided into five mutually exclusive and exhaustive categories and each category is used to define a binary predictand, then the set of five regression equations (all with the same predictors) will give a set of probabilities, p_i, where $\sum_i p_i = 1$ $(i = 1, \ldots, 5)$.

An important property of this model is that it minimizes the P-score defined by Brier (1950), which has certain desirable characteristics (Brier, op. cit.; Murphy, 1974) and which is frequently used in probabilistic forecast verification (see Chapter 10). Unfortunately, the individual p_i's are not constrained to the zero-one interval. Selection of predictors can be made by choosing next the predictor that contributed most to the R^2 of any one of the categories. Significance tests based on assumptions of normality are not appropriate. Experience has shown that, generally, a larger sample is required to obtain stable results when the predictand is binary than when it is continuous.

5.5 Binary Predictors

The screening regression model can be used when one or more of the predictors are binary. All predictors were binary in the first applications of REEP. An early reference to the use of binary variables in regression is Suits (1957), and Neter and Wasserman (1974) give a good discussion of the subject.

A binary variable (sometimes called an indicator or dummy variable) can arise naturally. For example, an observation may be made in this format, as in the case of rain or no rain. In addition, a continuous variable can also be "dummied"; this process is usually carried out in one of the two ways indicated in Table 3. (Variables are never really continuous, since only discrete values are used in practice. However, temperature measured to the nearest degree is quasi-continuous and will be considered to be continuous in this chapter.) In transformation 1, each dummy variable indicates whether or not the original variable has a value corresponding to its particular defining interval. In transformation 2, each dummy variable indicates whether or not the original variable has a value less than the upper limit of its particular defining interval. Note that any one of the four dummy variables for transformation 1 is redundant with the other three and, although all four could be screened (a

Table 3. Two methods of transforming a continuous variable into binary variables. In any particular column, ones and zeroes can be interchanged without affecting predictive capability.

Original variable category	Binary variable transformation 1				Binary variable transformation 2		
	1	2	3	4	1	2	3
1	1	0	0	0	0	0	0
2	0	1	0	0	1	0	0
3	0	0	1	0	1	1	0
4	0	0	0	1	1	1	1

fourth would never be selected), only three (any three) can be included in a regression equation. For transformation 2, only three meaningful variables are possible; the fourth would always have the same value. Dummy variable No. 1 corresponds to No. 1, and No. 4 to No. 3, for transformations 1 and 2, respectively. However, Nos. 2 and 3 for transformation 1 have no match in transformation 2.

Any combination of three dummy variables for transformation 1 will give the same reduction of variance as the three dummy variables for transformation 2. However, No. 2 for transformation 2 may be better than any single predictor for transformation 1. When several dummy variables are created from a continuous variable, careful consideration should be given to which transformation to use. A predictor such as No. 2 for transformation 1 treats a certain category of the original variable one way and the categories on both sides of it another way. This procedure may be appropriate if the predictand is continuous and is a quadratic function of the original (undummied) predictor or if the predictand is binary and was dummied by transformation 1. For instance, suppose a binary predictand represents ceiling height from 1000 to 1900 feet at 1200 GMT. Then a binary predictor representing that same ceiling interval at 0900 GMT would be a good predictor. However, if the predictand is continuous, then a few binary predictors defined by transformation 2 will usually yield better results than the same number of predictors defined by transformation 1.

Although dummy predictors have the potential of accommodating a nonlinear relationship between the predictand and predictors, some information is lost since all values within a defining interval are treated the same way (unless each value is represented by a different dummy variable). Also, it takes several binary predictors to provide about the same information as one continuous predictor. The number of binary variables to be defined for a given predictor is usually quite arbitrary as is the interval to associate with each variable. Traditional statistical significance tests are even less applicable in this case, due both to the binary nature of the variables and to the unknown number of degrees of freedom used in choosing, say, five out of eight possible dummy variables created from a single predictor.

The use of all binary variables permits much more efficient computer use, since one value need occupy only one bit rather than a complete word and, in addition, faster logical rather than arithmetic operations can be used in obtaining sums of squares and cross products. However, realization of this advantage usually requires considerable programming effort and use of such a program would have to be rather extensive before the effort would be worthwhile.

5.6 Computed Predictors

Although regression as presented above is a linear model, nonlinear relationships can be incorporated through special computations. For instance, divergence is not observed but can be calculated from wind observations. In addition, a predictor can be "linearized" in various ways. That is, it can be transformed in such a way that it has a more nearly linear relationship with the predictand than did the original variable. Consider again the histogram example shown in Figure 2. If we want to use several predictors, including 1000-500 mb thickness, then we could include a transformed or computed predictor representing the heights of the bars in Figure 2. The transformation could be exactly as indicated in Figure 2, or a curve could be fit that would undoubtedly provide a variable that would be more robust on independent data.

5.7 Orthogonal Predictors

We may have a problem in which we want to distill most of the linear predictive information from a large set of variables without using a large number of degrees of freedom. For instance, 1000-500 mb thickness values may be available at each of 25 stations surrounding a station for which we wish to predict max temperature. The predictors are highly correlated and we wouldn't want to include all 25 in a regression equation. We could screen the 25 and select, say, five. Another alternative is to transform the 25 variables into another set of variables that are more efficient in terms of retaining the large scale predictive information and discarding the small scale "noise." Orthogonal functions can be used for this purpose.

Assume that pressure values are available at each of m points and at each of n times. At each point the mean over time can be found and the deviations from the means put into an nxm matrix $\underline{P}$. The element P_{ii} on the diagonal of $\frac{1}{n}(\underline{P}'\underline{P})$ is the variance of the pressure at the ith point and the element P_{ij} is the covariance of the pressures at the ith and jth points. The time series of k new variables, represented by the nxk matrix $\underline{U}$ can be found by

$$\underline{U} = \underline{P}\,\underline{T},$$

where $\underline{T}$ is an mxk matrix of coefficients of k functions at m points. The matrix $\underline{T}$ is an efficient transformation matrix if the columns are orthogonal. Also, the total variance of the columns of $\underline{U}$ is equal to the total variance of the columns of $\underline{P}$ if the columns of $\underline{T}$ are orthonormal and k = m. Then

$$\underline{T}'\underline{T} = \underline{I}$$

($\underline{I}$ is a kxk identity matrix) and

$$\mathrm{tr}(\underline{U}'\underline{U}) = \mathrm{tr}(\underline{P}'\underline{P}),$$

where tr represents the trace of a matrix. The original pressure deviations at the m points can be approximately reproduced from the new functions by

$$\hat{\underline{P}} = \underline{U}\,\underline{T}',$$

and, if k = m,

$$\hat{\underline{P}} = \underline{P}.$$

Several authors, including Wadsworth (1948), White et al. (1958), and Jorgensen (1959), have adapted orthonormal Tschebyscheff functions for this use. The m points occur in a rectangular array and functions of degree zero through r and zero through s are used in the two dimensions, respectively. The columns of $\underline{T}$ are then made up of cross products of two functions. The function composed of the function of degree zero in both directions represents the mean of the m points. Functions composed of one function of degree zero and one function of degree not zero represent patterns which vary in only one direction. In general, the low degree functions represent large scale features of the map, whereas the high degree functions represent small scale features.

Much of the variance of pressure, and of many other meteorological variables, is explained by large scale components. On the other hand, it is the very small scale components that contain most of the observational error and are the least predictable. If very small scale features in the pressure map furnish much predictive information for other variables, then it is usually not beneficial to represent those features in terms of orthogonal functions.

Even though the transforming functions - the columns in $\underline{T}$ - are orthogonal, the new variables - the columns in $\underline{U}$ - are not necessarily orthogonal. Therefore, $\underline{T}$ is not as efficient as it might be, and the regression constants which relate a predictand to the new variables must be determined by considering the covariances between those new variables as well as the variances. That is, it is necessary to examine the complete matrix

$$\frac{1}{n}(\underline{U}'\underline{U}).$$

The most efficient way of representing the linear information in a set of data is through principal components. These functions were introduced into meteorology by Lorenz (1956) who called them Empirical Orthogonal Functions (EOFs). EOFs have been used to study meteorological data and for predictive purposes by several

researchers, including Gilman (1957), White et al. (1958), Glahn (1962), and Grimmer (1963). In addition to

$$\underline{T}'\underline{T} = \underline{I},$$

the condition

$$\frac{1}{n}(\underline{U}'\underline{U}) = \underline{D}$$

is also imposed here, where $\underline{D}$ is an mxm diagonal matrix. Since

$$\underline{U} = \underline{P}\,\underline{T},$$

substitution can be made to yield

$$\frac{1}{n}(\underline{T}'\underline{P}'\underline{P}\,\underline{T}) = \underline{D}.$$

The matrix

$$\frac{1}{n}(\underline{P}'\underline{P}) = \underline{R}$$

is the covariance matrix of the original variables. Therefore,

$$\underline{T}'\underline{R}\,\underline{T} = \underline{D}.$$

The columns of $\underline{T}$ are the characteristic vectors and the correspondi diagonal elements of $\underline{D}$ are the roots of the matrix $\underline{R}$. The k column of $\underline{T}$ which correspond respectively to the k largest diagonal elemen of $\underline{D}$ explain a larger fraction of the total variance of the origina variables, $\frac{1}{n}\mathrm{tr}(\underline{P}'\underline{P})$, than any other k linear combinations of those variables.

Regression estimates $\hat{\underline{y}}$ of a predictand time series $\underline{y}$ (in terms of deviations from the mean) can be found by

$$\hat{\underline{y}} = \underline{U}\,\underline{A},$$

where $\underline{A}$ is the kx1 vector of regression coefficients corresponding k of the new variables. The vector $\underline{A}$ is now easily determined from

$$\underline{A} = \frac{1}{n}(\underline{D}^{-1}\underline{U}\,\underline{y}),$$

since the inversion of $\underline{D}$ is trivial.

It is worth noting that if the original variables are normaliz to unit variance, the characteristic vectors and roots will not in general be the same as with the nonnormalized variables. For instance, the pressure at a point in middle latitudes has a larger variance than it does in low latitudes. If points from both region

are included without normalization, the columns of $\underline{U}$ having the largest variances will be dominated by the middle latitude pressures.

An advantage in the use of EOF's is that the points do not have to be in any organized pattern spatially. Therefore, observations taken at stations can be used directly rather than requiring the field to be specified in terms of a grid before the orthogonal functions are applied.

5.8 Normalized Variables

Although regression can be applied to variables with practically any distribution, it is the optimum model if the variables have a multivariate normal distribution. Boehm (1976) suggests that each variable used in the analysis should be "transnormalized." That is, some method such as histogram analysis or curve fitting should be applied to each variable separately, both predictand and predictors, so that the resulting transformed variable will have a (near) normal distribution. Boehm (op. cit.) uses the term transnormalized to highlight the fact that this transformation is not just subtraction of the mean and division by the standard deviation. It is the same transformation discussed by Panofsky and Brier (1958, p. 41). After the regression analysis is performed on the normalized variables, predictions of the normalized dependent variable can be made. Then these values must be transformed back to the original variable.

6. DISCRIMINANT ANALYSIS

Certain meteorological variables do not lend themselves well to prediction by linear regression due to their nonnumerical nature, highly nonnormal distribution, or nonlinear relationships to the predictors. In such cases, multiple discriminant analysis (MDA) provides a useful tool that has been applied extensively to weather forecasting problems by Miller (1962) and others at the Travelers Research Center, Inc.

Discriminant analysis was conceived by Fisher (1936) and first brought into the literature by Barnard (1935). MDA refers specifically to the Fisher analysis on more than two predictand groups. Barnard (op. cit.) used the analysis on 4 groups, but she considered only one discriminant function. Hotelling (1935) and others (Fisher, op. cit.; Brier, 1940) evidently appreciated the possibility of more than one function and mentioned the determinantal equation involved, but the burden of calculations forbade extensive use of MDA until the computational scheme of Bryan (1950) or electronic computers became available.

For a given problem, there is a maximum of p or G-1 (whichever is smaller) discriminant functions, where p is the number of predictors and G the number of groups. These functions are mutually uncorrelated and are found through the solution of the equations:

$$(\underline{W}^{-1}\underline{B} - \lambda_j \underline{I})\underline{V}_j = \underline{0}, \quad [j = 1, \ldots, \min(p, G-1)]$$

where $\underline{W}$ and $\underline{B}$ are respectively the matrices of within and between groups sums of squares of the predictors, $\underline{I}$ is the identity matrix, the $\underline{V}_j$'s (eigenvectors) are the coefficients in the discriminant functions, and the λ_j's are the roots (eigenvalues) of the determinantal equation

$$|\underline{W}^{-1}\underline{B} - \lambda \underline{I}| = 0.$$

In the special case involving only two groups, only one function is possible and its coefficients are proportional to those derived by regression. Therefore, for this special case, the two analyses are equivalent.

A significance test, which is a generalization of Mahalanobis' D^2, for the predictand-predictors relationship based on large sample theory has been developed by Rao (1952) and uses the statistic

$$V_{pG} = n(\mathrm{tr}\ \underline{W}^{-1}\underline{B}) = n \sum_{j=1}^{G-1} \lambda_j,$$

where n is the sample size and it is assumed that G-1 roots exist. V_{pG} is distributed as χ^2 with p(G-1) degrees of freedom provided that the predictors are multivariate normal within each group and that the covariance matrices for each group are identical.

The importance of each discriminant function $\underline{V}_j$ is indicated by its associated root λ_j. Since the number of discriminant functions may be less than the minimum of p and G-1, the significance of each root can be tested by an approximate procedure due to Bartlett (1934). For each nonzero root, the test statistic

$$[n - (1/2)(p + G)] \ln (1 + \lambda_j)$$

is computed. This statistic is approximately distributed as χ^2 with p+G-2j degrees of freedom.

The selection of variables for MDA by screening has also been described by Miller (1962). In the same way that the selection for regression maximizes the F-statistic, the selection for MDA maximizes

V_{pG}. At each step, after p-1 predictors have been selected, the quantity

$$V_{pG} - V_{(p-1)G} = n[\mathrm{tr}(\underline{W}_p^{-1}\,\underline{B}_p) - \mathrm{tr}(\underline{W}_{p-1}^{-1}\,\underline{B}_{p-1})]$$

is evaluated for each remaining possible predictor and the largest value indicates the preferred variable. This statistic is considered to be distributed as χ^2 with G-1 degrees of freedom (Rao, 1952), and Miller (1962) suggests, as with regression, that the critical value $\chi^2_{1-\alpha/(m-p+1)}$ be used since the selection is not random.

MDA as used in prediction can be considered to be a linear transformation from a p-dimensional predictor space to a G-1 dimensional discriminant space (assuming G-1 ≤ p), such that the sample points plotted in the discriminant space exhibit as much clustering according to predictand category and as little dispersion from their respective cluster centers as possible. There will be, then, G regions in the discriminant space, one for each predictand category.

Let us consider the effect of the relationship between one of the predictors and the predictand on this transformation and the desirability of using this predictor. Three possibilities can be mentioned:

(a) It may be that a particular value of the predictor will indicate only one predictand category, and that category will be indicated by no other value of the predictor. This situation is the most desirable. It does not matter how the predictand categories are arranged on the predictor scale; group 1 could be between group 4 and group 5 just as well as anywhere else (see Figure 6). Thus, this type of nonlinearity is accommodated by MDA, since no numerical scale is associated with the predictand categories themselves.

(b) It may be that a low predictor value will indicate two different predictand categories. In this case the regions in the discriminant space containing the points representing these two predictand categories will be superimposed (unless the effect of other predictors can spread them apart), and the predictor will be worthless in discriminating between these two predictand categories. However, the predictor could still be very useful in separating these two predictand categories from all the rest and could also distinguish between those remaining categories. No transformation of the predictor before its inclusion in the analysis would be useful in separating the superimposed groups. See groups 3 and 4 in Figure 6 for an example of this situation.

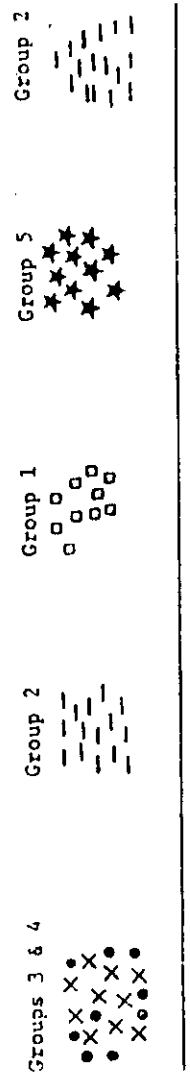

Figure 6. Hypothetical relationships of predictand groups 1 through 5 to the predictor x. See text for explanation.

(c) It could happen that both low and high values of the predictor would indicate the same predictand category. This situation is very undesirable since the predictor would tend to spread the points in the discriminant space representing this predictand category and these points might group themselves into two distinct regions (see group 2 in Figure 6). This type of nonlinearity is not accommodated by MDA, and a nonlinear transformation of the predictor is indicated if it is to be used. In a screening procedure where the raw variable is a possible predictor, it will probably be overlooked as its nonlinear relationship to the predictand will not be recognized.

Since the criterion for selecting the variables and for determining the functions themselves is to maximize the between to within groups variance ratio, groups with many cases will highly influence the results. This state of affairs is generally detrimental, except in those cases in which the costs of misclassification are all equal and the concern is only for the number of correct forecasts. Miller (1962) has attempted to counteract the large group effect in predictor selection by making the size of all groups equal to that of the smallest group during the selection process and then using the complete sample to determine the discriminant functions and probabilities of misclassification.

Discriminant analysis can be considered to be complete when the functions and their associated roots have been found. However, the problem of how to use these functions in probabilistic prediction of the predictand groups still remains unspecified. Miller (1962) used Bayes' theorem to find the a posteriori group probabilities from the a priori group probabilities, which are estimated from the sample and the assumed multivariate normal distribution of the discriminant functions within each group. For the sample of meteorological data considered by Miller (op. cit.), the multivariate normal assumption proved to be untenable.

If the data do not appear to justify the use of a completely parametric model, discriminant analysis can be used to map the p-dimensional space into a G-1 or less dimensional space, and the conditional probability distributions can be determined by other means. If the dimensionality of the discriminant space is only 1 or 2, a scatter diagram can be employed to determine how the G groups are distributed within this space.

Miller (1962) found a nonparametric method described by Fix and Hodges (1951) to be useful in determining the a posteriori probabilities directly from the discriminant function values. At any point Y' within the discriminant space, the probability of each group can be estimated by the relative frequency of that group occurring in the k sample points closest to Y'. The value of k should be relatively large but small compared to the sample size. The k closest points can be defined in terms of the Euclidean distance. This procedure is

essentially a smoothing process, and precautions must be taken to ensure that the importance of the discriminant functions is taken into account in this process. Specifically, an arbitrary metric has been used to transform the discriminant space into a space in which each function has zero mean and variance equal to λ_j/λ_1. The distance between a point Y' and a sample point Y in this space can be defined as

$$D = \left[\frac{\lambda_1}{\lambda_1}\left(\frac{y_1 - y_1'}{\hat{\sigma}_{y_1}}\right)^2 + \frac{\lambda_2}{\lambda_1}\left(\frac{y_2 - y_2'}{\hat{\sigma}_{y_2}}\right)^2 + \ldots + \frac{\lambda_j}{\lambda_1}\left(\frac{y_j - y_j'}{\hat{\sigma}_{y_j}}\right)^2\right]^{1/2},$$

where $\hat{\sigma}_{y_j}$ is the standard deviation of the jth discriminant function. This procedure usually produces extreme smoothing over the least important functions but retains the predictive information in the more important functions.

The most extensive use and thorough testing of the MDA technique in meteorology has been in the short range prediction of visibility and ceiling height undertaken at the Travelers Research Center, Inc. (e.g., see Enger et al., 1964).

7. CANONICAL CORRELATION

Canonical correlation, first developed by Hotelling (1936), is a technique for finding orthogonal relationships between two sets of variables. Consider a situation involving n observations of each of p variables X_i (i = 1, 2, ..., p) and of q variables Y_i (i = 1, 2, ..., q). These observations represent points in a p+q dimensional space and can be arranged in the nxp matrix $\underline{X}$ and the nxq matrix $\underline{Y}$. The variables have means $\bar{X}_i$ and $\bar{Y}_i$, respectively, and deviations from the mean are given by $x_i = X_i - \bar{X}_i$ and $y_i = Y_i - \bar{Y}_i$. New variables $\underline{x}\,\underline{A}_i$ and $\underline{y}\,\underline{B}_i$ (i = 1, 2, ..., r), where r is less than or equal to the smaller of p and q, can be formed such that their means are zero and

$$\underline{A}'\underline{x}'\underline{x}\,\underline{A} = n\,\underline{I}, \tag{6}$$

$$\underline{B}'\underline{y}'\underline{y}\,\underline{B} = n\,\underline{I}, \tag{7}$$

$$\underline{A}'\underline{x}'\underline{y}\,\underline{B} = n\,\underline{\Lambda}, \tag{8}$$

where $\underline{I}$ is an rxr identity matrix,

$$\underline{\Lambda} = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_r \end{bmatrix}, \tag{9}$$

and $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_r$. Eqs. (6) and (7) state that the variance of each of the new variables is unity and each is uncorrelated with all others in its respective set. Eqs. (8) and (9), together with Eqs. (6) and (7), state that each $\underline{x}\,\underline{A}_i$ is uncorrelated with each $\underline{y}\,\underline{B}_j$ except when $i = j$ and then the correlation is λ_i.

It can be shown [for instance, see Anderson (1958)] that the $\underline{A}_i$ ($i = 1, 2, \ldots, r$) can be found from

$$(\underline{S}_{11}^{-1}\,\underline{S}_{12}\,\underline{S}_{22}^{-1}\,\underline{S}_{21} - \lambda_i^2\,\underline{I})\underline{A}_i = \underline{0}$$

(providing $\underline{S}_{11}$ and $\underline{S}_{22}$ are not singular), where the λ_i satisfy the determinantal equation

$$|\underline{S}_{11}^{-1}\,\underline{S}_{12}\,\underline{S}_{22}^{-1}\,\underline{S}_{21} - \lambda^2\,\underline{I}| = 0$$

and where

$$\underline{S}_{11} = \frac{1}{n}\,\underline{x}'\underline{x},$$

$$\underline{S}_{12} = \underline{S}_{21}' = \frac{1}{n}\,\underline{x}'\underline{y},$$

and

$$\underline{S}_{22} = \frac{1}{n}\,\underline{y}'\underline{y}$$

are the variance-covariance matrices. Then the $\underline{B}_i$ can be found from

$$\underline{B} = \underline{S}_{22}^{-1}\,\underline{S}_{21}\,\underline{A}\,\underline{\Lambda}^{-1}$$

Alternatively, we could use

$$(\underline{S}_{22}^{-1}\,\underline{S}_{21}\,\underline{S}_{11}^{-1}\,\underline{S}_{12} - \lambda_i^2\,\underline{I})\,\underline{B}_i = \underline{0},$$

$$|\underline{S}_{22}^{-1}\,\underline{S}_{21}\,\underline{S}_{11}^{-1}\,\underline{S}_{12} - \lambda^2\,\underline{I}| = 0,$$

and

$$\underline{A} = \underline{S}_{11}^{-1}\,\underline{S}_{12}\,\underline{B}\,\underline{\Lambda}^{-1}.$$

The latter equations are to be preferred if $q < p$, because the matrix that must be diagonalized is then of a lesser dimension.

The "first" pair of functions, defined by the first column of each $\underline{A}$ and $\underline{B}$, have as large a correlation λ_1 as any other possible pair of functions, each composed of a linear combination of the original variables. Also, the "second" function pair has as large a correlation λ_2 as any other possible pair of functions, each being composed of a linear combination of the original variables and each being uncorrelated with both members of the first pair.

Either set of new variables can be predicted in a least-squares sense by the new variables in the other set. The prediction equations are

$$(\hat{\underline{y}}\,\underline{B}) = \underline{x}\,\underline{A}\,\underline{\Lambda}$$

and

$$(\hat{\underline{x}}\,\underline{A}) = \underline{y}\,\underline{B}\,\underline{\Lambda}.$$

In addition, the original variables in one set can be predicted in a least-squares sense by the new variables in the other set; for example, by

$$\hat{\underline{y}} = \underline{x}\,\underline{A}\,\underline{\Lambda}\,\underline{B}'\underline{S}_{22}. \tag{10}$$

In the case that $r = q$, Eq. (10) can be written as

$$\hat{\underline{y}} = \underline{x}\,\underline{A}\,\underline{\Lambda}\,\underline{B}^{-1}. \tag{11}$$

Similar equations can be written for predicting $\underline{x}$.

Eq. (10) represents the prediction equation for each of the $\hat{y}_i$ in terms of all of the x_i. One may want to relate one set of variables to the other set but include only a portion of the correlations in $\underline{\Lambda}$, perhaps those k correlations that are judged to be significantly different from zero. An equation corresponding to Eq. (10) can be written as

$$\hat{\underline{y}} = \underline{x}\,\underline{A}\,\underline{\Lambda}^*\underline{B}'\underline{S}_{22}, \tag{12}$$

where the rxr matrix $\underline{\Lambda}^*$ has only k nonzero elements, the others having been set equal to zero. Eq. (12) has the effect of including a contribution only from those k columns of $\underline{A}$ and k rows of $\underline{B}'$ corresponding to the k nonzero correlations.

The prediction equations can be expressed in terms of the original variables X_i and Y_i. For instance, Eq. (11) becomes

$$\underline{Y} = \underline{X}\,\underline{A}\,\underline{\Lambda}\,\underline{B}^{-1} - \underline{\bar{X}}\,\underline{A}\,\underline{\Lambda}\,\underline{B}^{-1} + \underline{\bar{Y}}. \qquad (13)$$

Discriminant analysis and multiple regression, including REEP, are special cases of canonical correlation. For instance, when q = 1 Eq. (13) is the same as the least squares regression equation for a single predictand [see Eq. (5)]. In addition, if the predictands represent group membership in the same manner as for REEP, then Eq. (13) is the same as the set of REEP equations (see Glahn, 1968). Therefore, canonical correlation has little to offer in a purely predictive sense over the simpler regression or discriminant analysis (this statement is not meant to imply that canonical correlation is not useful in studying relationships between sets of variables). One possibility for canonical correlation in prediction does exist, and that involves the use of Eq. (12). As stated previously, defining EOFs on a set of predictors and using only those functions that explain a non-trivial portion of the predictor variance as new predictors in a regression equation filters out predictor "noise" (of course, one must be careful to ensure that it really is noise and not good predictive information). Eq. (12) seems to provide a way of filtering noise out of both the predictor and predictand sets and could provide more stable prediction equations. However, we know of no case where this possibility has been investigated. For further discussion of canonical correlation and an example using meteorological data, see Glahn (1968).

8. LOGIT MODEL

The logit model (Brelsford and Jones, 1967; Jones, 1968) provides a means of fitting a sigmoid or S-shaped curve to data when the dependent variable is binary and the independent variable is continuous. From this model, the probability of the binary variable Y having the value of one can be expressed as follows:

$$P(Y = 1|X) = \frac{1}{1 + \exp(\alpha + \beta X)} \qquad (14)$$

The model can also be extended to several independent variables and to several, rather than two, categories of a dependent variable. Determination of the parameters [α and β in Eq. (14)] is usually more difficult than determination of coefficients in a regression equation. Iterative procedures can be used or, if each specific value of X in the sample is repeated and the relative frequency of the dependent event is neither zero or one, then a more direct method of solution can be used for the following linearized form of Eq. (14):

$$\ln\left(\frac{1 - p}{p}\right) = \alpha + \beta X.$$

Usually, in meteorological applications, several predictors are to be included, and this method of solution would require enough replications of each combination of predictor values to estimate the relative frequency of the predictand for that combination. We know of no meteorological application where this method of solution for multiple predictors has been used. For a discussion of this method, see Neter and Wasserman (1974).

9. MAP TYPING

The concept of weather types arose early in the history of meteorology. The aim is to define a partition of weather maps (or sequences of maps), so that the differences between the maps (or sequences) of one type are small compared to the differences between maps (or sequences) of another type. Once a set of weather types has been defined, it can be used in various ways to forecast specific weather elements. Early work was done by Bowie and Weightman (1914) who stratified storms by their movement. Average tracks, expected direction, and average speeds were then computed. An historic development of weather types was done by Irving P. Krick at the California Institute of Technology (1943), leading to the identification of the so-called CIT types.

The determination of map types can be accomplished in many ways. Initially, the methods employed were largely subjective, and even in the application phase, the user had to "decide" what type existed on a given day. More recently, with the advent of the electronic computer and the desire to process large quantities of data, objective methods of classification have been developed. One such method that has been rather extensively applied was developed by Lund (1963). The example he used to explain the method involved the classification of wintertime sea level pressure maps over the northeastern U.S. The steps involved in this method are:

Step 1. Correlate the sea level pressures on each map with the corresponding pressures on all of the other maps in the sample. That is, if each of 500 maps had 25 values of pressure (which could be reported values at stations or at grid points arrived at by an analysis of station values), then each of the 500 maps would be correlated with 499 other maps, the computations of the correlation coefficient involving 25 pairs of values.

Step 2. Select the map which has the most correlation coefficients ≥ 0.7 and designate it as Type A.

Step 3. Remove all of the cases that are correlated ≥ 0.7 with the Type A map, and select from the remaining maps the one with the most correlations ≥ 0.7. Designate this map as Type B.

Step 4. Remove Type B cases, and repeat the process until only a few cases with correlations $\geq$ 0.7 remain.

In application, a map is classified according to the type with which it correlates most highly. Of course, the 0.7 correlation criterion stated above can be modified as desired.

A problem in the use of map types is that a particular map may not classify well into any of the defined types. In the Lund method, after the definition of several types, a few cases generally will remain that are not very similar to any other map in the sample.

Many times only one variable, such as sea level pressure, is used to define the types. However, the evolution of weather systems and the correspondence to predictand variables depend on more than that one element. Other variables can be included in the definition of types, but finding "good" types - that is, maps which resemble others in terms of all the considered variables - is then more difficult.

A possible forecast aid employing map types is to define the conditional precipitation probability at a station given that a particular map type exists. This procedure could involve a lag relationship, in which case the application would probably be to existing maps. On the other hand, it could be a concurrent relationship, in which case the application would probably be to numerical forecasts of the variable(s) used to define the types. This latter approach would be a perfect prog application. Augulis (1969) describes a forecast aid developed along these lines; it is still in use in the western U.S.

10. ANALOGUES

The term analogue can be defined as follows: "In synoptic meteorology, a past large-scale synoptic weather pattern which resembles a given (usually current) situation in its essential characteristics. The use of analogues as an aid in forecasting is based upon the assumption that two similar synoptic weather patterns will retain similarity through at least a short period of further development" (American Meteorological Society, 1959). Analogues were investigated comprehensively by Wadsworth (1948) and their use has been discussed realistically by Willett (1951).

Selecting an analogue is much like selecting a weather type - the idea is to choose one or more maps which are very similar to other maps or to a particular map. Generally, map types of, say, sea level pressure or 500 mb height are employed to forecast other variables such as temperature or precipitation at specific points. However, analogues of, say, sea level pressure and/or 500 mb height may be used to forecast future states of those same variables over

the areas for which the analogues are defined. In these days of numerical prediction models, analogues appear to be of very limited use.

11. PRESENT STATUS

The most concentrated effort today in statistical weather forecasting is at the Techniques Development Laboratory (TDL) of the U.S. National Weather Service. TDL's objective systems, implemented by the National Meteorological Center, produce about 600,000 forecasts daily from about 90,000 regression and logit equations (as of 1 April 1981). These forecasts are disseminated by teletypewriter and facsimile to civilian and military weather stations and to non-government users throughout the United States (Glahn, 1976). The elements being forecast include probability of precipitation, precipitation type, precipitation amount (Bermowitz, 1975), surface wind at land stations (Carter, 1975) and at marine stations and over the Great Lakes (Feit and Pore, 1978), surface temperature and dew point (Dallavalle et al. 1980), severe convective weather (Reap and Foster, 1979; Charba, 1979), cloud amount (Carter and Glahn, 1976), ceiling height and visibility (Globokar, 1974), storm surge (Pore, 1976; Richardson and Pore 1969), and beach erosion. Some of these current applications are discussed briefly in this section.

11.1 Probability of Precipitation

MOS probability of precipitation (PoP) forecasts have been produced operationally by the REEP model for several years. This statistical product replaced the subjectively produced NMC product in January of 1972. The developmental sample was divided into 2 seasons - April through September, the summer season, and October through March, the winter season. The event is defined to be 0.01 inches or more of measurable liquid equivalent precipitation in a 12 h period at a point, represented by a station rain gauge. Separate equations were developed for the 12-24, 36-48, and 48-60 h projections (i.e., lead times) and for each of the initial data times of 0000 and 1200 GMT. Data for several stations within a region were pooled, and one set of equations was developed that applied to all stations within that region. Details of the evolution of the PoP forecasting system are given by Lowry and Glahn (1976).

In terms of the P-score, the MOS PoP's improved upon the climatological relative frequency (defined by month and by station) by about 48%, 33%, and 34% for the 12-24, 24-36, and 36-38 h projections, respectively, for the 1979-80 winter. The corresponding 1979 summer improvements were about 29%, 22%, and 19%. Using the MOS PoP's as guidance, the local forecasters were able to improve upon them by 8.1%, 1.4%, and 2.3% for the three periods, respectively, during the 1979-80 winter.

11.2 Precipitation Type

TDL's system for predicting the conditional probability of precipitation type (PoPT), conditional on the occurrence of precipitation, gives forecasts for three categories: frozen (snow or ice pellets), freezing (freezing rain or drizzle), and liquid (rain or mixed types) (Bocchieri, 1979). The PoPT system evolved from the conditional probability of frozen precipitation (PoF) system (Glahn and Bocchieri, 1975; Bocchieri and Glahn, 1976), which had been operational since November 1972. In PoF, explicit probability forecasts of freezing precipitation were not available. In the PoPT system, one logit equation was developed for each initial data time and each projection. Although data from about 200 stations were used in the development of each equation, the predictors were defined to be departures from 50% values. As an example, consider the 850 mb temperature as a predictor. For each station, the value that specifies a 50% conditional probability of frozen precipitation was found empirically. (This value was actually found by determining a one predictor logit equation for each station.) Then, the 850 mb temperature minus the unique 50% station value was used as a predictor in the multipredictor logit equation.

Heidke skill scores for the 1979-80 winter guidance forecasts were 0.88, 0.86, and 0.84 for 18, 30, and 42 h forecasts, respectively. These scores were computed only for those cases when the local PoP forecasts were greater than or equal to 30%.

11.3 Surface Wind

MOS surface wind forecasts for stations throughout the conterminous United States have been produced since May 1973 (Carter, 1975). Three regression equations are determined for each station for each projection - one for the U component, one for the V component, and one for speed. All three equations have the same predictors to ensure greater consistency between the three forecasts. Forecasts of the U and V components are used to determine direction. A separate equation is used for speed because speeds determined from regression estimates of the U and V components are biased toward zero (Glahn, 1970).

Verification of the MOS forecasts for the 1973-74 through 1979-80 winter shows a definite improving trend. Mean absolute errors in direction for the 1979-80 winter were 26, 30, and 35 degrees for the 18, 30, and 42 h projections, respectively. Corresponding skill scores for speed were 0.35, 0.34, and 0.26. The speed forecasts have been inflated (Klein et al., 1959) since 1975 in order to make a larger number of forecasts of strong winds.

11.4 Surface Temperature

Statistical forecasts of maximum and minimum temperature have been made and disseminated operationally by the National Weather Service since 1965 -longer than any other weather element. Initially the forecasts were made by the perfect prog technique (Klein and Lewis, 1970), but the MOS approach was adopted in August 1973 after considerable testing showed that MOS furnished better forecasts (Annett et al., 1972; Klein and Hammons, 1975). The forecasts are made from regression equations developed for individual stations, one for each of the initial data times and for each projection. A continuing evaluation has shown that MOS improves on the perfect prog forecasts by about 0.5°F in mean absolute error at 24 and 36 h projections. These statistical forecasts have shown a consistent improvement since 1973. The mean absolute error for 24 h maximum temperature forecasts was 3.5°F for the 1979-80 winter period. The forecasters are able to improve on the guidance by a few tenths of a degree Fahrenheit.

11.5 Extratropical Storm Surge

Storm surge is defined to be the piling up of water on the shore due to meteorological conditions. TDL's statistical systems forecast this surge at specific points on the Atlantic (Pore, 1976) and Great Lakes (Richardson and Pore, 1969) coasts due to extratropical storms. The perfect prog technique is used to develop regression equations that relate the surge to concurrent values of sea level pressure at grid points surrounding the forecast points. Since a very good physical basis exists for the dependence of surge on pressure gradients, the forecasts are quite good, and their skill depends mainly on the skill of the numerical model used to provide the pressure forecasts (Pore, 1972). Surge forecasts became operational for Buffalo and Toledo on Lake Erie in October 1969 and for Atlantic coastal stations in October 1971.

11.6 Thunderstorms and Severe Convective Weather

Medium-range (24 h projection) probability forecasts of thunderstorms and severe convective weather have been operationally available since the spring of 1972 (Reap and Foster, 1979). In addition, short-range (2-6 h projection) probability forecasts of the same variables were implemented in the spring of 1974 (Charba, 1977, 1979). The medium-range forecasts are provided by REEP equations developed by the MOS technique. The predictand is defined by radar echoes within a specified time period and within an area approximately 75x75 km. These forecasts of severe convective weather are conditional probabilities. That is, given a thunderstorm within the defined area and time period, the forecast specifies the probability of the occurrence of severe weather. It is interesting to note that reliable forecasts of 30 to 40% can be made even though the climatological relative frequency is only 6%.

The short-range forecasts are also provided by REEP equations, but the equations contain more predictors derived from recent observations (of surface atmospheric variables and radar echoes) than from model output. Therefore, this technique is a blend of the classical and MOS approaches. In addition, the severe storm probabilities as well as the thunderstorm probabilities are unconditional. Reliable probabilities approaching 100% are forecast for both predictands, although for severe storms the climatological frequency for a 4 h period is only about 2%.

12. FUTURE OF STATISTICAL WEATHER FORECASTING

Stochastic-dynamic prediction is a term used to describe models that combine statistics and dynamics and produce output in probability form. These models show some promise and may be the models of the future. However, they require considerably more in the way of computer resources than conventional numerical models, and much more research is required before they can compete with present operational models. Also, like present models, they do not produce forecasts of many weather elements for which forecasts are required - ceiling height, cloud amount, minimum temperature, etc. So it is likely that MOS will be used for many years to translate numerical model forecasts into other needed products. The perfect prog technique may find increased use for medium-range projections if numerical models become accurate enough so that the perfect prog assumption is reasonably satisfied.

More efficient methods of processing large quantities of data, better statistical models, and better use of present models will help to improve and to extend the application of statistical forecasting in the future.

REFERENCES

Allen, R. A., and E. M. Vernon, 1951: Objective weather forecasting. Compendium of Meteorology (T. F. Malone, Ed.). Boston, Mass., American Meteorological Society, pp. 796-801.

American Meteorological Society, 1959: Glossary of Meteorology. Boston, Mass, AMS.

Annett, J. R., H. R. Glahn, and D. A. Lowry, 1972: The use of model output statistics (MOS) to estimate daily maximum temperatures. Silver Spring, Md., NOAA, National Weather Service, Technical Memorandum NWS TDL-45, 14 pp.

Anderson, T. W., 1958: An Introduction to Multivariate Statistical Analysis. New York, John Wiley and Sons.

Augulis, R. P., 1969: Precipitation probabilities in the Western Region associated with winter 500 mb map types. Salt Lake City, Utah, ESSA, National Weather Service, Technical Memorandum WBTM WR 45-1, 91 pp.

Barnard, M., 1935: The secular variations of skull characters in four series of Egyptian skulls. Annals of Eugenics, 6, 352-371.

Bartlett, M. S., 1934: The vector representation of a sample. Proceedings of the Cambridge Philosophical Society, 30, 327-340.

Bermowitz, R. J., 1975: An application of model output statistics to forecasting quantitative precipitation. Monthly Weather Review, 103, 149-153.

Besson, L., 1905: Essai de prevision methodique du temps. Observatorie Municipal de Monsouris, Annals, 6, 473-495.

Bocchieri, J. R., 1979: A new operational system for forecasting precipitation type. Monthly Weather Review, 107, 637-649.

Bocchieri, J. R., and H. R. Glahn, 1972: Use of model output statistics for predicting ceiling height. Monthly Weather Review, 100, 869-879.

Bocchieri, J. R., and H. R. Glahn, 1976: Verification and further development of an operational model for forecasting the probability of frozen precipitation. Monthly Weather Review, 104, 691-701.

Boehm, A. R., 1976: Transnormalized regression probability. Scott Air Force Base, Ill., USAF, Air Weather Service, Technical Report 75-259, 52 pp.

Bowie, E. H., and R. H. Weightman, 1914: Types of storms of the United States and their average movements. Monthly Weather Review, Washington Supplement No. 1, 147 pp.

Brelsford, W. M., and R. H. Jones, 1967: Estimating probabilities. Monthly Weather Review, 95, 570-576.

Brier, G. W., 1940: The discriminant function. Washington, D.C., George Washington University, M.A. Thesis, 34 pp.

Brier, G. W., 1946: A study of quantitative precipitation forecasting in the TVA basin. Washington, D.C., U.S. Weather Bureau, Research Paper No. 26, 40 pp.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. Monthly Weather Review, 79, 1-3.

Bryan, J. G., 1944: Special techniques in multiple regression. Cambridge, Massachusetts Institute of Technology, unpublished manuscript, 17 pp.

Bryan, J. G., 1950: A method for the exact determination of the characteristic equation and latent vectors of a matrix with applications to the discriminant function for more than two groups. Cambridge, Mass., Harvard University, Ed. D. Dissertation, 290 pp.

California Institute of Technology, 1943: Synoptic weather types of North America. Pasadena, Calif., Department of Meteorology, Report, 237 pp.

Carter, G. M., 1975: Automated prediction of surface wind from numerical model output. Monthly Weather Review, 103, 866-873.

Carter, G. M., and H. R. Glahn, 1976: Objective prediction of cloud amount based on model output statistics. Monthly Weather Review, 105, 1565-1572.

Charba, J. P., 1977: Operational system for predicting thunderstorms two to six hours in advance. Silver Spring, Md., NOAA, National Weather Service, Technical Memorandum NWS TDL-64, 24 pp.

Charba, J. P., 1979: Two to six hour severe local storm probabilities: an operational forecasting system. Monthly Weather Review, 107, 268-282.

Dallavalle, J. P., J. S. Jensenius, Jr., and W. H. Klein, 1980: Improved surface temperature guidance from the limited-area fine mesh model. Preprints, Eighth Conference on Weather Forecasting and Analysis (Denver). Boston, Mass., American Meteorological Society, pp. 1-8.

Enger, I., J. A. Russo, Jr., and E. L. Sorenson, 1964: A statistical approach to 2-7 hr prediction of ceiling and visibility, volumes I and II. Hartford, Conn., Travelers Research Center, Inc., Contract No. CWB-10704, 48 pp. and 195 pp., respectively.

Feit, D. M., and N. A. Pore, 1978: Objective wind forecasting and verification on the Great Lakes. Journal of Great Lakes Research, 4, 10-18.

Fisher, R. A., 1936: The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, Part II, 179-188.

Fix, C., and J. L. Hodges, Jr., 1951: Discriminatory analysis, nonparametric discrimination: consistency properties. Randolph Field, USAF, School of Aviation Medicine, Report No. 4.

Freeman, M. H., 1961: A graphical method of objective forecasting derived by statistical techniques. Quarterly Journal of the Royal Meteorological Society, 87, 393-400.

Gilman, D. L., 1957: Empirical orthogonal functions applied to thirty-day forecasting. Cambridge, Massachusetts Institute of Technology, Department of Meteorology, Contract No. AF19 (604)-1283, Scientific Report No. 1, 129 pp.

Glahn, H. R., 1962: An experiment in forecasting rainfall probabilities by objective methods. Monthly Weather Review, 90, 59-67.

Glahn, H. R., 1965: Objective weather forecasting by statistical methods. The Statistician, 15, 111-142.

Glahn, H. R., 1968: Canonical correlation and its relationship to discriminant analysis and multiple regression. Journal of Atmospheric Sciences, 25, 23-31.

Glahn, H. R., 1970: A method for predicting surface winds. Silver Spring, Md., ESSA, National Weather Service, Technical Memorandum WBTM TDL 29, 18 pp.

Glahn, H. R., 1976: Progress in the automation of public weather forecasts. Monthly Weather Review, 104, 1505-1512.

Glahn, H. R., and J. R. Bocchieri, 1975: Objective estimation of the conditional probability of frozen precipitation. Monthly Weather Review, 103, 3-15.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. Journal of Applied Meteorology, 11, 1203-1211.

Globokar, F. T., 1974: Computerized ceiling and visibility forecasts. Preprints, Fifth Conference on Weather Forecasting and Analysis (St. Louis). Boston, Mass., American Meteorological Society, pp. 228-233.

Grimmer, M., 1963: The space-filtering of monthly surface temperature anomaly data in terms of pattern, using empirical orthogonal functions. Quarterly Journal of the Royal Meteorological Society, 89, 395-408.

Gringorten, I. I., 1955: Methods of objective weather forecasting. Advances in Geophysics, Vol. II. New York, Academic Press, Inc., pp. 57-92.

Hotelling, H., 1935: The most predictable criterion. Journal of Educational Psychology, 26, 139-142.

Hotelling, H., 1936: Relations between two sets of variates. Biometrika, 28, 321-377.

Jones, R. H., 1968: A nonlinear model for estimating probabilities of k events. Monthly Weather Review, 96, 383-384.

Jorgensen, D. L., 1959: Prediction of hurricane motion with use of orthogonal polynomials. Journal of Meteorology, 16, 21-29.

Klein, W. H., 1969: The computer's role in weather forecasting. Weatherwise, 22, 195-218.

Klein, W. H., and G. A. Hammons, 1975: Maximum/minimum temperature forecasts based on model output statistics. Monthly Weather Review, 103, 796-806.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperature during winter. Journal of Meteorology, 16, 672-682.

Klein, W. H., and F. Lewis, 1970: Computer forecasts of maximum and minimum temperatures. Journal of Applied Meteorology, 9, 350-359.

Lorenz, E. N., 1956: Empirical orthogonal functions and statistical weather prediction. Cambridge, Massachusetts Institute of Technology, Department of Meteorology, Scientific Report No. 1, 49 pp.

Lowry, D. A., and H. R. Glahn, 1976: An operational model for forecasting probability of precipitation - PEATMOS POP. Monthly Weather Review, 104, 221-232.

Lubin, A., and A. Summerfield, 1951: A square root method of selecting a minimum set of variables in multiple regression: I. The method. Psychometrika, 16, 271-284.

Lund, I. A., 1955: Estimating the probability of a future event from dichotomously classified predictors. Bulletin of the American Meteorological Society, 36, 325-328.

Lund, I. A., 1963: Map-pattern classification by statistical methods. Journal of Applied Meteorology, 2, 56-65.

Miller, R. G., 1958: The screening procedure. In Studies in Statistical Weather Prediction (B. Shorr, Ed.). Hartford, Conn., Travelers Research Center, Inc., Contract No. AF19 (604)-1590, Final Report, pp. 86-95.

Miller, R. G., 1962: Statistical prediction by discriminant analysis. Meteorological Monographs, 4, No. 25, 54 pp.

Miller, R. G., 1964: Regression estimation of event probabilities Hartford, Conn., Travelers Research Center, Contract Cwb-107(Technical Report No. 1, 153 pp.

Miller, R. G., Ed., 1977: Selected topics in statistical meteoro ogy. Scott Air Force Base, Ill., USAF, Air Weather Service, AWS-TR-77-273, 164 pp.

Mook, C. P., 1948: An objective method of forecasting thunderstor for Washington, D.C., in May. Washington, D.C., U.S. Weathe Bureau, unpublished manuscript.

Murphy, A. H., 1974: A sample skill score for probability foreca Monthly Weather Review, 102, 48-55.

Neter, J., and W. Wasserman, 1974: Applied Linear Statistical Models. Homewood, Ill., Richard D. Irwin, Inc.

Panofsky, H. A., and G. W. Brier, 1958: Some Applications of Statistics to Meteorology. University Park, Pennsylvania State University, College of Mineral Industries.

Pore, N. A., 1972: Marine conditions and automated forecasts for Atlantic coastal storm of February 18-20, 1972. Monthly Wea Review, 101, 363-370.

Pore, N. A., 1976: Automated forecasting of extratropical storm surges. Proceedings, Fifteenth Coastal Engineering Conferen (Honolulu), Vol. 1, pp. 906-913.

Pore, N. A., and W. S. Richardson, 1969: Second interim report o sea and swell forecasting. Silver Spring, Md., ESSA, Nation Weather Service, Technical Memorandum WBTM TDL 17, 17 pp.

Rao, C. R., 1952: Advanced Statistical Methods in Biometric Research. New York, John Wiley and Sons.

Reap, R. M., and D. S. Foster, 1979: Automated 12-36 hour probab ity forecasts of thunderstorms and severe local storms. Jou of Applied Meteorology, 18, 1304-1315.

Richardson, W. S., and N. A. Pore, 1969: A Lake Erie storm surge forecasting technique. Silver Spring, Md., ESSA, National Weather Service, Technical Memorandum WBTM TDL 24, 23 pp.

Shuman, F. G., and J. B. Hovermale, 1968: An operational six-lay primitive equation model. Journal of Applied Meteorology, 7 525-547.

Suits, D. B., 1957: Use of dummy variables in regression equatic Journal of the American Statistical Association, 52, 548-551