2. COMMON VERIFICATION METHODS

This chapter describes in detail the verification measures described in Figure 1.1 of Chapter 1. The description often includes examples and the enumeration of attributes.

2.1 SCATTER PLOTS

The scatter plot is probably the simplest verification tool. Intended for continuous elements such as temperature or wind, it normally consists of all the cases of the sample plotted on a graph as forecast vs. observed. The ordinate and the abscissa should have the same scale, in which case perfection is represented by any point on the 45 degree line for which forecast=observed. The 45 degree line is usually drawn to facilitate interpretation of the scatter plot.

Figure 2.1 is an example of a scatter plot for marine wind forecasts. Both observations and forecasts are expressed to the nearest knot, and each "x" represents at least one occurrence of a particular observation-forecast pair. A quick glance at the plot immediately reveals some characteristics: winds below 4 knots are never forecast; the forecast is reasonably good (points do tend to cluster along the 45 degree line); and higher wind speeds tend to be underforecast. On each graph, the other line is the least squares regression line fit to the data. If the forecasts were perfect, this line would coincide with the 45 degree line. Correspondence between the regression line and the 45 degree line is simply the measure of reliability. A comparison of the orientation of the regression line and the 45 degree line gives a visual representation of the relative quality of the forecasts. As the quality decreases (compare the analysis and 36 hour forecast plots), the regression line tends more toward the horizontal. A horizontal line means no skill; the observations are independent of the forecast, or, put another way, the mean of the observations (climatology) is just as good a predictor as the forecast method.

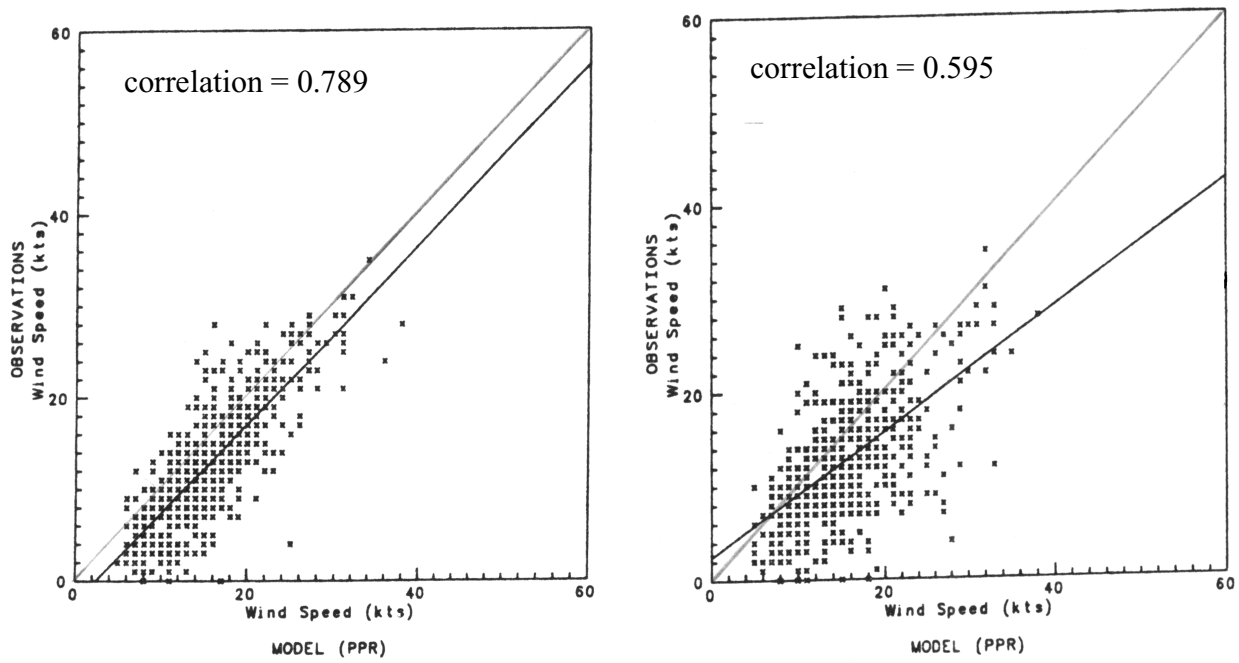

Figure 2.1 Scatter plots for marine wind forecasts produced by a perfect prog forecast technique, using the CMC (Canadian Meteorological Center) spectral model output data. Left: 0 hour forecasts. Right: 36 hour forecasts. Verification data are from NOAA buoy 44011 for the period January 1, 1985 to April 1, 1986. There are 575 cases.

Another verification measure that can be estimated from a scatter plot is the mean error or bias. Figure 2.2 is another scatter plot, for objective (model output statistics) temperature forecasts made for 11 Atlantic Region stations during the Canadian Atlantic Storms Project (CASP). The sample size is 701 cases. In this plot, the numbers on the graph represent the number of cases with the observed and forecast temperatures corresponding to the plotting point. For example, there were 7 cases with a forecast of -5°C degrees and observed temperature of -3°C degrees. Resolution is to the nearest degree Celsius.

It is possible to see from the plot that there is a slight tendency for the plotted values to lie above the 45 degree line, indicating a slight negative bias; temperatures are forecast too low (underforecasting). Note that a visual estimate of bias must include a visual estimate of the cumulative distance of the points from the line, and that multiple occurrences at a single plotting point must be taken into account. The bias in this case works out to -0.55 degrees Celsius.

A scatter plot represents the ultimate stratification since all the cases are represented separately. This allows considerable flexibility for analyzing the data. Suppose -20 degrees and +10 degrees Celsius are chosen as thresholds for extremes. The performance for these extreme temperatures may be analyzed visually from the graph. All points where the observed temperature is within the range for extremes and there is no forecast temperature are called "missed events", and all points where the forecast lies within the range for extremes and the observation doesn't are called false alarms. The false alarm and missed events areas are outlined on the graph. It can be seen from the graph that all the highest temperatures were missed (forecast below the threshold) while for the low temperatures there were occurrences of both missed events and false alarm.

Stratification on the basis of the observation permits identification of missed (extreme) events ONLY. In the example, the stratification on the basis of the observation is represented by the heavy horizontal lines drawn at the threshold values. Stratification on the basis of the forecast permits identification of false alarms ONLY. This stratification is represented by the vertical lines drawn at the threshold. It is important to use BOTH types of stratification to obtain complete information about the product being verified. For example, all but one of the missed extreme low temperatures could have been caught simply by forecasting 10 degrees too low. (Imagine moving all the points on the graph to the left 10 degrees) But the cost of that adjustment is a dramatic increase in the false alarms. (On the graph, all points for which the observed temperature lies above -20 degrees and the forecast is between -10 and -20 would now also lie in the false alarm region). This cost would not be seen by a stratification on the basis of observation alone. It is always possible to correct for missed extreme events by adjusting the bias. However, the inevitable cost of increased false alarms is a reduction in the utility (a user dependent attribute) of the forecast.

Considering again stratification by observation, a distribution of forecast temperatures can be identified for each value of the observed temperature. Figure 2.3 is an example for an observed temperature of -3°C. These are also called conditional distributions, namely in this case, the distributions of forecast temperatures given an observation. Ideally these distributions should be smooth and centered about the observation value. (Bias is indicated if they are not), and they should be sharp, with as little dispersion as possible. Most importantly, the distributions should be centered in different locations for different observed values. If they are not, it means that the forecasting system cannot discriminate between different observed values of the element and therefore has no skill. In terms of the graph, there would be no skill if the plotting points were clustered about a vertical line rather than about the 45 degree line. Figure 2.4 is a plot of two distributions of forecasts, one for observations of -3°C and the other for observations of -7°C. It is possible to see the separation between the two distributions, especially in the tails. Observations of -7°C tend to be associated with more forecasts of low temperatures while observations of -3°C tend to be associated with forecasts of higher temperatures. However, there is considerable overlap, which obscures the discriminating ability of the forecasts. Objective measures of the discrimination are possible, based on the separation of means, and on the overlap, as represented by a dispersion measure of the distributions.

Stratification on the basis of the forecast is perhaps more useful for understanding the real meaning of the forecast. Figure 2.5 is an example of the conditional distribution of observations given a forecast of -3°C. Such a distribution can be translated directly into information about the meaning of the forecast: if one is presented with a forecast of -3°C, the actual temperature may lie between -13°C and +3°C, with, fortunately, the most likely value being -3°C. The median is -3.5°C which means the verifying temperature is just as likely to be lower than the forecast value than equal to or higher. The mean is -4.3°C, which means that the verifying temperature tends to be more in error on the cold side than on the warm side. All this information applies only to cases where -3°C is forecast. Detailed informa-

tion on the meaning of the forecast for other temperatures can thus be obtained from a series of conditional distributions for each forecast temperature or for ranges of temperature.

One cautionary note: The greater the number of categories for stratification, the smaller the sample for each, and therefore the confidence in statistics calculated is lower on the partitioned sample. It is always important to keep track of the sample size, and to avoid the temptation of drawing more inferences than the sample can support.

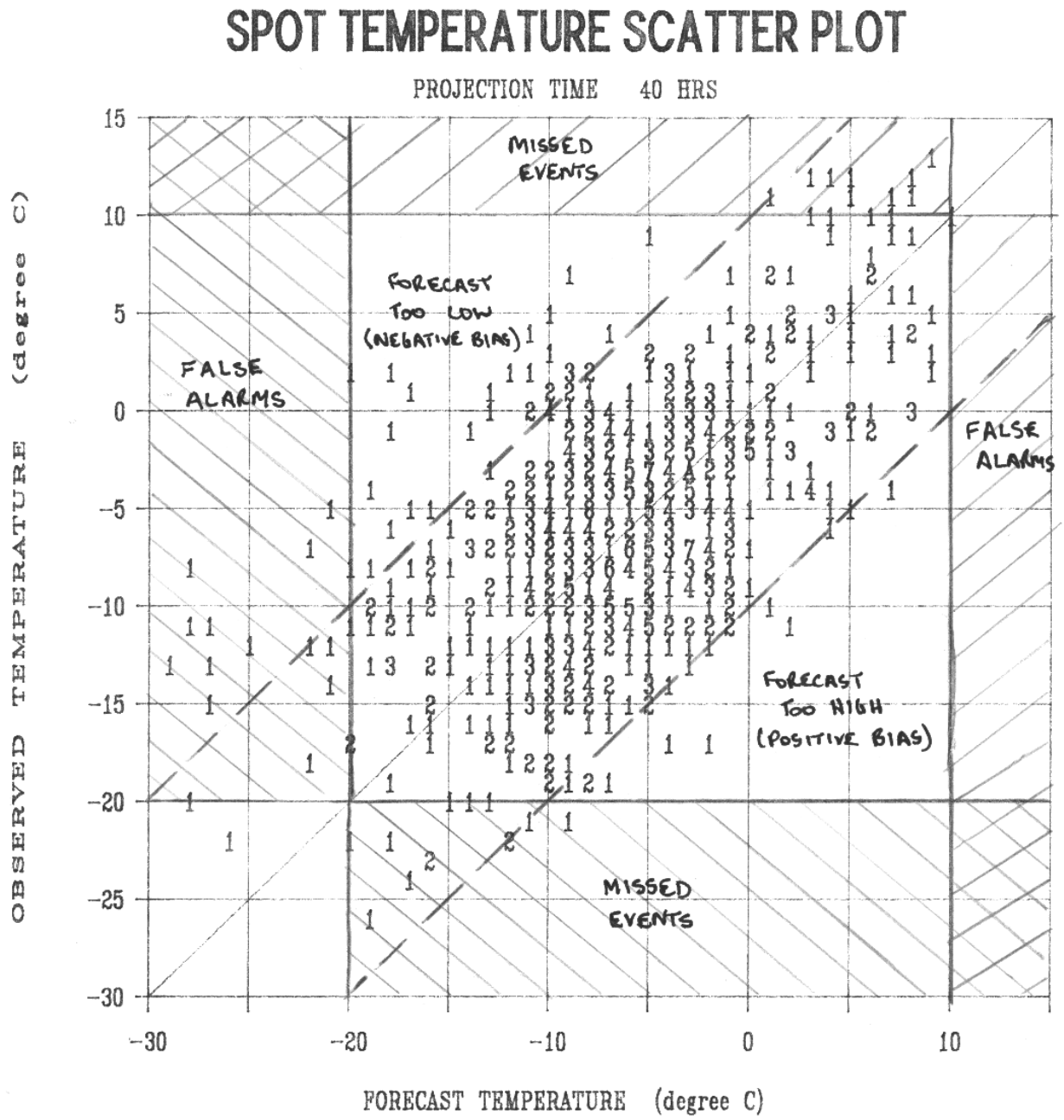

Figure 2.2 Scatter plot of observed and forecast spot temperatures.

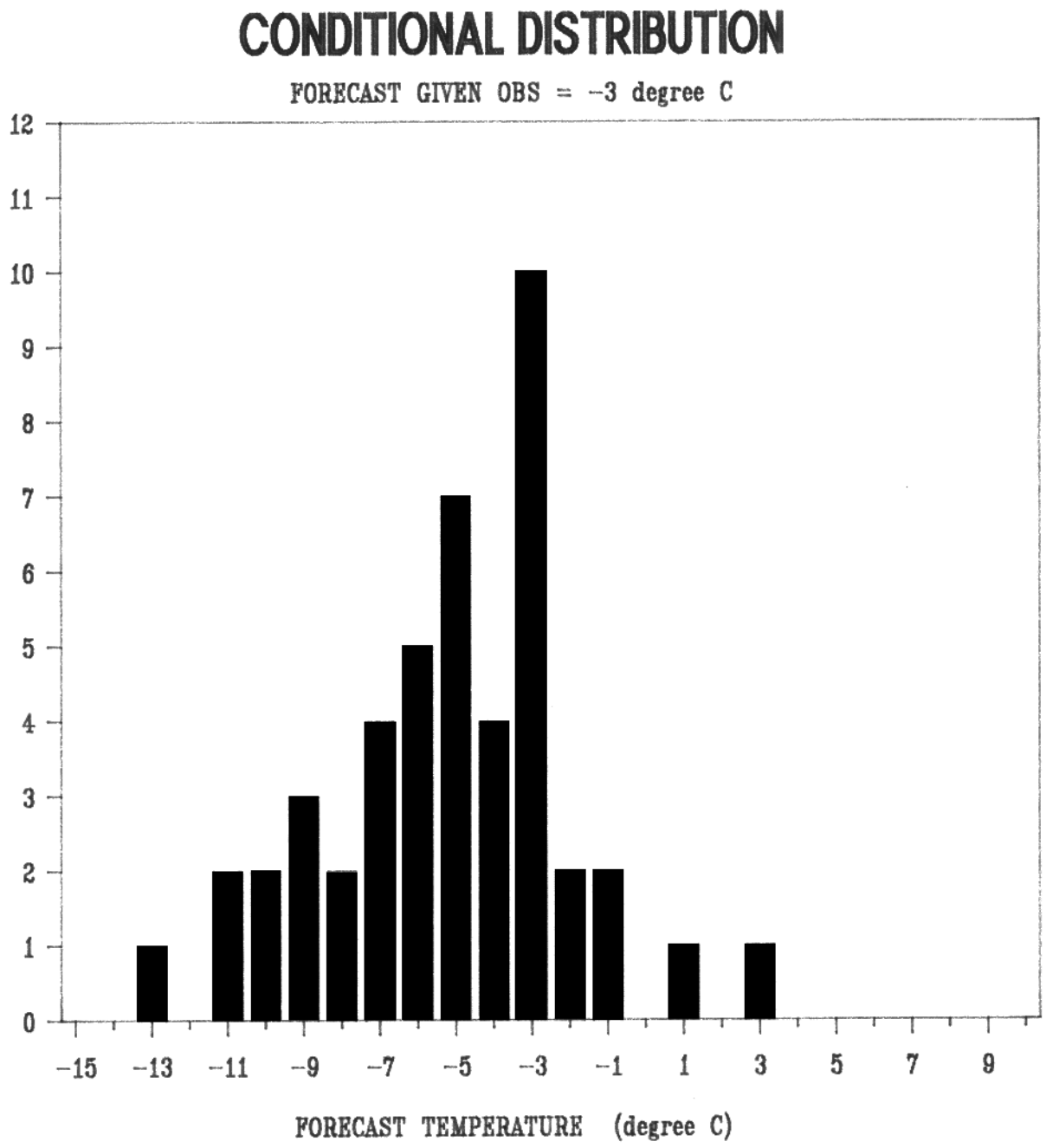

Figure 2.3 Histogram of forecast temperature given an observation of -3 ◦ C

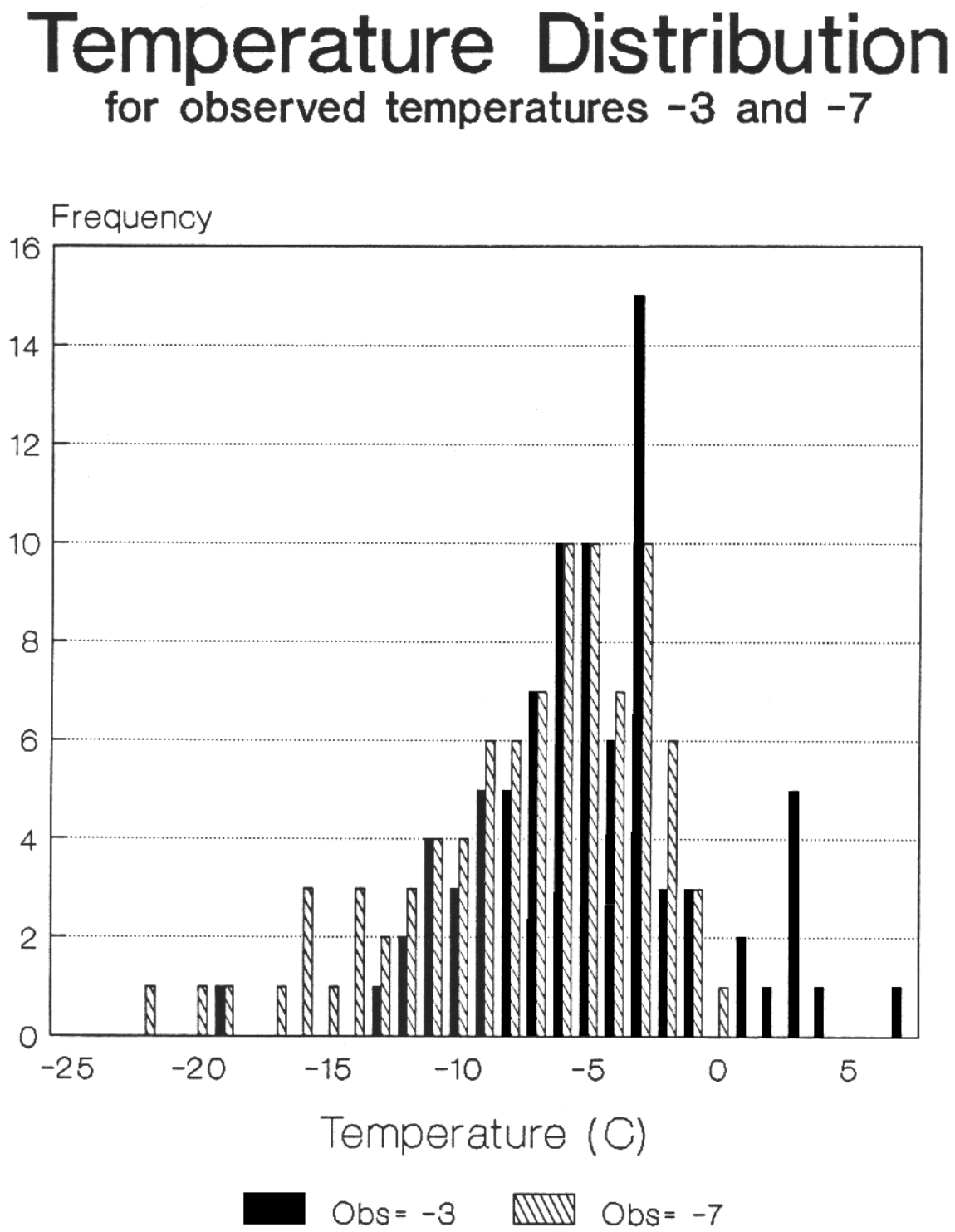

Figure 2.4. Histogram of forecast temperatures given an observed temperature of –3° C and –7° C. 11 Atlantic region stations for the period 1/86 to 3/86. Sample size 701 cases.

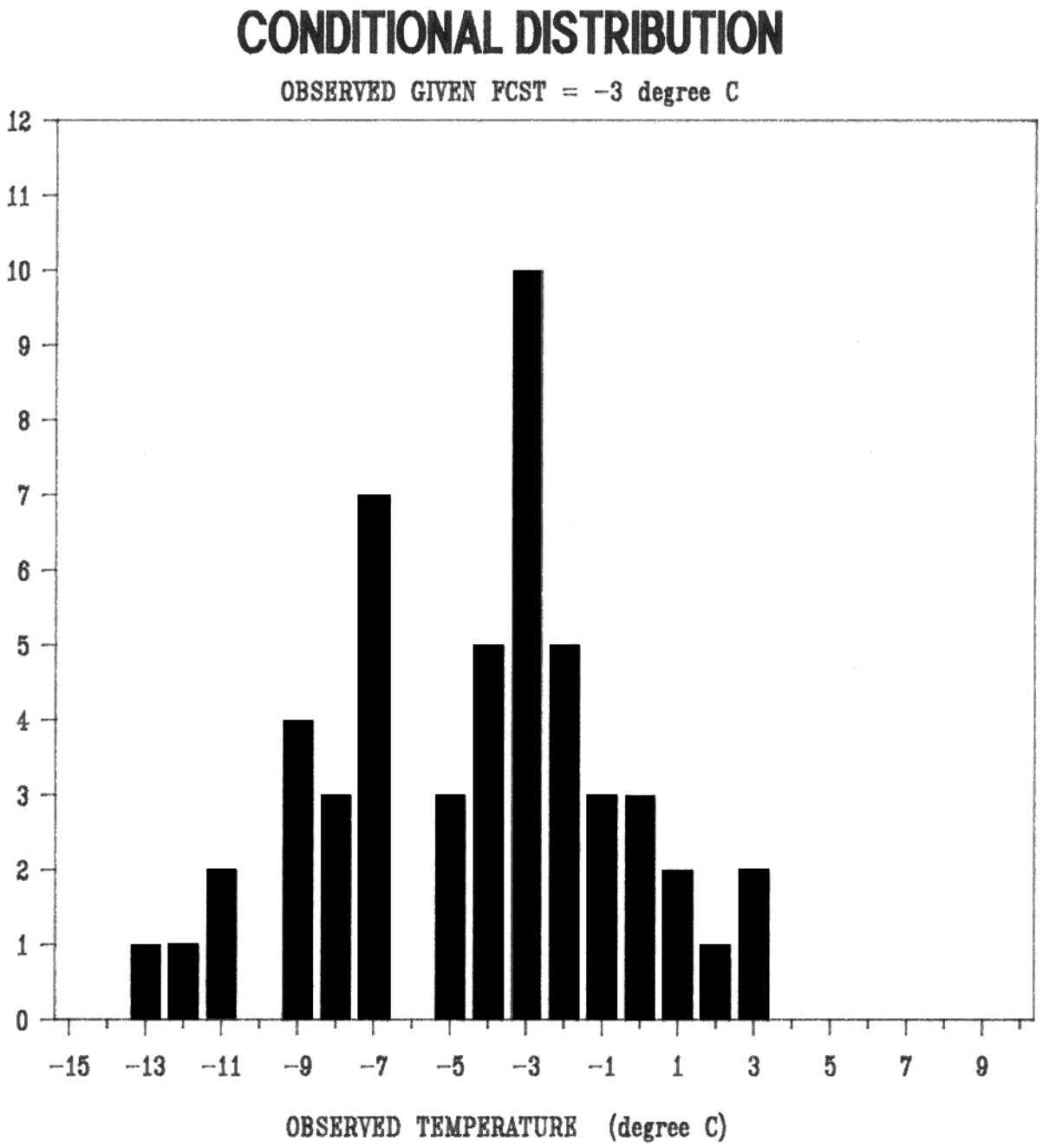

Figure 2.5 Histogram of observed temperatures given a forecast temperature of –3° C.

2.2 BIAS OR MEAN (ALGEBRAIC) ERROR MEASURE OF OVERALL RELIABILITY

If in a series of **N** forecasts, F_i represents the i-th forecast and O_i the corresponding observation, the bias is given by:

$$\mathbf{BIAS} = \frac{1}{N}\left[\sum_{i=1}^{N}(F_i - O_i)\right]$$

The error convention is (Forecast - Observed), although the reverse is sometimes seen.

The bias indicates the average direction of the deviation from observed values, but may not reflect the magnitude of the error. A positive bias indicates that the forecast value exceeds the observed value on the average, while a negative bias corresponds to underforecasting the observed value on the average. The bias may be expressed as the mean forecast value minus the mean observed value for the verification sample. As the mean is independent of order, one could for example, interchange the order of Winter and Summer forecasts for maximum temperature, and not change the bias for the year.

The bias should never be presented as a stand alone score. As a minimum the bias with respect to climatology for both the forecast and observation sample should accompany any calculation of forecast bias. Without comparative biases, there is no way to determine whether the forecast deviation is within acceptable limits and reflects the deviations occurring within the verification sample.

If bias statistics are to be used as a correction to forecasts, one has to be careful that the sample on which the bias correction is based is consistent with the application of the correction. For example, it would be inappropriate to apply a station bias as a correction to individual forecasters; each forecaster has his/her own bias which contributes to the overall station bias. These subjective biases result from the mix of a particular situation with surrounding influences and are not reproducible with certainty. However, a bias correction could be applied to numerical models insofar as recent changes have not affected the calculated bias. Numerical models approach each event in a consistent manner, it is hard coded in the computer program. As another example, it might be undesirable to apply a correction based on an annual average bias if the bias changes considerably from one season to another.

Note that the term "bias" is also used to refer to the frequency bias calculated from the marginal sums of the contingency table. This is a totally different score (one is a ratio of forecast over observed, while the other is the difference of forecast minus observed) and is described in section 2.6.

Table 2.1: Maximum Temperature, Forecast vs Observed (C)

Day:	1	2	3	4	5	6	7	8	9	10
Forecast:	5	10	9	15	22	13	17	17	19	23
Observation:	-1	8	12	13	18	10	16	19	23	24
$F_i - O_i$:	6	2	-3	2	4	3	1	-2	-4	-1

Bias = (8/10) °C

i.e. a tendency based on a limited sample to overforecast the maximum temperature. The larger the bias, the stronger this tendency.

Table 2.2: Repeat of the above example with a larger error magnitude

Day:	1	2	3	4	5	6	7	8	9	10
Forecast:	8	13	3	18	25	16	20	13	15	19
Observation:	-1	8	12	13	18	10	16	19	23	24
$(F_i - O_i)$:	9	5	-9	5	7	6	4	-6	-8	-5

Bias = (8/10) °C

i.e. the bias is the same as the previous example, but the error magnitude is larger here.

2.3 MEAN ABSOLUTE ERROR (MAE) MEASURES OVERALL ACCURACY

In a series of **N** forecasts where $\mathbf{F}_i$ represents the i-th forecast and $\mathbf{O}_i$ the corresponding observation, the mean absolute error is given by:

$$\mathbf{MAE} = \frac{1}{N}\left[\sum_{i=1}^{N} |F_i - O_i|\right]$$

This is a linear score which gives the `average' magnitude of the errors, but not the direction of the deviation.

Table 2.3: Calculation of Mean Absolute Error using data from Table 2.1 Maximum Temperature Forecast and Observation (°C)

Day:	1	2	3	4	5	6	7	8	9	10
Forecast:	5	10	9	15	22	13	17	17	19	23
Observation:	-1	8	12	13	18	10	16	19	23	24
$(F_i - O_i)$:	6	2	-3	2	4	3	1	-2	-4	-1
$\lvert F_i - O_i \rvert$:	6	2	3	2	4	3	1	2	4	1

Mean Absolute Error = (28/10) = 2.8 °C

i.e the average magnitude of the error is 2.8 °C

The MAE can be used in a skill score formulation, following the general format given in Chapter 1. This is sometimes done to provide a skill measure for continuous elements that is based on a linear score. This form of the skill score may be written:

$$\mathbf{SKILL} = \frac{\sum_i |FC_i - O_i| - \sum_i |F_i - O_i|}{\sum_i |FC_i - O_i|}$$

where: FC_i forecast using a control, such as climatology

F_i forecast value

O_i observed value

(the summation is from i=1, to i=N events)

Climatology is usually used as the standard, but persistence or chance could be used if appropriate.

To minimize the MAE one could use the median value as a forecast, but of course this would be playing the score and not forecasting.

2.4 ROOT MEAN SQUARE ERROR (RMSE) MEASURES OVERALL ACCURACY

The Root Mean Square Error for a series of **N** forecasts is given by:

$$RMSE = \left[\frac{1}{N}\sum_{i=1}^{N}(F_i - O_i)^2\right]^{\frac{1}{2}}$$

where: F_i is the i^{th} forecast

O_i is the i^{th}observation

This is a quadratic scoring rule which gives the `average' magnitude of errors, weighted according to the square of the error. Like the MAE, it does not indicate the direction of the deviation. Compared to the mean absolute error, the RMSE gives greater weight to large errors than to small errors in the average. It is therefore a more appropriate statistic to use when large errors are particularly undesirable. Its sensitivity to large errors also means that it may not give stable estimates of error if small samples are used. This measure, as with other quadratic scoring rules, can encourage conservative forecasting (forecasting close to the climatological mean). If extreme events are rarely forecast, it is hoped that the frequency of large errors will be reduced and thus minimize the RMSE.

Table 2.4: RMSE calculation using data from Table 2.1 Maximum Temperature Forecast and Observation.

Day:	1	2	3	4	5	6	7	8	9	10
Forecast:	5	10	9	15	22	13	17	17	19	23
Observation:	-1	8	12	13	18	10	16	19	23	24
$(F_i - O_i)$:	6	2	-3	2	4	3	1	-2	-4	-1
$(F_i - O_i)^2$:	36	4	9	4	16	9	1	4	16	1

RMSE = (100/10)1/2 =3.16 °C

i.e. the root mean square error is 3.16 °C

When compared to the MAE, the RMSE gives an indication of the error variance, i.e. if RMSE >> MAE, this would indicate a high error variance; while RMSE = MAE if and only if all errors have the same magnitude. The RMSE can never be smaller than the MAE.

2.5 REDUCTION OF VARIANCE (RV)

The reduction of variance is given by:

$$RV = 1 - \left[\frac{\sum_{i=1}^{N} (F_i - O_i)^2}{\sum_{i=1}^{N} (M - O_i)^2} \right] = \frac{MSE_c - MSE}{MSE_c}$$

where F_i is the i^{th} forecast

O_i is the i^{th} observation

M is the mean of the sample or the climate mean, with the climate mean being preferred...however the standard depends upon the (known) purpose.

and

MSE_c is the mean squared error with respect to climatology

MSE is the mean squared error of the forecast.

The reduction of variance is a "skill" measure in the sense that it compares the mean squared error of the forecast to the mean squared error (variance) of the unskilled estimate M, based on climatology.

The range of RV is $-\infty$ to +1. A negative value implies that the climate mean is a better predictor than the forecast value, or the converse if RV is positive. If the number of events verified is small, a negative value is more likely to occur due to the greater weight given large errors and by minimizing the averaging process.

RV = 1 implies that there is no error in the forecast value, i.e. forecast equals observation.

RV = 0 implies that the forecast is no better that the climate mean.

It is not necessary that M be the climate mean. It could be the mean of the sample. The difference here is the same as the difference between skill with respect to long-term climatology, and the skill with respect to ad hoc (independent sample) climatology. Long-term climatology is preferred for purely practical reasons. When the forecaster is compared against long-term climatology he always has the option of using the unskilled climatology forecast when making his forecast. Alternatively, if the chosen standard is the mean of the independent cases, the standard is not known until after the independent sample is completely collected.

On the other hand, one could remove the variance between the sample mean and long term climatology first, thus permitting a realistic comparison of skill between different samples. This is an important issue because one can always make the skill look better by removing the variance of the sample. It is therefore important to know whether the bias component has been removed when evaluating the results of verification using either the RMSE or the RV. A suggested exercise is to prove that the RMSE with respect to the sample mean is always smaller than the RMSE with respect to the climatological mean.

The Multiple Correlation Coefficient (R) measures the association between observed and predicted values, and is related to the reduction of variance:

$$R = (RV)^{\frac{1}{2}}$$

where R = 1 implies perfect linear correlation,
R = 0 implies no linear correlation.

2.6 CONTINGENCY TABLES AND ASSOCIATED SCORES STRATIFICATION OF DATA

A useful summary of the forecast and observed weather events can be presented in the form of a contingency

table. Such a table does not constitute a verification method itself, but provides the basis from which a number of useful scores can be obtained. A contingency table is nothing more nor less than a scatter plot for a categorical variable. As with scatter plots, the entries of the table simply represent a convenient presentation of the raw verification dataset from which many statistical inferences can be drawn, and different scores can be computed.

The contingency table summary may be archived for later reference and the final verification objective directly influences the archive design. Verification for scientific reasons requires that both the forecasts and observations be saved in their raw form along with date and time information. Verification for administrative reasons on the other hand would probably require that only the resultant contingency table elements be saved, but this option also prevents one from modifying the verification goals at a later date. All too often experience expands rather than diminishes the requirement for information. Note also that the contingency table summaries can be conditional, i.e. a comparison of forecast and observed values given a specific time period or weather event. The utility of verification databases often suffers more from a lack of information saved than from the actual storage space required.

The contingency table is formed in either of two ways:

1) for categorical forecasts....each event adds 1 to the grid element of contingency table according to the intersection of the forecast category and observed category

2) for probability forecasts...each event adds the probability forecast for all categories to the row elements corresponding to the observed category, i.e. if the forecast probabilities are: 50% 30% and 20%, and category 2 is observed, the three probabilities would then be added to elements 'd', 'e', and 'f' of Table 2.5. Since this method generates table entries that are harder to understand and interpret in terms of the scores discussed below, it is not recommended. It is better to categorize the forecast first using some categorization procedure, then form the table. Note also, other definitions for contingency tables involving probabilistic forecasts are possible.

Table 2.5: The Contingency Table

		FORECAST CATEGORY			
		1	2	3	Total
OBSERVATION	1	a	b	c	J
	2	d	e	f	K
	3	g	h	i	L
	Total	M	N	O	T

Since the creation of a contingency table represents a stratification of the dataset into categories, this often creates sample size problems. The selection of categories usually is a compromise between significance to users of the verification information and the statistical goal of having representative (large enough) samples in each category of the table. A weighting scheme might be used to correct for the relative importance of various forecasts (Burrows, 1989).

The following subsections describe scores often computed from the entries of a contingency table. Table 2.5 is used for demonstration purposes.

2.6.1 Percent Correct

Percent correct (PC) is the summation of the diagonal elements divided by the total number of events:

$$PC = \frac{a + e + i}{T} \times 100$$

where: a, e, i are the number of correct forecasts in each category.
T is the total number of forecast

When probabilities are used in place of categorical forecasts, the percent correct is interpreted as the percentage time correct.

The overall percentage correct is greatly influenced by the common (most frequent) categories, and can give misleading information. For example, in a 2x2 Yes/No contingency table for rain, the no-rain events may greatly outnumber the rain events, and the percentage correct is dominated by the no-rain frequency.

When dealing with Yes/No forecasts, and especially severe weather forecasts, any presentation of the overall percent correct will most likely initiate controversy. A common complaint will be that one could ignore all weather maps and simply predict a priori the most likely event and obtain a large percentage correct (this method is called "Play the winner strategy"). The simple fact is that no single score, measure or index has been found to satisfy this verification controversy. The controversy can be defused somewhat by being aware of the climatological frequencies when interpreting the percent correct, or by calculating the percent correct by category. When dealing with important rare categories in a table, for example, the percent correct for the rare category is sometimes referred to as the "hit rate".

One other related complaint that applies to any of the summary scores of the table is that they compress the information contained in the elements of the table into one number, resulting in a loss of information.

Ideal situation.....large percentage correct, greater than post agreement (defined next) for the most common category.

2.6.2 Post Agreement (PAG), False Alarm Ratio (FAR)

Post agreement is the number correct divided by the number forecast for each category.

$$PAG = \frac{a}{M}, \frac{e}{N}, \frac{i}{O}$$ for the three categories

Ideal situation....value of one.

FAR falls into the category of verification measures that imply stratification by forecast, and therefore, as the name implies, is sensitive only to false predictions of the severe event, not to missed events. The term False Alarm Ratio (FAR) is usually used when referring to a contingency table for severe weather forecasts; the ratio is $(1 - PAG)$ (of the severe event). This score can always be increased by underforecasting the number of severe events, but only at the cost of more missed events, which are not seen by the score.

2.6.3 Prefigurance or Probability of Detection (POD)

This is the number correct divided by the number observed in each category. It is a measure of the ability to correctly forecast a certain category, and is sometimes referred to as "hit rate", especially when applied to severe weather verification.

$$POD = \frac{a}{J}, \frac{e}{K}, \frac{i}{L}$$ for the three categories.

Ideal situation.....value of one. Like the FAR, it is not a complete score. It is in the class of verification measures that imply stratification by observation, and thus is sensitive only to missed events, not false alarms. Thus the POD can be increased by issuing a larger number of forecasts on the assumption that a greater number will then be correct, usually at the cost of more false alarms. This is similar to a rifleman using a shotgun instead of a rifle to increase his target accuracy. In short, the POD can always be increased by crying wolf.

2.6.4 Bias or Frequency Bias

Bias is the number forecast divided by the number observed for each category. It measures the ability to forecast events at the same frequency as found in the sample without regard to forecast accuracy.

$$BIAS = \frac{M}{J}, \frac{N}{K}, \frac{O}{L}, \quad \text{for the three categories.}$$

where: Bias = 1 implies no Bias
Bias > 1 implies overforecasting the event
Bias < 1 implies underforecasting the event

2.6.5 Threat Score or Critical Success Index (CSI)

The threat score is the number correct divided by the number of cases forecast and/or observed for that category. It is a measure of relative accuracy.

The threat score for each of the three categories is:

$$CSI = \frac{a}{M + J - a}, \frac{e}{N + K - e}, \frac{i}{O + L - i}$$

and the range is from zero to 1.

Ideal situation... values close to one preferred. The advantage of the threat score over the FAR and the POD is that it is sensitive to both false alarms and missed events. Thus it gives a more representative idea of real accuracy both in situations where rare events are involved and in situations where the climatological frequencies of the categories are nearly equal.

Example: Aviation Forecasts: Pacific Region January 1975, 12-hour forecasts based on Ceiling and Visibility Ranges.

Table 2.6: A contingency Table example and associated scores using ceiling and visibility categories. Category 6 corresponds to unlimited ceiling and good visibility, while category 1 refers to very low ceilings and limited visibility. Ceiling and Visibility ranges are defined in Table 2.18

		FORECAST						
		1	2	3	4	5	6	TOTAL
OBSERVATION	1	2	0	0	1	10	3	16
	2	1	0	0	1	8	4	14
	3	2	0	0	0	7	10	20
	4	7	1	0	8	112	108	236
	5	0	6	0	2	40	158	206
	6	0	5	0	12	85	894	996
		12	12	0	25	262	1117	1488

Total Correct = 944
Percent Correct = (944/1488) * 100 = 63.4%

Note that the statistics calculated from contingency tables change if categories are combined. By combining categories 5 and 6 into a single category, for example, more weight is given to the predominant observations (unlimited ceiling and good visibility) and the percentage correct always rises because of the addition of the off-diagonal elements between the combined categories to the diagonal. For this reason, it is always hazardous to compare contingency table statistics from tables of different dimensions.

Percent correct diagonal only (944/1488) * 100 = 63.4%
Percent correct with 5 & 6 combined (1187/1488) * 100 = 79.8%
Percent correct within one category (1302/1488) * 100 = 87.5%
Percent correct forecast only category 6 (894/1488) * 100 = 60.1%
Percent correct fcst only 5&6 combined (1177/1488) * 100 = 79.1%

Note also that "percent correct" is sometimes calculated "within one category". This has the effect of making the percent correct look better by inflating the numerical value. We can think of no other reason for doing this.

Table 2.7: Contingency-Table and associated scores continued for Table 2.6. Summary of Statistics from Contingency-Table (Fractional and Decimal)

Category	1	2	3	4	5	6
Post Agreement	2/12 0.17	0/12 0.00	0/0 0.00	8/25 0.32	40/262 0.15	894/1177 0.76
Prefigurance	2/16 0.12	0/14 0.00	0/20 0.00	8/236 0.01	40/206 0.19	894/996 0.90
Bias	12/16 0.75	12/14 0.86	0/20 0.00	25/236 0.11	262/206 1.27	1177/996 1.18
Threat	2/26 0.08	0/26 0.00	0/20 0.00	8/253 0.03	40/428 0.09	894/1279 0.70

Normally, only the decimal values are displayed.

Any interpretation of verification results should first describe the constraints of the verification. In the Pacific region results for 1975, the verification results are derived from the 124 12-hours forecasts that were issued at 6-hour intervals during the month of January. The sample size is too small for both scientific purposes and administrative purposes, however the temptation to make broad and general statements is always present. There is an overforecasting bias in categories 5 and 6, and the low threat score for category 5 indicates too much optimism. Stratification by issue time may have narrowed the cause of the optimism, i.e. early clearing of morning fog, etc.. The low values for post agreement and prefigurance for category 4 suggest a possible problem worth inquiry. Verification results for categories 1 to 3 are based on too few events to be significant, and may be attributable to outside influences such as staff changes, equipment failure etc. Finally, consider the numbers themselves. Category 1, listed for 16 hours in the table, was actually observed for only 8 hours (i.e. although the forecast period is for 12 hours, new forecasts are issued every 6 hours, hence each hour is repeated twice in the contingency table). Hence Category 1 was observed 0.01% during the month (24hours x 31 days = 744 hours and 8/744 = 0.0108%). Moreover, these 8 hours may be eight individual events or one eight-hour event and any verification interpretation on such a small sample would be meaningless.

2.6.6 SKILL SCORE (HEIDKE SKILL SCORE)

The Heidke skill score is uniquely associated with the contingency table-related scores, and is computed from

the entries of the table. The Skill score(SS) is in the standard format shown in Chapter 1 and may be expressed as:

$$SS = \frac{R - E}{T - E}$$

where: R is the number of correct forecasts
T is the total number of forecasts
E is the number expected to be correct based on some standard such as chance, persistence, climatology or some other forecast.

The range of score values is from $-\infty$ to +1; a perfect SS has a value of +1. If the number of correct forecasts equals the expected number correct, the SS equals zero. If the control forecasts are perfect, SS equals minus infinity; negative values indicate negative skill with reference to the standard.

When chance is the standard, the assumption is made that there is no correlation between forecast and observed values. Hence, for any position (i,j) (i.e. row,column) of a contingency table, the number of forecast-observation pairs expected to be in the (i,j) position of the table is:

$$E_{ij} = (R_i C_j)/T$$

where: Eij is the number of events by chance at (i,j)
Ri is the total number of observations of category i (i.e. total of row i)
Cj is the total number of forecasts of category j (i.e. total of column j)
T is the total number of forecasts

Thus from the contingency table (Table 2.5):

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

The skill score thus defined expresses as a decimal fraction the percentage of forecasts which are correct after eliminating those forecasts which would have been correct on the basis of chance. A weighting matrix can be applied to the contingency table before the skill score is calculated (see Muller, 1944; p51). This would increase the importance of certain forecasts relative to others and involves user-related considerations; otherwise the score assumes all forecasts are of equal value.

The term "Heidke" skill score in particular is most often associated with **chance** as the standard of comparison, and is a popular verification statistic. We speculate that its popularity is due either to the fact that 'chance' is the most unskilled standard (you need only guess) and is therefore easy to show positive scores against or because no information other than the contingency table elements is needed to calculate the score. The use of persistence or climatology requires the creation of another table for the standard forecast. The irony of the "Heidke" skill score is that Heidke (1926) had stated a preference for persistence and climatology as standards for comparison, and **not** chance.

Some characteristics of skill scores are:

1) The better the standard score, the more difficult it is to have a large positive value, especially if the period of verification is short. Since the skill score is a primarily a summary score, the period of verification should be sufficiently long to allow the opportunity to evaluate skill under all conditions.

2) The skill score is not strictly proper, it can be played by increasing the variability (Glahn, 1976).

3) Appleman (1960) has shown that the skill score **with respect to chance may not satisfy** either one of the following statements:

* a positive skill score indicates that the skilled technique will give better results than can be expected from any available unskilled forecast procedure;

* where two different skilled forecast techniques are applied to a representative set of data, the one yielding the larger positive score will be the better technique.

This disturbing evidence results from the fact that the random chance standard is in fact a function of the forecasts issued. Indeed, with no information at all (and assume only 2 categories), a forecast accuracy of 50% is obtainable by random chance with a sufficiently large sample. Similar to flipping a coin, pure chance results in half the answers being correct.

2.6.7 TRUE SKILL STATISTIC (TSS)

Discussion on the true skill statistic (TSS) (Flueck, 1987) is included as an example of the requirement to investigate attributes of a "new" and "improved" measure before inclusion in a verification scheme. All too often we accept and use verification measures without truly understanding their limitations. Without careful and detailed study, what may have been described as an exquisite diamond, may in reality be an excellent glass copy. As in business, let the user (buyer) beware. The true skill statistic was intended to be both understandable and relevant. An evaluation follows the TSS development.

The TSS was developed to address problems encountered with Yes/No forecasts and to include the following desirable properties;

1) to provide a measure of association between the forecasts and observations,

2) to be easily understandable, and be in the fixed range -1 to +1,

3) to use all the relevant information contained in the observations and forecasts,

4) to permit estimation of the probability that the observation/forecast association is real,

5) to permit estimation of the variability of the verification measure.

Verification of tornado forecasts was included with the TSS development to help focus attention on the drawbacks of the current scores. In March 1884, Sergeant John Finley initiated twice daily tornado forecasts for eighteen regions in the United States, east of the Rocky Mountains. Issued at 7:00 am and 3:00 pm EST, the first forecast was valid for 16 hours and the second for 8 hours. Forecasts were based on the surface weather charts of the same time period. Finley published the first three months results and he claimed 95.61 to 98.65% overall verification results, depending on time and district. Some districts claimed to have 100% verification for all 3 months. A critic of the results pointed out that a 98.18% verification score could be had by merely forecasting no tornado, and thus the need was established for a new and improved verification score.

Table 2.8: Finley's Tornado(1884) Contingency Table

		FORECAST		Total
		Tornado	No Tornado	
OBSERVED	**Tornado**	28	23	51
	No Tornado	72	2680	2752
	Total	100	2703	2803

The calculation of the TSS uses the components of a 2x2 contingency table as follows:

		Forecast	
		Yes	No
Obs	Yes	A	B
	No	C	D

$$TSS = \frac{AD - BC}{(A + B)(C + D)}$$

where $-1 \le TSS \le 1$
perfect = **+1**
worst=**-1**

This score is also known as Hanssen and Kuipers' Discriminant (Hanssen and Kuipers, 1965) and its derivation has been claimed by many others, including Pierce, Clayton, Dobrysham, Bontrond, and Flueck. Enthusiasm for this score is unfortunately misplaced. The historical criticism of this score has been on the grounds that it unduly weights prefigurance. (The score can be reduced to the prefigurance of the event, plus the prefigurance of the non-event -1, or alternatively, POD-POFD where POFD refers to the probability of false detection, C/(C+D) in the table). This characteristic is attributable to the fact that the score emphasizes stratification of data according to the observation, since both its components POD and POFD imply stratification according to the observation. Thus, if a certain event occurs a few times and is not forecast, the score turns negative, even though the overall forecast accuracy is high, that is the score is sensitive to missed events and not false alarms. In persistent situations, the forecaster also shows no skill by deciding that the weather pattern will NOT change.

The two components of this score, POD and POFD are the same as for signal detection theory. See Chapter 4.

The final comment on this score is that in over one hundred years it is not well known although derived independently by many. In the selection of any verification score, the score's characteristics should be well known by all, otherwise the decisions made upon the score could have their foundations in sand.

2.6.8 Examples of Contingency Tables

Several examples of contingency tables are shown in this section, along with some of the associated scores. These 5 tables were generated using the same dataset of probability forecasts and observations, but the forecasts were converted to categorical forecasts before creation of the tables. Each table represents the results of a specific objective strategy for categorization of the probability forecasts. By comparison of the tables it is possible to glean useful information on the performance of the different strategies, and to select the best for the intended use of the forecasts.

The 5 categorization strategies are: 1. Maximum Probability; 2. Maximum increment over climatology; 3. Unit Bias Method; 4. Maximum Threat Score (for strategies 1 to 4: Miller and Best, 1979); and 5. Binary Tree Method. A complete description of these procedures is beyond the scope of this document, but it is worth noting that some of the procedures use contingency table scores as part of the decision rule for category selection, and all involve the establishment of threshold probabilities which are compared with sets of forecast probabilities for decision-making. Categorization of probability forecasts effectively changes the forecast set of probabilities to a set of "zeros" and "ones" for all categories. It is most controversial because it implies loss of information, but is often done to assist in the interpretation of probability forecasts.

The following contingency tables and scores are verification of the probability of precipitation type (POPT) forecasts for Montreal during the winter of 1985-86 and use an independent sample, that is, separate from the sample used to develop the objective forecast procedure. The POPT forecasts are conditional upon the event of precipitation, i.e. given that precipitation is observed, the probability of precipitation type is computed for liquid, frozen or freezing categories (labelled as rain, snow and frzg). The probabilities are generated every six hours and valid for a 6-hour forecast period. Forecasts here are 6 hour forecasts valid at 18Z. The determination of whether or not precipitation will occur is left to the forecaster, hence POPT forecasts are issued even with cloudless skies. Only precipitation

events are included in the contingency table. The probabilities are MOS (Model Output Statistics) derived and driven by the Canadian Spectral model.

Table 2.9: Contingency Table using Maximum Probability as Threshold

Observation	Forecast		
	Rain	Snow	Frzg
Rain	21	7	0
Snow	1	43	1
Frzg	2	1	2
Bias (by category)	0.857	1.133	0.600
Threat Score	0.677	0.811	0.333

Percent Correct...84.62

Table 2.10: Contingency Table using Climatology Thresholds (threshold order: rain,snow, frzg)

Observation	Forecast		
	Rain	Snow	Frzg
Rain	21	6	1
Snow	3	33	9
Frzg	3	0	2
Bias (by category)	0.964	0.866	2.400
Threat Score	0.688	0.647	0.133

Percent Correct...71.79
Thresholds: 0.350(rain), 0.90(snow)
Climatological frequencies: 0.35(rain), 0.55(snow), 0.10(frzg)

Table 2.11: Contingency Table using Climatology Thresholds (threshold order: frzg, snow, rain)

Observation	Forecast		
	Rain	Snow	Frzg
Rain	19	4	5
Snow	2	34	9
Frzg	0	0	5
Bias (by category)	0.750	0.844	3.800
Threat Score	0.760	0.694	0.263

Percent Correct...74.36
Threshold values: 0.10(frzg), 0.65(snow)
Climatological frequencies: .10(frzg), .55(snow), .35(rain)

Table 2.12: Contingency Table using the Unit Bias Threshold

Observation	Forecast		
	Rain	Snow	Frzg
Rain	23	5	0
Snow	3	39	3
Frzg	3	1	1
Bias (by category)	1.036	1.000	0.800
Threat Score	0.676	0.765	0.125

Percent Correct...80.77
Threshold values: 0.26(rain), 0.69(snow)

Table 2.13: Contingency Table using Maximum Threat Score Threshold

Observation	Forecast		
	Rain	Snow	Frzg
Rain	20	7	1
Snow	0	43	2
Frzg	0	3	2
Bias (by category)	0.714	1.178	1.000
Threat Score	0.714	0.782	0.250

Percent Correct...83.33
Threshold value: 0.67(rain), 0.68(snow)

Table 2.14: Contingency Table using a Binary Tree Classification

Observation	Forecast		
	Rain	Snow	Frzg
Rain	20	0	0
Snow	7	45	1
Frzg	0	0	4
Bias (by category)	0.714	1.178	1.000
Threat Score	0.714	0.849	0.667

Percent Correct...88.46

Threshold rule: if snow probability greater than 0.325, then forecast snow, otherwise if rain probability greater than .680 then forecast rain or if rain probability less than or equal to 0.680 then forecast frzg. Classification using binary tree rules was a new statistical method acquired by MSRB in May 1988

Table 2.15: Comparison of observed and climatological frequencies (in percent) for precipitation type for Montreal.

	Climatology	Observed during Verification period
Rain	35	36
Snow	55	58
Freezing	10	6

During the verification period freezing precipitation was rarer than expected according to climatology, while snow events were more frequent. The maximum probability threshold model, which has a tendency to ignore the rare event and overforecast the common event, benefitted from the departure from climatology. The climatological model on the other hand, regardless of threshold order, greatly overforecast the freezing category and obtained the lowest

scores for the set of methods. The unit bias model was able to establish thresholds that resulted in category biases that were close to unity, but no improvement in accuracy followed. The maximum threat model showed results very similar to the maximum probability model with one striking difference: The maximum threat model had a higher threat score for rain, resulting in no false alarms. The maximum probability threshold model had difficulty resolving all three categories, while the maximum threat model could separate rain from the other two categories. The problem of significance remains yet to be answered. The binary tree classification method was the most successful of the threshold methods, forecasting 4 of the 5 freezing events and having high threat scores for the snow and rain categories. The binary tree method, or CART (Classification and Regression Trees) (Breiman et al., 1984). is relatively new in meteorological applications, but shows promise for use in categorical prediction of weather elements. The conclusion based on this small sample is that there are number of ways of determining threshold probabilities, with varying degrees of accuracy.

2.7 RELIABILITY DIAGRAMS MEASURES RELIABILITY AND SHARPNESS

A reliability diagram is a verification tool specifically designed to give a graphical depiction of the reliability of the probability forecast (the extent to which a probability forecast matches the actual frequency of occurrence of the event).

The diagram is formed by stratifying a set of probability forecasts according to deciles of probability. Deciles are either centered on each 10%, in which case the extreme categories are half-size, or they may be centered on 5%, 15% etc. The former is preferred when the forecast set consists of probabilities expressed to the nearest even 10%. For each probability category, the frequency of observation of the event is calculated. Observed frequencies are plotted against the set of forecast probability categories, the 45 degree line is included to represent perfect reliability, and the sample size for each category may be plotted next to the points on the graph. A horizontal line representing the long term or sample climatological probability may also be included.

The reliability diagram is in some ways analogous to scatter plots. For reliability tables, the data is already stratified by the forecast into 11 categories (usually), and there are only 11 plotting points. Reliability diagrams thus represent stratification conditioned on the forecast and can be expected to give information on the real meaning of the forecast. One other difference is that with reliability tables the observations can take on only two values, 1 or 0. That is, the forecast event either occurred or it didn't.

2.7.1 Interpretation

Both reliability and sharpness attributes can be assessed (Hsu,1986) from the reliability diagram. Reliability is indicated by the proximity of the plotted curve to the 45 degree line. If the curve lies below the line, this indicates overforecasting (probabilities too high); points above the line indicate underforecasting. Reliability is analogous to bias for continuous forecasts, and is evaluated in much the same way on reliability diagrams as bias is on scatter plots.

Sharpness may be inferred from the distribution of sample sizes in the probability categories. The sharper the forecast, the greater the sample sizes in the extreme categories near 0 and 100%. Conservative (not sharp) forecasting is indicated by large percentages of the sample occurring in the probability categories near the climatological probability. Sharpness is related to the dispersion (variance, for example) of continuous forecasts, but is a more important attribute of probability forecasts because the observations are two-valued. A perfectly sharp probability forecast is also two-valued, that is, a categorical forecast. Categorical forecasts are always unreliable to a greater or lesser degree, unless they are accurate. A perfectly reliable and perfectly sharp probability forecast is one for which only 0 and 100% is forecast (categorical), and for which all the 0's correspond to non-occurrences of the event and all 100's correspond to occurrences of the event. On a reliability table, this would show up as two points at (0,0) and (100,100)

The following schematics show some typical types of curves seen on reliability diagrams:

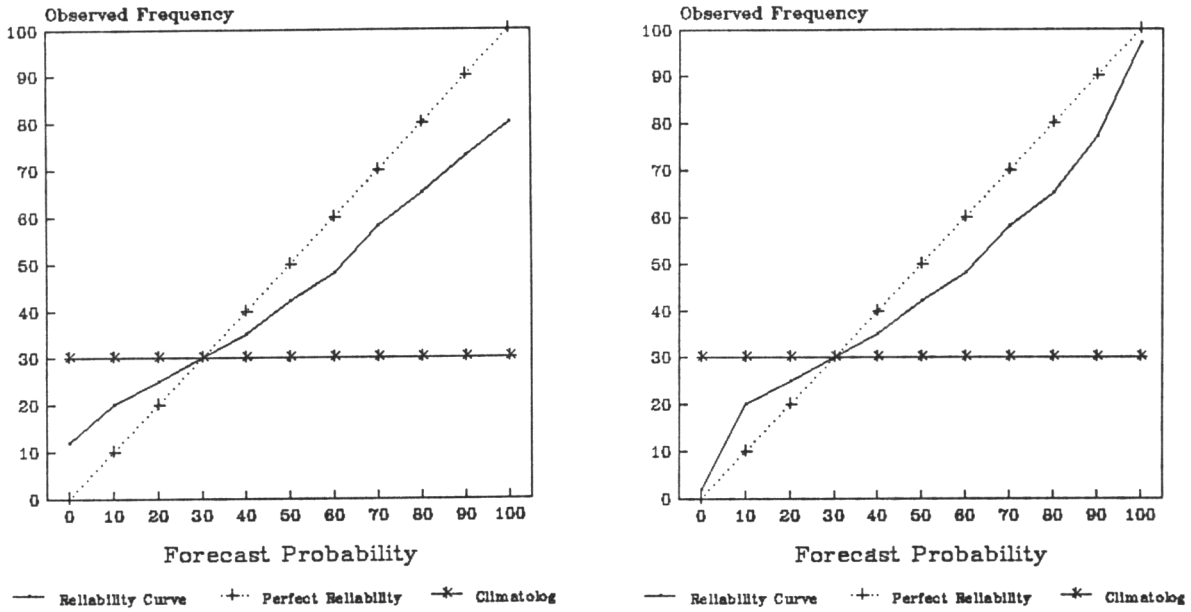

Figure 2.6 **Left:** An example of a method which has resolution and sharpness, but at the expense of reliability. High probabilities are overforecast and low probabilities are underforecast. If this characteristic is present by itself, the curve will cross the 45 degree line at or near the climatological probability. The two attributes of reliability and resolution can be traded against each other in any forecasting system, but for a given sample with an overall accuracy as defined for example by the Brier score, it is not possible to increase either reliability or resolution without decreasing the other. If the forecasts are perfectly reliable, then resolution and sharpness are identical. The degree of sharpness cannot be determined from the curve alone; usually the frequency that each category is forecast is included when plotting the curve. Thus if the 90% category was forecast 20 times by one procedure and 1000 times by another procedure, we would say the later procedure is sharper than the former. **Right:** A slight variation on the left hand curve. Reliability curves will sometimes curve back towards the 45 degree line near 0% and 100%. This is thought to be caused by the fact that the probabilities are bounded by 0 and 100, while the forecast system that generates them may use unbounded variables. Cases where the system tries to forecast out of range are slotted into the extreme categories.

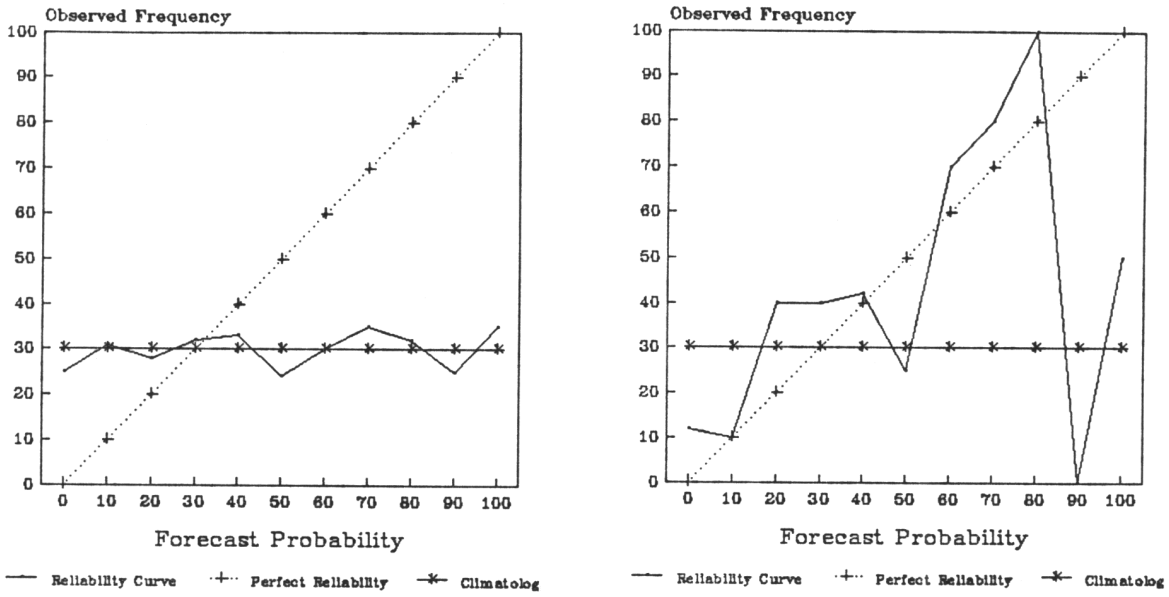

Figure 2.7 **Left:** A forecast which completely lacks resolution. Whatever the forecast probability, the climatological frequency occurs. Plenty of sharpness, but little reliability. **Right:** Effect of sample sizes that are too small. The large variations in the curve do not have statistical significance. The curves will usually settle down when there are over 50 cases in each probability category.

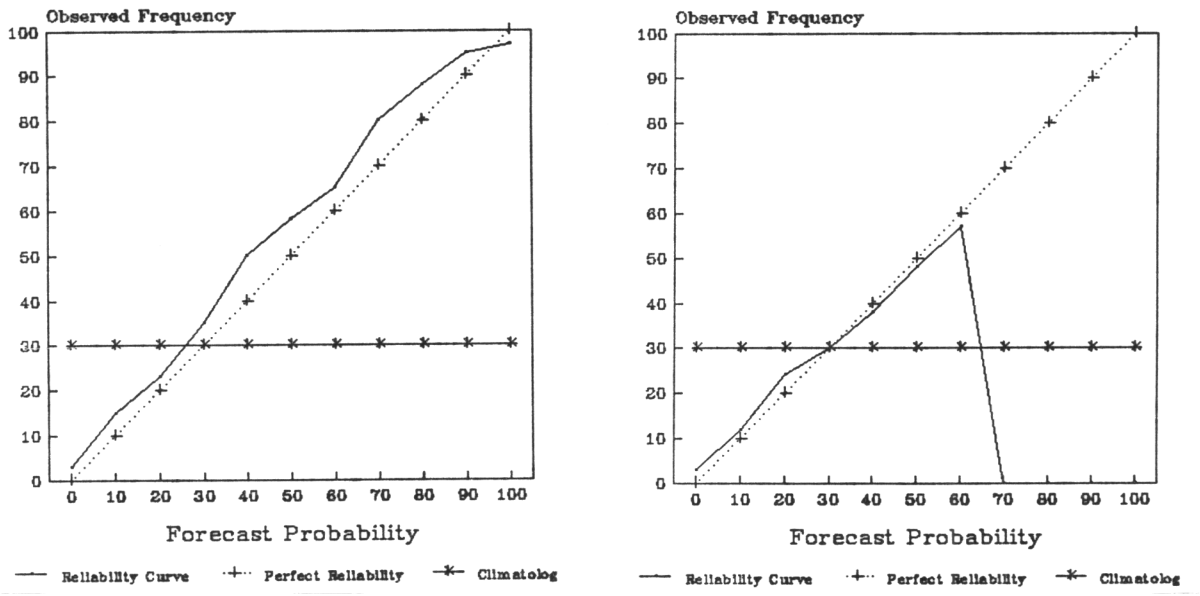

Figure 2.8 **Left:** A forecast with a bias. When the entire curve or most of it lies on one side of the 45 degree line, consistent under- or overforecasting is indicated. This is usually due to biases in the forecast system. **Right:** A conservative forecast system. Reliable, but does not attempt to forecast high probabilities of the event. This type of curve often occurs in reliability diagrams used to verify forecasts of rare events.

As shown above, a horizontal curve represents sharpness with little reliability, one type of useless forecast. Another type of unskilled forecast is represented by the case of perfect reliability but no sharpness given by a climatological forecast. A climatological forecast is represented on a reliability diagram by a point at the intersection of the climatological probability line and the 45 degree line.

Figure 2.9 shows two reliability tables for one year of forecasts for 72 Canadian stations, representing approximately 50,000 cases. The forecast event is the occurrence of heavy precipitation (> 10mm) within a 12 hour period ending at T+24 hours. Forecasts were made with two different statistical methods (Multiple Discriminant Analysis (MDA) and Regression Estimate of Event Probability (REEP) respectively), but used the same predictor set, and the verification sample is the same for both. Differences in sharpness and reliability are clearly evident from the figure. The MDA forecasts are sharper, even attempting 49 forecasts of near 100% probability of heavy precipitation, but are not particularly reliable. All probabilities over 20% are overforecast. In fact, the curve is nearly horizontal above 60%, meaning that all forecasts over 60% achieve about the same hit rate of 30%. The REEP forecasts are very reliable, but not sharp, never attempting a forecast of greater than 50%. Note, however, that the climatological frequency of the event is less than 5%. In light of this, even a 50% forecast would be significant. Which of the two methods would be preferred probably depends mostly on the application, but this case illustrates that different characteristics of forecast systems can be well-documented by reliability tables.

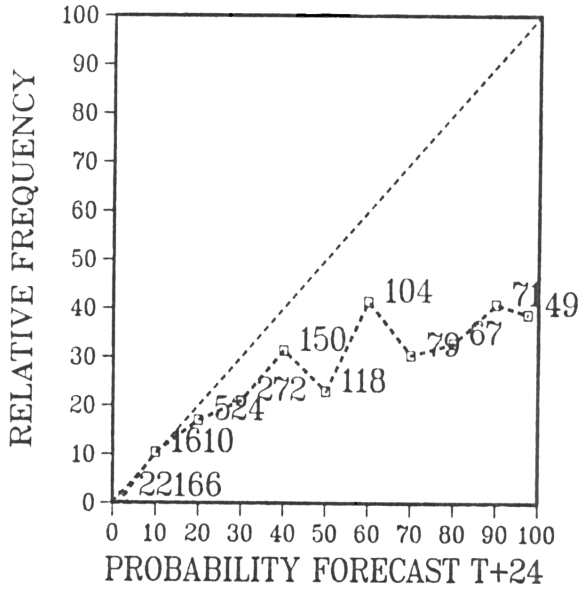

Figure 2.9 a) Reliability table for heavy precipitation category forecast, statistical method MDA.

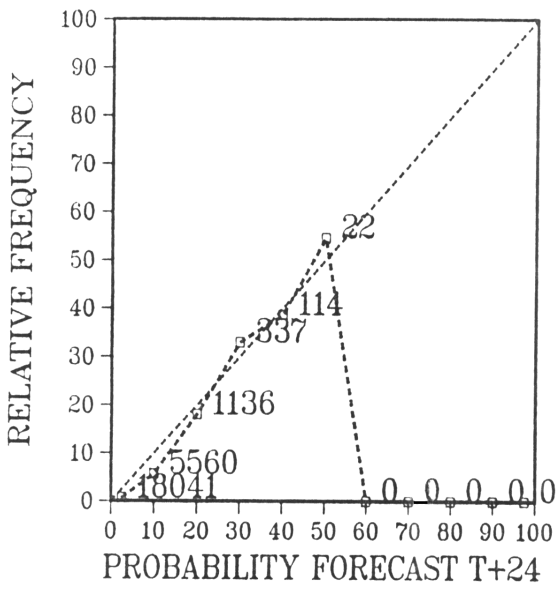

Figure 2.9. b) Reliability table for heavy precipitation category forecast, statistical method REEP.

Figure 2.10 shows another reliability table. In addition to plotting the category sample sizes next to the points on the graph, the distribution of forecasts is also given as a bar graph for ease of interpretation. This is a reliability table verification of one year (May, 1987 to April, 1988) of operational model output statistics POP forecasts for 23 Canadian stations. The curves both show an overforecasting tendency for all forecast probabilities except the lowest. This is probably due to a bias in the driving model that is not represented in the equations. The shortest forecast range actually shows a slightly greater overforecasting tendency than the longest, which may be related to bias characteristics in the driving model (the Canadian Spectral model). The two distributions of forecasts clearly show the tendency towards less sharpness in the forecasts with longer projections. For the 0 to 12 h range, more forecasts of the extreme probabilities 0 and 100% are made, while at longer ranges, the decrease in accuracy of the model renders such definitive forecasts unwise. Note that the frequencies for the longer range forecasts tend to show the greatest increase near the climatological mean probability of 26 %.

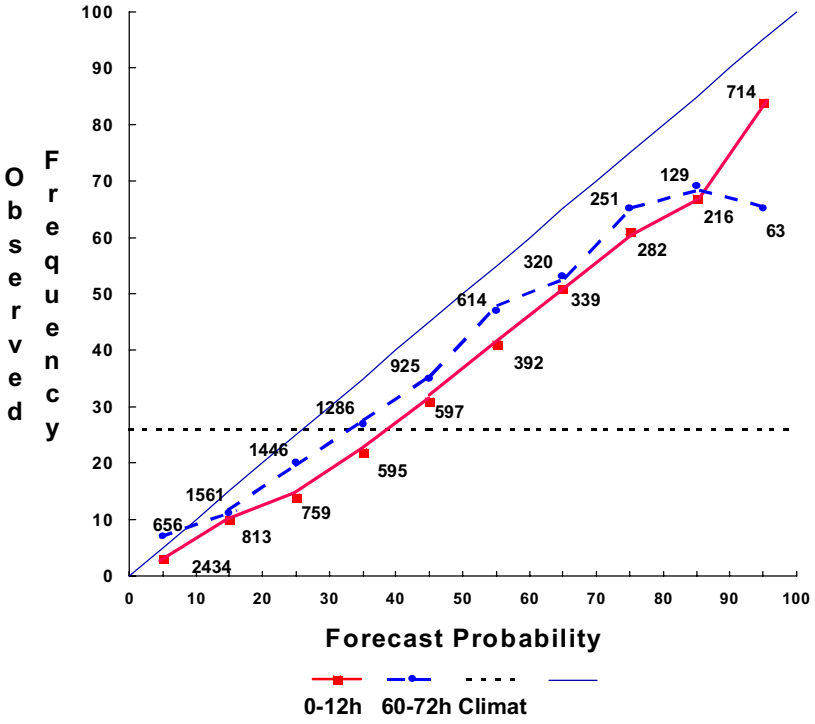

Figure 2.10 a) Reliability table verification of operational POP forecasts for 23 Canadian stations, 00UTC model run.

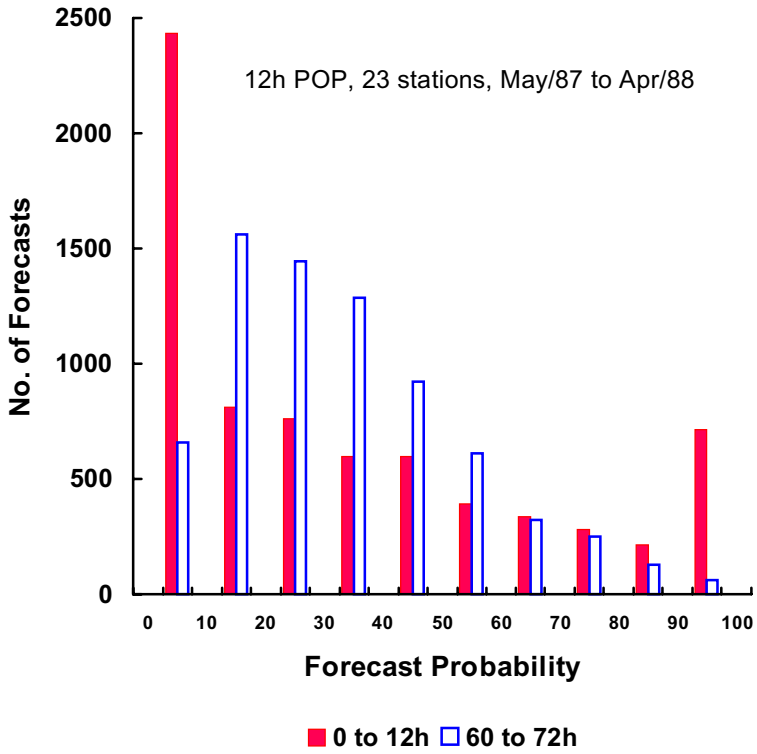

Figure 2.10 b) Bar graph representation of reliability table found in figure 2.10a.

Finally, when the reliability curve traces a saw-tooth type pattern this is often indicative of insufficient data and interpretation should be withheld until the volume of data points has been increased.

2.8 THE BRIER SCORE (PS), BRIER SKILL SCORE (BSS) SUMMARY SCORE

The Brier Score (Brier, 1950) is the mean square probability error and can be expressed most simply as:

$$PS = \frac{1}{N}\sum_{i=1}^{N}(f_i - O_i)^2$$

for a set of N forecasts of a two category (dichotomous) element, where f_i is the forecast probability of occurrence of the event and O_i is one if the event occurred and zero if it didn't. In this form, the score has a range of 0 to 1 and is negatively oriented, that is, zero represents a perfect score and 1 represents the worst possible score. If a positive orientation is desired (usually preferred by administrators and politicians), the score value can be subtracted from 1.

The Brier Score can be used for either categorical or probability forecasts. In all cases, the O_i will be all zeros and ones. It is also possible to use values of the O_i in between 0 and 1 to represent percentages of time each category occurred over a finite valid period, as long as the f_i are valid for the same period.

The Brier Score should not be used for multiple-category events, because it does not take into account the proximity of the observed category to the forecast category. Nevertheless it is sometimes used for multiple category events, in the form:

$$PS = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{k}(f_{ij} - O_{ij})^2$$

for k categories and N events. This form of the Brier Score has a range of 0 to 2, negatively oriented, and is usually divided by 2 to make the range 0 to 1. The following example illustrates the potential for misleading results when the Brier Score is used for multiple category events:

```
     Forecast A: 4 categories, probabilities:  0   90   10    0

     Forecast B: 4 categories, probabilities:  0   30   30   40

     Forecast C: 4 categories, probabilities:  0   10    0   90

     Occurrence: The first category

     Brier Scores according to above formula, divided by 2:
                 Forecast A:  0.91
                 Forecast B:  0.67
                 Forecast C:  0.91

Comment: All scores show a poor forecast because the verifying event was
assigned 0 probability in all cases.  However, Forecast A can be considered
to be closer because high probability is assigned to the next category.  The
score says that B is better (lower score).  This is because the probabilities
are spread out among the categories, and smaller magnitude errors are penal-
ized relatively less than large magnitude errors in a quadratic scoring rule.
Since the Brier Score is not sensitive to distance, each forecast receives
the same score for a given set of probability magnitudes, no matter how they
are distributed among the non-verifying categories.  Forecast C illustrates
this.
```

The Brier Score may be used with any two-category element; the most common being precipitation occurrence. The following table illustrates the use of the score for 10 cases.

Table 2.16: Data for sample calculation of the Brier Score

Occasion (i)	Forecast Probability (f_i)	Observed (O_i)
1	0.7	0
2	0.9	1
3	0.8	1
4	0.4	1
5	0.2	0
6	0.0	0
7	0.0	0
8	0.0	0
9	0.0	0
10	0.1	0

$$PS = 1/10[(0.7-0.0)^2 + (0.9-1.0)^2 + (0.8-1.0)^2 + (0.4-1.0)^2 + \ldots + (0.1-0.0)^2] = 0.095$$

2.8.1 Characteristics of the Brier Score:

A) Relationship to climatological frequency

The Brier Score is strongly influenced by the climatological frequency of the event in the verification sample. If the climatological frequency of rain, say, is 30%, then in the absence of any forecasting skill, the best strategy is to forecast 30% POP on all occasions. The expected value of the Brier Score (i.e., for a sample with a frequency of occurrence that equals the climatological frequency) is then:

$$E(PS) = 0.7(0.3)^2 + 0.3(0.7)^2$$
$$= 0.21$$

for the 70% of occasions where rain doesn't occur and the 30% of occasions where rain occurs. If the event is rarer, for example thunderstorms with a climatological frequency of 5%, the expected value of the Brier Score is,

$$E(PS) = 0.95(0.05)^2 + 0.05(0.95)^2$$
$$= 0.047$$

This means that, the rarer the event, the better the Brier Score that can be obtained, without showing any skill in the forecast. Furthermore, the frequency of the event in the verifying sample has a similar effect on the Brier Score, despite the forecast skill. For this reason, it is hazardous to compare Brier scores based on different samples, or for different elements since the sample climatology cannot be controlled by the forecaster. Table 2.17 shows this effect for a slightly different type of unskilled forecast, "conditional persistence". This forecast is generated by using the climatological probability of the event, conditioned on the six-hour previous observation. Taking the first element fog as an example, if fog occurred six hours ago, the forecast probability is given by the frequency of fog occurrences in all cases where fog occurred six hours previous. If fog didn't occur six hours ago, the forecast probability is derived from all cases where fog didn't occur six hours previous. The table gives the Brier Scores for these unskilled forecasts, based on 28 years of hourly weather data. Conditional Persistence is more skilled than climatology at six hours, and can be expected to have lower scores, but the variability due to the climatological frequency is still dominant. Thus, freezing drizzle is the rarest event in this sample and fog is the most common. The weather elements

listed in the table are all treated as dichotomous variables.

Table 2.17: Brier Scores for unskilled forecasts for weather elements at Toronto and Trenton (combined), Spring season. Scored on three year independent sample, but forecasts based on 28 year climatology.

Element	Conditional Persistence
Fog	.08385
Haze	.02492
Blowing Snow	.00102
Drizzle	.01454
Rain	.04091
Rain Shower	.04565
Snow	.02485
Snow Shower	.02014
Freezing Drizzle	.00078
Freezing Rain	.00338
Thunder	.00307

B) Murphy's partitioning of the Brier Score

Murphy (1973) proposed a three-way partitioning of the Brier Score which contributes to an understanding of the score. Suppose the forecast probability is allowed to take on specific discrete values, for example intervals of 10% from 0 to 100%. Then the N events of the sample are divided into 11 categories according to the forecast probability. The Brier Score for the N forecasts may be written

$$PS = \frac{1}{N} \sum_{k=1}^{T} \sum_{i=1}^{M_k} (f_k - O_{ki})^2$$

$$PS = \frac{1}{N} \sum_{k=1}^{T} \sum_{i=1}^{M_k} [(f_k - \bar{O}_k) + (\bar{O}_k - \bar{O}) + (\bar{O} - O_{ki})]^2$$

where O_k bar is the average frequency of occurrence for the kth category and O bar is the average frequency of occurrence over the whole sample. This can be rewritten as:

$$PS = \frac{1}{N} \sum_{k=1}^{T} \sum_{i=1}^{M_k} [(f_k - \bar{O}_k)^2 + (\bar{O}_k - \bar{O})^2 + (\bar{O} - O_{ki})^2]$$

$$= (\text{RELIABILITY} + \text{RESOLUTION} + \text{UNCERTAINTY})$$

RELIABILITY is the degree to which the forecast probabilities agree with the observed frequency of occur-

rence in the sample, as defined in Chapter 1.

RESOLUTION is a measure of the degree to which the relative frequencies of the event in the T subsamples differ from the relative frequency of the event in the whole sample, again as defined in Chapter 1. Note that the resolution term is negative. The greater the resolution, the lower (better) the score.

The Brier Score can be partitioned to yield a different measure of resolution (Sanders, 1963, 1967)).

UNCERTAINTY. This is simply the variance of the observations in the sample, as defined in Chapter 1. It is the uncertainty term that gives the Brier Score its dependence on the climatological frequency of the event and makes it difficult to compare scores for samples with different frequencies. This must be viewed as a drawback of the score since the uncertainty term cannot be controlled by the forecaster. Alternately, the uncertainty term represents an internal standard of reference, leading to a sample skill score.

For a given set of forecasts, the uncertainty factor will be fixed. Reliability and resolution can be traded by altering the probability forecasts. The numerical contribution of reliability to the score is greater than the resolution contribution. Therefore, forecasting systems that emphasize sharpness (categorical forecasts for an extreme event) will usually be met with poorer Brier Scores because the reliability penalty will increase at the expense of resolution. (The resolution reward (negative contribution) must exceed the reliability penalty (positive contribution) for forecasts to exhibit positive skill. Although sharpness is not included in the three-way partition, the PS is sensitive to sharpness by the interaction of reliability, resolution and sharpness). In our experience it is often worthwhile to sacrifice some reliability for the sake of sharpness or resolution, even if the Brier Score is lowered (i.e.poorer). Sharp forecasts act as an alert signal, which is of use particularly in situations where forecasts are used as guidance in the preparation of other forecasts. If the Brier score is used to compare forecast systems, it will generally favour those systems that are reliable.

The Brier Score is not enormously sensitive to variations in forecast accuracy in the sense that a score of greater than 0.30 in most cases represents a poor forecast. Scores tend to lie in the range 0.10 to 0.25 for most elements. Scores for forecasts of rare events tend to be better, and will usually lie below 0.10.

2.8.2 Brier Skill Score (BSS)

The Brier Skill Score is the skill score based on the Brier Score. It is in the usual format described in Chapter 1,

$$\mathbf{BSS} = \frac{\mathbf{PS_c - PS_f}}{\mathbf{PS_c}}$$

where PS_f is the Brier Score for the forecast, PS_c is the Brier Score for the standard forecast, usually climatology, and the perfect forecast is a Brier Score of zero. For simplicity, the numerator is reversed because the Brier Score is negatively oriented. Although climatology is usually used as a standard, any unskilled or even a skilled forecast could be used. The Brier skill score expresses the percentage improvement over the standard, and has a range of $-\infty$ to 1.

Characteristics of the Brier Skill Score:

1. The better the standard score, the smaller the skill score for a given set of forecasts. The smaller the set of forecasts, the more likely the standard forecast will, by chance, be very good and hard to beat.

2. It is NOT strictly proper. This means that it is theoretically possible for forecasters to improve their skill score by forecasting probability values which differ from their true belief. However, it is difficult to conceive of how this might be done on a day-to-day basis since the skill score that a forecaster obtains is based on an ensemble of forecasts, and also depends on the skill of the climatological forecast, which cannot be known until the set of forecasts is complete. Strictly properness is not an important consideration for the BSS.

3. It is a highly non-linear score. The non-linearity is due to the use of the standard score in the denominator. As the skill of the standard approaches the perfect score, the denominator goes to zero and any difference between the forecast and standard forecast Brier scores (the numerator) is amplified considerably making the score unstable. This

means that it is very important that the BSS be computed on a sufficiently large sample, one for which the sample climatology of the event is representative of the long term climatology. Otherwise, the score is likely to oscillate through large variations over a period of time. The rarer the event, the longer the sample needed to stabilize the skill score.

The BSS is inherently a summary score. For best results, the scores should be computed on the whole sample, then inserted into the formula for the skill score. This gives a more stable result and prevents undesired weighting due to variations in the denominator. For this reason, skill scores computed on several samples should not be averaged. Rather, the skill should be recomputed for the aggregated sample.

Figures 2.11 and 2.12 show Brier skill scores for two Canadian stations for POP forecasts. The scores are computed on a monthly basis, with respect to long term climatology. There is obviously considerable variance in the scores from month to month. Note particularly Saskatoon, where, in January, 1985, a score of 45% better than climatology was obtained, while the next month, the skill was -65%, 65% worse than climatology. Oscillations such as this are unlikely to be due to real changes in forecast skill, but rather to changes in the skill of the standard forecast, due to changes in climatology on a monthly basis. A comparison with figure 2.13 shows this to be true: the Brier Scores over the same period varied over a much narrower range. One month is too short a time period for calculation of skill scores on a single station basis, especially for Saskatoon and other stations with a low climatological frequency of precipitation.

Large variances such as shown in Figure 2.11 make it difficult to use the scores to discern trends in forecast skill, since they are masked by the variability. Visually, it is possible to imagine a regression line on Figures 2.11 and 2.12 which has a slightly positive slope as a function of time, but the regression coefficient might not pass a significance test because of the variance in the data values. Without waiting for more data, more useful graphs would be obtained by recalculating the skill scores based on a running year of data.

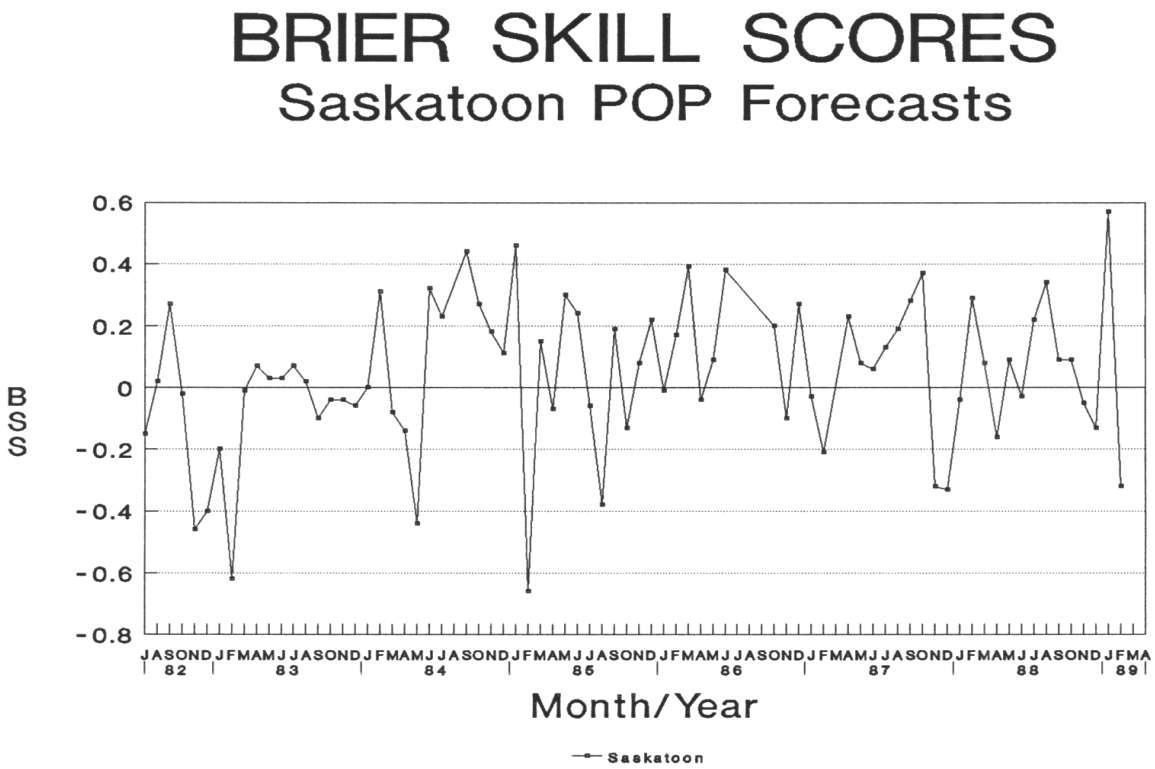

Figure 2.11. Monthly Brier skill scores for POP forecasts at Saskatoon, tomorrow period (36 h).

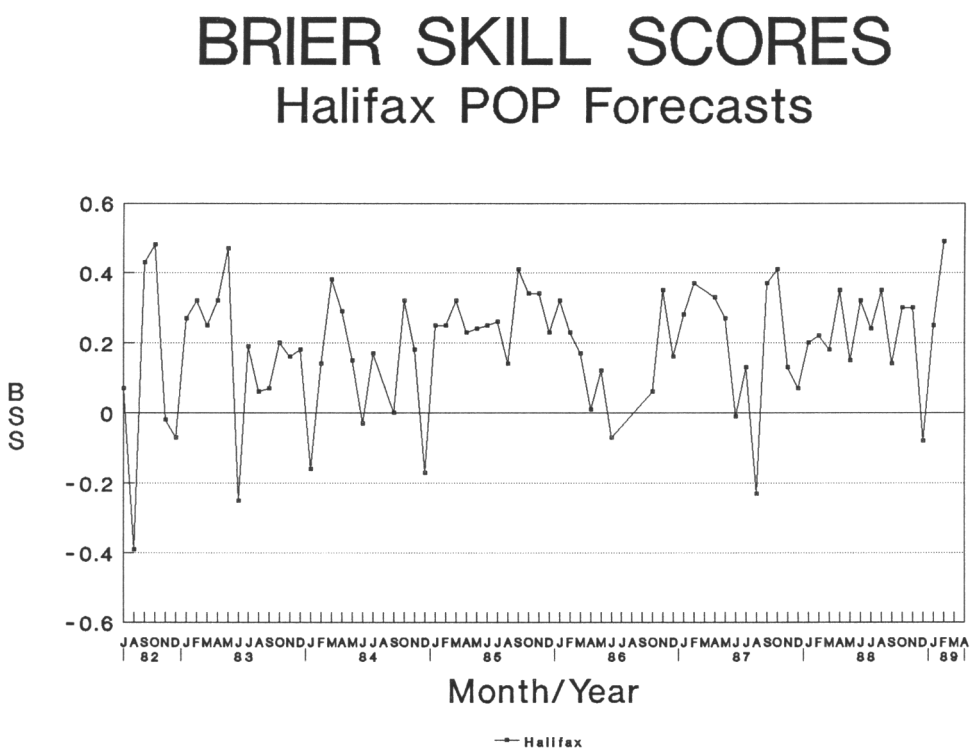

Figure 2.12. Monthly Brier skill scores for POP forecasts at Halifax, tomorrow period (36 h).

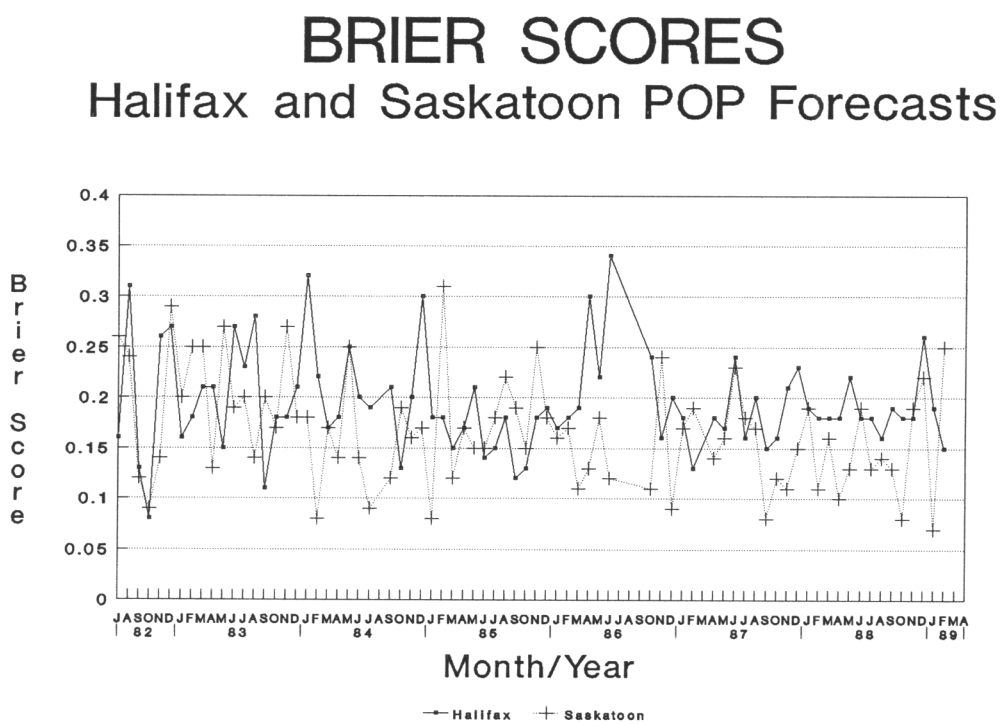

Figure 2.13. Monthly Brier scores for POP forecasts at both Saskatoon and Halifax, tomorrow period (36 h).

Some examples of the behaviour of the BSS are presented below to illustrate its sensitivity in verification of a rare event on a moderate size sample. In all cases, a sample of 250 events is assumed, representing about 4 months of forecasts for one station (2 per day). The first case shows the effect of sharpness in a situation where the expected number of rare events occurs, and the second case shows how a change in the **sample** climatology can affect the score. The assumed climatological frequency is .02 which would apply to heavy precipitation events, or perhaps some precipitation frequencies at prairie stations.

CASE 1

Occurrence: 5 cases of heavy precipitation
245 cases of anything else

Forecasts: Forecaster A: 20% probability forecast 5 times, gets one of them right, i.e. 1 of the 5 forecasts coincides with heavy precipitation occurrence. Assume A forecasts the climatology probability of heavy precipitation for the rest of the cases (2%)

Forecaster B: 60% probability of heavy precipitation forecast 5 times (he is surer of himself), gets one of them right. B forecasts climatological probability for the rest.

The BSS numerator is the difference of the squared errors between the forecast and climatological forecast.

For *forecaster A*:
numerator of BSS = 1 x (.9604-.64) + (245x0) + 4x(.0004-.04)
= .3204-.1584
= .1620

denominator of BSS is the climatology brier score, in this case,
= 245x(.0004) + 5x(.9604)
= 4.90 for 250 cases

BSS= .1620/4.90 = 3.3% overall

i.e. A achieves a 3.3% improvement over climatology by forecasting 20% probability against a climatology of 2%, 5 times, once correctly (4 false alarms) and no skill in other cases. Note that for the purpose of this argument it is necessary only that the skill of the other 245 cases average to 0 for simplicity.

For *forecaster B*:
numerator of BSS = 1x(.9604-.16) + 245x0 + 4x(.0004-.36)
= .8004-1.4384
= -.6380

denominator is the same as above. Therefore,

BSS = -.6380/4.90 = -15.0%

i.e. B gets the same event correct, calls the same 4 false alarms, but calls them more sharply (60%). This alone causes a difference in skill of 18% or so on a 4 month sample. It should be noted that the B forecast is an overforecast. That is, if 60% is forecast, 3 of the 5 should be correct. The break-even point in the score is 2 correct. That is, if A gets 1 correct with a 20% forecast (reliable), B will score as well or better with his 60% forecast if 2 or more are correct.

This case does show that a difference of 40% in forecast probability, in only 5 cases out of 250 can lead to substantial differences in the value of the BSS.

CASE 2 - Effect of sample climatology

Occurrence: 1 case of heavy precipitation
249 other cases

Forecasts: Exactly as for case 1, i.e. A forecasts 20% probability of heavy rain 5 times, gets the one occurrence. There are 4 false alarms. B forecasts 60% probability of heavy rain 5 times, gets the one occurrence.

Scoring: because the forecasts and their correspondence with events are all the same, the numerator of the skill score remains the same as in case 1. However, the sample now contains an abnormally low number of occurrences of the rare event. (It is rarer than expected, as in a dry period). Thus, the climatology score changes.

For *A*:
numerator of BSS = .1620, as above

denominator = (249x.0004) + 1x(.9604)
= 1.06 for 250 cases

BSS = .1620/1.06 = 15.3%

That is, the skill has been amplified to 15% improvement over climatology. The one event was caught (20% forecast) and there were 4 false alarms.

For *B*:
numerator of BSS = -.6380 as before

denominator = 1.06

BSS = -.6380/1.06 = -60%

Negative skill is also amplified. At first glance, catching a single heavy precipitation event with a forecast of 60% vs. a climatology of 2% would seem valuable, but the skill is overwhelmingly offset by the false alarms, 4 cases when climatology, with a score of .0004 is extremely hard to beat.

These examples show that the BSS can oscillate through wide variations when it is applied to forecasts of rare events. Differences of skill of 50% or more can easily be generated by slight differences in forecast probabilities when the common, non-event occurs. Furthermore, with relatively small samples, the score values are highly dependent on the sample frequency of occurrence of the rare event. The rarer the event that is being verified, the larger the sample required to stabilize the score. For heavy precipitation, thunderstorms, or perhaps any precipitation on the prairies, several years of data may be required before the score could be relied on.

The broader underlying question is whether a quadratic scoring rule is the most appropriate measure of skill on a summary basis. Perhaps it should be one of several measures, each of which answers part of the skill question. The deliberate overforecasting of a rare event could certainly be justified for its potential to alert, even at the cost of false alarms. Such practices are not encouraged by any quadratic scoring rule, leading to the suggestion that other verification methods be adopted, especially for rare events.

2.9 RANKED PROBABILITY SCORE (RPS) and SKILL SCORE (RPSS) SUMMARY SCORE

The ranked probability score (RPS) was developed by Ed Epstein in 1969 and modified by Allan Murphy in 1971. The Brier score (PS) is intended for binary predictands while the RPS satisfies the requirement to verify multi-category probability forecasts. Although not as widely used as the PS, the RPS was chosen as the principal verification measure in the 1982 Canadian National Terminal Forecast Verification system. Comments on the results from the National scheme can be found in Chapter 5.

Ranked and unranked (ordered and unordered) variables are those variables for which the concept of numerical distance between the values, or states, is and is not meaningful, respectively. Examples of ranked variables are temperature, and precipitation amount. An example of an unranked variable is the 3-category predictand, "no precipitation", "rain", and "snow".

The ranked probability score (RPS) of the ordered forecast $\mathbf{P} = (P_1, P_2, \ldots\ldots, P_k)$ for a certain time is defined as:

$$\mathbf{RPS(p,d)} = 1 - \frac{1}{K-1}\left[\sum_{i=1}^{K}\left(\sum_{n=1}^{i} P_n - \sum_{n=1}^{i} d_n\right)^2\right]$$

for K mutually exclusive and collectively exhaustive classes or states. The vector $d = (d_1, d_2, \ldots d_k)$ represents the observation vector such that d_n equals one if class n occurs, and zero otherwise. The RPS has a range of zero to one and is positively oriented (the higher the value, the better the forecast).

For the special case where there are only two classes, i.e. rain/no rain, VFR/IFR*, non-severe/severe etc., the RPS and Brier score are the same. The Brier score however must be normalized (i.e. divided by two) and oriented positively (subtract from one) to be numerically equal to the RPS, for the 2 category case. Actually, the original RPS as defined by Epstein has negative orientation like the Brier score. We (in Canada) tend to use the positively oriented version of the RPS.

* VFR, Visual Flight Rules; IFR, Instrument Flight Rules

2.9.1 Characteristics of the RPS:

1) The RPS has a positive orientation in that the more accurate forecasts receive the higher score (between zero and one). A perfect categorical forecast always receives a score of one. The worst possible categorical forecast receives a score of zero. With multi-category probabilistic forecasts, the minimum possible score is larger than zero, and the maximum is less than one. The maximum score decreases as the forecast moves away from categorical forecasts.

2) A forecast which assigns equal probability to two or more categories will not necessarily receive an identical score for the occurrence of each of these categories because the RPS considers both the categories to which the bulk of the probability is assigned, and also the 'expected' category implied by the distribution of probabilities.

3) The RPS is sensitive to the distance or magnitude of the error so that more credit is given to a forecast which concentrates its probability about the event that occurs.

2.9.2 RPS Skill Score

The RPS may also be used in a skill score formulation (RPSS) in the format given in Chapter 1:

$$\mathbf{RPSS} = \frac{\mathbf{RPS(forecast) - RPS(standard)}}{\mathbf{1 - RPS(standard)}}$$

The RPSS measures skill with respect to the standard, and ranges from +1 (perfect forecast) to $-\infty$. Negative RPSS values indicate that the forecast has less accuracy than the standard. Previous comments on the Brier skill scores apply equally to the RPSS. It is unstable when applied to small datasets which may result in large changes in the score's value when one dataset is compared to another. Interpretation of the RPSS is similar to the interpretation of the BSS.