Verification of Precipitation Forecasts: A Survey of Methodology Part II: Verification of Probability Forecasts at Points

Barbara G. Brown[1]
National Center for Atmospheric Research
Boulder CO U.S.A.

1. Introduction

Probabilistic forecasts provide a consistent approach to presenting the information in weather forecasts, as expressed by Murphy (1993):

> *"...the widespread practice of ignoring uncertainty when formulating and communicating forecasts represents an extreme form of inconsistency and generally results in the largest possible reduction in quality and value."*

While probabilistic forecasts represent a very special type of forecast (especially from the users' perspective), many aspects of the evaluation of probabilistic forecasts are similar to verification of other types of forecasts. In many cases, related measures are utilized for evaluation of both types of forecasts. In fact, the dichotomous forecasts discussed by Wilson (2002; this volume) can be considered to be probabilistic forecasts in which only two probability values (0 and 1) are allowed.

Probabilistic forecasts come in a variety of different "flavors." Most commonly, probabilistic forecasts are associated with a dichotomous (e.g., Yes/No) predictand. However, it also is possible to formulate multi-category probabilistic forecasts, in which a consistent set of probability values is assigned to several categories of the predictand. Quantitative precipitation forecasts in which probabilities are assigned to several categories of precipitation amount are an example of this type of forecast. Ensemble forecasts represent a third, related, type of probabilistic forecast, in which the full distribution of values is represented.

This paper extends the framework described by Wilson (2002) to the case of probabilistic forecasts. Some common methods for verification of probabilistic forecasts are described, as well as approaches based on signal detection theory (SDT). While most of the paper focuses on dichotomous forecasts, some attention also is given to multi-category forecasts. Finally, special consideration is given to methods for verification of ensemble forecasts, and a brief discussion is presented regarding the connection between verification measures and estimates of economic value.

2. Background and basics

2.1 Forecasts and observations

As noted in the previous section, probabilistic forecasts come in a variety of flavors. The primary distinction among them is the number of categories of the predictand. Most commonly,

1. Author contact information: Barbara G. Brown, NCAR, P.O. Box 3000, Boulder, CO, U.S.A. E-mail: bgb@ucar.edu

the predictand is dichotomous (i.e., has two categories). Two examples of dichotomous predictands are (a) the occurrence or non-occurrence of precipitation and (b) the exceedance or the non-exceedance of a temperature threshold. Examples of multi-category events include probabilistic quantitative precipitation forecasts (QPFs) in which probabilities associated with several categories of precipitation amount (e.g., 0, >0 - 0.5 mm; 0.5 - 0.75 mm; 0.75 - 1 mm, etc.) are predicted. As noted by Wilson (2002), variables can be stratified into two categories by setting one threshold and, in general, into n categories by setting n-1 thresholds. However, the forecast categories also may be associated with nominal characteristics (e.g., precipitation type, such as rain, snow, freezing rain, etc.).

The explicit quantification of forecast uncertainty is the aspect of probabilistic forecasts that sets them apart from other types of forecasts. In fact, standard categorical forecasts can be considered to be completely confident probabilistic forecasts, in which only probability values of 0 and 1 are used. For example, a forecast of “rain today” implies a probability of 1 that rain will occur today, and a probability of 0 that no rain will occur today. In contrast, a 40% probability of rain today expresses the uncertainty in the forecast, indicating there is a 40% chance of rain occurrence and a 60% chance of no rain.

Probabilistic forecasts can be generated objectively [e.g., through model output statistics (MOS) or ensemble forecasts] or they can be formulated subjectively (i.e., by human forecasters). In the former case, a forecast equation has been fitted to the occurrence of the event by minimizing (or maximizing) some function; or the event distribution has been represented by an ensemble of model runs. In the latter case, a human forecaster subjectively assesses the basic uncertainty associated with his/her forecast. While interpretation of forecasts formulated using these three approaches may be somewhat different, the verification approaches are basically the same. In the case of ensemble forecasts, however, more information is generally available than is the case with the other types of forecasts. In particular, ensemble forecasts provide a distribution of possible forecasts, as opposed to a single probability associated with a particular outcome.

2.2 Statistical framework

The statistical framework for verification, first identified by Murphy and Winkler (1987) and described by Wilson (2002), applies directly to verification of probabilistic forecasts, with some simplifications in the case of a dichotomous predictand. In particular, the joint distribution of forecasts and observations can be represented as $p(x,f)$, where f represents the forecasts and x represents the observations, and $p(f,x)$ is the joint probability of f and x. This joint distribution can be factored in two different ways: (a) as the calibration-refinement (CR) factorization,

$$p(f, x) = p(x|f)p(f), \tag{1}$$

where $p(x|f)$ is the conditional distribution of observations given the forecast, and $p(f)$ is the marginal (i.e., unconditional) distribution of the forecasts; and (b) using the likelihood-base rate (LBR) factorization,

$$p(f, x) = p(f|x)p(x), \tag{2}$$

where $p(f|x)$ is the conditional distribution of forecasts given the observation, and $p(x)$ is the marginal distribution of the observations. These factorizations are more completely described by Wilson (2002), Wilks (1995), and Murphy and Winkler (1987).

Of most importance is recognition that each factorization involves the combination of a conditional distribution and a marginal distribution. The CR factorization (1) involves the conditional distribution of *observations given forecasts* (called *"calibration"* by Murphy and Winkler) and the marginal distribution of *forecasts* (called *"refinement"* by Murphy and Winkler). The likelihood-base rate factorization (2) involves the conditional distribution of *forecasts given observations* (called the "*likelihood*") and the marginal distribution of *observations* (called the "*base rate*" by Murphy and Winkler).

In the case of probabilistic forecasts of a dichotomous event, the verification framework is greatly simplified because there are only two possible observations. In particular, let $x=1$ when the event occurs and $x=0$ when it doesn't occur. Often, the forecast probabilities also are limited to a particular set of values. This simplification is very common for human-generated forecasts; objective probabilistic forecasts frequently are rounded to the nearest "whole" probability value as well. Thus, the number of possible forecasts may be limited to, say, 13 discrete values (e.g., 0, 0.05, 0.1, 0.2, ..., 0.8, 0.9, 0.95, 1). In this case, the joint distribution would have 26 elements (2 possible observations and 13 possible forecasts). Because the sum of all the joint probabilities must equal one, the entire joint distribution can be specified in this case using 25 numbers. This number is the *dimension* of the verification problem.

The LBR factorization is greatly simplified in the case of probabilistic forecasts of a dichotomous predictand. In particular, the conditional distribution of forecasts given observations, $p(f|x)$, is based on only two distributions, $p(f|x=0)$ and $p(f|x=1)$. Moreover, the marginal distribution of observations simply consists of two probabilities, $p(x=0)$ and $p(x=1)$. By definition,

$$p(x = 0) + p(x = 1) = 1 . \tag{3}$$

Thus, the probability distribution of the observations is completely specified by only one of these probabilities. It is important to note that $p(x=1)$ is the *sample climatological probability of the event*.

Note that all of the distributions can be characterized using standard statistical measures. For example, the expected values are represented by the means, μ_f, μ_x, $\mu_{f|x}$, and $\mu_{x|f}$. The variances can be represented by σ_f^2, σ_x^2, $\sigma_{f|x}^2$, and $\sigma_{x|f}^2$. For dichotomous events, some of the summary statistics are related in a simple way to the distribution probabilities. For example, $\mu_x = p(x = 1)$. Similarly, $\mu_{x|f} = p(x = 1|f)$. In the case of the CR factorization (1), it is easy to see that

$$p(x = 0|f) = 1 - p(x = 1|f) = 1 - \mu_{x|f}, \tag{4}$$

where $\mu_{x|f}$ is the expected (i.e., mean) value of x given a particular forecast, f. Thus, only one number is needed to specify the distribution $p(x|f)$ for each f.

2.3 Attributes

The most common and important attributes of forecasts are described in Wilson (2002). These attributes also are presented by Murphy and Winkler (1992) in a slightly different form in the context of the joint distribution and its factorizations. The attributes of most interest for probabilistic forecasts are listed in Table 1. Some of these attributes are described in more detail in the paragraphs below.

Table 1: Attributes of forecast quality for probabilistic forecasts (adapted from Murphy 1997 and Murphy and Winkler 1992)

Attribute	Definition	Basic distribution(s)	Graphs and measures
Sharpness (refinement)	Degree to which probability forecasts approach zero and 1; "spread" of distribution of forecasts	$p(f)$	• Histogram of $p(f)$ • Variance of forecasts, σ_f^2
Resolution	Difference between $\mu_{x\|f}$ and μ_x, considered over all values of f	$p(x\|f)$, $p(x)$	• Resolution component of Brier Score • Attributes diagram
Discrimination	Degree to which forecasts discriminate between occasions when $x=1$ and occasions when $x=0$	$p(f\|x)$	• Discrimination diagram (plot of likelihood functions) • Difference in conditional means: $\mu_{f\|x=1} - \mu_{f\|x=0}$
Bias	Difference between mean forecast and mean observation	$p(f)$, $p(x)$	Mean Error (ME): $\mathrm{ME} = \mu_f - \mu_x$
Reliability (Calibration)	Degree of correspondence between conditional relative frequencies, $p(x\|f)$ and f, considered for all values of f	$p(x\|f)$	• Reliability diagram • Attributes diagram • Reliability measure from Brier score decomposition
Accuracy	Average degree of correspondence between f and x	$p(f,x)$	• Brier score = MSE • Other scores
Skill	Accuracy of forecasts relative to accuracy of forecasts based on a standard of comparison (e.g., climatology)	$p(f,x)$	• Brier skill score, BSS • Correlation, $\rho_{f,x}$, measures potential skill • ROC Area

Sharpness measures the "spread" or variability in the forecasts. Probability forecasts can vary between 0 and 1, but perfect forecasts only include the two end points, 0 and 1. Thus, sharper forecasts will more consistently include values close to 0 and 1, and sharpness measures the degree to which the forecasts approach these extremes. Forecasts with more variability (e.g., as measured by the standard deviation) are sharper forecasts. The basic shape of the distribution of forecasts also can provide clues about the relative sharpness of a set of forecasts. If the histogram of forecast relative frequencies is "bell-shaped" or flat, the forecasts are not very sharp. In contrast, if the histogram has a U shape, with most or all of the frequency at 0 or 1, the forecasts are relatively sharp. The histogram for perfect forecasts has two spikes, one at 0 and one at 1. Of course, sharpness is only one aspect of forecast quality: it is easy to have perfectly sharp forecasts that are very inaccurate.

Resolution measures how well the observations are "sorted" among the different forecasts. One would expect the mean observation to be different for different forecasts, and to be different from the overall mean observation. This attribute is measured using the RES statistic, which is described in Section 4.2.

Discrimination measures how well the forecasts discriminate between the events and non-events. Ideally, the distribution of forecasts for cases when the event occurs differs from the distribution when the event does not occur.

Bias measures the overall (average) error in the forecasts. It is simply measured by the difference between the mean forecast and the mean observation.

Reliability measures how well the forecast probabilities correspond to the conditional frequency of occurrence of the event. For example, over all occasions when a probability forecast of 0.20 is issued, the event would be expected to occur 20% of the time. The degree of this correspondence is a measure of reliability.

Accuracy measures the overall correspondence between the forecasts and observations. For probability forecasts, it can be measured using the Brier score (see Section 4.2), the correlation coefficient, or the ROC area (an SDT measure - see Section 4.4).

Skill measures relative accuracy by comparing the accuracy of the forecasts to the accuracy of some standard of comparison, such as climatology or persistence.

It is clear from this discussion and from Table 1 that a variety of different attributes are of interest when considering the quality of a set of forecasts. By itself, no single attribute or measure can provide a complete picture of the characteristics of the forecasts. Thus, it is critical - with probability forecasts as with other types of forecasts - to consider all the characteristics that are of interest and to measure a variety of different attributes to obtain a relatively complete picture of forecast quality.

3. Completely confident probabilistic forecasts

As noted earlier, dichotomous (e.g., Yes/No) forecasts of the occurrence or non-occurrence of the event of interest (e.g., precipitation occurrence) can be thought of as "completely confident" forecasts. That is, a Yes forecast can be thought of as a probability forecast with $f=1$, and a No forecast can be considered a probability forecast with $f=0$. Hence, these forecasts can

be treated as categorical forecasts as described in Wilson (2002). Table 2 shows a typical contingency table for verification of these types of forecasts, in the context of the joint distribution of forecasts and observations. In this table, each cell represents the number of times that each forecast-observation pair occurred. Note that Table 2 could be transformed into tables of the joint, conditional, and marginal probabilities by dividing each count by the appropriate overall, row, or column total.

Table 2: Contingency table for verification of dichotomous "completely confident" probability forecasts.

Observation	Forecast		Total
	Yes (f=1)	No (f=0)	
Yes (x=1)	n_{11}	n_{01}	$n_{11} + n_{01}$
No (x=0)	n_{10}	n_{00}	$n_{10} + n_{00}$
Total	$n_{11} + n_{10}$	$n_{01} + n_{00}$	$T = n_{11} + n_{10} + n_{01} + n_{00}$

The counts in Table 2 can be used to compute a variety of different measures. Some of these statistics are shown in Table 3, which includes a few additional measures not included in Wilson (2002). Table 3 also includes the definitions of these measures in the context of the distributional framework for verification. Some of the measures are of particular interest here because they form the basis for SDT, discussed in Section 4.4.

Many of the attributes listed in Table 1 can be evaluated using the measures in Table 3 as well as some other measures. For example, POD and POFD are related to discrimination, and FAR is related to reliability. Unfortunately, because only three numbers are required to specify the joint distribution of forecasts and observations in this case (i.e., the dimensionality of the completely confident dichotomous forecast verification situation is three), the measures are also strongly related, in sometimes complex ways. Improvements in one measure (e.g., POD) generally are associated with degradations in another measure (e.g., POFD, FAR). Thus, it is critical to consider a variety of measures when evaluating these types of forecasts, despite their apparent simplicity. One particularly important dependency is the strong relationship of FAR, CSI, and other measures to the climatological probability, $p(x=1)$ (Brown and Young 2000; Mason, 1989). This relationship makes it inappropriate to compare forecasts for situations with different climatological probabilities, and also limits use of these measures for certain types of observations (Brown and Young 2000).

The choice of verification measures seems complicated even for a verification problem as simple as this one, particularly in light of the relationships among the measures. Additionally, it is well known that improvements in any one of the measures do not necessarily lead to increases in the value of the forecasts to users (e.g., Murphy 1993). Fortunately, it is possible to determine combinations of measures that can indicate superiority of one set of forecasts over another, by applying the statistical concept of *sufficiency* (Ehrendorfer and Murphy, 1988). Superiority means that one set of forecasts would be preferred by all users of the forecasts. In the case of dichotomous, completely confident forecasts, certain pairs of statistics can be shown to meet this criterion; for forecasts and events with higher dimensions, the situation is much more complex, and it

Table 3: Some verification measures for dichotomous completely confident probability forecasts. n_{ab} represents the counts from the verification contingency table, where a is the index for the forecast and b is the index for the observation.

Statistic	Definition	Description
POD	$\frac{n_{11}}{n_{11} + n_{01}}$	• Probability of detection of "Yes" observations • Estimate of $p(f=1\|x=1)$ • Proportion of "Yes" observations that were correctly forecasted • Also called the Hit Rate (HR) in SDT
POFD	$\frac{n_{00}}{n_{10} + n_{00}}$	• Probability of False Detection = Probability of Detection of "No" observations • Estimate of $p(f=0\|x=0)$ • Proportion of "No" observations that were correctly forecasted • 1-POFD = "False Alarm Rate" in SDT
FAR	$\frac{n_{10}}{n_{10} + n_{11}}$	• False Alarm Ratio • Estimate of $p(x=0\|f=1)$ • Proportion of "Yes" forecasts that were incorrect
CSI	$\frac{n_{11}}{n_{11} + n_{10} + n_{01}}$	• Critical Success Index • Also known as "threat score" • Number of "hits" relative to number of Yes forecasts and observations
H-K	POD + POFD - 1	• Hanssen-Kuiper's Discrimination statistic • Also known as Pierce score or True Skill Statistic
Frequency Bias	$\frac{n_{11} + n_{10}}{n_{11} + n_{01}}$	• A measure of over- or under-forecasting • Estimate of $\frac{p(f = 1)}{p(x = 1)}$

is difficult to find combinations that satisfy the requirements for this designation. Examples of pairs of sufficient statistics for the dichotomous completely confident case include (POD, FAR), and (POD, POFD). POD and POFD also are components of the Hanssen-Kuipers discriminant statistic, which has been shown to be an equitable score (i.e., it does not reward over- or under-forecasting; Gandin and Murphy 1992).

4. Explicit probability forecasts

In this section, standard verification approaches and measures are considered for forecasts in which the forecast uncertainty is stated explicitly. As before, dichotomous events (e.g., precipitation occurrence or non-occurrence) are of interest. Many of the verification measures and approaches are demonstrated using a set of experimental probability forecasts described in Brown et al. (1999). These forecasts are subjective outlooks (12-18 hour) and short-term (1-8 hour) forecasts of in-flight icing conditions in the vicinity of several large U.S. cities. While they are not precipitation forecasts, they represent a level of quality that is similar to the quality of short-term precipitation forecasts. The forecasts were formulated by two forecasters who were experienced in forecasting the existence of icing conditions.The forecasters were allowed to use

a limited set of 13 forecast probabilities, including 0, 0.02, 0.10, 0.20, ..., 0.80, 0.90, 0.98, and 1. They were discouraged from using the extreme values, 0 and 1.

4.1 Summary measures

As described in Section 2, the characteristics of the joint distribution of forecasts and observations can be summarized using standard measures such as the mean and standard deviation. Some of these summary measures for the example forecasts are shown in Table 4. The mean forecast values ($\bar{f}$) indicate that slightly larger probabilities were used - on average - for the outlooks than for the short-term forecasts. The mean observation, $\bar{x}$ [= $p(x=1)$], represents the frequency that the event (icing conditions, in this case) was observed to occur. This frequency is equivalent to the mean forecast value for the outlooks, and is notably smaller than the mean forecast for the short-term forecasts. In addition, there is a slight difference between the mean observations for the outlooks and short-term forecasts, due to differences in the forecast periods and days included for the two types of forecasts. The long-term climatological frequencies are 0.41 for the outlooks and 0.37 for the short-term forecasts. Thus, the outlook event occurred somewhat more frequently than normal.

Table 4: Summary statistics for sample forecasts. Mean values are represented by the overbar; s is the standard deviation associated with different distributions.

Forecast type	$\bar{f}$	s_f	$\bar{x}$	s_x	$\bar{f}$ given x=0	$\bar{f}$ given x=1
Outlook	0.52	0.18	0.52	0.50	0.45	0.59
Short-term	0.46	0.25	0.35	0.48	0.37	0.66

The standard deviations characterize the variability of the forecasts and observations. The standard deviation of the forecasts is computed using the usual formula for the variance (see Wilks 1995); the standard deviation is the square-root of the variance. For dichotomous events, the variance reduces to a simple formula,

$$\mathrm{Var}(x) = \mu_x[1-\mu_x], \tag{5}$$

which can be estimated by

$$s_x^2 = \bar{x}(1-\bar{x}) = p(x)[1-p(x)]. \tag{6}$$

The standard deviation of the observations is simply the square-root of s_x^2. For the example (Table 4), the standard deviations of the forecasts are much smaller than the standard deviations of the observations. This result is a common characteristic of probability forecasts. The larger standard deviation for the short-range forecasts indicates that these forecasts are somewhat sharper than the forecasts for the outlooks.

Table 4 also includes summary information for the conditional distributions of f given x, in the form of means. The mean forecast value when the event did not occur is smaller than the

mean forecast value when the event occurred, for both the outlooks and short-term forecasts. This result is desirable, and indicates that the forecasts have at least some ability to discriminate between the occurrence and non-occurrence of the event. Conditional standard deviations also could be examined, but are not included here.

4.2 Performance measures

Several performance measures are of interest, including the Bias, a measure of discrimination, measures of accuracy, and measures of skill. *Bias* is simply defined as

$$\text{Bias} = \mu_f - \mu_x \tag{7}$$

and is computed as

$$\text{Bias} = \overline{f} - \overline{x}. \tag{8}$$

Bias is the *average error* in the forecasts. A negative value indicates the probabilities were too small on average; a positive value indicates that they were too large.

One measure of *discrimination* is

$$\text{DIS} = \mu_{(f|x=1)} - \mu_{(f|x=0)}, \tag{9}$$

which is estimated using the conditional mean forecast values for each subsample (i.e., based on the occurrence or non-occurrence of the event). Conceptually, discrimination is consistent with the idea that the distribution of forecasts when the event occurs should be different from the forecast distribution when it doesn't occur. DIS measures this difference in terms of the mean forecast values for each outcome.

The most common measure of *accuracy* for probabilistic forecasts is the Brier score (Brier 1950). This performance measure is defined in practice as

$$\text{BS} = \frac{1}{n}\sum_{k=1}^{n}(f_k - x_k)^2, \tag{10}$$

where n is the number of forecasts. BS is analogous to the mean-squared error (MSE), commonly used in verification of forecasts of continuous predictands such as temperature. Like the MSE, the best value of BS is 0. The Brier score is most commonly converted to a skill score, using the standard formula for a skill score:

$$\text{BSS} = \frac{\text{BS} - \text{BS}_{\text{ref}}}{0 - \text{BS}_{\text{ref}}}. \tag{11}$$

This skill score can range in value between 0 and 1, and is analogous to a skill score based on the MSE. Commonly, the reference forecast is the sample climatology [i.e., the value of $p(x=1)$

from the sample forecasts], which can be a relatively difficult standard of comparison to beat. Another common standard of comparison is the long-term climatology.

The Brier score can be decomposed into several components that are relevant for interpretation of the sources of errors in the forecasts (Murphy 1973):

$$\mathrm{BS} = \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i(f_i - \bar{x}_i)^2}_{\mathrm{REL}} - \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i(\bar{x}_i - \bar{x})^2}_{\mathrm{RES}} + \underbrace{\bar{x}(1-\bar{x})}_{\mathrm{UNC}}. \tag{12}$$

For this decomposition, it is assumed that there is a discrete number of forecast possibilities, I, and the forecasts and observations have been sorted by the forecast value. Each of the terms in (12) can be interpreted in the context of attributes of forecast quality. The first term on the right-hand-side is a measure of reliability (REL): it measures the difference between the forecast and the mean observation associated with that forecast value, over all of the forecasts. Recall that the conditional mean observation also is the relative frequency of occurrence of the event, and thus ideally also will be close in value to f_i when f_i is forecast. The second term is a measure of resolution (RES): it measures the differences between the average observations for particular forecast values and the overall average observation; ideally, the observations for different forecasts will differ from the overall average observation. Finally, the third term is the familiar estimate of the variance of the observations [see Eq. (6)], and is called the “uncertainty” (UNC). This component can be thought of as measuring the difficulty of the forecasting situation; note that the largest value of UNC is associated with $\bar{x} = 0.5$. Since smaller values of the Brier score are desirable, small values of REL and UNC are best, along with large values of RES. Using this decomposition of the Brier score, the Brier skill score simplifies to

$$\mathrm{BSS} = \frac{\mathrm{RES} - \mathrm{REL}}{\mathrm{UNC}}, \tag{13}$$

assuming that the standard of comparison is the sample climatology.

Some other measures of accuracy are also available. One common measure is the correlation coefficient. The correlation coefficient measures association between the forecasts and observations, but ignores biases in the forecasts, and thus might be best thought of as a measure of potential skill (Murphy 1995). Murphy and Daan (1985) identify some additional performance measures for probabilistic forecasts that are less commonly used.

Table 5 shows some performance measures for the two sets of sample forecasts. The Bias values indicate the outlooks were completely unbiased, but the short-term forecasts were over-forecast, on average, by 0.11. Discrimination is moderate for the outlooks and much larger for the short-term forecasts. Decompositions of the Brier scores indicate that the REL component contributed relatively little to both scores, while RES was also fairly small and UNC was quite large, indicating these were fairly difficult forecasting situations. The BSS values indicate that the forecasts have positive skill, and that the skill is greater for the short-term forecasts. The correlation coefficients are fairly large, with the larger value for the short-term forecasts indicating again that these forecasts are somewhat more skilful than the outlooks.

Table 5: Performance statistics for sample forecasts. r is the sample correlation coefficient between the forecasts and observations.

Forecast type	Bias	DIS	BS	REL	RES	UNC	BSS	r
Outlook	0.01	0.14	0.21	0.01	0.05	0.25	0.17	0.41
Short-term	0.11	0.29	0.17	0.02	0.07	0.23	0.24	0.54

4.3 Graphical approaches

A number of characteristics of the forecasts can be represented very effectively using graphical techniques. One useful diagram is the forecast histogram, which illustrates the sharpness attribute. Figure 1 shows histograms of the forecast probabilities for the outlooks and short-term forecasts. These diagrams indicate that the outlook forecast probabilities most commonly were close to the climatological frequency of the event (i.e., around 0.50), whereas the distribution for the short-term forecasts exhibits greater spread in the values used. In fact, both large and small values were used for the short-term forecasts. Thus, the short-term forecasts are sharper than the outlooks.

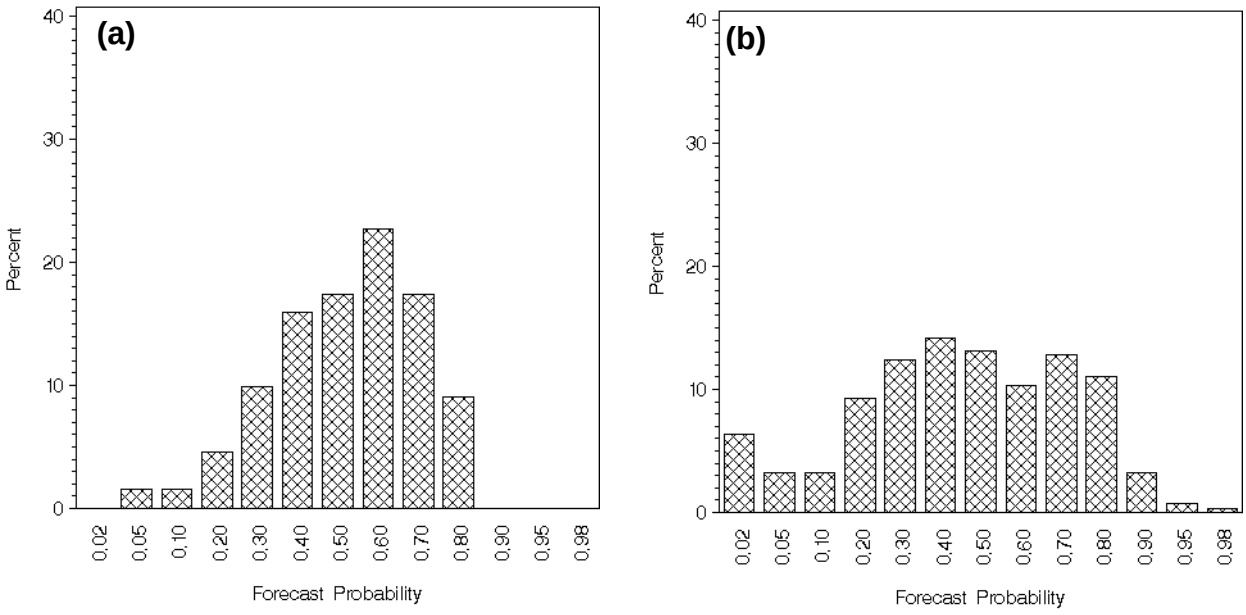

Figure 1: Histograms of forecast probabilities for sample forecasts, for (a) outlooks and (b) short-term forecasts.

The ability of the forecasts to discriminate between events and non-events can be illustrated in a discrimination diagram, which shows the conditional distributions of the forecasts given the observations [i.e., $p(f|x=0)$ and $p(f|x=1)$]. Examples of discrimination diagrams for the sample forecasts are shown in Figure 2. These diagrams indicate that the outlooks have some ability to discriminate between events and non-events, and this discrimination is somewhat better for the short-term forecasts. In particular, the distributions for the outlooks are somewhat separated, although they overlap to a large extent. Even greater separation is apparent for the short-

term forecasts, which suggests that the forecasters (not surprisingly) were better able to discriminate between icing and no-icing conditions on the shorter time scale.

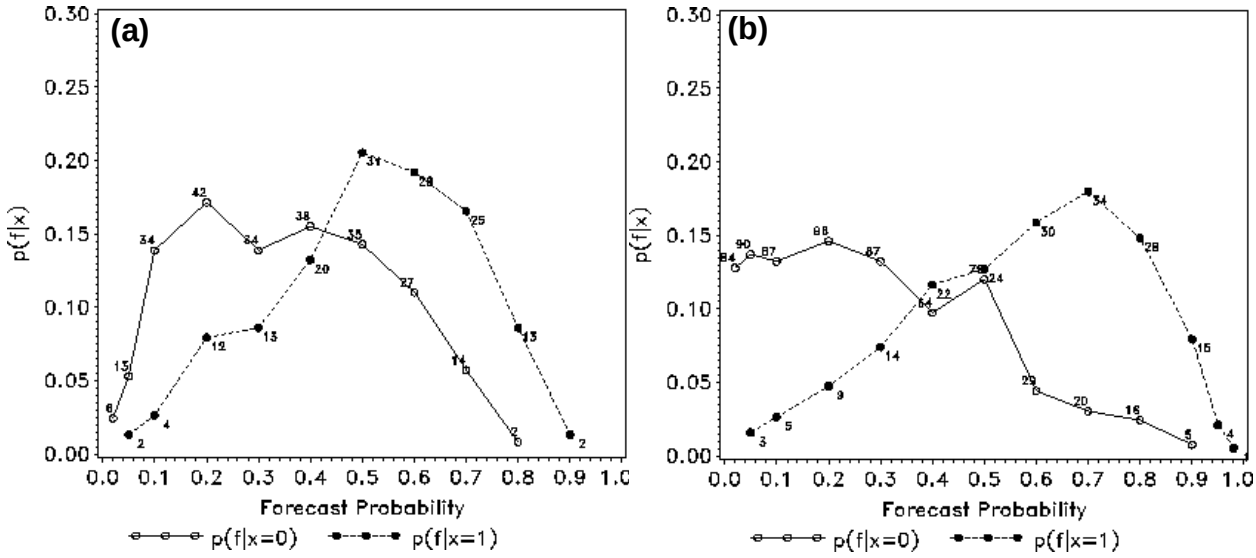

Figure 2: Discrimination diagrams for sample forecasts, for (a) outlooks and (b) short-term forecasts.

Reliability diagrams provide a graphical view of how well the conditional probabilities, $p(x|f)$, match the forecast probability, f. Reliability diagrams for the sample forecasts are shown in Figure 3. The diagram for the outlooks indicates that these forecasts were quite reliability. In particular, the points lie close to the diagonal line indicating 1:1 correspondence between the forecasts and the conditional observation probabilities. In contrast, the diagram for the short-term forecasts indicates that these forecasts generally were too large. For example, over all occasions when the forecast was 0.5, the event actually only occurred about 25% of the time. These results are consistent with the overall bias values shown in Table 5, which were close to 0 for the outlooks and fairly large for the short-term forecasts.

Another type of diagram, the attributes diagram, includes a variety of informations about the quality of the forecasts (Hsu and Murphy 1986). The basis for an attributes diagram is a reliability diagram. In addition, the average forecast and observation are shown, along with information about resolution and identification of regions where the forecasts are contributing positively toward forecast skill. Wilks (1995) includes a good example of an attributes diagram.

It is clear that verification of probability forecasts requires a great deal of analysis and interpretation of many numbers. It would be desirable to somehow condense those numbers into a more manageable set of information. One step in this direction has been taken by Murphy and Wilks (1998). They propose using regression models and statistical distributions to summarize characteristics of the forecasts and their quality. In particular, they illustrate the use of a linear regression model to represent the reliability diagram and a beta distribution to model the forecast distribution. Examining the “raw” reliability and sharpness diagrams requires consideration of many numbers - say, 13 for the distribution of forecasts, and 13 for the reliability diagram. In contrast, using the modeling approach only requires four parameters to describe both distributions.

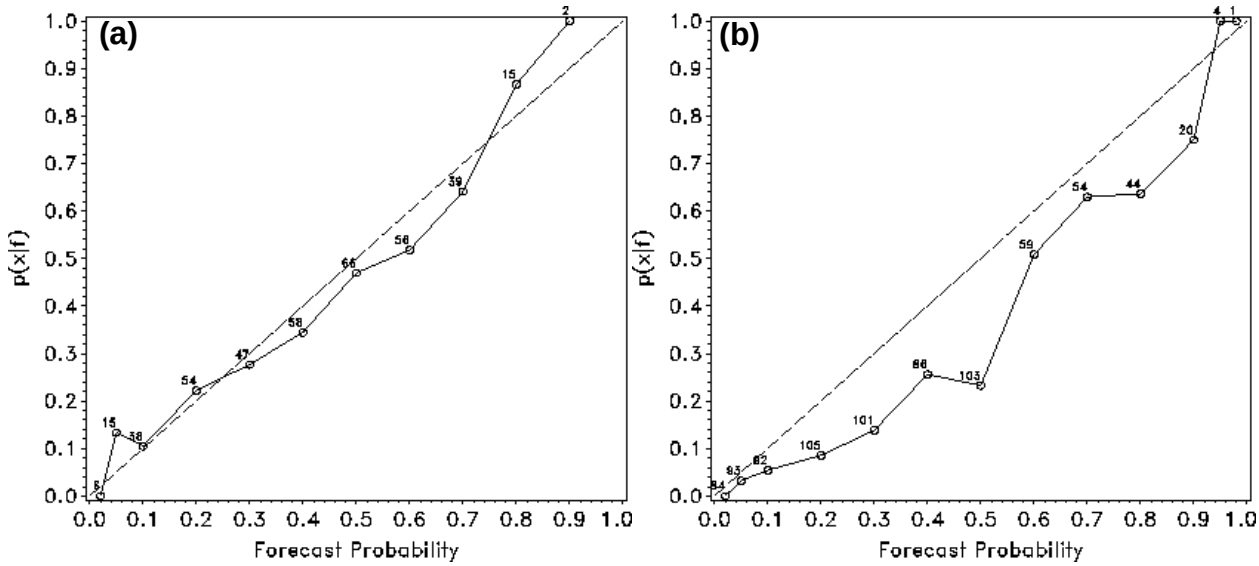

Figure 3: Reliability diagrams for sample forecasts, for (a) outlooks and (b) short-term forecasts.

4.4 Signal detection theory

Signal detection theory (SDT) is becoming a well-accepted approach for evaluation of probabilistic forecasts, following its introduction to the meteorological community in the early 1980s by Ian Mason (1982). This approach essentially focuses on the LB factorization of the joint distribution of forecasts and observations, and thus on the ability of the forecasts to discriminate between events and non-events. SDT's premise is to provide information to users to help them identify appropriate probability thresholds for use in making decisions.

The basis for verification using SDT is the Relative Operating Characteristic (ROC) curve. This curve is formed by setting a series of thresholds to transform the probability forecasts into Yes/No forecasts - a Yes forecast if the forecast probability exceeds the threshold; a No forecast if the forecast probability is less than the threshold. For each set of thresholded forecasts, two statistical measures are computed: the Hit rate (HR= POD) and the False Alarm Rate (FAR= 1 - POFD). These measures are then plotted against each other, with FR on the x-axis and HR on the y-axis. The ROC diagrams for the sample forecasts are shown in Figure 4.

Ideally, an ROC curve will lie toward the upper left corner of the diagram. The diagonal line represents no skill, and curves that fall below the line represent forecasts with negative skill. The curves in Figure 4 suggest that the short-term forecasts are more skillful than the outlooks, which is consistent with the results presented in earlier sections. It is of interest to note that the shape of the ROC curve is related to the forecasts' discrimination, as measured by the separation between the two conditional distributions of forecasts, given event occurrence and non-occurrence (e.g., Mason 1982).

The area under the ROC curve is a measure of skill. Thus, an area value of 0.5 represents a forecast with no skill. Sometimes, however, the areas are measured relative to no skill (i.e., 0.5 is subtracted) so that the skill can range from -0.5 to 0.5, with negative skill represented

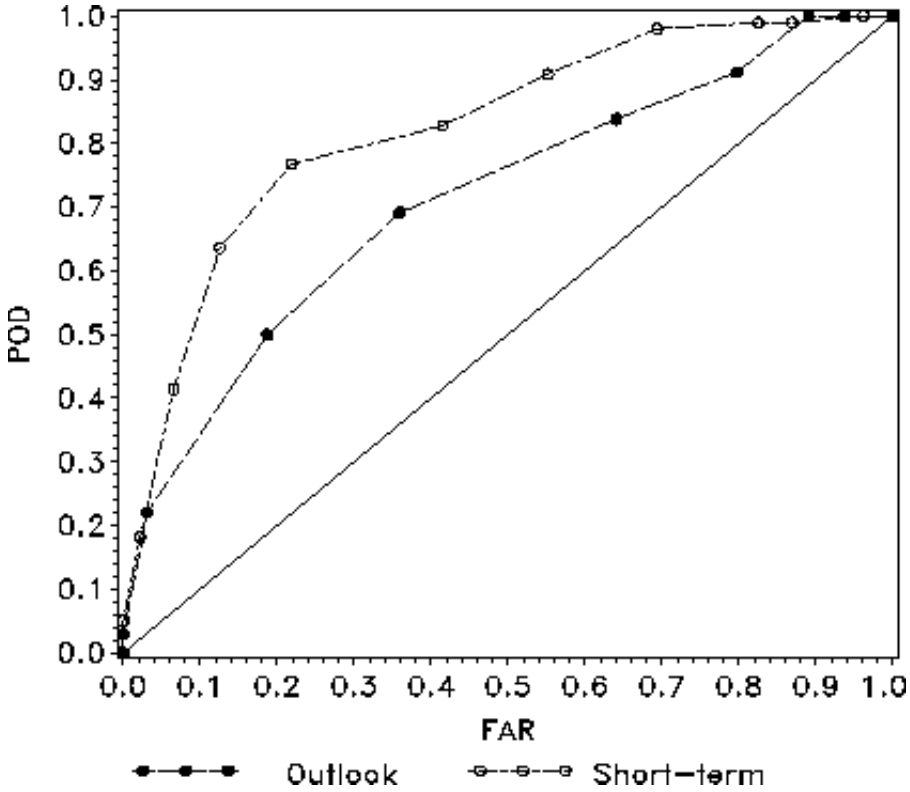

Figure 4: ROC diagrams for sample forecasts.

by a negative value. The ROC area can be estimated using a geometric approach. However, such measurements are very sensitive to the number of points on the curve, and are inappropriate when comparing sets of forecasts with different numbers of thresholds. An alternative approach described by Mason (1982) and Harvey et al. (1992), which is in common use, is to model the conditional probabilities using the normal distribution. This approach removes issues related to the number of thresholds, and provides a relatively simple method for estimating the ROC areas.

One issue with the SDT approach, particularly when it is used in isolation, is that it ignores calibration or reliability. Thus, forecasts that are very unreliable could be very skillful in the SDT context. Some work has been done to provide an analogue to the ROC in the calibration-refinement factorization (Mason and Graham 1999). However, perhaps the best approach is to consider the ROC as one more tool in the arsenal for verification of probability forecasts. In fact, the four graphical approaches that have been focused on so far - the *sharpness*, *discrimination*, *reliability*, and *ROC* diagrams - provide a relatively complete picture of the quality of the forecasts, and the best approach may be to use them all.

4.5 Extensions to multiple categories

Verification of probability forecasts becomes somewhat more complex when the forecasts are for multiple rather than dichotomous categories. This verification problem can be handled in a few different ways. One approach would be to simplify the forecasts into a set of dichotomous

forecasts, by combing categories. This approach simplifies the verification analyses, but also loses or ignores some information associated with the set of probabilities assigned to all the categories. A second approach is to use scores designed to evaluate multi-category events.

The most common measure used to evaluate probability forecasts of multiple categories is the *Ranked Probability Score* (RPS). This measure is analogous to the BS, and has the form

$$\mathrm{RPS} = \frac{1}{J-1}\left[\sum_{m=1}^{J}\left(\sum_{j=1}^{m} P_j - \sum_{j=1}^{m} d_j\right)^2\right], \tag{14}$$

where J is the number of event categories, P_j is the probability assigned to category j, and d_j is the observation for category j. As implied by (14), this computation is made on the basis of forecast and observation values that are accumulated as the categories become more extreme.

Table 6: Artificial example of computation of RPS.

j	m	P_j	d_j	$\sum_{j=1}^{m} P_j$	$\sum_{j=1}^{m} d_j$
1	1	0.20	0	0.20	0
2	2	0.33	1	0.53	1
3	3	0.47	0	1	1

Table 6 shows an artificial example of how the RPS is computed. In this example, there are three forecast categories. The first category has been assigned a probability of 0.20, the second category has been assigned 0.33, and the third category has been assigned a probability of 0.47; thus, the forecast values sum to 1. The accumulated forecast values are shown in the fifth column. The event that occurred was in the second category, as shown in the fourth column, and the accumulated observation values are shown in the sixth column. From the numbers in Table 6, the RPS for this forecast is computed as

$$\mathrm{RPS} = \frac{1}{2}[(0.2-0)^2 + (0.53-1)^2 + (1-1)^2] = 0.13\,. \tag{15}$$

For a set of forecasts, the average RPS can be computed, as shown in Wilks (1995), as

$$\overline{\mathrm{RPS}} = \frac{1}{n}\sum_{k=1}^{n} \mathrm{RPS}_k \tag{16}$$

and the RPS skill score, analogous to the BSS, can be computed as

$$\overline{\mathrm{RPSS}} = \frac{\overline{\mathrm{RPS}_S} - \overline{\mathrm{RPS}_F}}{\overline{\mathrm{RPS}_S}}, \tag{17}$$

where $\overline{\mathrm{RPS}_S}$ is the mean RPS associated with the standard of comparison and $\overline{\mathrm{RPS}_F}$ is the mean RPS associated with the forecasts. It also is worth noting that the RPS can be decomposed into components that are analogous to the decomposition of the Brier score.

Recent extensions to verification approaches for multi-category probability forecasts include a version of the RPS that is appropriate for continuous forecasts (i.e., without discrete categories), along with the appropriate decompositions (e.g., Hersbach 2000). Hamill (1997) also proposed a multi-category calibration diagram, which takes advantage of the accumulation of forecast and observed values across categories, to reduce the impacts of small numbers of event occurrences in extreme categories (Hamill 1997).

5. Extensions

5.1 Ensemble forecast verification

Recently, verification of ensemble forecasts has become a major area of interest within both the modeling and verification communities. In general, the methods used to verify ensemble forecasts are the same as those used to verify other types of probability forecasts, along with a few additional approaches. For example, both the Brier score and the ROC are commonly used to verify ensemble forecasts that have been reduced to probability forecasts of a particular event.

One approach that is unique to ensemble forecasts is the use of the rank histogram (also commonly known as the “Talagrand” diagram). This diagram shows, over a large set of forecasting occasions, where in the ranking of forecasts the actual observation “landed.” Ideally (e.g., if the ensembles are functioning appropriately, and they represent the distribution of possible outcomes), the distribution of observation ranks should be “flat.” That is, no rank should be more common than another. A non-uniform rank distribution might have a peak at the extremes, which would indicate that the observations are commonly more extreme than any of the ensemble members. However, Hamill (2000) has shown that non-uniform ranks also can occur for a variety of other reasons.

A number of other approaches for verifying ensemble forecasts have recently been developed, or are being developed. For example, Wilson et al. (1999) outlines a method in which a distribution is fit to the ensemble, and the probability associated with the observed value is computed using the fitted distribution function. This approach eliminates some of the noise associated with using the observed distributions, and provides a measure of how far the extreme observations are from the main part of the distribution. As ensemble forecasts become increasingly a central approach to forecasting, the development of appropriate verification methods will also progress.

5.2 Connections to value

It is well known that measuring the quality of forecasts is not equivalent to measuring their value, although these two aspects of forecast goodness are intimately related (Murphy 1993). Unfortunately, that relationship commonly takes very complex forms (Katz and Murphy 1997).

Recently, however, measures of forecast value have been developed that are associated with verification measures for probability forecasts. These approaches have relied on a simple decision-making model known as the Cost-Loss model. This model is illustrated in Table 7.

Table 7: The cost-loss ratio decision-making model.

		Adverse weather?	
		Yes	**No**
Protect?	Yes	C	C
	No	L	0

In the cost-loss decision-making situation, the decision maker has the option to implement some protection against the effects of adverse weather. It is assumed that the cost of protection is C and that there will be no loss if the decision-maker chooses to protect. However, if he/she chooses to not protect and adverse weather occurs, the decision-maker will suffer a loss L. It turns out that the ratio C/L is a critical factor in determining optimal actions in this decision-making problem; hence the name, “cost-loss ratio situation.”

Wilks (2001) and Richardson (2000) have both developed approaches to evaluate the impacts of forecasts in this situation. In both cases, equations for the value of forecasts were developed that take into account measures of the quality of the forecasts, to provide a measure of forecast value as a function of C/L. The main difference between the approaches is that the forecasts are not penalized for lack of calibration in the Richardson approach since this approach is based on the ROC. Both approaches measure value according to the decrease in expected expense associated with the forecasts, over the expected expense associated with climatological forecasts.

For illustration, Figure 5 shows Wilks’ value score (VS) for the sample icing forecasts. This figure suggests that both the outlooks and short-term forecasts have potential value to decision-makers. For the short-term forecasts, the greatest potential value would be for users with a cost-loss ratio of about 0.3, and for the outlooks, the maximum value would accrue to users with a cost-loss ratio near 0.5 (it is not a coincidence that these are approximately equal to the sample climatological values). Although the cost-loss ratio situation represents a relatively simple decision-making problem, and a number of other factors have been ignored (e.g., impacts of multiple users), these approaches show promise for extending verification into true evaluation of the forecasts.

6. Summary, issues, and conclusions

This paper has shown the wide variety of measures that are available for verification of probabilistic forecasts, and the importance of making use of these approaches to gain a broad understanding of the quality of the forecasts. In some respects, verification of probability forecasts is actually less complex than verification of some non-probabilistic forecasts, due to the simplifications afforded in the case of dichotomous events. The basis for verification of probability forecasts was established 30-50 years ago by thoughtful scientists such as Glen Brier, Ed Epstein, and Allan Murphy. Progress was stimulated by the development of the framework for verification in 1987, and has been further stimulated recently by the development and wide application of ensemble forecasts.

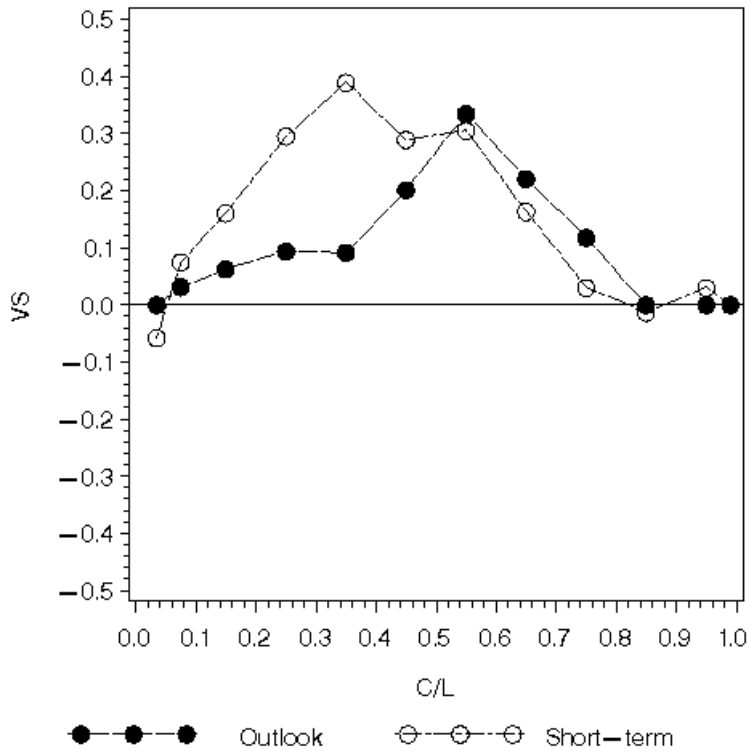

Figure 5: Value Score (VS) for sample forecasts as a function of cost-loss ratio, C/L.

Some issues still remain, including (a) development of improved approaches for verifying multi-category forecasts; (b) development of appropriate methods for evaluating the uncertainty in the verification measures themselves; and (c) developing improved ways to represent and interpret observations, especially in the case of QPF where the observations are extremely variable and subject to a variety of interpretations. With regard to (a), it appears that multi-category forecasts will become more and more important, especially as ensemble forecasts become more common. Current methods are somewhat difficult to apply and to interpret. With regard to (b), measuring the uncertainty in verification measures has long been ignored, due to some of the characteristics of the forecasts and observations (e.g., lack of independence, non-normal distribution). Fortunately, some effort has been put forward in this area in recent years, using standard approaches to estimate confidence intervals (Seaman et al. 1996) and applying modern re-sampling approaches (e.g., Kane and Brown 2000). Finally, with respect to the third issue, it should be possible to make use of information about the uncertainty associated with the observations in the process of forecast verification. However, research is needed in this area, both to determine how to quantify the uncertainty, and to determine how to use that information in verification analyses.

References

Brier, G.W., 1950: Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review*, **78**, 1-3.

Brown, B.G., B.C. Bernstein, F. McDonough, and T.A.O. Bernstein, 1999: Probability forecasts of in-flight icing conditions. *Preprints, 8th Conference on Aviation, Range, and Aerospace Meteorology*, Dallas, 10-15 January. American Meteorological Society (Boston), 433-437.

Brown, B.G., and G.S. Young, 2000: Verification of icing and turbulence forecasts: Why some verification statistics can't be computed using PIREPs. *Preprints, 9th Conference on Aviation, Range, and Aerospace Meteorology*, 11-15 September, Orlando, FL, American Meteorological Society (Boston), 393-398.

Ehrendorfer, M., and A.H. Murphy, 1988: Comparative evaluation of weather forecasting systems: Sufficiency, quality, and accuracy. *Monthly Weather Review*, **116**, 1757-1770.

Gandin, L.S., and A.H. Murphy 1992: Equitable skill scores for categorical forecasts. *Monthly Weather Review*, **120**, 361-370.

Hamill, T.M., 1997: Reliability diagrams for multicategory probabilistic forecasts. *Weather and Forecasting*, **12**, 736-741.

Hamill, T.M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, **129**, 550-560.

Harvey, L.O., Jr., K.R. Hammond, C.M. Lusk, and E.F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Monthly Weather Review*, **120**, 863-883.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15**, 559-570.

Hsu, W.-R., and A.H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, **2**, 285-293.

Kane, T.L., and B.G. Brown, 2000: Confidence intervals for some verification measures - a survey of several methods. *Preprints, 15th Conference on Probability and Statistics in the Atmospheric Sciences*, 8-11 May, Asheville, NC, U.S.A., American Meteorological Society (Boston), 46-49.

Katz, R.W., and A.H. Murphy, Editors, 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press.

Mason, I., 1982: A model for assessment of weather forecasts. *Australian Meteorological Magazine*, **30**, 291-303.

Mason, I., 1989: Dependence of the critical success index on sample climate and threshold probability. *Australian Meteorological Magazine*, **37**, 75-81.

Mason, S., and N.E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather and Forecasting*, **14**, 713-725.

Murphy, A.H., 1973: A new vector partition of the probability score. *Journal of Applied Meteorology*, **12**, 595-600.

Murphy, A.H., 1993: What Is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281-293.

Murphy, A.H., 1995: The coefficients of correlation and determination as measures of performance in forecast verification. Weather and Forecasting, **10**, 681-688.

Murphy, A.H., 1997: Forecast verification. *Economic value of Weather and Climate Forecasts*, R.W. Katz and A.H. Murphy, editors. Cambridge University Press, 19-74.

Murphy, A.H., and H. Daan, 1985: Forecast evaluation. In *Probability, Statistics, and Decision-making in the Atmospheric Sciences*, A.H. Murphy and R.W. Katz, editors. Westview Press, Boulder, CO, 379-437.

Murphy, A.H., and D.S. Wilks, 1998: A case study of the use of statistical models in forecast verification: Precipitation probability forecasts. *Weather and Forecasting*, **13**, 795-810.

Murphy, A.H., and R.L. Winkler, 1987: A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330-1338.

Murphy, A.H., and R.L. Winkler, 1992: Diagnostic verification of probability forecasts. *International Journal of Forecasting*, **7**, 435-455.

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **126**, 649-667.

Seaman, R., I. Mason, and F. Woodcock, 1996: Confidence intervals for some performance measures of Yes-No forecasts. *Australian Meteorological Magazine*, **45**, 49-53.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences*, Academic Press, San Diego, CA, 467 pp.

Wilks, D.S., 2001: A skill score based on economic value for probability forecasts. *Meteorological Applications*, **8**, 209-219.

Wilson, L.J., 2002: Verification of precipitation forecasts: A survey of methodology. Part I: General Framework and verification of continuous variables. This volume.

Wilson, L.J., W.R. Burrows, and A. Lanzinger, 1999: A strategy for verification of weather element forecasts from an ensemble prediction system. *Monthly Weather Review*, **127**, 956-970.