551.509.325.2
ISSN 0110-5965

NEW ZEALAND METEOROLOGICAL SERVICE

# TECHNICAL NOTE NO.256

# STATISTICAL FORECASTING OF PROBABILITY OF PRECIPITATION

J.A. Renwick

Issued for limited distribution by:

The Director
New Zealand Meteorological Service
P.O. Box 722
WELLINGTON

14 July 1983

# Statistical Forecasting of Probability of Precipitation

J.A. Renwick

ABSTRACT

The problem of objectively forecasting probability of precipitation (PoP) for a number of stations around New Zealand has been approached using two different statistical methods, discriminant analysis and linear regression. Both methods use output from a numerical weather prediction (NWP) scheme as predictands.

Linear regression predictions have properties which are more desirable for operational use than those calculated by discriminant analysis, although results are similar for both methods in terms of fit to dependent data. It is concluded that objective forecasts of PoP may be a useful forecasting tool, especially over relatively short prognosis periods.

## 1. BACKGROUND

### 1.1 Introduction

Currently in the National Weather Forecasting Centre, forecasts issued to the general public expressing uncertainty about any weather element are expressed grammatically, with terms such as 'possible', 'likely', etc. Forecasts are not issued in terms of probability of occurrence. Comparing the two ways of expressing uncertainty, it seems the public understand numerical probability forecasts as well as qualitative terms (Murphy, *et al*., 1980). Forecasts issued in terms of probability of occurrence are especially desirable to some users, for example in weather dependent industry, since a definite cost/penalty system can be applied to these forecasts. This allows strategies to be calculated for minimization of expected loss due to unsuitable weather conditions, etc.

As a first step in the preparation of operational PoP forecasts, it is useful to have some form of objectively calculated parameter for use as guidance. One way of objectively calculating a forecast of any weather element over a short time scale is to relate output from a numerical weather prediction (NWP) model to the predictand weather element by some form of statistical scheme. The weather element studied here is 'occurrence of precipitation', which has been taken as a total recorded precipitation of 0.1 millimetres or more during the period under consideration.

Two possible techniques may be applied when relating NWP fields and weather elements. First, one may rely on the "perfect prognosis" assumption, where relationships are derived between predictands and analyses of NWP fields. To obtain a forecast operationally, relevant prognosed values from the NWP model are substituted, giving an estimate of the predictand at some future time. The assumption here is that the prognosis used will match perfectly with its verifying analysis. Alternatively, one may relate prognoses directly with predictand values, obtaining a separate relationship for each prognosis used. This is known as the Model Output Statistics (MOS) technique. The main advantage of this technique is that as the prognosis period increases, the prediction equations become more conservative because less correlation can be found between the predictors and predictand. This means the long period forecasts tend toward climatology, reflecting the increased uncertainty in the predictor fields. Also any systematic bias in the prognosed NWP fields will be taken into account in the derived relationships. The MOS technique has been used throughout this work.

## 1.2 Statistical Methods

### (a) Linear Discriminant Analysis.

Given some set of predictors split into G ( $\geq 2$) groups based on a predictand variable, a discriminant analysis calculates new predictors (known as discriminant functions) which maximise the ratio of variance between the given groups to variance within these groups. There may be up to G-1 new predictors, which are uncorrelated linear combinations of the original predictors. Assuming some distribution for the discriminant functions (or using some non-parametric technique), one may calculate the probability that a new observation belongs to any of the groups by measuring the 'nearness' of the new observation to each group of the dependent sample.

That is, if the assumed density function for predictors $y_1, \ldots y_t$ in group g is $f(y_1, \ldots y_t | g)$ and the *a priori* probability of group g occurring is $\Omega_g$ (the climatological probability or relative frequency from the dependent sample), Bayes Theorem may be used to give the probability that a new observation of the predictors belongs to group g (Miller, 1962):

$$f(g | y_1, \ldots, y_t) = \frac{f(y_1, \ldots y_t | g) \cdot \Omega_g}{\sum_{k=1}^{G} f(y_1, \ldots y_t | k) \cdot \Omega_k} \qquad (1)$$

where G is the total number of groups and $f(g | y_1, \ldots y_t)$ is the *a posteriori* probability that the observation $y_1, \ldots y_t$ belongs to group g.

In the work that follows, the data used have been split into two groups (G=2) for each station, depending on whether or not measurable precipitation fell at that station near to the valid time of the predictors. Having only two groups, only one discriminant function is calculable. The appropriate density functions have been assumed to be normal throughout.

(b) Multiple Linear Regression.

The method of multiple linear regression involves the estimation by least squares of a linear relationship between a set of predictors and some predictand (Ostle and Mensing, 1975). In classical linear regression, the predictand is assumed to be continuous and normally distributed. However, for this application, the predictand is not continuous and may take only two values; 0 when precipitation did not occur at the station and 1 when precipitation did occur. Hence a predictand of this sort is known as 'binary', either 'on' (set at 1) or 'off' (set at 0). The estimates yielded from an equation on such a predictand are viewed as the estimated probability of occurrence of the event in question.

## 1.3 Predictors Used

Predictors were taken from archived prognoses of fields produced by the New Zealand Meteorological Service numerical weather prediction model (Trenberth, 1973). The grid for this model covers a large area of the Pacific as well as the Australia and New Zealand region. All predictors used were extracted from a small subsection of that grid which covers the New Zealand area, as used for numerical forecasts of maximum and minimum temperatures (Renwick, 1980).

Fields used were 1000, 700 and 500 hectopascal (hPa)* geopotential height, 600hPa vertical motion and 1000-500hPa mean relative humidity for analysis times of 0000 and 1200GMT and prognosis periods of 12,34,36 and 48 hours. Archives of all these fields have been made routinely since the beginning of 1979. The dependent data set was taken from archives covering the period February 1979 to December 1981 inclusive.

Predictor variables used were principal components of each field as well as averaged grid point values of the vertical motion and mean relative humidity fields at the grid point nearest each station forecast for. The principal components are coefficients of empirical orthogonal functions (EOFs), or eigenvectors, of the cross covariance matrix of each field over all the grid points used. They were calculated separately for each prognosis period. For each field, out of a possible thirty EOFs, the first six were taken as predictors.

No clear-cut tests for significance of EOFs exist, but a crude $\chi^2$ test on the eigenvalues suggested that only the first three for each height field and the first four, or in some cases five, for the vertical motion and relative humidity fields, were significantly different from zero. Six principal components of the height fields were included as possible predictors, since work on statistical temperature forecasting suggests that the fourth, fifth and sixth height field components contribute useful information about the flow patterns, even though they do not appear to be nominally significant. A cutoff of six components was taken arbitrarily for the other two fields for programming simplicity.

Figures for explained variance for each of the chosen EOFs are as in Table 1.

---

*1 hectopascal (hPa) = 1 mb (millibar)

Table 1:

Explained Variance (Percent) for First Six EOFs of each Field

| Field type and EOF number | | PROGNOSIS PERIOD (HR) 12 | 24 | 36 | 48 |
|---|---|---|---|---|---|
| 1000 hPa height | 1 | 68 | 67 | 63 | 63 |
| | 2 | 14 | 14 | 17 | 17 |
| | 3 | 11 | 11 | 10 | 10 |
| | 4 | 3 | 3 | 4 | 4 |
| | 5 | 2 | 2 | 3 | 3 |
| | 6 | 1 | 1 | 1 | 1 |
| | total | 99 | 98 | 98 | 98 |
| 700 hPa height | 1 | 74 | 74 | 75 | 76 |
| | 2 | 13 | 14 | 14 | 14 |
| | 3 | 8 | 7 | 7 | 6 |
| | 4 | 2 | 2 | 2 | 2 |
| | 5 | 1 | 1 | 1 | 1 |
| | 6 | 1 | 1 | 0.5 | 0.5 |
| | total | 99 | 99 | 99.5 | 99.5 |
| 500 hPa height | 1 | 76 | 78 | 79 | 80 |
| | 2 | 12 | 11 | 12 | 12 |
| | 3 | 7 | 7 | 5 | 4 |
| | 4 | 2 | 2 | 2 | 1 |
| | 5 | 1 | 1 | 1 | 1 |
| | 6 | 1 | 0.5 | 0.5 | 0.5 |
| | total | 99 | 99.5 | 99.5 | 98.5 |
| 600 hPa vert. motion | 1 | 25 | 29 | 30 | 28 |
| | 2 | 23 | 24 | 26 | 28 |
| | 3 | 11 | 11 | 11 | 10 |
| | 4 | 8 | 7 | 7 | 7 |
| | 5 | 6 | 6 | 6 | 5 |
| | 6 | 4 | 4 | 4 | 3 |
| | total | 77 | 81 | 84 | 81 |
| Mean rel. humidity | 1 | 29 | 30 | 31 | 32 |
| | 2 | 22 | 24 | 25 | 26 |
| | 3 | 12 | 12 | 13 | 13 |
| | 4 | 10 | 9 | 8 | 9 |
| | 5 | 6 | 6 | 7 | 6 |
| | 6 | 4 | 4 | 3 | 3 |
| | total | 83 | 85 | 87 | 89 |

For the height fields, the first EOF is obviously the most significant and the time series of its coefficients is largely a seasonal trend - typically about fifty percent of the variance in the time series is explained by a simple seasonal trend equation. The trends were subtracted from these series and the remaining anomalies were used as predictors.

For the vertical motion and mean relative humidity fields, however, explained variance is spread much more evenly over the first two or three EOFs and no significant seasonal trend is evident in their coefficients. Figure 1 illustrates the first six EOFs for each field for the 12 hour prognosis period. Corresponding EOF patterns are similar for the three other prognosis periods.

Apart from principal components, actual values of the magnitude of vertical motion and relative humidity fields at thirteen of the NWP model gridpoints were also used as possible predictors. These values were averages of the magnitude of either field, taken over the nine grid points centered on and weighted towards the grid point closest to each of the stations. Principal components effectively describe gradients of the fields used, so that components of the height fields used describe the wind flows associated with those fields. Preliminary work suggested that while measures of the gradients of the height fields were more useful for prediction than actual point values of geopotential, the same was not the case for vertical motion and relative humidity. Hence the inclusion of grid point values as well as principal components of vertical motion and relative humidity.

## 1.4 Predictands

### 1.4.1 Stations and Time Periods Used

Rainfall records from thirteen stations around New Zealand were obtained from the hourly rainfall tape files of the New Zealand Meteorological Service held at the Trentham Computer Centre. Records were made available for the period of the complete NWP archive, January 1979 to mid 1982. Prediction equations were calculated on the basis of rainfall totals over two twelve hour periods; 6 am to 6pm

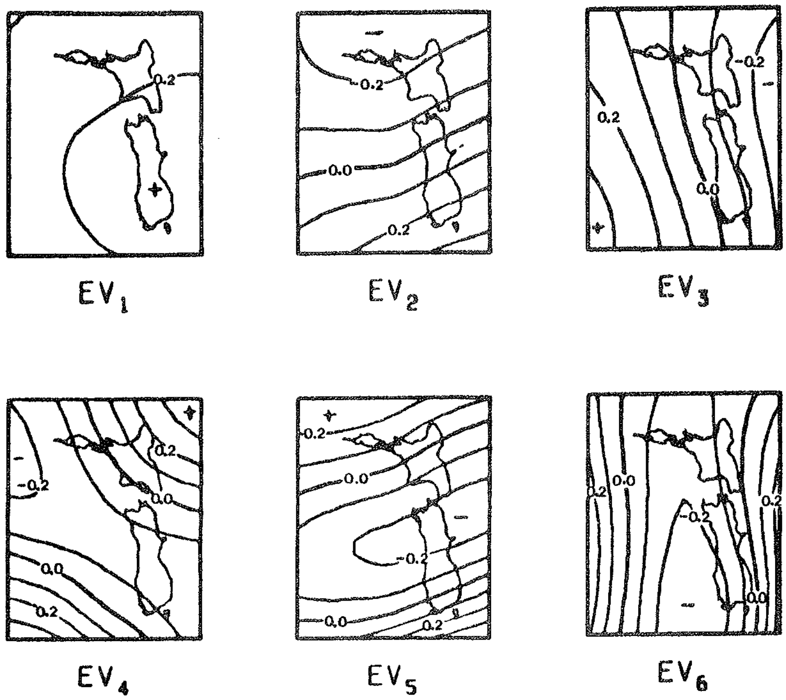


FIGURE 1(a) 1000 hPa HEIGHT EIGENVECTORS.
12 HOUR PROGNOSIS PERIOD.

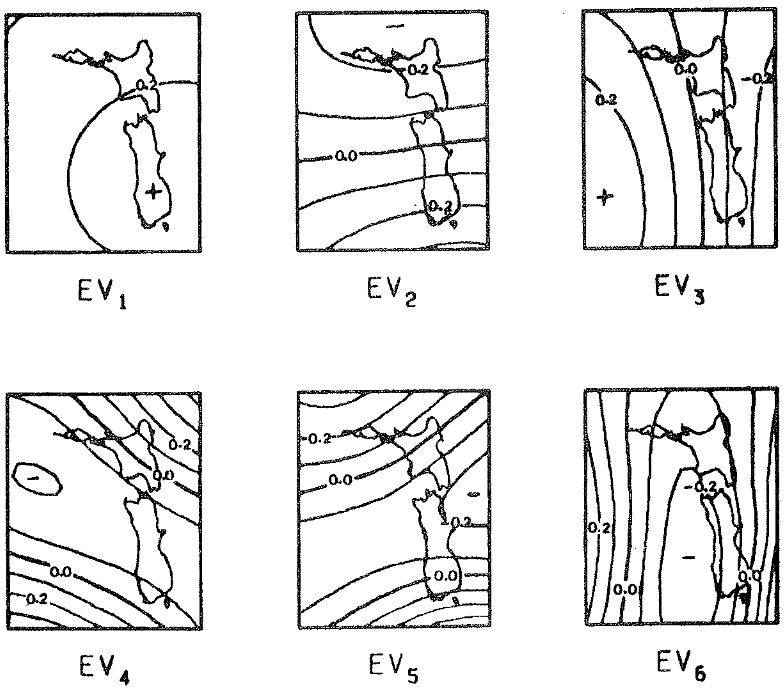


FIGURE 1(b) 700 hPa HEIGHT EIGENVECTORS.
12 HOUR PROGNOSIS PERIOD.

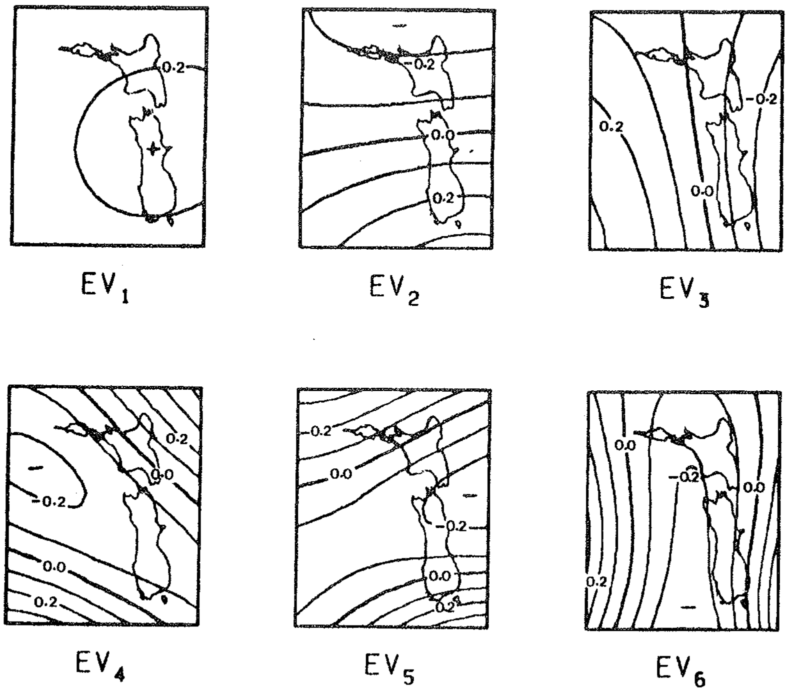


FIGURE 1(c) 500 hPa HEIGHT EIGENVECTORS.
12 HOUR PROGNOSIS PERIOD.

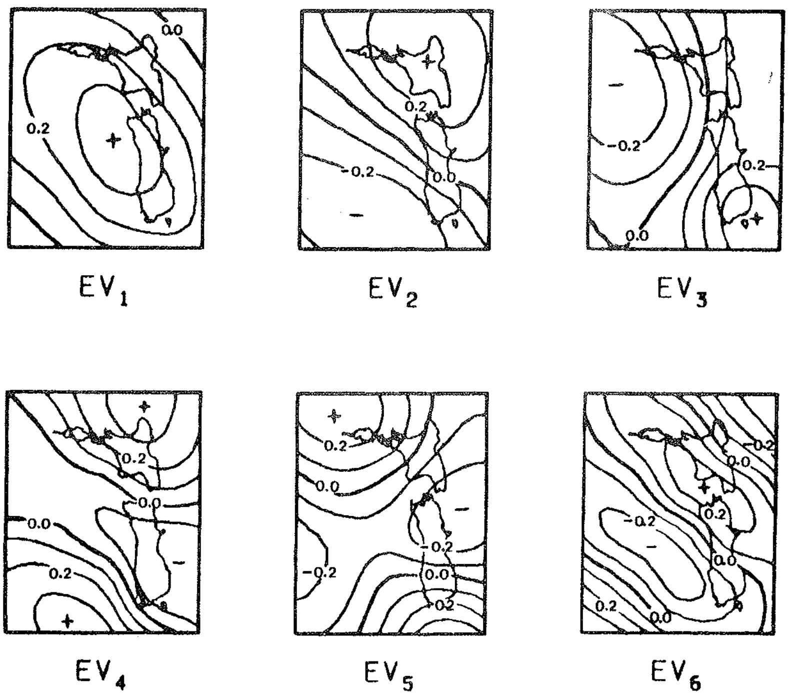


FIGURE 1(d) 600 hPa VERTICAL MOTION EIGENVECTORS.
12 HOUR PROGNOSIS PERIOD.

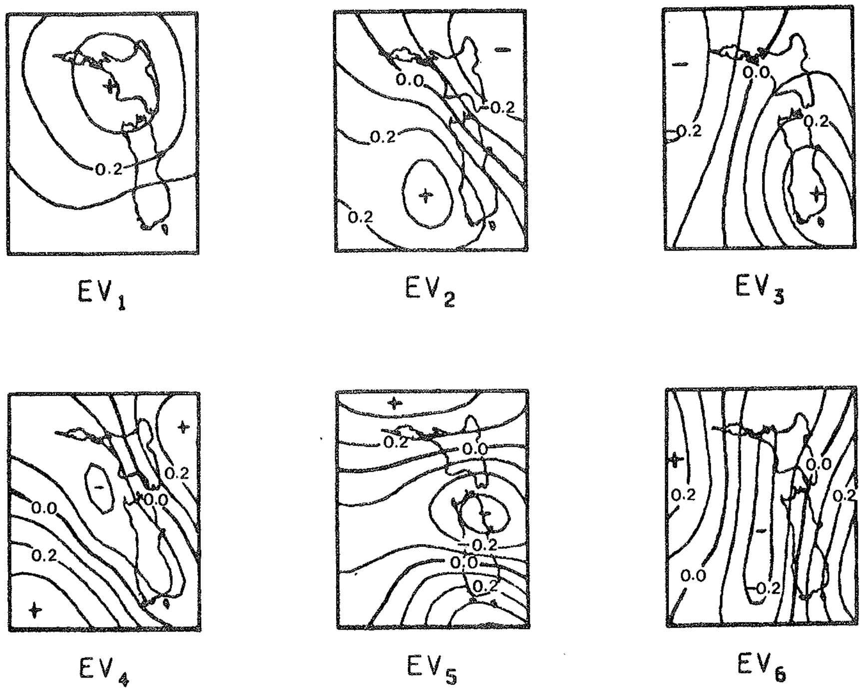


FIGURE 1 (e) 1000 - 500 hPa MEAN RELATIVE HUMIDITY EIGENVECTORS. 12 HOUR PROGNOSIS PERIOD.

('day') and 6 pm to 6am ('night'). The dependent sample covered the period February 1979 to December 1981 inclusive and comprised approximately 900 observations for each twelve hour period for each station. Predictors valid at 0000GMT were used for the 'day' period and those valid at 1200GMT were used for the 'night' period. The list of stations used along with the relative frequency of occurrence of precipitation (for both twelve hour periods combined) may be seen in Table 2.

Table 2:

Stations used in PoP Forecasting

| | Station Name | Relative frequency (%) of precipitation (from dept. sample) |
|---|---|---|
| AK | Auckland airport | 38 |
| RO | Rotorua " | 34 |
| GS | Gisborne " | 30 |
| NP | New Plymouth " | 41 |
| OH | Ohakea airforce base | 33 |
| PP | Paraparaumu " | 36 |
| KL | Kelburn | 34 |
| NS | Nelson " | 26 |
| HK | Hokitika " | 46 |
| KI | Kaikoura " | 23 |
| CH | Christchurch " | 24 |
| DN | Dunedin " | 32 |
| NV | Invercargill " | 42 |

### 1.4.2 Seasonal Stratification

Apart from being split into two sections for day and night, prediction equations were calculated season by season as well as for all seasons combined. Prediction equations calculated by season rely on only two hundred to two hundred and fifty observations each. The results obtained varied markedly through the seasons and gave very poor fit in spring and summer for some stations, especially using the discriminant approach. Hence most results derived have been over the year as a whole. Over the period of the dependent sample, probability of occurrence of rain exhibited a slight seasonal variation at all of the stations used (of the order of two percent of the total variance in the predictand probabilities). These variations were subtracted from each predictand series, as for the principal component series. Results presented have been based on analysis of all seasons combined.

## 2. RESULTS

### 2.1 Derivation of Prediction Equations

(a) Discriminant analysis

For the two group case, a discriminant function defines one 'best' predictor from those available, that is, the linear combination of original predictors which maximises between group to within group variance. A screening technique was employed (Miller, 1962) to determine the most significant initial predictors out of a possible thirty-two, i.e. the thirty principal component values plus the grid point values of vertical motion and mean relative humidity nearest each station. The screening technique uses a Chi-squared test on the increase in between group variance to within group variance for each new predictor entered. The significance level used was 0.05 (95%), adjusted for degrees of freedom. Since the humidity and vertical motion fields show a more rapid deterioration in accuracy with increasing prognosis period than the geopotential height fields, the geopotential height predictors became relatively more strongly correlated with the predictands than humidity and vertical motion predictors over the longer prognosis periods. Typically, about eight predictors were chosen, with the five most important being (in order of frequency of choice):

Relative humidity at the nearest grid point:

1st 1000 hPa principal component
5th 1000 hPa principal component
2nd 1000 hPa principal component
1st vertical motion principal component.

The potential usefulness of the discriminant functions may be judged by two parameters:

$$E^2 = \frac{\text{Sum of squares accounted for between groups}}{\text{Total sum of squares}} \qquad (2)$$

which may be compared to the 'variance reduction', $R^2$, of regression analysis. Values of $E^2$ vary between 0 and 100 percent.

and

(ii) The ratio

$$\lambda = \frac{\text{Sum of squares accounted for between groups}}{\text{Sum of squares accounted for within groups}} \qquad (3)$$

which is a measure of the discriminatory value of the derived equation. This ratio is preferably 'large' and may be very large if within group sum of squares is small. For the two group case, these two parameters are related as follows:

$$E^2 = \frac{100\,\lambda}{1 + \lambda} \qquad (4)$$

when $E^2$ is expressed as a percentage. In Table 3, the values of $E^2$ are displayed (for twelve and forty-eight hour prognosis periods), as they may be most easily compared with variance reductions in regression analysis.

Table 3:
Correlation Ratio $E^2$ (percent) for Discriminant Analysis

| STATION | 12 HOUR PROGNOSIS DAY | 12 HOUR PROGNOSIS NIGHT | 48 HOUR PROGNOSIS DAY | 48 HOUR PROGNOSIS NIGHT |
|---|---|---|---|---|
| AK | 34.1 | 32.1 | 14.0 | 11.5 |
| RO | 38.6 | 37.9 | 22.3 | 19.2 |
| GS | 30.3 | 33.8 | 14.8 | 19.3 |
| NP | 38.6 | 33.7 | 17.6 | 15.8 |
| OH | 34.4 | 32.2 | 19.6 | 12.8 |
| PP | 29.0 | 28.7 | 13.4 | 13.4 |
| KL | 32.6 | 36.5 | 12.1 | 17.7 |
| NS | 33.3 | 28.2 | 15.3 | 14.3 |
| HK | 49.7 | 48.7 | 29.7 | 24.3 |
| KI | 34.8 | 36.6 | 13.3 | 12.6 |
| CH | 30.2 | 29.5 | 15.0 | 7.4 |
| DN | 23.4 | 19.9 | * | 7.8 |
| NV | 26.8 | 24.4 | 14.8 | 12.1 |

* No significant discriminant function found.

Table 3 demonstrates the deterioration with time (prognosis period) of the predictive value of relationships derived. While data for some stations are fitted relatively well at the 12 hour prognosis period (those whose a priori probability of occurrence of precipitation is close to 50%), all results at the 48 hour period are well below what may be considered 'useful', with no significant discriminant function being found for 'day' at Dunedin airport at this prognosis period.

(b) Multiple Linear Regression

A forward stepwise screening regression has been used to pick the most significant predictors out of the possible thirty-two. A standard analysis of variance F-test on increase in regression sum of squares, at the 0.05 level, yielded about 9 predictors for each station. The number of significant predictors tended to decrease with increasing prognosis period, reflecting the decrease in accuracy of the NWP prognoses. Results were similar to those for discrimination, with the five most significant predictors overall being (in order of frequency of choice):

1st 1000 hPa principal component
Mean relative humidity at nearest grid point
3rd 1000 hPa principal component
5th 1000 hPa principal component
2nd 1000 hPa principal component

The main parameter quoted in regression analysis is the variance reduction, defined as:

$$R^2 = \frac{\text{Sum of squares accounted for by the regression}}{\text{Total sum of squares}} \quad (5)$$

Values of the variance reduction, $R^2$, are similar to those of $E^2$ for discrimination. Table 4 shows values of variance reduction at 12 and 48 hour prognosis periods.

Table 4:
Variance Reduction $R^2$ (Percent) for Regressions

| STATION | 12 HOUR PROGNOSIS DAY | 12 HOUR PROGNOSIS NIGHT | 48 HOUR PROGNOSIS DAY | 48 HOUR PROGNOSIS NIGHT |
|---|---|---|---|---|
| AK | 34.7 | 33.8 | 16.0 | 13.8 |
| RO | 40.7 | 39.9 | 22.9 | 22.7 |
| GS | 32.0 | 34.7 | 17.3 | 20.9 |
| NP | 39.5 | 34.9 | 20.9 | 18.5 |
| OH | 36.8 | 34.4 | 20.3 | 14.3 |
| PP | 31.3 | 30.9 | 13.5 | 17.3 |
| KL | 35.0 | 36.3 | 15.3 | 15.5 |
| NS | 36.4 | 29.9 | 17.3 | 14.8 |
| HK | 49.9 | 48.0 | 31.3 | 25.4 |
| KI | 34.8 | 37.2 | 14.1 | 16.2 |
| CH | 32.5 | 30.5 | 15.6 | 11.5 |
| DN | 23.4 | 20.0 | 10.7 | 8.1 |
| NV | 27.0 | 26.5 | 16.3 | 13.4 |

As for discrimination, variance reduction decreases markedly with increasing prognosis period. Only a few of the stations seem to have usable results for 12 hour prognoses (also for 24 hour prognoses), while none appears useful at 26 and 48 hours.

## 2.2 Skill Scores

Apart from tests of significance applied to equations derived on the dependent data, there are various skill scores used to test the value of probability forecasting schemes. Perhaps the most useful and most easily interpreted is the Brier Score (Brier and Allen, 1951) which may be defined for a binary predictand as

$$\overline{B} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \qquad (6)$$

where $y_i$ is the observed probability (0 or 1) and $y_i$ is the forecast probability. N is the number of points in the sample The score defined in (6) is one half of the score originally proposed by Brier and Allen. This score may be thought of as the 'mean residual variance', that amount of variation in the predictand which is unaccounted for by the prediction scheme. Perfect forecasts have a score of zero and forecasts perfectly negatively correlated with the predictand have a score of 1. Most of the following results are in terms of this score.

For regression analysis, $R^2$ is related to the Brier score of the regression on the dependent data as follows;

$$\bar{B} = q(1-q)(1-R^2) \qquad (7)$$

where q is the 'climatological probability' of precipitation (i.e. the relative frequency from the dependent sample). A regression estimate of a predictand which is either 0 or 1 does not necessarily lie within those bounds. Any estimate outside those bounds may be truncated, however, which will decrease (improve) the Brier score slightly. Any constant forecast (e.g. climatology) of a probability 'p' receives a score of $q(1-p)^2 + (1-q)p^2$, where q is the true climatological probability of the event occurring. This is minimised when p is chosen to be equal to q.

The only other score presented here is the Hanssen Score (Hanssen and Kuipers, 1965), defined as

$$I = \frac{\text{number of rain events correctly forecast}}{\text{total number of rain events}} + \frac{\text{number of non-rain events correctly forecast}}{\text{total number of non-rain events}} - 1 \qquad (8)$$

It has a value of 1 for perfect forecasts, 0 for forecasts uncorrelated with the predictand and -1 for forecasts perfectly negatively correlated with the predictand. This score involves a categorization of the probability forecasts into yes/no forecasts and hence may be varied depending on the threshold chosen for categorizing the forecasts (it is maximised when the cutoff for categorization is taken as the climatological probability of occurrence). For this reason, it does not seem as suitable as the Brier score for marking probability forecasts.

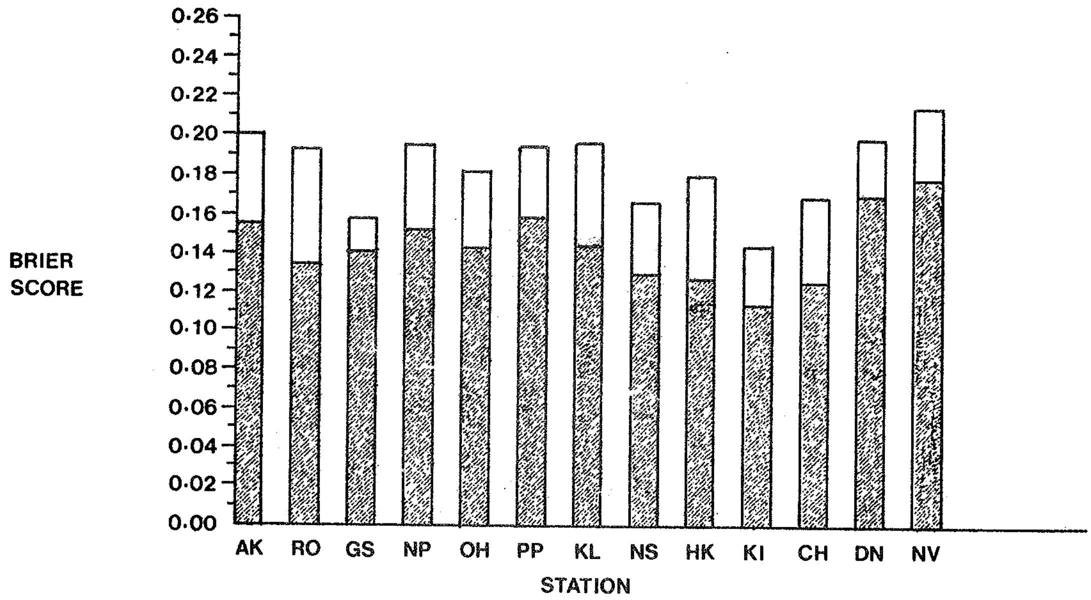


FIGURE 2a: BRIER SCORES - DEPENDENT DATA. 12 HOUR PROGNOSES. REGRESSION VERSUS DISCRIMINATION.
REGRESSION: DISCRIMINATION:

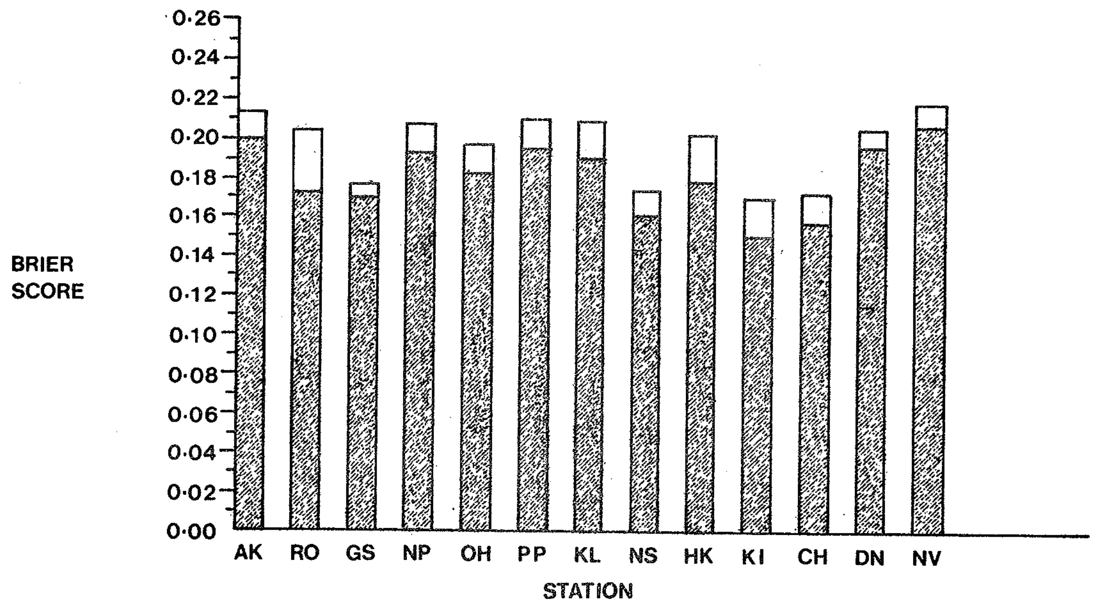


FIGURE 2b: BRIER SCORES - DEPENDENT DATA. 48 HOUR PROGNOSES. REGRESSION VERSUS DISCRIMINATION.
REGRESSION: DISCRIMINATION:

### 2.2.1 Dependent Data

Figure 2 shows Brier scores for both statistical methods on the dependent sample for 12 and 48 hour prognoses. The values for regressions have been calculated using equation (7). For the 12 hour forecasts, regression estimates give a consistently lower Brier score. This is also apparent at longer prognosis periods, but the difference is not so marked. These figures highlight the greater operational usefulness of regression estimates. The basis of a forecast by discriminant analysis is to calculate the probability that the given observation belongs to each possible group. The essential problem is that any observation bears at least some small 'resemblance' to each group of the dependent data; hence it is very unlikely that an extreme forecast (that of 0 or 100%) will ever be produced. In fact, if the function used does not discriminate very well (as is the case here), the forecast probabilities tend towards climatological probabilities (relative frequencies from the dependent sample). In other words, unless the original groups are very sharply defined in terms of the predictors used, the resulting forecast probabilities will tend to be 'hedged' heavily towards climatology.

Estimates calculated from a regression equation are not so heavily biased towards climatology Since there are more points in the dependent sample associated with 'dry' periods than with 'wet' periods, there is a tendency for regression estimates to be biased slightly low, especially at long prognosis periods. By truncating estimated probabilities to the allowable range (0 to 100%), results are much 'sharper' (more definite) than those for discriminant analysis. The Brier score reflects the sharpness of the predictions made, with perfectly sharp forecasts scoring zero if correct.

Because of the tendency for discrimination to underforecast rainfall occurrence more than regression, discrimination was consistently worse than regression in terms of the Hanssen score, on the dependent data.

Stations which gave the best results on dependent data were Nelson, Hokitika, Kaikoura and Christchurch. These stations all have quite clear-cut preferred wind direction ranges for wet and dry conditions, due mostly to the effect of the Southern Alps. Hokitika is the most extreme of these, where the equations derived imply any flow with a westerly component is almost invariably wet and any flow with an easterly component is dry. The same can be said for Kaikoura and Christchurch, although in reverse. Nelson shows similar characteristics, applied to flows with either a northerly or southerly component. North Island stations which gave the best results were Rotorua and Gisborne, both of which tend on average to be wet in easterlies and dry in westerlies. Results were worst for Dunedin and Invercargill where there is no strongly preferred flow for either wet or dry weather, based on the data used here.

As the statistical techniques used were both linear, results must be best where the relationship between rainfall occurrence and wind flow, mean relative humidity and vertical motion is most linear. Stations which have scored the worst may be described as having the most non-linear relationships between rainfall occurrence and the overall flow patterns. These stations should benefit most from a non-linear approach to PoP forecasting.

### 2.2.2 Independent Data

Due to the poor results obtained for discriminant analysis on the dependent data, no tests of the discriminant equations have been carried out on independent data, apart from a few trial operational runs. These suggested that discriminant-derived predictions would rarely fall more than twenty percent either side of the climatological probability for each station.

The independent data set comprised approximately 650 points per station, being the rainfall occurrences for both day and night periods for the months January to November 1982 inclusive. Verifications are presented for the twelve hour and forty-eight hour prognosis predictions only.

(a) Brier Scores.

On the whole, Brier scores for both forecast periods were lower (better) for the independent data than for the dependent data. The only exceptions were Kelburn, Kaikoura and Invercargill for 12 hour predictions and Rotorua, Gisborne, Kelburn and Invercargill for forty-eight hours. Mean scores over all stations were 0.138 and 0.174 for 12 and 48 hours respectively, compared with 0.144 and 0.181 for the dependent data. 1982 was a relatively low-rainfall year for many places.

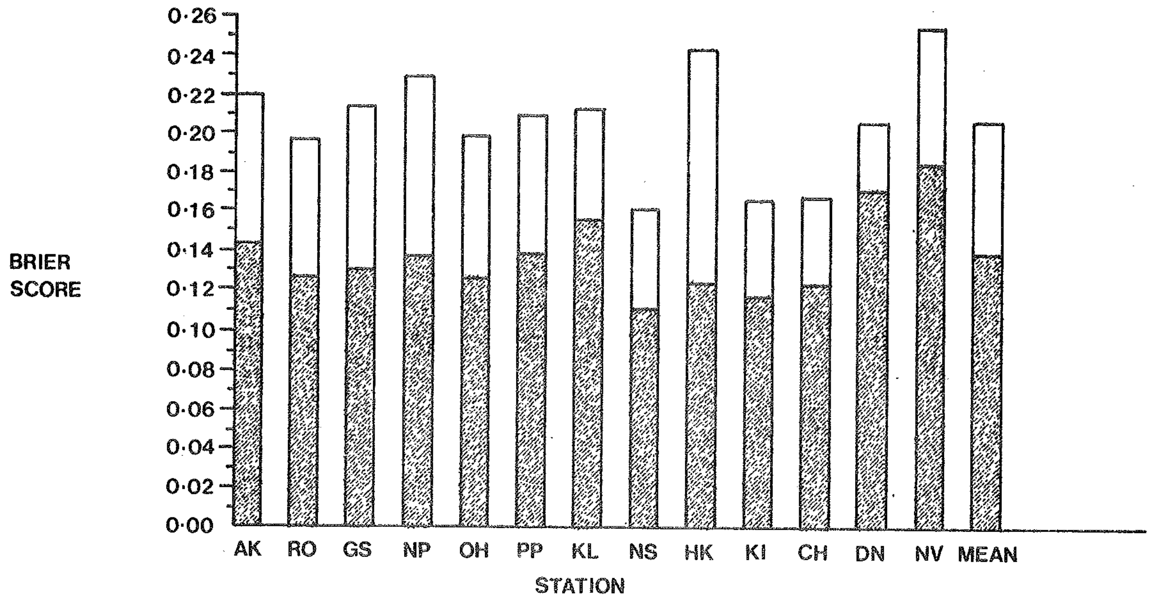


FIGURE 3a: BRIER SCORES - INDEPENDENT DATA. 12 HOUR PROGNOSES. REGRESSION VERSUS CLIMATOLOGY.
REGRESSION: ▨ CLIMATOLOGY: □

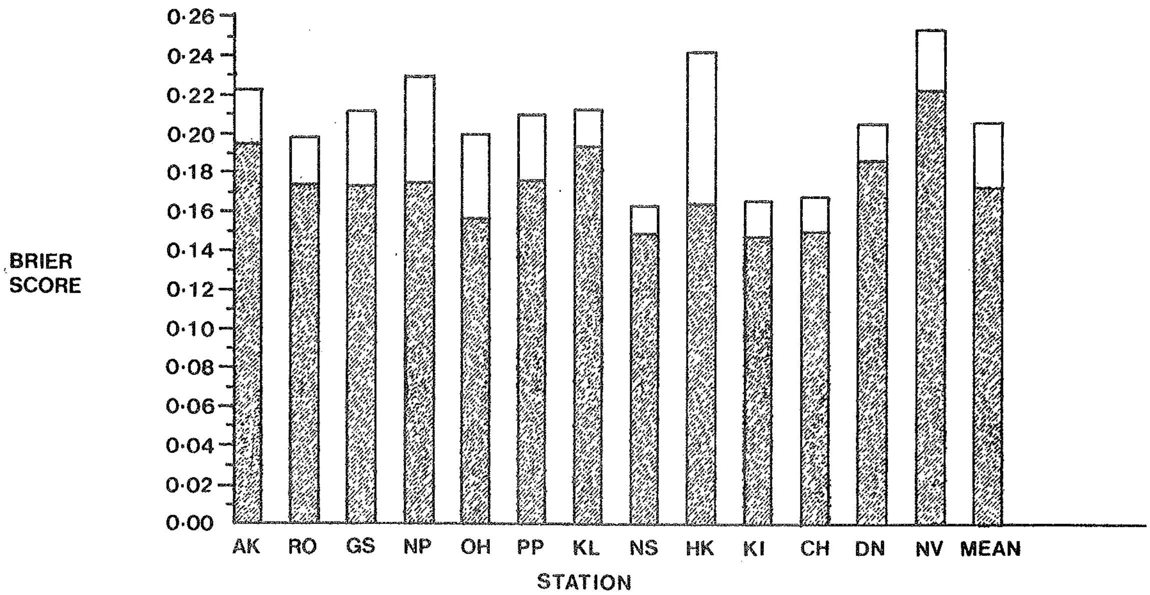


FIGURE 3b: BRIER SCORES - INDEPENDENT DATA. 48 HOUR PROGNOSES. REGRESSION VERSUS CLIMATOLOGY.
REGRESSION: ▨ CLIMATOLOGY: □

As the regression estimates are biased slightly low, this would assist in giving good verifications. Scores for all stations may be seen in Fig.3, which also includes the comparable scores obtained by forecasting 'climatology' (calculated from seasonal trend equations) every period. As Fig.3(b) shows, while the accuracy of predictions becomes quite low at most stations for forty-eight hour forecasts, it still remains better than a forecast of climatology. One may apply a T-test to the differences in Brier scores obtained by two different forecasting methods, using the following assumed distribution (Miller 1962).

$$\frac{(\overline{B}_X - \overline{B}_Y)\sqrt{M(M-1)}}{\sqrt{\sum_{m=1}^{M}(B_{Xm} - B_{Ym})^2 - \dfrac{\left[\sum_{m=1}^{M}(B_{Xm} - B_{Ym})\right]^2}{M}}} \sim t(M-1) \qquad (9)$$

where $\overline{B}_X$ and $\overline{B}_Y$ represent the mean Brier scores for the two forecasting schemes X and Y, while $B_{Xm}$ and $B_{Ym}$ represent individual Brier scores for each scheme on occasion m. Using a 95 percent significance level, regression scores were significantly lower then those for climatology for twelve hour forecasts at all stations as well as for forecasts taken together over all the stations. With increasing prognosis period, however, improvements over climatology tend not to be significant, since the MOS technique hedges towards climatology more and more as prognosis period increases. At the forty-eight hour prognosis period, the improvement over climatology was not significant at seven stations; Rotorua, Wellington, Nelson, Kaikoura, Christchurch, Dunedin and Invercargill, and was only marginally significant at all other stations, apart from Hokitika. Based on the 1982 verifications, regression equations appear at least to give stable predictions over independent samples.

(b) Hanssen Scores

Using a threshold of fifty percent to signify a forecast of 'rain', the average scores over all stations for twelve and forty-eight hour predictions were 0.492 and 0.301 respectively. Using the relative frequencies of occurrence of rain from the dependent sample ('climatology') as the threshold, these scores increased to 0.534 and 0.352. As a comparison with manual forecasts, Table 5 gives the individual scores for each station, using a threshold of fifty percent, along with the scores obtained by Thompson (1968) for manual forecasts, at the appropriate stations.

Table 5:
Hanssen Scores for Regression Estimates on Independent data

| STATION | 12 HOUR PROGNOSIS | 48 HOUR PROGNOSIS | MANUAL FORECAST (24 hours) |
|---|---|---|---|
| AK | 0.508 | 0.289 | 0.450 |
| RO | 0.467 | 0.289 | |
| GS | 0.524 | 0.267 | |
| NP | 0.577 | 0.416 | |
| OH | 0.475 | 0.318 | |
| PP | 0.485 | 0.267 | |
| KL | 0.430 | 0.253 | 0.418 |
| NS | 0.488 | 0.134 | |
| HK | 0.652 | 0.517 | |
| KI | 0.346 | 0.136 | |
| CH | 0.342 | 0.115 | 0.314 |
| DN | 0.294 | 0.122 | |
| NV | 0.421 | 0.261 | 0.320 |

The scores obtained for twelve hour forecasts are slightly higher than those for manual forecasts, while skill generally becomes very low as the prognosis period increases to forty-eight hours with severe underforecasting of precipitation occurrence becoming apparent.

(c) Reliability

A useful measure of the skill of probability forecasts is that of 'reliability', which is a comparison between forecast likelihood and the actual likelihood of occurrence. For example, with a perfectly reliable forecasting scheme, precipitation should occur on twenty percent of occasions when the forecast probability of occurrence was twenty percent, and so on.

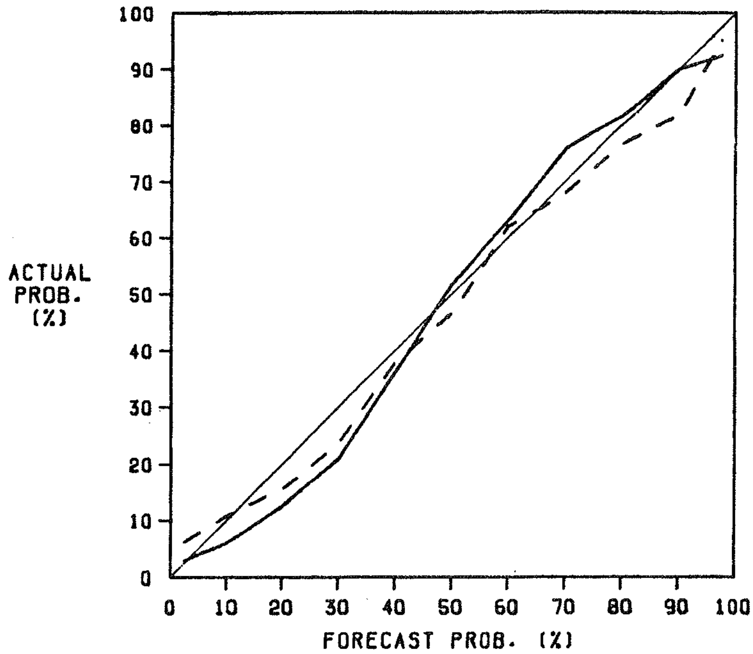


Figure 4 Reliability of regression estimates.
Independent data.
Twelve hour predictions: ______
Forty-eight hour predictions: __ __

Figure 4 shows a plot of forecast versus actual relative frequencies taken over all stations, for both forecast periods considered. Forecasts were divided into eleven categories, centered on 2.5, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 97.5 percent. As can be seen from the figures, the twelve hour forecast probability is high and too high when the actual probability is low. This tendency is reversed near either end of the probability scale, reflecting non-linearities inherent in the real relationships between the predictors and the predictand. For forty-eight hour prognoses, almost all categories fall below the 'perfect fit' line, suggesting a consistent low bias in the forecasts.

Probabilities most frequently forecast are those near the climatological probabilities of occurrence, especially for the longest prognosis period, as may be expected. This leads to the greatest deviation from 'perfect' reliability, at forecast probabilities of about thirty percent.

## 3. CONCLUSIONS

Given the data used here, a linear regression approach to probability of precipitation forecasting produces more useful results than linear discriminant analysis. Using the NWP data available at present, only short forecast period predictions show a useful level of skill. Use of output from a more advanced and higher resolution numerical model should produce better estimates over the longer (more than twenty-four hours) forecast periods. The level of skill obtained by regression estimation appears sufficiently high to allow these estimates to be used as guidance for the forecaster, at least over the short range, if rainfall forecasts are to be issued operationally in terms of probability of occurrence.

Some form of non-linear technique may produce more satisfactory results for at least some of the stations and further work needs to be done on the statistical techniques used for probability forecasting.

## ACKNOWLEDGEMENTS

The author wishes to thank Dr J.W. Kidson for provision of computer routines for discriminant analysis; Dr N.D. Gordon and Mr S. W. Goulter for useful help and advice. Also the staff of E.D.P. section for provision of rainfall data.

## REFERENCES

Brier, G.W. and Allen, R.A., 1951: Verification of weather forecasts. Compendium of Meteorology, Ed. T.F. Malone, Boston, American Meteorological Society; pp 841-848.

Hanssen, A.W. and Kuipers, W.J.A., 1965: On the relationship between the frequency of rain and various meteorological parameters. Mededelingen en Verhandelingen, No.81. Netherlands Meteorological Institute.

Miller, R.G., 1962: Statistical prediction by discriminant analysis. American Meteorological Society Meteorological Monograph 4, 25.

Murphy, A.H., Lichtenstein, S., Fischhoff, B. and Winkler, R.L., 1980: Misinterpretations of precipitation probability forecasts. Bulletin of the American Meteorological Society, 61 (7): pp 695-701.

Ostle, B. and Mensing, R.W., 1975: Statistics in research. Iowa State University Press.

Renwick, J.A., 1980: Numerical forecasts of maximum and minimum temperatures. New Zealand Meteorological Service Technical Note No. 241.

Thompson, D.C., 1968: Skill scores achieved in N.W.F.C. precipitation forecasting. New Zealand Meteorological Service Technical Note No. 177.

Trenberth, K.E., 1973: A five layer numerical prediction model. New Zealand Meteorological Service Technical Note No. 222.