Chapter 9

Forecast Verification

9.1. BACKGROUND

9.1.1. Purposes of Forecast Verification

Forecast verification is the process of assessing the quality of forecasts. This process perhaps has been most fully developed in the atmospheric sciences, although parallel developments have taken place within other disciplines as well, such as finance and economics (e.g., Armstrong, 2001; Clark and McCracken, 2013), medical diagnosis (e.g., Pepe, 2003), and statistics (e.g., Gneiting, 2011a; Gneiting and Raftery, 2007; Schervish, 1989; Winkler, 1996). In the literature of those disciplines the activity is more typically called validation, or assessment, or evaluation. Verification of weather forecasts has been undertaken since at least 1884 (Muller, 1944; Murphy, 1996), and use of the term "verification" to mean evaluation of forecasts appears to have originated with the seminal paper of Finley (1884). In addition to this chapter, other reviews of forecast verification can be found in Jolliffe and Stephenson (2012a), Murphy (1997), and Stanski et al. (1989). Open-source software resources are also available (Pocernich, 2007).

Perhaps not surprisingly, there can be differing views of what constitutes a good forecast (Murphy, 1993). Many forecast verification procedures exist, but all involve statistics characterizing the relationship between a forecast or set of forecasts, and the corresponding observation(s) of the predictand. Any forecast verification method thus necessarily involves comparisons between matched pairs of forecasts and the observations to which they pertain.

On a fundamental level, forecast verification involves investigation of the properties of the *joint distribution of forecasts and observations* (Murphy and Winkler, 1987). That is, any given verification data set consists of a collection of forecast/observation pairs whose joint behavior can be characterized in terms of the relative frequencies of the possible combinations of forecast/observation pairs. A parametric joint distribution such as the bivariate normal (Section 4.4.2) can sometimes be useful in representing this joint distribution for a particular data set, but the empirical joint distribution of these quantities (more in the spirit of Chapter 3) usually forms the basis of forecast verification measures. Ideally, the association between forecasts and the observations to which they pertain will be reasonably strong, but in any case the nature and strength of this association will be reflected in their joint distribution.

Objective evaluations of forecast quality are undertaken for a variety of reasons. Brier and Allen (1951) categorized these as serving administrative, scientific, and economic purposes. In this view, administrative use of forecast verification pertains to ongoing monitoring of operational forecasts. For example, it is often of interest to examine trends of forecast performance through time (e.g., Stern and Davidson, 2015). Rates of forecast improvement, if any, for different predictands, locations, or lead times can be compared. Verification of forecasts from different sources for the same events can also be compared. Here forecast verification techniques allow comparison of the relative merits of competing forecasters or forecasting systems. For example, this is the purpose to which forecast verification is often put in scoring student forecast contests at colleges and universities.

Statistical Methods in the Atmospheric Sciences. https://doi.org/10.1016/B978-0-12-815823-4.00009-2

Analysis of verification statistics and their components can also help in the assessment of specific strengths and weaknesses of forecasters or forecasting systems. Although classified by Brier and Allen as scientific, this application of forecast verification is perhaps better regarded as *diagnostic verification* (Murphy et al., 1989; Murphy and Winkler, 1992). Here specific attributes of the relationship between forecasts and the subsequent events are investigated, highlighting strengths and deficiencies in a set of forecasts. Diagnostic verification allows human forecasters to be given feedback on the performance of their forecasts in different situations, which hopefully will lead to better forecasts in the future. Similarly, forecast verification measures can point to problems in forecasts produced by objective means, possibly leading to better forecasts through methodological improvements.

Ultimately, the justification for any forecasting enterprise is that it supports better decision making. The usefulness of forecasts to support decision making clearly depends on their error characteristics, which are elucidated through forecast verification methods. Thus the economic motivations for forecast verification are to provide the information necessary for users to derive full economic value from forecasts and to enable estimation of that value. However, since the economic value of forecast information in different decision situations must be evaluated on a case-by-case basis (e.g., Katz and Murphy, 1997a), forecast value cannot be computed from the verification statistics alone. Although it is sometimes possible to guarantee the economic superiority of one forecast source over another for all forecast users on the basis of a detailed verification analysis, which is a condition called *sufficiency* (Ehrendorfer and Murphy, 1988; Krzysztofowicz and Long, 1990, 1991; Murphy, 1997; Murphy and Ye, 1990), superiority with respect to a single verification measure does not necessarily imply superior forecast value for all users. Furthermore, actual as opposed to potential forecast value depends on psychosocial factors as well as purely economic ones (Millner, 2008; Stewart, 1997).

9.1.2. The Joint Distribution of Forecasts and Observations

The joint distribution of the forecasts and observations is of fundamental interest with respect to the verification of forecasts. In most practical settings, both the forecasts and observations are discrete variables. That is, even if the forecasts and observations are not already discrete quantities, they are typically rounded operationally to one of a finite set of values. Denote the forecast by y_i, which can take on any of the I values $y_1, y_2, \ldots, y_I$; and the corresponding observation as o_j, which can take on any of the J values $o_1, o_2, \ldots, o_J$. Then the joint distribution of the forecasts and observations is denoted

$$p(y_i, o_j) = \Pr\{y_i, o_j\} = \Pr\{y_i \cap o_j\}; \quad i = 1, \ldots, I; \; j = 1, \ldots, J. \tag{9.1}$$

This is a discrete bivariate probability distribution function, associating a probability with each of the $I \times J$ possible combinations of forecast and observation.

Even in the simplest cases, for which $I = J = 2$, this joint distribution can be difficult to use directly. From the definition of conditional probability (Equation 2.10) the joint distribution can be factored in two ways that are informative about different aspects of the verification problem. From a forecasting standpoint, the more familiar and intuitive of the two is

$$p(y_i, o_j) = p(o_j | y_i)\, p(y_i); \quad i = 1, \ldots, I; \; j = 1, \ldots J; \tag{9.2}$$

which is called the *calibration-refinement factorization* (Murphy and Winkler, 1987). One part of this factorization consists of a set of the I conditional distributions, $p(o_j \mid y_i)$, each of which consists of probabilities for all the J outcomes o_j, given one of the forecasts y_i. That is, each of these conditional

distributions specifies how often each possible weather event occurred on those occasions when the single forecast y_i was issued, or how well each forecast y_i is calibrated. The other part of this factorization is the unconditional (marginal) distribution $p(y_i)$, which specifies the relative frequencies of use of each of the forecast values y_i, or how often each of the I possible forecast values were used. This marginal distribution is called the *refinement distribution* of the forecasts. The refinement of a set of forecasts refers to the dispersion of the distribution $p(y_i)$. A refinement distribution with a large spread implies refined forecasts, in that different forecasts are issued relatively frequently, and so have the potential to discern a broad range of conditions. Conversely, if most of the forecasts y_i are the same or very similar, $p(y_i)$ is narrow, which indicates a lack of refinement. This attribute of forecast refinement often is referred to as *sharpness*, in the sense that refined forecasts are called sharp.

The *likelihood-base rate factorization* (Murphy and Winkler, 1987) is the other factorization of the joint distribution of forecasts and observations,

$$p\left(y_i, o_j\right) = p\left\{y_i \middle| o_j\right\} p\left\{o_j\right\}; \quad i = 1, \ldots, I; \; j = 1, \ldots, J. \tag{9.3}$$

Here the conditional distributions $p(y_i \mid o_j)$ express the likelihoods that each of the allowable forecast values y_i would have been issued in advance of each of the observed weather events o_j. Although this concept may seem logically reversed, it can reveal useful information about the nature of forecast performance. In particular, these conditional distributions relate to how well a set of forecasts is able to discriminate among the events o_j, in the same sense of the word used in Chapter 15. The unconditional distribution $p(o_j)$ consists simply of the relative frequencies of the J weather events o_j in the verification data set. This distribution usually is called the sample climatological distribution or simply the *sample climatology*.

Both the likelihood-base rate factorization (Equation 9.3) and the calibration-refinement factorization (Equation 9.2) can be calculated from the full joint distribution $p(y_i, o_j)$. Conversely, the full joint distribution can be reconstructed from either of the two factorizations. Accordingly, the full information content of the joint distribution $p(y_i, o_j)$ is included in either of the pairs of distributions, Equation 9.2 or Equation 9.3. Forecast verification approaches based on these distributions are sometimes known as *distributions-oriented* (Murphy, 1997) or *diagnostic verification* (Murphy et al., 1989; Murphy, 1991; Murphy and Winkler, 1992) approaches, in distinction to potentially incomplete summaries based on one or a few scalar verification measures, known as *measures-oriented* approaches. The name diagnostic verification is appropriate because portraying the full joint distribution of forecasts and observations allows detailed strengths and weaknesses of a set of forecasts to be diagnosed, potentially providing insights into the optimal use and value of the forecasts (e.g., Epstein, 1966), and into particular avenues for future forecast improvement (Murphy et al., 1989; Murphy and Winkler, 1992).

Although the two factorizations of the joint distribution of forecasts and observations can help organize the verification information conceptually, neither reduces the dimensionality (Murphy, 1991), or degrees of freedom, of the verification problem. That is, since all the probabilities in the joint distribution (Equation 9.1) must add to 1, it is completely specified by any $(I \times J) - 1$ of these probabilities. The factorizations of Equations 9.2 and 9.3 express this information differently and informatively, but $(I \times J) - 1$ distinct probabilities are still required to completely specify each factorization.

9.1.3. Scalar Attributes of Forecast Performance

Even in the simplest case of $I = J = 2$, complete specification of forecast performance requires a $(I \times J) - 1 = 3$-dimensional set of verification quantities. This minimum level of dimensionality is

already sufficient to make understanding and comparison of forecast evaluation statistics less than straightforward. The difficulty is compounded in the many verification situations where $I > 2$ and/or $J > 2$, and such higher-dimensional verification situations may be further complicated if the sample size is not large enough to obtain good estimates for all of the required $(I \times J) - 1$ probabilities. As a consequence, it is traditional to summarize forecast performance using one or several scalar (i.e., one-dimensional) verification measures. Many of the scalar summary statistics have been found through analysis and experience to provide very useful information about forecast performance, but some of the information in the full joint distribution of forecasts and observations is inevitably discarded when the dimensionality of the verification problem is reduced.

The following is a partial list of scalar aspects, or attributes, of forecast quality. These attributes are not uniquely defined, so that each of these concepts may be characterized using different functions of a verification data set.

(1) *Accuracy* refers to the average correspondence between individual forecasts and the events they predict. Scalar measures of accuracy are meant to summarize, in a single number, the overall quality of a set of forecasts. Several of the more common measures of accuracy will be presented in subsequent sections. The remaining forecast attributes in this list can often be interpreted as components, or aspects, of accuracy.

(2) *Bias*, or *unconditional bias*, or systematic bias, measures the correspondence between the average forecast and the average observed value of the predictand. This concept is different from accuracy, which measures the average correspondence between individual pairs of forecasts and observations. Temperature forecasts that are consistently too warm or precipitation forecasts that are consistently too wet both exhibit bias, whether or not the forecasts are otherwise reasonably accurate or quite inaccurate.

(3) *Reliability*, or *calibration*, or *conditional bias*, pertains to the relationship of the forecast to the distribution of observations, for specific values of (i.e., conditional on) the forecast. Reliability statistics sort the forecast/observation pairs into groups according to the value of the forecast variable, and characterize the conditional distributions of the observations given the forecasts. Thus measures of reliability summarize the I conditional distributions $p(o_j \mid y_i)$ of the calibration-refinement factorization (Equation 9.2). This attribute is also referred to as *validity* in some older literature (Bross, 1953; Sanders, 1963).

(4) *Resolution* refers to the degree to which the forecasts sort the observed events into groups that are different from each other. It is related to reliability, in that both are concerned with the properties of the conditional distributions of the observations given the forecasts, $p(o_j \mid y_i)$. Therefore resolution also relates to the calibration-refinement factorization of the joint distribution of forecasts and observations. However, resolution pertains to the differences between the conditional averages of the observations for different values of the forecast, whereas reliability compares the conditional distributions of the observations with the forecast values themselves. If average temperature outcomes following forecasts of, say, 10°C and 20°C are very different, the forecasts can resolve these different temperature outcomes, and are said to exhibit resolution. If the temperature outcomes following forecasts of 10°C and 20°C are nearly the same on average, the forecasts exhibit almost no resolution. Resolution can be regarded as a measure of the information content of forecasts, in the sense of characterizing the reduction in uncertainty about the predictand (e.g., Jolliffe and Stephenson, 2012b).

(5) *Discrimination* is the converse of resolution, in that it pertains to differences between the conditional distributions of the forecasts for different values of the observation. Measures of discrimination

summarize the J conditional distributions of the forecasts given the observations, $p(y_i \mid o_j)$, in the likelihood-base rate factorization (Equation 9.3). The discrimination attribute reflects the ability of the forecasting system to produce different forecasts for those occasions having different realized outcomes of the predictand. If a forecasting system forecasts $y =$ "snow" with equal frequency when $o =$ "snow" and $o =$ "sleet," the two conditional probabilities of a forecast of snow are equal, and the forecasts are not able to discriminate between snow and sleet events.

(6) *Sharpness*, or refinement, is an attribute of the forecasts alone, without regard to their corresponding observations. Measures of sharpness characterize the unconditional distribution (relative frequencies of use) of the forecasts, $p(y_i)$ in the calibration-refinement factorization (Equation 9.2). Forecasts that rarely deviate much from the climatological value of the predictand exhibit low sharpness. In the extreme, forecasts consisting only of the climatological value of the predictand exhibit no sharpness. By contrast, forecasts that are frequently much different from the climatological value of the predictand are sharp. Sharp forecasts exhibit the tendency to "stick their neck out." Sharp forecasts will be accurate only if they also exhibit good reliability, or calibration, and an important goal is to maximize sharpness without sacrificing calibration (Brier, 1950; Bross, 1953; Sanders, 1963; Gneiting et al., 2007; Murphy and Winkler, 1987; Winkler, 1996). Anyone can produce sharp forecasts, but the difficult task is to ensure that these forecasts correspond well to the subsequent observations.

9.1.4. Forecast Skill

Forecast *skill* refers to the relative accuracy of a set of forecasts, with respect to some set of standard *reference forecasts*. Common choices for the reference forecasts are climatological values of the predictand, persistence forecasts (values of the predictand in the previous time period), or random forecasts (with respect to the climatological relative frequencies of the observed events o_j). Yet other choices for the reference forecasts can be more appropriate in some cases. For example, when evaluating the performance of a new forecasting system, it could be appropriate to compute skill relative to the forecasts that this new system might replace.

Forecast skill is usually presented as a *skill score*, which is often interpreted as a percentage improvement over the reference forecasts. In generic form, the skill score for forecasts characterized by a particular measure of accuracy A, with respect to the accuracy A_{ref} of a set of reference forecasts, is given by

$$\text{SS}_{\text{ref}} = \frac{A - A_{\text{ref}}}{A_{\text{perf}} - A_{\text{ref}}} \times 100\%, \tag{9.4}$$

where A_{perf} is the value of the accuracy measure that would be achieved by perfect forecasts. Note that this generic skill score formulation gives consistent results whether the accuracy measure has a positive (larger values of A are better) or negative (smaller values of A are better) orientation. If $A = A_{\text{perf}}$ the skill score attains its maximum value of 100%. If $A = A_{\text{ref}}$ then $\text{SS}_{\text{ref}} = 0\%$, indicating no improvement over the reference forecasts. If the forecasts being evaluated are inferior to the reference forecasts with respect to the accuracy measure A, $\text{SS}_{\text{ref}} < 0\%$. Skill scores are known as U-statistics in the economics literature (Campbell and Diebold, 2005).

The use of skill scores often is motivated by a desire to equalize effects of intrinsically more or less difficult forecasting situations, when comparing forecasters or forecast systems. For example, forecasting precipitation in a very dry climate is generally relatively easy, since forecasts of zero, or the

climatological average (which will be very near zero), will exhibit good accuracy on most days. If the accuracy of the reference forecasts (A_{ref} in Equation 9.4) is relatively high, a higher accuracy A is required to achieve a given skill level than would be the case in a more difficult forecast situation, in which A_{ref} would be smaller. Some of the effects of the intrinsic ease or difficulty of different forecast situations can be equalized through use of skill scores such as Equation 9.4, but unfortunately skill scores have not been found to be fully effective for this purpose (Glahn and Jorgensen, 1970; Winkler, 1994, 1996).

When skill scores are averaged over nonhomogeneous forecast-observation pairs (e.g., for a single location across a substantial fraction of the annual cycle, or for multiple locations with different climates), care must be taken to compute the skill scores consistently, so that credit is not given for correctly "forecasting" mere climatological differences (Hamill and Juras, 2006; Juras, 2000). For example, correctly forecasting summers being warmer than winters is not credited as a contribution to skill. When computing averaged skill scores, each of the three quantities on the right-hand side of Equation 9.4 should be computed separately for each homogeneous subset of the forecast-observation pairs, with the summary average skill calculated as the weighted average of the resulting component skills.

9.2. NONPROBABILISTIC FORECASTS FOR DISCRETE PREDICTANDS

Forecast verification is perhaps easiest to understand in the context of nonprobabilistic forecasts for discrete predictands. Nonprobabilistic indicates that the forecast consists of an unqualified statement that a single outcome will occur. Nonprobabilistic forecasts contain no expression of uncertainty, in distinction to probabilistic forecasts. A discrete predictand is an observable variable that takes on one and only one of a finite set of possible values. This is in distinction to a scalar continuous predictand, which (at least conceptually) may take on any value on the relevant portion of the real line.

Verification for nonprobabilistic forecasts of discrete predictands has been undertaken since the 19th century (Murphy, 1996), and during this considerable time a variety of sometimes conflicting terminology has been used. For example, nonprobabilistic forecasts have been called categorical, in the sense their being firm statements that do not admit the possibility of alternative outcomes. This usage of the term appears to date from early in the 20th century (Liljas and Murphy, 1994). However, more recently the term categorical has come to be understood as relating to a predictand belonging to one of a set of MECE categories, that is, a discrete variable. In an attempt to minimize confusion, the term categorical will be avoided here, in favor of the more explicit terms, nonprobabilistic and discrete. Other instances of the multifarious nature of forecast verification terminology will also be noted in this chapter.

9.2.1. The 2×2 Contingency Table

There is usually a one-to-one correspondence between allowable nonprobabilistic forecast values and values of the discrete observable predictand to which they pertain. In terms of the joint distribution of forecasts and observations (Equation 9.1), $I = J$. The simplest possible situation is for the dichotomous $I = J = 2$ case, or verification of nonprobabilistic yes/no forecasts. Here there are $I = 2$ possible forecasts, either that the event will ($i = 1$, or y_1) or will not ($i = 2$, or y_2) occur. Similarly, there are $J = 2$ possible outcomes: either the event subsequently occurs (o_1) or it does not (o_2). Despite the simplicity of this verification setting, a surprisingly large body of work on the 2×2 verification problem has developed.

Conventionally, nonprobabilistic verification data are displayed in an $I \times J$ contingency table of absolute frequencies, or counts, of the $I \times J$ possible combinations of forecast and event pairs. If these counts are transformed to relative frequencies, by dividing each tabulated entry by the sample size (total number of forecast and event pairs), the (sample estimate of the) joint distribution of forecasts and observations (Equation 9.1) is obtained. Figure 9.1 illustrates the essential equivalence of the contingency table and the joint distribution of forecasts and observations for the simple $I = J = 2$ case. The boldface portion in Figure 9.1a shows the arrangement of the four possible combinations of forecast and event pairs as a square contingency table, and the corresponding portion of Figure 9.1b shows these counts transformed to joint relative frequencies.

In terms of Figure 9.1, the event in question was successfully forecast to occur a times out of n total forecasts. These a forecast-observation pairs usually are called *hits*, and their relative frequency, a/n, is the sample estimate of the corresponding joint probability $p(y_1, o_1)$ in Equation 9.1. Similarly, on b occasions, called *false alarms*, the event was forecast to occur but did not, and the relative frequency b/n estimates the joint probability $p(y_1, o_2)$. There are also c instances of the event of interest occurring despite not being forecast, called *misses*, the relative frequency of which estimates the joint probability $p(y_2, o_1)$; and d instances of the event not occurring after a forecast that it would not occur, sometimes called a *correct rejection* or *correct negative*, the relative frequency of which corresponds to the joint probability $p(y_2, o_2)$.

It is also usual to include what are called *marginal totals* with a contingency table of counts. These are simply the row and column totals yielding, in this case, the numbers of times each yes or no forecast, or observation, respectively, occurred. These are shown in Figure 9.1a in normal typeface, as is the sample size,

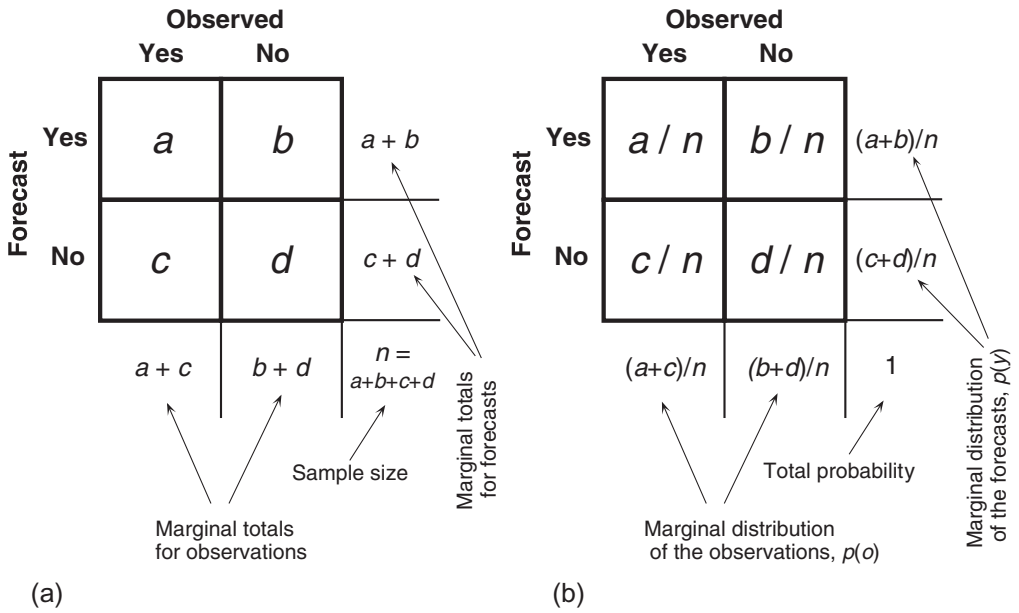

FIGURE 9.1 Relationship between counts (letters *a*–*d*) of forecast and event pairs for the dichotomous nonprobabilistic verification situation as displayed in a 2×2 contingency table (bold, panel a), and the corresponding joint distribution of forecasts and observations [$p(y, o)$] (bold, panel b). Also shown are the marginal totals, indicating how often each of the two events was forecast and observed in absolute terms; and the marginal distributions of the observations [$p(o)$] and forecasts [$p(y)$], which indicates the same information in relative frequency terms.

$n = a+b+c+d$. Expressing the marginal totals in relative frequency terms, again by dividing through by the sample size, yields the *marginal distribution* of the forecasts, $p(y_i)$, and the marginal distribution of the observations, $p(o_j)$. The marginal distribution $p(y_i)$ is the refinement distribution, in the calibration-refinement factorization (Equation 9.2) for the 2×2 joint distribution in Figure 9.1b. Since there are $I = 2$ possible forecasts, there are two calibration distributions $p(o_j \mid y_i)$, each of which consists of $J = 2$ probabilities. Therefore in addition to the refinement distribution $p(y_1) = (a+b)/n$ and $p(y_2) = (c+d)/n$, the calibration-refinement factorization in the 2×2 verification setting consists of the conditional probabilities

$$p(o_1 \mid y_1) = a/(a+b) \tag{9.5a}$$

$$p(o_2 \mid y_1) = b/(a+b) \tag{9.5b}$$

$$p(o_1 \mid y_2) = c/(c+d) \tag{9.5c}$$

and

$$p(o_2 \mid y_2) = d/(c+d). \tag{9.5d}$$

In terms of the definition of conditional probability (Equation 2.10), Equation 9.5a (for example) would be obtained as $[a/n]/[(a+b)/n] = a/(a+b)$.

Similarly, the marginal distribution $p(o_j)$, with elements $p(o_1) = (a+c)/n$ and $p(o_2) = (b+d)/n$, is the base-rate (i.e., sample climatological) distribution in the likelihood-base rate factorization (Equation 9.3). That factorization is completed by the four conditional probabilities

$$p(y_1 \mid o_1) = a/(a+c) \tag{9.6a}$$

$$p(y_2 \mid o_1) = c/(a+c) \tag{9.6b}$$

$$p(y_1 \mid o_2) = b/(b+d) \tag{9.6c}$$

and

$$p(y_2 \mid o_2) = d/(b+d). \tag{9.6d}$$

9.2.2. Scalar Attributes of the 2×2 Contingency Table

Even though the 2×2 contingency table summarizes verification data for the simplest possible forecast setting, its dimensionality is 3. That is, the forecast performance information contained in the contingency table cannot fully be expressed with fewer than three sample statistics. It is perhaps not surprising that a wide variety of these scalar attributes have been devised and used to characterize forecast performance, over the long history of the verification of forecasts of this type. Unfortunately, a similarly wide variety of nomenclature also has appeared in relation to these attributes, both within and outside the atmospheric sciences. This section lists scalar attributes of the 2×2 contingency table that have been most widely used, together with much of the synonymy associated with them. The organization follows the general classification of attributes in Section 9.1.3.

Accuracy

Accuracy statistics reflect correspondence between pairs of forecasts and the events they are meant to predict. Perfectly accurate forecasts in the 2×2 nonprobabilistic forecasting situation will clearly exhibit

all "yes" forecasts for the event followed by the event and all "no" forecasts for the event followed by nonoccurrence, so that $b = c = 0$. For real, imperfect forecasts, accuracy measures characterize degrees of this correspondence. Several scalar accuracy measures are in common use, with each reflecting somewhat different aspects of the underlying joint distribution.

Perhaps the most direct and intuitive measure of the accuracy of nonprobabilistic forecasts for discrete events is the *proportion correct* proposed by Finley (1884). This is simply the fraction of the n forecast occasions for which the nonprobabilistic forecast correctly anticipated the subsequent event or nonevent. In terms of the counts Figure 9.1a, the proportion correct is given by

$$PC = \frac{a+d}{n}. \tag{9.7}$$

The proportion correct satisfies the principle of *equivalence of events*, since it credits correct "yes" and "no" forecasts equally. As Example 9.1 will show, however, this is not always a desirable attribute, particularly when the "yes" event is rare, so that correct "no" forecasts can be made fairly easily. The proportion correct also penalizes both kinds of errors (false alarms and misses) equally. The worst possible proportion correct is zero. The best possible proportion correct is one. Sometimes PC in Equation 9.7 is multiplied by 100%, and referred to as the *percent correct*, or percentage of forecasts correct. Because the proportion correct does not distinguish between correct forecasts of the event, a, and correct forecasts of the nonevent, d, this fraction of correct forecasts has also been called the hit rate. However, in current usage the term hit rate usually is reserved for the discrimination measure given in Equation 9.12.

The *threat score* (TS) or *critical success index* (CSI) is an alternative to the proportion correct that is particularly useful when the event to be forecast (as the "yes" event) occurs substantially less frequently than the nonoccurrence (the "no" event). In terms of Figure 9.1a, the threat score is computed as

$$\mathrm{TS} = \mathrm{CSI} = \frac{a}{a+b+c}, \tag{9.8}$$

which is the number of correct "yes" forecasts divided by the total number of occasions on which that event was forecast and/or observed. It can be viewed as a proportion correct for the quantity being forecast, after removing correct "no" forecasts from consideration. The worst possible threat score is zero, and the best possible threat score is one. When originally proposed (Gilbert, 1884) it was called the *ratio of verification*, and denoted as V, and so Equation 9.8 is sometimes called the Gilbert Score (as distinct from the Gilbert Skill Score, Equation 9.20). In the ecology literature TS is known as the Jaccard coefficient (Janson and Vegelius, 1981). Very often each of the counts in a 2×2 contingency table pertains to a different forecasting occasion (as illustrated in Example 9.1), but the threat score (and the skill score based on it, Equation 9.20) may also be used to assess simultaneously issued spatial forecasts, for example, heavy precipitation warnings (e.g., Ebert and McBride, 2000; Stensrud and Wandishin, 2000). In this setting, a represents the intersection of the areas over which the event was forecast and subsequently occurred, b represents the area over which the event was forecast but failed to occur, and c is the area over which the event occurred but was not forecast to occur. The threat score is convenient in this setting because often the relevant "no" forecast area is arbitrary or unknown.

A third approach to characterizing forecast accuracy in the 2×2 situation is in terms of odds, or the ratio of a probability to its complementary probability, $p/(1-p)$. In the context of forecast verification the ratio of the conditional odds of a hit, given that the event occurs, to the conditional odds of a false alarm, given that the event does not occur, is called the *odds ratio*,

$$\theta = \frac{p(y_1 | o_1)/[1 - p(y_1 | o_1)]}{p(y_1 | o_2)/[1 - p(y_1 | o_2)]} = \frac{p(y_1 | o_1)/p(y_2 | o_1)}{p(y_1 | o_2)/p(y_2 | o_2)} = \frac{a\,d}{b\,c}. \tag{9.9}$$

The conditional distributions making up the odds ratio are all likelihoods from Equation 9.6. In terms of the 2×2 contingency table, the odds ratio is the product of the numbers of correct forecasts divided by the product of the numbers of incorrect forecasts. Clearly, larger values of this ratio indicate more accurate forecasts. No-information forecasts, for which the forecasts and observations are statistically independent (i.e., $p(y_i, o_j) = p(y_i)\, p(o_j)$, Equation 2.12), yield $\theta = 1$. The odds ratio was introduced into meteorological forecast verification by Stephenson (2000), although it has a longer history of use in medical statistics.

Bias

The bias, or comparison of the average forecast with the average observation, usually is represented as a ratio for verification of contingency tables. In terms of the 2×2 table in Figure 9.1a the bias ratio is

$$B = \frac{a + b}{a + c}. \tag{9.10}$$

The bias is simply the ratio of the number of "yes" forecasts to the number of "yes" observations. Unbiased forecasts exhibit $B = 1$, indicating that the event was forecast the same number of times that it was observed. Note that bias provides no information about the correspondence between the individual forecasts and observations of the event on particular occasions, so that Equation 9.10 is not an accuracy measure. Bias greater than one indicates that the event was forecast more often than observed, which is called *overforecasting*. Conversely, bias less than one indicates that the event was forecast less often than observed, or was *underforecast*.

Reliability and Resolution

Equation 9.5 shows four calibration attributes for the 2×2 contingency table. That is, each quantity in Equation 9.5 is a conditional relative frequency for event occurrence or nonoccurrence, given either a "yes" or "no" forecast, in the sense of the calibration distributions $p(o_j \mid y_i)$ of Equation 9.2. Actually, Equation 9.5 indicates two calibration distributions, one conditional on the "yes" forecasts (Equations 9.5a and 9.5b), and the other conditional on the "no" forecasts (Equations 9.5c and 9.5d). Each of these four conditional probabilities is a scalar reliability statistic for the 2×2 contingency table, and all four have been given names (e.g., Doswell et al., 1990). By far the most commonly used of these conditional relative frequencies is Equation 9.5b, which is called the *false alarm ratio* (FAR). In terms of Figure 9.1a, the false alarm ratio is computed as

$$\text{FAR} = \frac{b}{a + b}. \tag{9.11}$$

That is, FAR is the fraction of "yes" forecasts that turn out to be wrong, or that proportion of the forecast events that fail to materialize. The FAR has a negative orientation, so that smaller values of FAR are to be preferred. The best possible FAR is zero, and the worst possible FAR is one. The FAR has also been called the false alarm *rate* (Barnes et al., 2009 sketch a history of the confusion), although this rather similar term is now generally reserved for the discrimination measure in Equation 9.13. 1–FAR is known as the *positive predictive value* in medical statistics.

Discrimination

Two of the conditional probabilities in Equation 9.6 are used frequently to characterize 2×2 contingency tables, although all four of them have been named (e.g., Doswell et al., 1990). Equation 9.6a is commonly known as the *hit rate*,

$$H = \frac{a}{a+c}. \tag{9.12}$$

Regarding only the event o_1 as "the" event of interest, the hit rate is the ratio of correct forecasts to the number of times this event occurred. Equivalently this statistic can be regarded as the fraction of those occasions when the forecast event occurred on which it was also forecast, and so is also called the *probability of detection* (POD). In medical statistics this quantity is known as the *true-positive fraction*, or the *sensitivity*.

Equation 9.6c is called the *false alarm rate*,

$$F = \frac{b}{b+d}, \tag{9.13}$$

which is the ratio of false alarms to the total number of nonoccurrences of the event o_1, or the conditional relative frequency of a wrong forecast given that the event does not occur. The false alarm rate is also known as the *probability of false detection* (POFD). Jointly the hit rate and false alarm rate provide both the conceptual and geometrical basis for the signal detection approach for verifying probabilistic forecasts (Section 9.4.6). In medical statistics this quantity is known as the *false-positive fraction*, or 1 minus the *specificity*.

Noting that many summary statistics degenerate to trivial values as the climatological probability, or base rate,

$$s = p(o_1) = (a+c)/n \tag{9.14}$$

becomes small, Ferro and Stephenson (2011) proposed the extremal dependence index,

$$EDI = \frac{\ln(F) - \ln(H)}{\ln(F) + \ln(H)} \tag{9.15}$$

to address 2×2 verification for extreme (and therefore rare) events. This measure takes on values between ± 1, with positive values indicating better performance than random forecasts.

9.2.3. Skill Scores for 2×2 Contingency Tables

Forecast verification data in contingency tables are often characterized using relative accuracy measures or skill scores in the general form of Equation 9.4. A large number of such skill scores have been developed for the 2×2 verification situation, and many of these have been presented by Muller (1944), Mason (2003), Murphy and Daan (1985), Stanski et al. (1989), and Woodcock (1976). Some of these skill measures date from the earliest literature on forecast verification (Murphy, 1996) and have been rediscovered and (unfortunately) renamed on multiple occasions. In general the different skill scores perform differently and sometimes inconsistently. This situation can be disconcerting if we hope to choose among alternative skill scores, but should not really be surprising given that all of these skill scores are scalar measures of forecast performance in what is intrinsically a higher-dimensional setting. Scalar skill scores are used because they are conceptually convenient, but they are necessarily incomplete representations of forecast performance.

One of the most frequently used skill scores for summarizing square contingency tables was originally proposed by Doolittle (1888), but because it is nearly universally known as the Heidke Skill Score (Heidke, 1926) this latter name will be used here. The *Heidke Skill Score* (HSS) follows the form of Equation 9.4, based on the proportion correct (Equation 9.7) as the basic accuracy measure. Thus perfect forecasts receive HSS = 1, forecasts equivalent to the reference forecasts receive zero scores, and forecasts worse than the reference forecasts receive negative scores.

The reference accuracy measure in the Heidke score is the proportion correct that would be achieved by random forecasts that are statistically independent of the observations. In the 2×2 situation, the marginal probability of a "yes" forecast is $p(y_1) = (a+b)/n$, and the marginal probability of a "yes" observation is $s = \mathrm{p}(o_1) = (a+c)/n$. Therefore the probability of a correct "yes" forecast by chance is

$$p(y_1)\,p(o_1) = \frac{(a+b)}{n}\frac{(a+c)}{n} = \frac{(a+b)(a+c)}{n^2}, \tag{9.16a}$$

and similarly the probability of a correct "no" forecast by chance is

$$p(y_2)\,p(o_2) = \frac{(b+d)}{n}\frac{(c+d)}{n} = \frac{(b+d)(c+d)}{n^2}. \tag{9.16b}$$

Thus following Equation 9.4, for the 2×2 verification setting the Heidke Skill Score is

$$\begin{aligned} \mathrm{HSS} &= \frac{(a+d)/n - [(a+b)(a+c)+(b+d)(c+d)]/n^2}{1 - [(a+b)(a+c)+(b+d)(c+d)]/n^2} \\ &= \frac{2(ad-bc)}{(a+c)(c+d)+(a+b)(b+d)}, \end{aligned} \tag{9.17}$$

where the second equality is easier to compute. The HSS is referred to as Cohen's kappa (Cohen, 1960) in the social science literature. Hyvärinen (2014) provides an alternative, Bayesian derivation for HSS that does not invoke the notion of naive random forecasts.

Another popular skill score for contingency-table forecast verification has been rediscovered many times since being first proposed by Peirce (1884). The *Peirce Skill Score* is also commonly referred to as the *Hanssen-Kuipers discriminant* (Hanssen and Kuipers, 1965) or *Kuipers' performance index* (Murphy and Daan, 1985), and is sometimes also called the *true skill statistic* (TSS) (Flueck, 1987). It is known as *Youden's* (1950) index in the medical statistics literature. *Gringorten's* (1967) *skill score* contains equivalent information, as it is a linear transformation of the Peirce Skill Score. The Peirce Skill Score is formulated similarly to the Heidke score, except that the reference hit rate in the denominator is that for random forecasts that are constrained to be unbiased. That is, the imagined random reference forecasts in the denominator of Equation 9.4 have a marginal distribution that is equal to the (sample) climatology, so that $p(y_1) = p(o_1)$ and $p(y_2) = p(o_2)$. Again following Equation 9.4 for the 2×2 situation of Figure 9.1, the Peirce Skill Score is computed as

$$\begin{aligned} \mathrm{PSS} &= \frac{(a+d)/n - [(a+b)(a+c)+(b+d)(c+d)]/n^2}{1 - \left[(a+c)^2+(b+d)^2\right]/n^2} \\ &= \frac{ad-bc}{(a+c)(b+d)}, \end{aligned} \tag{9.18}$$

where again the second equality is computationally more convenient. The PSS can also be understood as the difference between two conditional probabilities in the likelihood-base rate factorization of the joint distribution (Equation 9.6), namely, the hit rate (Equation 9.12) and the false alarm rate (Equation 9.13). That is, PSS $= H - F$. Perfect forecasts receive a score of one (because $b = c = 0$, or in an alternative view, $H = 1$ and $F = 0$), random forecasts receive a score of zero (because $H = F$), and forecasts inferior to the random forecasts receive negative scores. Constant forecasts (i.e., always forecasting one or the other of y_1 or y_2) are also accorded zero skill. Furthermore, unlike the Heidke score, the contribution made to the Peirce Skill Score by a correct "no" or "yes" forecast increases as the event is more or less likely, respectively. Thus forecasters are not discouraged from forecasting rare events on the basis of their low climatological probability alone.

The *Clayton* (1927, 1934) *Skill Score* can be formulated as the difference of the conditional probabilities in Equation 9.5a and 9.5c, relating to the calibration-refinement factorization of the joint distribution, that is,

$$\text{CSS} = \frac{a}{(a+b)} - \frac{c}{(c+d)} = \frac{ad - bc}{(a+b)(c+d)}. \tag{9.19}$$

The CSS indicates positive skill to the extent that the event occurs more frequently when forecast than when not forecast, so that the conditional relative frequency of the "yes" outcome given "yes" forecasts is larger than the conditional relative frequency given "no" forecasts. Clayton (1927) originally called this difference of conditional relative frequencies (multiplied by 100%) the percentage of skill, where he understood skill in the modern sense of accuracy relative to climatological expectancy. Perfect forecasts exhibit $b = c = 0$, yielding CSS $= 1$. Random forecasts (Equation 9.16) yield CSS $= 0$.

A skill score in the form of Equation 9.4 can also be constructed using the threat score (Equation 9.8) as the basic accuracy measure, using TS for random (Equation 9.16) forecasts as the reference. In particular, $\text{TS}_{\text{ref}} = a_{\text{ref}}/(a+b+c)$, where Equation 9.16a implies $a_{\text{ref}} = (a+b)(a+c)/n$. Since $\text{TS}_{\text{perf}} = 1$, the resulting skill score is

$$\text{GSS} = \frac{a/(a+b+c) - a_{ref}/(a+b+c)}{1 - a_{ref}/(a+b+c)} = \frac{a - a_{ref}}{a - a_{ref} + b + c}. \tag{9.20}$$

This skill score, called the *Gilbert Skill Score* (GSS) originated with Gilbert (1884), who referred to it as the *ratio of success*. It is also commonly (although erroneously, Hogan et al., 2010) called the *Equitable Threat Score* (ETS). Because the sample size n is required to compute a_{ref}, the GSS depends also on the number of correct "no" forecasts, unlike the TS.

The odds ratio (Equation 9.9) can also be used as the basis of a skill score,

$$Q = \frac{\theta - 1}{\theta + 1} = \frac{(ad/bc) - 1}{(ad/bc) + 1} = \frac{ad - bc}{ad + bc}. \tag{9.21}$$

This skill score originated with Yule (1900), and is called *Yule's Q* (Woodcock, 1976), or the *Odds Ratio Skill Score* (ORSS) (Stephenson, 2000). Random (Equation 9.15) forecasts exhibit $\theta = 1$, yielding $Q = 0$; and perfect forecasts exhibit $b = c = 0$, producing $Q = 1$. However, an apparently perfect skill of $Q = 1$ is also obtained for imperfect forecasts, if either one or the other of b or c is zero.

All the skill scores listed in this section depend only on the four counts a, b, c, and d in Figure 9.1 and are therefore necessarily related. Notably, HSS, PSS, CSS, and Q are all proportional to the quantity $ad - bc$. Some specific mathematical relationships among the various skill scores are noted in Hogan and Mason (2012), Mason (2003), Murphy (1996), Stephenson (2000), and Wandishin and Brooks (2002).

Example 9.1. The Finley Tornado Forecasts

The Finley tornado forecasts (Finley, 1884) are historical 2×2 forecast verification data that are often used to illustrate evaluation of forecasts in this format. John Finley was a sergeant in the U.S. Army who, using telegraphed synoptic information, formulated yes/no tornado forecasts for 18 regions of the United States east of the Rocky Mountains. The data set and its analysis were instrumental in stimulating much of the early work on forecast verification (Murphy, 1996). The contingency table for Finley's $n = 2803$ forecasts is presented in Table 9.1a.

Finley chose to evaluate his forecasts using the proportion correct (Equation 9.7), which for his data is PC = (28+2680)/2803 = 0.966. On the basis of this result, Finley claimed 96.6% accuracy. However, the proportion correct for this data set is dominated by the correct "no" forecasts, since tornados are relatively rare. Very shortly after Finley's paper appeared, Gilbert (1884) pointed out that always forecasting "no" would produce an even higher proportion correct. The contingency table that would be obtained if tornados had never been forecast is shown in Table 9.1b. These hypothetical forecasts yield a proportion correct of PC = (0+2752)/2803 = 0.982, which is indeed higher than the proportion correct for the actual forecasts.

Employing the threat score (Equation 9.8) gives a more reasonable comparison, because the large number of easy, correct "no" forecasts are ignored. For Finley's original forecasts, the threat score is TS = 28/(28+72+23) = 0.228, whereas for the obviously useless "no" forecasts in Table 9.1b the threat score is TS = 0/(0+0+51) = 0. Clearly the threat score would be preferable to the proportion correct in this instance, but it is still not completely satisfactory. Equally useless would be a forecasting system that always forecast "yes" for tornados. For constant "yes" forecasts the threat score would be TS = 51/(51+2752+0) = 0.018, which is small, but not zero. The odds ratio for the Finley forecasts is $\theta = (28)(2680)/(72)(23) = 45.3 > 1$, suggesting better than random performance for the forecasts in Table 9.1a. The odds ratio is not computable for the forecasts in Table 9.1b.

The bias ratio for the Finley tornado forecasts is $B = 1.96$, indicating that approximately twice as many tornados were forecast as actually occurred, much of which might be attributable to the sparseness of the observation network at the time. The false alarm ratio is FAR = 0.720, which expresses the fact that a fairly large fraction of the forecast tornados did not eventually materialize. On the other hand, the hit rate is $H = 0.549$ and the false alarm rate is $F = 0.0262$; indicating that more than half of the actually observed tornados were forecast to occur, whereas a very small fraction of the nontornado cases falsely warned of a tornado.

TABLE 9.1 Contingency Tables for Verification of the Finley Tornado Forecasts, from 1884

(a)

		Observed	
		Tornados	
		Yes	No
Forecast	Yes	**28**	**72**
	No	**23**	**2680**
			$n = 2803$

(b)

		Observed	
		Tornados	
		Yes	No
Forecast	Yes	**0**	**0**
	No	**51**	**2752**
			$n = 2803$

The forecast event is occurrence of a tornado, with separate forecasts for 18 regions of the United States east of the Rocky Mountains. (a) The table for the forecasts as originally issued; and (b) data that would have been obtained if "no tornados" had always been forecast.

The various skill scores yield a very wide range of results for the Finely tornado forecasts: HSS = 0.355, PSS = 0.523, CSS = 0.271, GSS = 0.216, EDI = 0.717, and $Q = 0.957$. Zero skill is attributed to the constant "no" forecasts in Table 9.1b by HSS, PSS, and GSS, but CSS, EDS, SEDS, and Q cannot be computed for $a = b = 0$. ◇

9.2.4. Which Score?

The wide range of skills attributed to the Finley tornado forecasts in Example 9.1 may be somewhat disconcerting, but should not be surprising. The root of the problem is that, even in this simplest of all possible forecast verification settings, the dimensionality (Murphy, 1991) of the problem is $I \times J - 1 = 3$, and the collapse of this three-dimensional information into a single number by any scalar verification measure necessarily involves a loss of information. Put another way, there are a variety of ways for forecasts to go right and for forecasts to go wrong, and mixtures of these are combined differently by different scalar attributes and skill scores. There is no single answer to the question posed in the heading for this section.

It is sometimes necessary to choose a single scalar summary of forecast performance, accepting that the summary will necessarily be incomplete. For example, competing forecasters in a contest must be evaluated in a way that produces an unambiguous ranking of their performances. Choosing a single score for such a purpose involves investigating and comparing relevant properties of competing candidate verification statistics, a process that is called *metaverification* (Murphy, 1996). Which property or properties might be most relevant may depend on the specific situation, but one reasonable criterion can be that a chosen verification statistic should be *equitable* (Gandin and Murphy, 1992). An equitable skill score rates random forecasts, and all constant forecasts (such as always forecasting "no tornados" in Example 9.1), equally. Usually this score for useless forecasts is set to zero, and equitable scores are scaled such that perfect forecasts have unit skill. Equitability also implies that correct forecasts of less frequent events (such as tornados in Example 9.1) are weighted more strongly than correct forecasts of more common events, which discourages distortion of forecasts toward the more common event in order to artificially inflate the resulting score.

The original Gandin and Murphy (1992) definition of equitability imposed the additional condition that any equitable verification measure must be expressible as a linear weighted sum of the elements of the contingency table, which leads to the use of the PSS (Equation 9.18) as the only equitable skill score for the 2×2 verification setting. However, Hogan et al. (2010) have argued persuasively that this second condition is not compelling, and if it is not required HSS (Equation 9.17) also equitable, in the sense of also yielding zero skill for random or constant forecasts. Interestingly Hogan et al. (2010) also show that GSS (also known as the "equitable" threat score, Equation 9.20) is not equitable because it does not yield zero skill for random forecasts. However, Hogan et al. (2010) find that GSS, CSS (Equation 9.19), and Q (Equation 9.21) are *asymptotically equitable*, meaning that they approach equitability as the sample size becomes very large.

Because the dimensionality of the 2×2 problem is 3, the full information in the 2×2 contingency table can be captured fully by three well-chosen scalar attributes. Using the likelihood-base rate factorization (Equation 9.6), the full joint distribution can be summarized by (and recovered from) the hit rate H (Equations 9.12 and 9.6a), the false alarm rate F (Equation 9.13 and 9.6c), and the base rate s (Equation 9.14). Similarly, using the calibration-refinement factorization (Equation 9.5), forecast performance depicted in a 2×2 contingency table can be fully captured using the false alarm ratio FAR (Equations 9.11 and 9.5b), its counterpart in Equation 9.5d, and the probability $p(y_1) = (a+b)/n$ defining

the refinement distribution. Other triplets of verification measures can also be used jointly to illuminate the data in a 2×2 contingency table (although not any three scalar statistics calculated from a 2×2 table will fully represent its information content). For example, Stephenson (2000) suggests use of H and F together with the bias ratio B, calling this the BHF representation. He also notes that, jointly, the likelihood ratio θ and Peirce Skill Score PSS represent the same information as H and F, so that these two statistics together with either s or B will also fully represent the 2×2 table. The joint characterization using H, F, and s is also sometimes used in the medical literature (Pepe, 2003). Stephenson et al. (2008a) and Brill (2009) analyze properties of various 2×2 performance measures in terms of H, B, and s.

Consistent with the ideas presented in Chapter 3, the understanding of multivariable statistics can be enhanced through well-designed graphics. Roebber (2009) suggests summarizing 2×2 verification tables in a 2-dimensional diagram whose axes are 1–FAR and H, an example of which is shown in Figure 9.2. In this diagram the solid contours show isolines of constant TS (Equation 9.8) and dashed radial lines indicate forecast bias (Equation 9.10). The short line segments in this particular example of the diagram connect results for forecasts before (circles) and after (triangles) human intervention, highlighting the resulting improvements in both H and TS. The Roebber (2009) performance diagram is not a complete 3-dimensional summary of the joint distribution, since it does not reflect the number of correct "no" forecasts, but may nevertheless be informative when summarizing performance of forecasts of rare events, or spatial forecasts where the quantity d in Figure 9.1 may be ambiguous or unknown.

Wilks (2016a) proposed two diagnostic verification diagrams for the 2×2 forecast setting in Figure 9.1, based on one of the factorizations of the joint distribution of forecasts and observations. The first, called the H–F diagram, plots verification results in the space defined by H (Equations 9.6a and 9.13) and F (Equations 9.6c and 9.13), in common with the ROC diagram (Section 9.4.6). Together with the base rate s (Equation 9.14) these quantities completely specify the likelihood-base rate factorization (Equation 9.3). Mason (2003) points out that isolines of the bias ratio B (Equation 9.10) exhibit slope

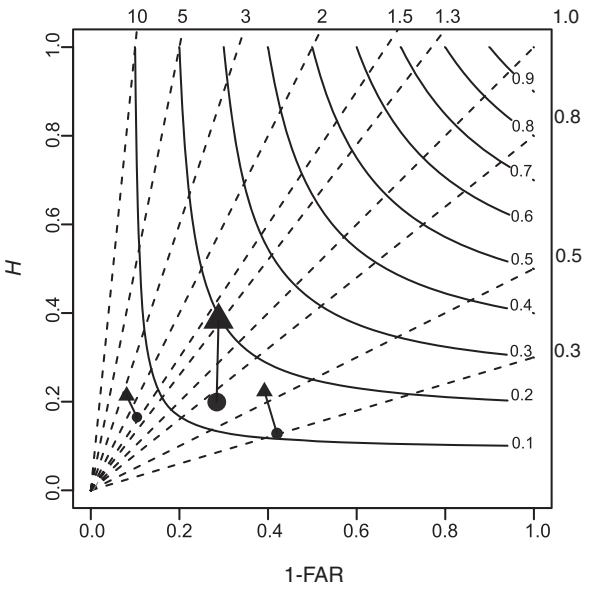

FIGURE 9.2 A Roebber (2009) performance diagram, showing paired results for convective storm initiation forecasts before (circles) and after (triangles) human intervention, as functions of 1–FAR and H. Curves show TS and dashed lines (with labels outside the figure frame) show B, as functions of the two coordinate axes. *Modified from Roberts et al. (2012). © American Meteorological Society. Used with permission.*

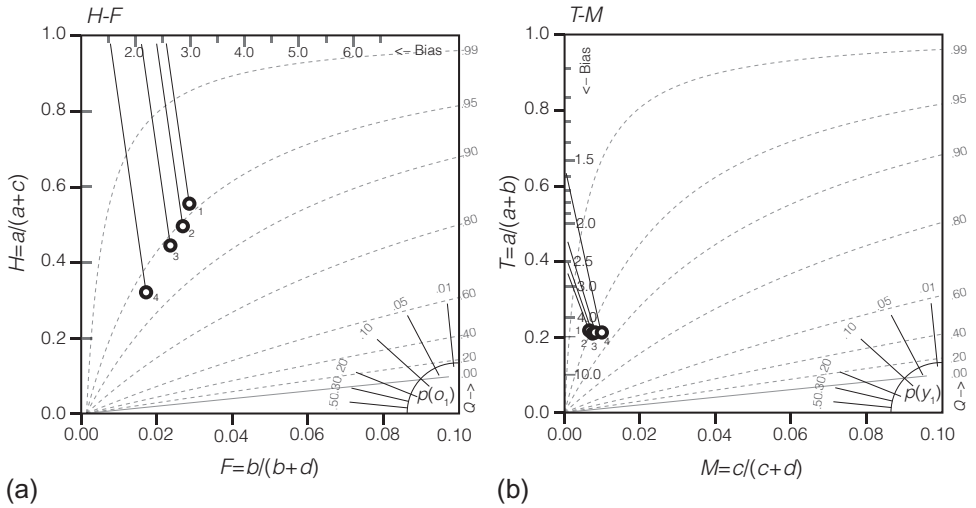

FIGURE 9.3 (a) H–F diagram, and (b) T–M diagram, for Storm Prediction Center convective outlooks, 2002–2012, at 6-h ("1"), 18.5-h ("2") 30-h ("3"), and 48-h ("4") lead times. Angles of vectors from the horizontal represent the event probability in panel (a), and relative frequencies of "yes" forecasts in panel (b), as indicated by the legends in the lower-right corners. Bias ratios are shown on the indicated diagram margins, and dashed curves show isopleths of the Odds Ratio Skill Score, Q (Equation 9.21). *From Wilks (2016a).*

$$\frac{dH}{dF} = \frac{-(1-s)}{s} \tag{9.22}$$

and intersect the H axis at $H = B$. Therefore s can be represented by the angle from the horizontal of a vector in the H–F space given by

$$\alpha = \tan^{-1}\left(\frac{1-s}{s}\right) \tag{9.23}$$

The full joint distribution for a set of forecasts can be represented as a point in the H–F plane locating the likelihood distribution probabilities, together with an arrow pointing upward and to the left at an angle defined by Equation 9.23. An example is shown in Figure 9.3a, where the legend in the lower-right corner indicates values of $s = p(o_1)$, and the upward arrows locate the values of B. The dashed contours show isolines of the Odds Ratio Skill Score Q (Equation 9.21), indicating that more skillful forecasts are located toward the upper-left portion of the diagram. Either *EDI* (Equation 9.15) or *PSS* (Equation 9.18) could be used instead since both may be expressed as functions of H and F only.

Figure 9.3b shows an example of the T–M diagram, which visually portrays the calibration-refinement factorization (Equation 9.2) for the 2×2 verification setting. In this diagram, a set of forecasts is located as a point defined by $T = a/(a+b)$ (Equation 9.5a) and $M = c/(c+d)$ (Equation 9.5c). Representation of the calibration-refinement factorization is completed by the relative frequency of "yes" forecasts, $p(y_1) = (a+b)/n$ which, analogously to Equations 9.22 and 9.23, can be represented by an angle in the T–M space. Arrows drawn at these angles from the plotted points locate the bias on the margin of the diagram, with the legend in the lower-right corner again indicating the correspondence between these directions and the frequency of "yes" forecasts. The dashed curves in Figure 9.3b show isolines of Q (Equation 9.21), indicating that more skillful forecasts are located toward the upper-left portion of the diagram.

9.2.5. Conversion of Probabilistic to Nonprobabilistic Forecasts

The MOS system from which the nonprobabilistic precipitation amount forecasts in Table 7.8 were taken actually produces probability forecasts for discrete precipitation amount classes. The publicly issued precipitation amount forecasts in the table were then derived by converting the underlying probabilities to the nonprobabilistic format by choosing one and only one of the possible categories. This unfortunate information degradation is distressing, but is sometimes advocated under the rationale that nonprobabilistic forecasts are easier to understand. However, the loss of information content is detrimental to the forecast users.

For a dichotomous predictand, the conversion from a probabilistic to a nonprobabilistic format requires selection of a threshold probability, above which the forecast will be "yes," and below which the forecast will be "no." This procedure seems simple enough; however, the proper threshold to choose depends on the user of the forecast and the particular decision problem(s) to which that user will apply the forecast. If a decision problem has undergone quantitative analysis, for example, as that described in Section 9.9 or by Manzato and Jolliffe (2017), then the choice of the appropriate threshold may be clear. However, different decision problems will require different threshold probabilities, and this is the crux of the information-loss issue. In a very real sense, the conversion from a probabilistic to a nonprobabilistic format amounts to the forecaster making decisions for the forecast users, but without knowing the particulars of their decision problems. Often the choice of threshold is made which optimizes the value of some verification score, but different scores will be optimized for different thresholds, and indeed a score can be constructed that corresponds to any chosen threshold (Mason, 1979). Necessarily, then, the conversion of a probabilistic forecast to a nonprobabilistic format is arbitrary.

Example 9.2. Effects of Different Thresholds on Conversion to Nonprobabilistic Forecasts
It is instructive to examine the procedures used to convert probabilistic to nonprobabilistic forecasts. Table 9.2 contains a verification data set of probability-of-precipitation forecasts, issued for the United States during the period October 1980 through March 1981. Here the joint distribution of the $I = 12$ possible forecasts and the $J = 2$ possible observations is presented in the form of the calibration-refinement factorization (Equation 9.2). For each allowable forecast probability, y_i, the conditional probability $p(o_1 \mid y_i)$ indicates the relative frequency of the event $j = 1$ (precipitation occurrence) for these $n = 12{,}402$ forecasts. The marginal probabilities $p(y_i)$ indicate the relative frequencies with which each of the $I = 12$ possible forecast values was used.

These precipitation occurrence forecasts were issued as probabilities. If it had been intended to convert them first to a nonprobabilistic rain/no rain format, a threshold probability would have been

TABLE 9.2 Verification data for subjective 12–24 h lead time probability-of-precipitation forecasts for the United States during October 1980–March 1981, expressed in the form of the calibration-refinement factorization (Equation 9.2) of the joint distribution of these forecasts and observations.

y_i	0.00	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
$p(o_1 \mid y_i)$	.006	.019	.059	.150	.277	.377	.511	.587	.723	.799	.934	.933
$p(y_i)$	.4112	.0671	.1833	.0986	.0616	.0366	.0303	.0275	.0245	.0220	.0170	.0203

There are $I = 12$ allowable values for the forecast probabilities, y_i, and $J = 2$ events ($j = 1$ for precipitation and $j = 2$ for no precipitation). The sample climatological relative frequency is 0.162, and the sample size is $n = 12{,}402$.
From Murphy and Daan (1985).

chosen in advance. There are many possibilities for this choice, each of which gives different results. The two simplest approaches are used rarely, if ever, in operational practice. The first procedure is to forecast the more likely event, which corresponds to selecting a threshold probability of 0.50. The other simple approach is to use the climatological relative frequency of the event being forecast as the threshold probability. For the data set in Table 9.2 this relative frequency is $\Sigma_i\, p(o_j|y_i)p(y_i) = 0.162$ (Equation 2.14), although in practice this probability threshold would need to have been estimated in advance using historical climatological data, and likely would have been estimated separately for the different locations whose data are aggregated in the table. Forecasting the more likely event turns out to maximize the expected values of both the proportion correct (Equation 9.7) and the Heidke Skill Score (Equation 9.17), and using the climatological relative frequency for the probability threshold maximizes the expected Peirce Skill Score (Equation 9.18) (Mason, 1979).

The two methods for choosing the threshold probability that are most often used operationally are based on the threat score (Equation 9.8) and the bias ratio (Equation 9.10) for 2×2 contingency tables. For each possible choice of a threshold probability, a different 2×2 contingency table, in the form of Figure 9.1a, results, and therefore different values of TS and B are obtained. When using the threat score to choose the threshold, that threshold producing the maximum TS is selected. When using the bias ratio, that threshold producing, as nearly as possible, no bias ($B = 1$) is chosen.

Figure 9.4 illustrates the dependence of the bias ratio and threat score on the threshold probability for the data given in Table 9.2. Also shown are the hit rates H and false alarm ratios FAR that would be obtained. The threshold probabilities that would be chosen according to the climatological

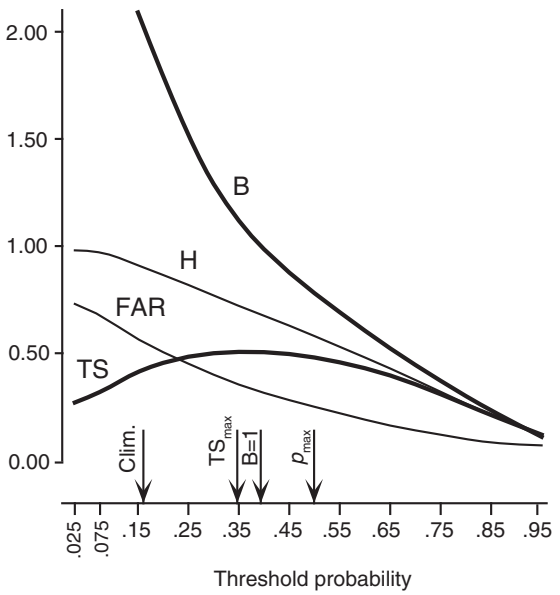

FIGURE 9.4 Derivation of candidate threshold probabilities for converting the probability-of-precipitation forecasts in Table 9.2 to nonprobabilistic rain/no rain forecasts. The Clim threshold indicates a forecast of rain if the probability is higher than the climatological probability of precipitation, TS_{max} is the threshold that would maximize the threat score of the resulting nonprobabilistic forecasts, the $B = 1$ threshold would produce unbiased forecasts, and the p_{max} threshold would produce nonprobabilistic forecasts of the more likely of the two events. Also shown (lighter lines) are the hit rates H and false alarm ratios FAR for the resulting 2×2 contingency tables.

relative frequency (Clim), the maximum threat score (TS_{max}), unbiased nonprobabilistic forecasts ($B = 1$), and maximum probability (p_{max}) are indicated by the arrows at the bottom of the figure. For example, choosing the overall relative frequency of precipitation occurrence, 0.162, as the threshold results in forecasts of PoP = 0.00, 0.05, and 0.10 being converted to "no rain," and the other forecasts being converted to "rain." This would have resulted in $n[p(y_1)+p(y_2)+p(y_3)] = 12{,}402[0.4112+0.0671+0.1833] = 8205$ "no" forecasts, and 12,402–8205 = 4197 "yes" forecasts. Of the 8205 "no" forecasts, we can compute, using the multiplicative law of probability (Equation 2.11), that the proportion of occasions that "no" was forecast but precipitation occurred was $p(o_1|y_1)p(y_1)+p(o_1|y_2)p(y_2)+p(o_1|y_3)p(y_3) = (.006)(.4112)+(.019)(.0671)+(.059)(.1833) = 0.0146$. This relative frequency is c/n in Figure 9.1, so that $c = (0.0146)(12{,}402) = 181$, and $d = 8205–181 = 8024$. Similarly, we can compute that, for this cutoff, $a = 12{,}402[(0.150)(0.0986)+\ldots+(0.933)(0.203)] = 1828$ and $b = 2369$. The resulting 2×2 table yields $B = 2.09$, and TS = 0.417. By contrast, the threshold maximizing *TS* is near 0.35, which also would have resulted in overforecasting of precipitation occurrence. ◇

9.2.6. Extensions for Multicategory Discrete Predictands

Nonprobabilistic forecasts for discrete predictands are not limited to the 2×2 format, although that simple situation is the most commonly encountered and the easiest to understand. In some settings it is natural or desirable to consider and forecast more than two discrete MECE events. The left side of Figure 9.5, in boldface type, shows a generic contingency table for the case of $I = J = 3$ possible forecasts and events. Here the counts for each of the nine possible forecast and event pair outcomes are denoted by the letters *r* through *z*, yielding a total sample size $n = r+s+t+u+v+w+x+y+z$. As before, dividing each of the nine counts in this 3×3 contingency table by the sample size yields a sample estimate of the joint distribution of forecasts and observations, $p(y_i, o_j)$.

Of the accuracy measures listed in Equations 9.7 through 9.9, only the proportion correct (Equation 9.7) generalizes directly to situations with more than two forecast and event categories. Regardless of the size of *I* and *J*, the proportion correct is still given by the number of correct forecasts

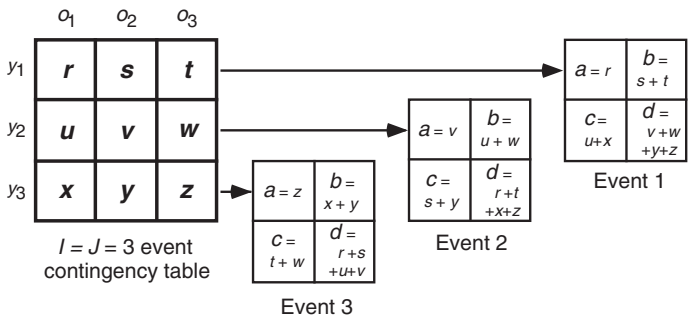

FIGURE 9.5 Contingency table for the $I = J = 3$ nonprobabilistic forecast verification situation (bold), and its reduction to three 2×2 contingency tables. Each 2×2 contingency table is constructed by regarding one of the three original events as "the" event being forecast, and the remaining two original events combined as complementary, that is, not the forecast event. For example, the 2×2 table for Event 1 lumps Event 2 and Event 3 as the single event "not Event 1." The letters *a*, *b*, *c*, and *d* are used in the same sense as in Figure 9.1a. Performance measures specific to the 2×2 contingency tables can then be computed separately for each of the resulting tables. This procedure generalizes easily to square forecast verification contingency tables with arbitrarily many forecast and event categories.

divided by the total number of forecasts, n. This number of correct forecasts is obtained by adding the counts along the diagonal from the upper left to the lower-right corners of the contingency table. In Figure 9.5, the numbers r, v, and z represent the numbers of occasions when the first, second, and third events were correctly forecast, respectively. Therefore in the 3×3 table represented in this figure, the proportion correct would be $\text{PC} = (r+v+z)/n$.

The other statistics listed in Section 9.2.2 pertain only to the dichotomous, yes/no forecast situation. In order to apply these to nonprobabilistic forecasts that are not dichotomous, it is necessary to collapse the $I = J > 2$ contingency table into a series of 2×2 contingency tables. Each of these 2×2 tables is constructed, as indicated in Figure 9.5, by considering "the" forecast event in distinction to the complementary, "not the forecast event." This complementary event simply is constructed as the union of the $J - 1$ remaining events. For example, in Figure 9.5 the 2×2 contingency table for Event 1 lumps Events 2 and 3 as "not Event 1." Thus the number of times Event 1 is correctly forecast is still $a = r$, but the number of times it is incorrectly forecast is $b = s+t$. From the standpoint of this collapsed 2×2 contingency table, whether the incorrect forecast of Event 1 was followed by Event 2 or Event 3 is unimportant. Similarly, the number of times "not Event 1" is correctly forecast is $d = v+w+y+z$, and includes cases where Event 2 was forecast but Event 3 occurred, and Event 3 was forecast but Event 2 occurred.

Attributes for 2×2 contingency tables can be computed for any or all of the 2×2 tables constructed in this way from larger square tables. For the 3×3 contingency table in Figure 9.5, the bias (Equation 9.10) for forecasts of Event 1 would be $B_1 = (r+s+t)/(r+u+x)$, the bias for forecasts of Event 2 would be $B_2 = (u+v+w)/(s+v+y)$, and the bias for forecasts of Event 3 would be $B_3 = (x+y+z)/(t+w+z)$.

Example 9.3. A Set of Multicategory Forecasts

The left-hand side of Table 9.3 shows a 3×3 verification contingency table for forecasts of freezing rain (y_1), snow (y_2), and rain (y_3). These are nonprobabilistic MOS forecasts, conditional on the occurrence of some form of precipitation, for the Eastern region of the United States, for October through March of 1983/1984 through 1988/1989. For each of the three precipitation types, a 2×2 contingency table can be constructed, following Figure 9.5, that summarizes the performance of forecasts of that precipitation type in distinction to the other two precipitation types together. Table 9.3 also includes forecast attributes from Section 9.2.2 for each 2×2 decomposition of the 3×3 contingency table. These are reasonably consistent with each other for a given 2×2 table, and indicate that the rain forecasts were slightly superior to the snow forecasts, but that the freezing rain forecasts were substantially less successful, with respect to most of these measures. ◇

The Heidke and Peirce Skill Scores can be extended easily to verification problems where there are more than $I = J = 2$ possible forecasts and events. The formulae for these scores in the more general case can be written most easily in terms of the joint distribution of forecasts and observations, $p(y_i, o_j)$, and the marginal distributions of the forecasts, $p(y_i)$ and of the observations, $p(o_j)$. For the Heidke Skill Score this more general form is

$$\text{HSS} = \frac{\sum_{i=1}^{I} p(y_i, o_i) - \sum_{i=1}^{I} p(y_i)p(o_i)}{1 - \sum_{i=1}^{I} p(y_i)p(o_i)}, \tag{9.24}$$

and the higher-dimensional generalization of the Peirce Skill Score is

TABLE 9.3 Nonprobabilistic MOS Forecasts for Freezing Rain (y_1), Snow (y_2), and Rain (y_3), Conditional on Occurrence of Some Form of Precipitation, for the Eastern Region of the United States During Cool Seasons of 1983/1984 Through 1988/1989

Full 3 × 3 Contingency Table				Freezing Rain			Snow			Rain		
	o_1	o_2	o_3		o_1	Not o_1		o_2	Not o_2		o_3	Not o_3
y_1	50	91	71	y_1	50	162	y_2	2364	217	y_3	3288	259
y_2	47	2364	170	Not y_1	101	6027	Not y_2	296	3463	Not y_3	241	2552
y_3	54	205	3288									
					TS = 0.160			TS = 0.822			TS = 0.868	
					$\theta = 18.4$			$\theta = 127.5$			$\theta = 134.4$	
					$B = 1.40$			$B = 0.97$			$B = 1.01$	
					FAR = 0.764			FAR = 0.084			FAR = 0.073	
					$H = 0.331$			$H = 0.889$			$H = 0.932$	
					$F = 0.026$			$F = 0.059$			$F = 0.092$	

The verification data are presented as a 3 × 3 contingency table on the left, and then as three 2 × 2 contingency tables for each of the three precipitation types. Also shown are scalar attributes from Section 9.2.2 for each of the 2 × 2 tables. The sample size is $n = 6340$. Data are from Goldsmith (1990).

$$\mathrm{PSS} = \frac{\sum_{i=1}^{I} p(y_i, o_i) - \sum_{i=1}^{I} p(y_i)p(o_i)}{1 - \sum_{j=1}^{J} \left[p(o_j)\right]^2}. \tag{9.25}$$

Equation 9.24 reduces to Equation 9.17, and Equation 9.25 reduces to Equation 9.18, for $I = J = 2$.

Using Equation 9.24, the Heidke score for the 3 × 3 contingency table in Table 9.3 would be computed as follows. The proportion correct, PC = $\Sigma_i p(y_i,o_i)$ = (50/6340)+(2364/6340)+(3288/6340) = 0.8994. The proportion correct for the random reference forecasts would be $\Sigma_i p(y_i)p(o_i)$ = (.0334)(.0238)+(.4071)(.4196)+(.5595)(.5566) = 0.4830. Here, for example, the marginal probability $p(y_1)$ = (50+91+71)/6340 = 0.0344. The proportion correct for perfect forecasts is of course one, yielding HSS = (.8944 − .4830)/(1 − .4830) = 0.8054. The computation for the Peirce Skill Score, Equation 9.25, is the same except that a different reference proportion correct is used in the denominator only. This is the unbiased random proportion $\Sigma_i[p(o_i)^2] = .0238^2 + .4196^2 + .5566^2 = 0.4864$. The Peirce Skill Score for this 3 × 3 contingency table is then PSS = (.8944 − .4830)/(1 − .4864) = 0.8108. The difference between the HSS and the PSS for these data is small, because the forecasts exhibit little bias.

There are many more degrees of freedom in the general $I \times J$ contingency table setting than in the simpler 2×2 problem. In particular $I \times J - 1$ elements are necessary to fully specify the contingency table, so that a scalar score must summarize much more even in the 3 x 3 setting as compared to the 2 x 2 problem. Accordingly, the number of possible scalar skill scores that are plausible candidates increases rapidly with the size of the verification table. The notion of *equitability* for skill scores describing performance of nonprobabilistic forecasts of discrete predictands was proposed by Gandin and Murphy (1992) to define a restricted set of these yielding equal (zero) scores for random or constant forecasts.

When three or more events having a natural ordering are being forecast, it is usually required in addition that multiple-category forecast misses are scored as worse forecasts than single-category misses. Equations 9.24 and 9.25 both fail this requirement, as they depend only on the proportion correct. Gerrity (1992) suggested a family of equitable (in the sense of Gandin and Murphy, 1992) skill scores that are also sensitive to distance in this way and appear to provide generally reasonable results for rewarding correct forecasts and penalizing incorrect ones (Livezey, 2003). The computation of Gandin-Murphy skill scores involves first defining a set of scoring weights $w_{i,j}$, $i = 1, \ldots, I$, $j = 1, \ldots, J$; each of which is applied to one of the joint probabilities $p(y_j, o_j)$, so that in general a *Gandin-Murphy Skill Score* is computed as linear weighted sum of the elements of the contingency table

$$\text{GMSS} = \sum_{i=1}^{I} \sum_{j=1}^{J} p\left(y_i, o_j\right) w_{i,j}. \tag{9.26}$$

As noted in Section 9.2.4 for the simple case of $I = J = 2$, when linear scoring weights are required as one of the equitability criteria in the 2 x 2 setting, the result is the Peirce Skill Score (Equation 9.18). More constraints are required for larger verification problems, and Gerrity (1992) suggested the following approach to defining the scoring weights based on the sample climatology $p(o_j)$. First, define the sequence of $J - 1$ odds ratios

$$D(j) = \frac{1 - \sum_{r=1}^{j} p(o_r)}{\sum_{r=1}^{j} p(o_r)}, \quad j = 1, \ldots, J-1, \tag{9.27}$$

where r is a dummy summation index. The scoring weights for correct forecasts are then

$$w_{j,j} = \frac{1}{J-1}\left[\sum_{r=1}^{j-1} \frac{1}{D(r)} + \sum_{r=j}^{J-1} D(r)\right], \quad j = 1, \ldots, J; \tag{9.28a}$$

and the weights for the incorrect forecasts are

$$w_{i,j} = \frac{1}{J-1}\left[\sum_{r=1}^{i-1} \frac{1}{D(r)} + \sum_{r=j}^{J-1} D(r) - (j-i)\right], \quad 1 \le i < j \le J. \tag{9.28b}$$

The summations in Equation 9.28 are taken to be equal to zero if the lower index is larger than the upper index. These two equations fully define the I x J scoring weights when symmetric errors are penalized

equally, that is, when $w_{i,j} = w_{j,i}$. Equation 9.28a gives more credit for correct forecasts of rarer events and less credit for correct forecasts of common events. Equation 9.28b also accounts for the intrinsic rarity of the J events, and increasingly penalizes errors for greater differences between the forecast category i and the observed category j, through the penalty term $(j-i)$. Each scoring weight in Equation 9.28 is used together with the corresponding member of the joint distribution $p(y_j, o_j)$ in Equation 9.26 to compute the skill score. When the weights for the Gandin-Murphy Skill Score are computed according to Equations 9.27 and 9.26, the result is sometimes called the *Gerrity skill score*.

Example 9.4. Gerrity Skill Score for a 3 x 3 Verification Table

Table 9.3 includes a 3 x 3 contingency table for nonprobabilistic forecasts of freezing rain, snow, and rain, conditional on the occurrence of precipitation of some kind. Since there is not an obvious ordering of these three categories, use of *HSS* (Equation 9.24) or *PSS* (Equation 9.25) might be preferred in this case, but for convenience these data will be used in this example to illustrate computation of the Gerrity skill score. Figure 9.6a shows the corresponding joint sample probability distribution $p(y_i, o_j)$, calculated by dividing the counts in the contingency table by the sample size, $n = 6340$. Figure 9.6a also shows the sample climatological distribution $p(o_j)$, computed by summing the columns of the joint distribution.

The Gerrity (1992) scoring weights for the Gandin-Murphy Skill Score (Equation 9.26) are computed from these sample climatological relative frequencies using Equations 9.27 and 9.28. First, Equation 9.27 yields the $J-1 = 2$ likelihood ratios $D(1) = (1-.0238)/.0238 = 41.02$, and $D(2) = [1-(.0238+.4196)]/(.0238+.4196) = 1.25$. The rather large value for $D(1)$ reflects the fact that freezing rain was observed rarely, on only approximately 2% of the precipitation days during the period considered. The scoring weights for the three possible correct forecasts, computed using Equation 9.28a, are

$$w_{1,1} = \frac{1}{2}(41.02 + 1.25) = 21.14, \tag{9.29a}$$

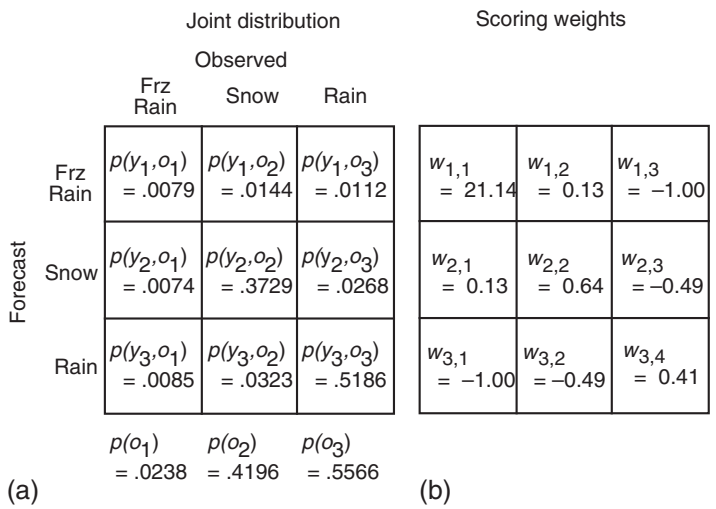

FIGURE 9.6 (a) Joint distribution of forecasts and observations for the 3×3 contingency table in Table 9.3, with the marginal probabilities for the three observations (the sample climatological probabilities). (b) The Gerrity (1992) scoring weights computed from the sample climatological probabilities.

$$w_{2,2} = \frac{1}{2}\left(\frac{1}{41.02} + 1.25\right) = 0.64, \tag{9.29b}$$

and

$$w_{3,3} = \frac{1}{2}\left(\frac{1}{41.02} + \frac{1}{1.25}\right) = 0.41; \tag{9.29c}$$

and the weights for the incorrect forecasts are

$$w_{1,2} = w_{2,1} = \frac{1}{2}(1.25 - 1) = 0.13, \tag{9.30a}$$

$$w_{2,3} = w_{3,2} = \frac{1}{2}\left(\frac{1}{41.02} - 1\right) = -0.49 \tag{9.30b}$$

and

$$w_{3,1} = w_{1,3} = \frac{1}{2}(-2) = -1.00. \tag{9.30c}$$

These scoring weights are arranged in Figure 9.6b in positions corresponding to the joint probabilities in Figure 9.6a to which they pertain.

The scoring weight $w_{1,1} = 21.14$ is much larger than the others in order to reward correct forecasts of the rare freezing rain events. Correct forecasts of snow and rain are credited with much smaller positive values, with $w_{3,3} = 0.41$ for rain being smallest because rain is the most common event. The scoring weight $w_{2,3} = -1.00$ is the minimum value according to the Gerrity algorithm, produced because the $(j - i) = 2$-category error (cf. Equation 9.28b) is the most severe possible when there is a natural ordering among the three outcomes. The penalty for an incorrect forecast of snow when rain occurs, or for rain when snow occurs (Equation 9.30b), is moderately large because these two events are relatively common. Mistakenly forecasting freezing rain when snow occurs, or vice versa, actually receives a small positive score because the frequency $p(o_1)$ is so small.

Finally, the Gandin-Murphy Skill Score in Equation 9.26 is computed by summing the products of pairs of joint probabilities and scoring weights in corresponding positions in Figure 9.6. That is, GMSS = (.0079)(21.14)+(.0144)(.13)+(.0112)(−1)+(.0074)(.13)+(.3729)(.64)+(.0268)(−.49)+(.0085)(−1)+(.0323)(−.49)+(.5186)(.41) = 0.57. ◇

The Gerrity skill score is one of an essentially infinite number of skill scores that can be constructed using the framework of Equation 9.26. For example, Rodwell et al. (2010) constructed such a skill score, for three-category precipitation forecasts. Defining p_1, p_2, and p_3 as the climatological probabilities of "dry," "light precipitation," and "heavy precipitation" categories, respectively, they construct the matrix of scoring weights in Equation 9.26 as

$$[W] = \frac{1}{2}\begin{bmatrix} 0 & \frac{1}{1-p_1} & \frac{1}{p_3} + \frac{1}{1-p_1} \\ \frac{1}{p_1} & 0 & \frac{1}{p_3} \\ \frac{1}{p_1} + \frac{1}{1-p_3} & \frac{1}{1-p_3} & 0 \end{bmatrix}, \tag{9.31}$$

in which $w_{i,j}$ is the entry in the ith row and jth column of [W]. Rodwell et al. (2010) recommend choosing the boundary between the two wet categories to yield $p_2/p_3 = 2$.

9.3. NONPROBABILISTIC FORECASTS FOR CONTINUOUS PREDICTANDS

A different set of verification measures generally is applied to forecasts of continuous variables. Continuous variables in principle can take on any value in a specified segment of the real line, rather than being limited to a finite number of discrete events. Temperature is an example of a continuous variable. In practice, however, forecasts and observations for continuous atmospheric variables are made using a finite number of discrete values. For example, temperature forecasts usually are rounded to integer degrees. It would be possible to deal with this kind of forecast verification data in discrete form, but there are usually so many allowable values of forecast and observation pairs that the resulting contingency tables would become unwieldy and possibly quite sparse. Just as discretely reported observations of continuous variables were treated as continuous quantities in Chapter 4, it is convenient and useful to treat the verification of (operationally discrete) forecasts of continuous quantities in a continuous framework as well.

Conceptually, the joint distribution of forecasts and observations is again of fundamental interest. This distribution will be the continuous analog of the discrete joint distribution of Equation 9.1. Because of the finite nature of the verification data, however, explicitly using the concept of the joint distribution in a continuous setting generally requires that a parametric distribution such as the bivariate normal (Equation 4.31) be assumed and fit. Parametric distributions and other statistical models occasionally are assumed for the joint distribution of forecasts and observations or their factorizations (e.g., Bradley et al., 2003; Katz et al., 1982; Krzysztofowicz and Long, 1991; Murphy and Wilks, 1998), but it is far more common that scalar performance and skill measures, computed using individual forecast/observation pairs, are used in verification of continuous nonprobabilistic forecasts.

9.3.1. Scalar Accuracy Measures

Scalar accuracy measures for evaluation of nonprobabilistic forecasts of continuous predictands are generally functions of the differences between pairs of forecasts and observations. It is important that the function chosen for evaluation of forecasts of this type be *consistent* (Murphy and Daan, 1985) with the process used to develop the forecast. In the case of a human forecaster, whose underlying judgment about the future weather condition will be characterized by a subjective probability distribution (Section 7.10), the extraction of a single real-valued forecast quantity from this distribution will depend on a *directive* (Murphy and Daan, 1985). Typical directives in this context are to forecast a quantity such as the mean, or the median or some other quantile, of the subjective distribution. A scoring function that is optimized by the forecasts derived from a particular directive is consistent with that directive (Gneiting, 2011a; Murphy and Daan, 1985). Conversely, once a scoring function has been decided upon, the forecaster should be informed (or the nonhuman forecasting algorithm should be constructed in light) of either its mathematical structure or of the directive with which it is consistent (Gneiting, 2011a).

Only two scalar measures of forecast accuracy for continuous predictands are in common use in meteorology and climatology, although Gneiting (2011a) lists some others that can be found in the statistics, econometrics, and nonmeteorological forecasting literature. The first of the two is the *mean absolute error*,

$$\text{MAE} = \frac{1}{n}\sum_{k=1}^{n} |y_k - o_k|. \tag{9.32}$$

Here (y_k, o_k) is the kth of n pairs of forecasts and observations. The MAE is the arithmetic average of the absolute values of the differences between the members of each pair. Clearly the MAE is zero if the forecasts are perfect (each $y_k = o_k$), and increases as discrepancies between the forecasts and observations become larger. The MAE can be interpreted as a typical magnitude for the forecast error in a given verification data set.

The MAE is consistent with the directive to forecast the median. It is a symmetric piecewise linear function, in that over- and underforecasts of the same magnitude are penalized equally. Penalizing these two types of errors differently, with an asymmetric piecewise linear function, would be consistent with forecasting directives involving different distribution quantiles (Gneiting, 2011b).

The MAE often is used for verification of temperature forecasts in the United States. Figure 9.7 shows seasonally stratified MAE for MOS minimum temperature forecasts over a portion of the country, during December 2004-November 2005, as functions of lead time. Winter temperatures are least accurate, summer temperatures are most accurate, and accuracy for the transition seasons is intermediate.

The *mean squared error* is the other common accuracy measure for continuous nonprobabilistic forecasts,

$$\text{MSE} = \frac{1}{n}\sum_{k=1}^{n} (y_k - o_k)^2. \tag{9.33}$$

The MSE is the average of the squared differences between the forecast and observation pairs. It is consistent with the directive to forecast the mean. Since the MSE is computed by squaring the forecast errors, it will be more sensitive to larger errors than will the MAE, and so will also be more sensitive to outliers. Squaring the errors necessarily produces positive terms in Equation 9.33, so the MSE

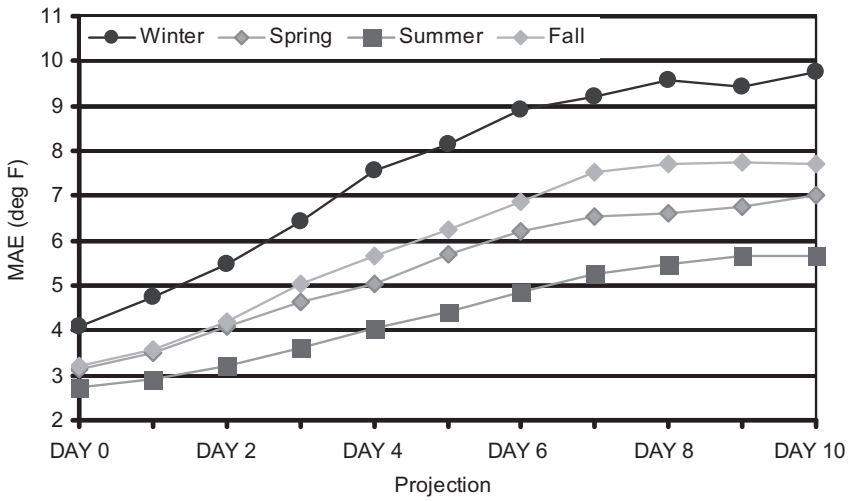

FIGURE 9.7 Seasonal MAEs for minimum temperature MOS forecasts over a portion of the United States, during December 2004 through November 2005, as functions of lead times. *From www.nws.noaa.gov/tdl/synop.*

increases from zero for perfect forecasts through larger positive values as the discrepancies between forecasts and observations become increasingly large. The similarity between Equations 9.33 and 3.6 indicates that forecasting the climatological mean on each of the n occasions being evaluated will yield MSE essentially equal to the climatological variance of the predictand o. On the other hand, forecasting a random draw from the climatological distribution yields MSE that is double the climatological variance (Hayashi, 1986). Sometimes the MSE is expressed as its square root, RMSE $= \sqrt{\text{MSE}}$, which has the same physical dimensions as the forecasts and observations, and like the MAE can also be thought of as a typical magnitude for forecast errors.

Initially, it might seem that the correlation coefficient (Equation 3.28) could be another useful accuracy measure for nonprobabilistic forecasts of continuous predictands. However, although the correlation does reflect linear association between two variables (in this case, forecasts and observations), it is sensitive to outliers, and is not sensitive to either conditional or unconditional biases that may be present in the forecasts. This latter problem can be appreciated by considering an algebraic manipulation of the MSE (Murphy, 1988):

$$\text{MSE} = (\overline{y} - \overline{o})^2 + s_y^2 + s_o^2 - 2\, s_y s_o r_{yo}. \tag{9.34}$$

Here $r_{y,o}$ is the product-moment correlation between the forecasts and observations, s_y and s_o are the standard deviations of the marginal distributions of the forecasts and observations, respectively, and the first term in Equation 9.34 is the square of the *mean error*,

$$\text{ME} = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k) = \overline{y} - \overline{o}. \tag{9.35}$$

The Mean Error is simply the difference between the average forecast and average observation, and therefore expresses the unconditional bias of the forecasts. Equation 9.35 differs from Equation 9.33 in that the individual forecast errors are not squared before they are averaged. Forecasts that are, on average, too large will exhibit ME > 0 and forecasts that are, on average, too small will exhibit ME < 0. It is important to note that the bias gives no information about the typical magnitude of individual forecast errors, and indeed the second equality in Equation 9.35 shows that it need not relate individual forecast and observation pairs. It is therefore not in itself an accuracy measure.

Returning to Equation 9.34, it can be seen that forecasts that are more highly positively correlated with the observations will exhibit lower MSE, other factors being equal. However, since the MSE can be written with the correlation $r_{y,o}$ and the bias (ME) in separate terms, we can imagine forecasts that may be highly correlated with the observations, but with sufficiently severe bias that they would be useless at face value. A hypothetical set of temperature forecasts could, for example, be exactly half of the subsequently observed temperatures, and so exhibit conditional bias. For convenience, imagine that these temperatures are nonnegative. A scatterplot of the observed temperatures versus the corresponding forecasts would exhibit all points falling perfectly on a straight line ($r_{y,o} = 1$), but the slope of that line would be 2. The bias, or mean error, would be ME $= n^{-1}\,\Sigma_k\,(f_k - o_k) = n^{-1}\,\Sigma_k\,(0.5\,o_k - o_k)$, or the negative of half of the average observation. This bias would be squared in Equation 9.34, leading to a very large MSE. A similar situation would result if all the forecasts were exactly 10 degrees colder than the observed temperatures, and so would be unconditionally biased. The correlation $r_{y,o}$ would still be one, the points on the scatterplot would fall on a straight line (this time with unit slope), the ME would be $-10°$, and the MSE would be inflated by $(10°)^2$. The definition of correlation (Equation 3.29) shows

clearly why these problems would occur: the means of the two variables being correlated are separately subtracted, and any differences in scale are removed by separately dividing by the two standard deviations, before calculating the correlation. Therefore any mismatches between either location or scale between the forecasts and observations are not reflected in the result. The Taylor diagram (e.g., -Figure 9.30) is an interesting graphical approach for separating the contributions of the correlation and the standard deviations in Equation 9.34 to the RMSE, when forecast biases are zero or are ignored.

9.3.2. Skill Scores

Skill scores, or relative accuracy measures of the form of Equation 9.4, can easily be constructed using the MAE, MSE, or RMSE as the underlying accuracy statistics. Usually the reference, or control, forecasts are provided either by the climatological values of the predictand or by persistence (i.e., the previous value in a sequence of observations). For the MSE, the accuracies of these two references are, respectively,

$$\mathrm{MSE}_{\mathrm{clim}} = \frac{1}{n}\sum_{k=1}^{n}(\overline{o} - o_k)^2 \tag{9.36a}$$

and

$$\mathrm{MSE}_{\mathrm{pers}} = \frac{1}{n-1}\sum_{k=2}^{n}(o_{k-1} - o_k)^2. \tag{9.36b}$$

Analogous equations can be written for the MAE, in which the squaring function would be replaced by the absolute value function.

In Equation 9.36a, it is implied that the climatological average value does not change from forecast occasion to forecast occasion (i.e., as a function of the index, k). If this implication is true, then $\mathrm{MSE}_{\mathrm{clim}}$ in Equation 9.36a is an estimate of the sample variance of the predictand (compare Equation 3.6). In some applications the climatological value of the predictand will be different for different forecasts. For example, if daily temperature forecasts at a single location were being verified over the course of several months, the index k would represent time, and the climatological average temperature usually would change smoothly as a function of the date. In this case the quantity being summed in Equation 9.36a would be $(c_k - o_k)^2$, with c_k being the climatological value of the predictand on day k. Failing to account for a time-varying climatology would produce an unrealistically large $\mathrm{MSE}_{\mathrm{clim}}$, because the correct seasonality for the predictand would not be reflected (Hamill and Juras, 2006; Juras, 2000). The MSE for persistence in 9.36b implies that the index k represents time, so that the reference forecast for the observation o_k at time k is just the observation of the predictand during the previous time period, o_{k-1}.

Either of the reference measures for accuracy in Equations 9.36a or 9.36b, or their MAE counterparts, can be used in Equation 9.4 to calculate skill. Murphy (1992) advocates use of the more accurate reference forecasts to standardize the skill. For skill scores based on MSE, Equation 9.36a is more accurate (i.e., is smaller) if the lag-1 autocorrelation (Equation 3.36) of the time series of observations is smaller than 0.5, and Equation 9.36b is more accurate when the autocorrelation of the observations is larger than 0.5. For the MSE using climatology as the control forecasts, the skill score (in proportion rather than percentage terms) becomes

$$SS_{clim} = \frac{MSE - MSE_{clim}}{0 - MSE_{clim}} = 1 - \frac{MSE}{MSE_{clim}}. \tag{9.37}$$

Notice that perfect forecasts have MSE or MAE = 0, which allows the rearrangement of the skill score in Equation 9.37. By virtue of this second equality in Equation 9.37, SS_{clim} based on MSE is sometimes called the *reduction of variance* (RV), because the quotient being subtracted is the average squared error (or residual, in the nomenclature of regression) divided by the climatological variance (cf. Equation 7.17). Equation 9.37 is known as the *Nash-Sutcliffe efficiency* in the hydrology literature.

The skill score for the MSE in Equation 9.37 can be manipulated algebraically in a way that yields some insight into the determinants of forecast skill as measured by the MSE, with respect to climatology as the reference (Equation 9.36a). Rearranging Equation 9.37, and substituting an expression for the Pearson product-moment correlation between the forecasts and observations, $r_{y,o}$ (Equation 3.29), yields (Murphy, 1988; Murphy and Epstein, 1989)

$$SS_{clim} = r_{y,o}^2 - \left[r_{y,o} - \frac{s_y}{s_o}\right]^2 - \left[\frac{\overline{y} - \overline{o}}{s_o}\right]^2. \tag{9.38}$$

Equation 9.38 indicates that the skill in terms of the MSE can be regarded as consisting of a contribution due to the correlation between the forecasts and observations, and penalties relating to the calibration (reliability, or conditional bias) and unconditional bias of the forecasts. The first term in Equation 9.38 is the square of the Pearson correlation coefficient, and is a measure of the proportion of variability in the observations that is (linearly) accounted for by the forecasts. Here the squared correlation is similar to the R^2 in regression (Equation 7.17), although least-squares regressions are constrained to be unbiased by construction, whereas forecasts in general are not.

The second term in Equation 9.38 is a measure of reliability, or calibration, or conditional bias, of the forecasts. This is most easily appreciated by imagining a linear regression between the observations and the forecasts. The slope, b, of that linear regression equation can be expressed in terms of the correlation and the standard deviations of the predictor and predictand as $b = (s_o/s_y)r_{y,o}$. This relationship can be verified by substituting Equations 3.6 and 3.25 into Equation 7.7a. If this slope is smaller than $b = 1$, then the predictions made with this regression are too large (positively biased) for smaller forecasts, and too small (negatively biased) for larger forecasts. However, if $b = 1$, there will be no conditional bias, and substituting $b = (s_o/s_y)r_{y,o} = 1$ into the second term in Equation 9.38 yields a zero penalty for conditional bias.

The third term in Equation 9.38 is the square of the unconditional bias, as a fraction of the standard deviation of the observations, s_o. If the bias is small compared to the variability of the observations as measured by s_o the reduction in skill will be modest, whereas increasing bias of either sign progressively degrades the skill.

Thus if the forecasts are completely reliable and unbiased, the second and third terms in Equation 9.38 are both zero, and the skill score is exactly $r_{y,o}^2$. To the extent that the forecasts are biased or not correctly calibrated (exhibiting conditional biases), then the square of the correlation coefficient will overestimate skill. Squared correlation is accordingly best regarded as measuring potential skill.

9.3.3. Conditional Quantile Plots

It is possible and quite informative to graphically represent certain aspects of the joint distribution of nonprobabilistic forecasts and observations for continuous variables. The joint distribution contains a

large amount of information that is most easily absorbed from a well-designed graphical presentation. For example, Figure 9.8 shows *conditional quantile plots* for a sample of daily maximum temperature forecasts issued during the winters of 1980/1981 through 1985/1986 for Minneapolis, Minnesota. Panel (a) illustrates the performance of objective (MOS) forecasts, and panel (b) illustrates the performance of the corresponding subjective forecasts. These diagrams contain two parts, representing the two factors in the calibration-refinement factorization of the joint distribution of forecasts and observations (Equation 9.2).

The conditional distributions of the observations given each of the forecasts are represented in terms of selected quantiles, in comparison to the 1:1 diagonal line representing perfect forecasts. These have been presented as slightly smoothed versions of the raw values, although they could instead be represented using quantile regressions (Section 7.7.3). Here it can be seen that the MOS forecasts (panel a) exhibit a small degree of overforecasting (the conditional medians of the observed temperatures are consistently colder than the forecasts), but that the subjective forecasts are essentially unbiased. The histograms in the lower parts of the panels represent the frequency of use of the forecasts, or $p(y_i)$. Here it can be seen that the subjective forecasts are somewhat sharper, or more refined, with more extreme temperatures being forecast more frequently, especially on the left tail.

Figure 9.8a shows the same data that is displayed in the glyph scatterplot in Figure 3.24, and the bivariate histogram in Figure 3.25. However, the two figures in Chapter 3 show the data in terms of their joint distribution, whereas the calibration-refinement factorization plotted in Figure 9.8a allows an easy visual separation between the frequencies of use of each of the possible forecasts, and the distributions of temperature outcomes conditional on each forecast. The conditional quantile plot is an example of a diagnostic verification technique, because it allows diagnosis of particular strengths and weakness of this set of forecasts through exposition of the full joint distribution of the forecasts and observations. In particular, Figure 9.8 shows that the subjective forecasts have improved over the MOS forecasts, both by correcting the overforecasting of the colder temperatures, and by exhibiting better sharpness for the coldest forecasts. Comparison of scalar scores such as MSE for the two forecast sources would indicate that the subjective forecasts were more accurate, but would yield no information about the specific nature of the improvements, or how even better subjective forecasts might be achieved in the future.

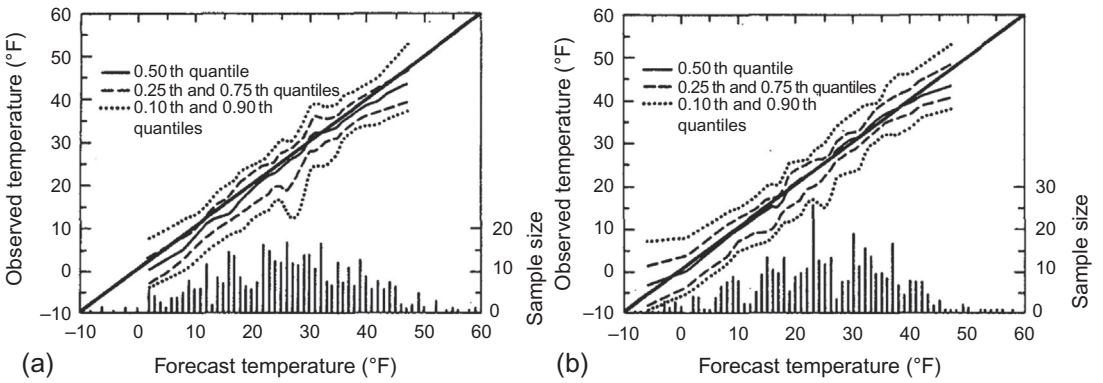

FIGURE 9.8 Conditional quantile plots for (a) objective and (b) subjective 24-h nonprobabilistic maximum temperature forecasts, for winter seasons of 1980 through 1986 at Minneapolis, Minnesota. Main body of the figures delineate smoothed quantiles from the conditional distributions $p(o_j \mid y_i)$ (i.e., the calibration distributions) in relation to the 1:1 line, and the lower parts of the figures show the unconditional distributions of the forecasts, $p(y_i)$ (the refinement distributions). *From Murphy et al. (1989).*

9.4. PROBABILITY FORECASTS FOR DISCRETE PREDICTANDS

9.4.1. The Joint Distribution for Dichotomous Events

Formulation and verification of probability forecasts for weather events have a long history, dating at least to Cooke (1906a) (Murphy and Winkler, 1984). Verification of probability forecasts is somewhat more subtle than verification of nonprobabilistic forecasts. Since nonprobabilistic forecasts contain no expression of uncertainty, it is clear whether an individual forecast is correct or not. However, unless a probability forecast is either 0.0 or 1.0, the situation is less clear-cut. For probability values between these two (certainty) extremes a single forecast is neither right nor wrong, so that meaningful assessments can only be made using collections of forecast and observation pairs. Again, it is the joint distribution of forecasts and observations that contains the relevant information for forecast verification.

The simplest setting for probability forecasts is in relation to dichotomous predictands, which are limited to $J = 2$ possible outcomes. The probability of precipitation (PoP) forecast is the most familiar example of probability forecasts for a dichotomous event. Here the event is either the occurrence (o_1) or nonoccurrence (o_2) of measurable precipitation. The joint distribution of forecasts and observations is more complicated than for the case of nonprobabilistic forecasts for binary predictands, however, because more than $I = 2$ probability values can allowably be forecast. In theory any real number between zero and one is an allowable probability forecast, but in practice the forecasts usually are rounded to one of a reasonably small number of values.

Table 9.4a contains a hypothetical joint distribution for probability forecasts of a dichotomous predictand, where the $I = 11$ possible forecasts might have been obtained by rounding continuous

TABLE 9.4 A Hypothetical Joint Distribution of Forecasts and Observations (a) for Probability Forecasts (Rounded to Tenths) for a Dichotomous Event, with (b) Its Calibration-Refinement factorization, and (c) Its Likelihood-Base Rate factorization

	(a) Joint Distribution		(b) Calibration-Refinement		(c) Likelihood-Base Rate	
y_i	$p(y_i, o_1)$	$p(y_i, o_2)$	$p(y_i)$	$p(o_1 \mid y_i)$	$p(y_i \mid o_1)$	$p(y_i \mid o_2)$
0.0	.045	.255	.300	.150	.152	.363
0.1	.032	.128	.160	.200	.108	.182
0.2	.025	.075	.100	.250	.084	.107
0.3	.024	.056	.080	.300	.081	.080
0.4	.024	.046	.070	.350	.081	.065
0.5	.024	.036	.060	.400	.081	.051
0.6	.027	.033	.060	.450	.091	.047
0.7	.025	.025	.050	.500	.084	.036
0.8	.028	.022	.050	.550	.094	.031
0.9	.030	.020	.050	.600	.101	.028
1.0	.013	.007	.020	.650	.044	.010
					$p(o_1) = .297$	$p(o_2) = .703$

probability assessments to the nearest tenth. Thus this joint distribution of forecasts and observations contains I x $J = 22$ individual probabilities. For example, on 4.5% of the forecast occasions a zero forecast probability was nevertheless followed by occurrence of the event, and on 25.5% of the occasions zero probability forecasts were correct in that the event o_1 did not occur.

Table 9.4b presents the same joint distribution in terms of the calibration-refinement factorization (Equation 9.2). That is, for each possible forecast probability, y_i, Table 9.4b presents the relative frequency with which that forecast value was used, $p(y_i)$, and the conditional probability that the event o_1 occurred given the forecast y_i, $p(o_1 | y_i)$, $i = 1, \ldots, I$. For example, $p(y_1) = p(y_1, o_1) + p(y_1, o_2) = .045 + .255 = .300$, and (using the definition of conditional probability, Equation 2.10) $p(o_1 | y_1) = p(y_1, o_1)/p(y_1) = .045/.300 = .150$. Because the predictand is binary it is not necessary to specify the conditional probabilities for the complementary event, o_2, given each of the forecasts. That is, since the two predictand values represented by o_1 and o_2 constitute a MECE partition of the sample space, $p(o_2 | y_i) = 1 - p(o_1 | y_i)$. Not all the $J = 11$ probabilities in the refinement distribution $p(y_i)$ can be specified independently either, since $\Sigma_j p(y_j) = 1$. Thus the joint distribution can be completely specified with I x $J - 1 = 21$ of the 22 probabilities given in either Table 9.4a or Table 9.4b, which is the dimensionality of this verification problem.

Similarly, Table 9.4c presents the likelihood-base Rate factorization (Equation 9.3) for the joint distribution in Table 9.4a. Since there are $J = 2$ MECE events, there are two conditional distributions $p(y_i | o_j)$, each of which includes $I = 11$ probabilities. Since these 11 probabilities must sum to 1, each conditional distribution is fully specified by any 10 of them. The refinement (i.e., sample climatological) distribution consists of the two complementary probabilities $p(o_1)$ and $p(o_2)$, and so can be completely defined by either of the two. Therefore the likelihood-base rate factorization is also fully specified by $10 + 10 + 1 = 21$ probabilities. The information in any of the three portions of Table 9.4 can be recovered fully from either of the others. For example, $p(o_1) = \Sigma_i p(y_i, o_1) = .297$, and $p(y_1 | o_1) = p(y_1, o_1)/p(o_1) = .045/.297 = .152$.

9.4.2. The Brier Score

Given the generally high dimensionality of verification problems involving probability forecasts even for dichotomous predictands (e.g., $I \times J - 1 = 21$ for Table 9.4), it is not surprising that such forecasts are often evaluated using scalar summary measures. Although attractive from a practical standpoint, such simplifications necessarily will give incomplete pictures of forecast performance. Multiple scalar accuracy measures for verification of probabilistic forecasts of dichotomous events exist (e.g., Bröcker, 2012a; Murphy and Daan, 1985), but by far the most common is the *Brier score* (BS). The Brier score is essentially the mean squared error of the probability forecasts, considering that the observation is $o_1 = 1$ if the event occurs, and that the observation is $o_2 = 0$ if the event does not occur. The score averages the squared differences between pairs of forecast probabilities and the subsequent binary observations,

$$\mathrm{BS} = \frac{1}{n} \sum_{k=1}^{n} (y_k - o_k)^2, \tag{9.39}$$

where the index k again denotes a numbering of the n forecast-event pairs. Comparing the Brier score with Equation 9.33 for the mean squared error, it can be seen that the two are completely analogous. As a mean-squared-error measure of accuracy, the Brier score is negatively oriented, with perfect forecasts exhibiting BS = 0. Less accurate forecasts receive higher Brier scores, but since individual

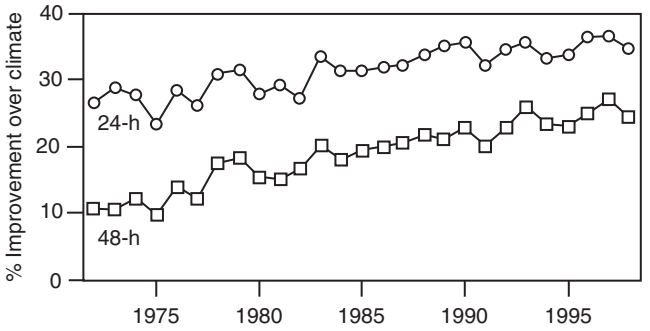

FIGURE 9.9 Trends in the skill of United States subjective PoP forecasts, measured in terms of the Brier score relative to climatological probabilities, April–September 1972–1998. *From www.nws.noaa.gov/tdl/synop.*

forecasts and observations are both bounded by zero and one, the score can take on values only in the range $0 \leq \mathrm{BS} \leq 1$.

The Brier score as expressed in Equation 9.39 is nearly universally used, but it differs from the score as originally introduced by Brier (1950) in that it averages only the squared differences pertaining to one of the two binary events. The original Brier score also included squared differences for the complementary (or non-) event in the average, with the result that Brier's original score is exactly twice that given by Equation 9.39. The confusion is unfortunate, but the usual present-day understanding of the meaning of Brier score is that given in Equation 9.39. In order to distinguish this from the original formulation, the Brier Score in Equation 9.39 sometimes is referred to as the *half-Brier score*.

Skill scores of the form of Equation 9.4 often are computed for the Brier score, yielding the *Brier Skill Score*

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{ref}}}{0 - \mathrm{BS}_{\mathrm{ref}}} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}, \tag{9.40}$$

since $\mathrm{BS}_{\mathrm{perf}} = 0$. This equation is analogous to Equation 9.37 for the MSE more generally. The BSS is the conventional skill-score form using the Brier score as the underlying accuracy measure. Usually the reference forecasts are the relevant climatological relative frequencies, which may vary with location and/or time of year (Hamill and Juras, 2006; Juras, 2000). Skill scores with respect to the climatological probabilities for subjective U.S. PoP forecasts during the warm seasons of 1972 through 1998 are shown in Figure 9.9. The labeling of the vertical axis as % improvement over climate indicates that it is the skill score in Equation 9.40, using climatological probabilities as the reference forecasts, that is plotted in the figure. According to this score, forecasts made for the 48-hour lead time in the 1990s exhibited skill comparable to 24-hour forecasts made in the 1970s.

9.4.3. Algebraic Decomposition of the Brier Score

An instructive algebraic decomposition of the Brier score (Equation 9.36) has been derived by Murphy (1973b). It relates to the calibration-refinement factorization of the joint distribution, Equation 9.2, in that it pertains to quantities that are conditional on particular values of the forecasts.

As before, consider that a verification data set contains forecasts taking on any of a discrete number, I, of forecast values y_i. For example, in the verification data set in Table 9.4, there are $I = 11$ allowable

forecast values, ranging from $y_1 = 0.0$ to $y_{11} = 1.0$. Let N_i be the number of times each forecast y_i is used in the collection of forecasts being verified. The total number of forecast-event pairs is simply the sum of these subsample, or conditional sample, sizes,

$$n = \sum_{i=1}^{I} N_i. \tag{9.41}$$

The marginal distribution of the forecasts—the refinement in the calibration-refinement factorization—consists simply of the relative frequencies

$$p(y_i) = \frac{N_i}{n}. \tag{9.42}$$

The first column in Table 9.4b shows these relative frequencies for the data set represented there.

For each of the subsamples delineated by the I allowable forecast values there is a relative frequency of occurrence of the forecast event. Since the observed event is dichotomous, a single conditional relative frequency defines the conditional distribution of observations given each forecast y_i. It is convenient to think of this relative frequency as the subsample relative frequency, or conditional average observation,

$$\overline{o}_i = p(o_1 | y_i) = \frac{1}{N_i} \sum_{k \in N_i} o_k, \tag{9.43}$$

where $o_k = 1$ if the event occurs for the kth forecast-event pair, $o_k = 0$ if it does not, and the summation is over only those values of k corresponding to occasions when the forecast y_i was issued. The second column in Table 9.4b shows these conditional relative frequencies. Similarly, the overall (unconditional) relative frequency, or sample climatology, of the observations is given by

$$\overline{o} = \frac{1}{n} \sum_{k=1}^{n} o_k = \frac{1}{n} \sum_{i=1}^{I} N_i \overline{o}_i. \tag{9.44}$$

After some algebra, the Brier score in Equation 9.39 can be expressed in terms of the quantities just defined as the sum of the three terms

$$\mathrm{BS} = \underbrace{\frac{1}{n} \sum_{i=1}^{I} N_i (y_i - \overline{o}_i)^2}_{\text{("Reliability")}} - \underbrace{\frac{1}{n} \sum_{i=1}^{I} N_i (\overline{o}_i - \overline{o})^2}_{\text{("Resolution")}} + \underbrace{\overline{o}\,(1 - \overline{o})}_{\text{("Uncertainty")}}. \tag{9.45}$$

As indicated in this equation, these three terms are known as Reliability, Resolution, and Uncertainty. Since more accurate forecasts are characterized by smaller values of BS, a forecaster would like the reliability term to be as small as possible, and the resolution term to be as large (in absolute value) as possible. Equation 9.44 shows that the uncertainty term depends only on the sample climatological relative frequency, and so is unaffected by the forecasts. The reliability and resolution terms in Equation 9.45 sometimes are used individually as scalar measures of these two aspects of forecast quality, and called REL and RES, respectively. If these two measures are normalized by dividing each by the uncertainty term, their difference equals the Brier skill score BSS, as in Equation 9.46.

The reliability term in Equation 9.45 summarizes the calibration, or conditional bias, of the forecasts. It consists of a weighted average of the squared differences between the forecast probabilities and the

relative frequencies of the observed event in each subsample. For forecasts that are perfectly reliable, the subsample relative frequency is exactly equal to the forecast probability in each subsample. The relative frequency of the event should be small on occasions when $y_1 = 0.0$ is forecast, and should be large when $y_I = 1.0$ is forecast. On those occasions when the forecast probability is 0.5, the relative frequency of the event should be near 1/2. For reliable, or well-calibrated forecasts, all the squared differences in the reliability term will be near zero, and their weighted average will be small.

The resolution term in Equation 9.45 summarizes the ability of the forecasts to discern subsample forecast periods with relative frequencies of the event that are different from each other. The forecast probabilities y_i do not appear explicitly in this term, yet it still depends on the forecasts through the sorting of the events making up the subsample relative frequencies (Equation 9.43). Mathematically, the resolution term is a weighted average of the squared differences between these subsample relative frequencies, and the overall sample climatological relative frequency. If the forecasts sort the observations into subsamples having substantially different relative frequencies than the overall sample climatology, the resolution term will be large. This is a desirable situation, since the resolution term is subtracted in Equation 9.45. Conversely, if the forecasts sort the events into subsamples with very similar event relative frequencies, the squared differences in the summation of the resolution term will be small. In that case the forecasts resolve the event only weakly, and the resolution term will be small.

The uncertainty term in Equation 9.45 depends only on the variability of the observations and cannot be influenced by anything the forecaster may do. This term is identical to the variance of the Bernoulli (binomial, with $N = 1$) distribution (see Table 4.3), exhibiting minima of zero when the climatological probability is either zero or one, and a maximum when the climatological probability is 1/2. When the event being forecast almost never happens, or almost always happens, the uncertainty in the forecasting situation is small. In these cases, always forecasting the climatological probability will give generally good results. When the climatological probability is closer to 1/2 there is substantially more uncertainty inherent in the forecasting situation, and the third term in Equation 9.45 is commensurately larger.

Equation 9.45 is an exact decomposition of the Brier score when the allowable forecast values are only the I probabilities y_i. This is the case when, for example, human forecasters are required to round their judgments to fixed fractions such as tenths. When a richer set of probabilities, derived perhaps from relative frequencies within a large forecast ensemble, or a logistic regression, have been rounded into I bins, Equation 9.45 will not balance exactly if BS on the left-hand side has been computed using the unrounded values. However, the resulting discrepancy can be quantified using two additional terms (Stephenson et al., 2008b). The three terms on the right-hand side of Equation 9.45 exhibit biases that may be appreciable if the subsample sizes N_i are not large (Bröcker, 2012b, 2012c), although corrections are available (Ferro and Fricker, 2012; Siegert, 2014).

The algebraic decomposition of the Brier score in Equation 9.45 is interpretable in terms of the calibration-refinement factorization of the joint distribution of forecasts and observations (Equation 9.2), as will become clear in Section 9.4.4. Murphy and Winkler (1987) also proposed a different three-term algebraic decomposition of the mean-squared error (of which the Brier score is a special case), based on the likelihood-base rate factorization (Equation 9.3), which has been applied to the Brier score for the data in Table 9.2 by Bradley et al. (2003).

9.4.4. The Reliability Diagram

Single-number summaries of forecast performance such as the Brier score can provide a convenient quick impression, but a comprehensive appreciation of forecast quality can be achieved only through

the full joint distribution of forecasts and observations. Because of the typically large dimensionality ($= I \times J - 1$) of these distributions their information content can be difficult to absorb from numerical tabulations such as those in Tables 9.2 or 9.4, but becomes conceptually accessible when presented in a well-designed graphical format. The *reliability diagram*, apparently introduced by Sanders (1963), is one such graphical device. It shows the full joint distribution of forecasts and observations for probability forecasts of a binary predictand, in terms of its calibration-refinement factorization (Equation 9.2). Accordingly, it is the counterpart of the conditional quantile plot (Section 9.3.3) for nonprobabilistic forecasts of continuous predictands. The fuller picture of forecast performance portrayed in the reliability diagram as compared to a scalar summary, such as BSS, allows diagnosis of particular strengths and weaknesses in a verification data set.

The two elements of the calibration-refinement factorization are the calibration distributions, or conditional distributions of the observation given each of the I allowable values of the forecast, $p(o_j|y_i)$; and the refinement distribution $p(y_i)$, expressing the frequency of use of each of the possible forecasts. Each of the calibration distributions is a Bernoulli (binomial, with $N = 1$) distribution, because there is a single binary outcome O on each forecast occasion, and for each forecast y_i the probability of the outcome o_1 is the conditional probability $p(o_1|y_i)$. This probability fully defines the corresponding Bernoulli distribution, because $p(o_2|y_i) = 1 - p(o_1|y_i)$. Taken together, these I calibration probabilities $p(o_1|y_i)$ define a *calibration function*, which expresses the conditional probability of the event o_1 as a function of the forecast y_i. In some settings, the forecasts to be evaluated have been rounded to a prespecified set of I probabilities before being issued. However, when the probability forecasts being evaluated are continuous, and so can take on any value on the unit interval, the number of bins I must be chosen in order to plot a reliability diagram. Bröcker (2008) suggests optimizing this choice by minimizing the cross-validated BS of the probability forecasts rounded to I discrete values. The forecasts y_i, $i = 1, \ldots, I$, corresponding to each bin are most consistently calculated as the average forecast probability within each bin (Bröcker, 2008; Bröcker and Smith, 2007b).

The first element of a reliability diagram is a plot of the calibration function, usually as I points connected by line segments for visual clarity. Figure 9.10a shows five characteristic forms for this portion of the reliability diagram, which allows immediate visual diagnosis of unconditional and conditional biases that may be exhibited by the forecasts in question. The center panel in Figure 9.10a shows the characteristic signature of well-calibrated forecasts, in which the conditional event relative frequency is essentially equal to the forecast probability. That is, $p(o_1|y_i) \approx y_i$, so that the I dots fall along the dashed 1:1 line except for deviations consistent with sampling variability. Well-calibrated probability forecasts "mean what they say," in the sense that subsequent event relative frequencies are essentially equal to the forecast probabilities. In terms of the algebraic decomposition of the Brier score (Equation 9.45), such forecasts exhibit excellent reliability, because the squared differences in the reliability term correspond to squared vertical distances between the dots and the 1:1 line in the reliability diagram. These distances are all small for well-calibrated forecasts, yielding a small reliability term, which is a weighted average of the I squared vertical distances.

The top and bottom panels in Figure 9.10a show characteristic forms of the calibration function for forecasts exhibiting unconditional biases. In the top panel, the calibration function is entirely to the right of the 1:1 line, indicating the forecasts are consistently too large relative to the conditional event relative frequencies, so that the average forecast is larger than the average observation (Equation 9.44). This pattern is the signature of overforecasting, or if the predictand is precipitation occurrence, a wet bias. Similarly, the bottom panel in Figure 9.10a shows the characteristic signature of underforecasting, or a dry bias, because the calibration function being entirely to the left of the 1:1 line indicates that the

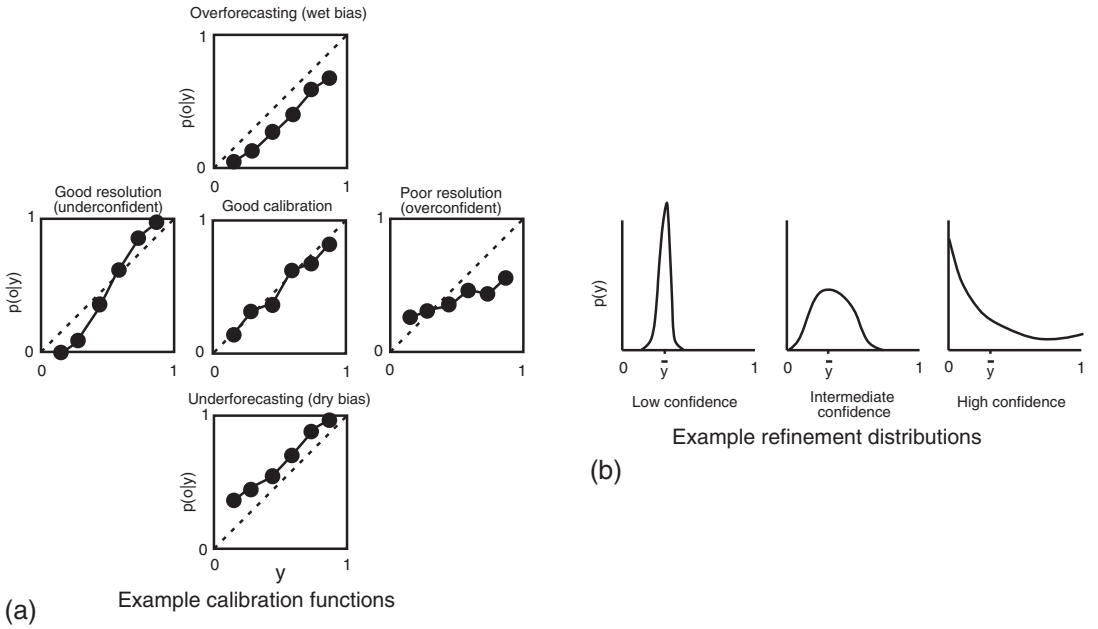

FIGURE 9.10 Example characteristic forms for the two elements of the reliability diagram. (a) Calibration functions, showing calibration distributions $p(o \mid y)$ (i.e., conditional Bernoulli probabilities), as functions of the forecast y. (b) Refinement distributions, $p(y)$, reflecting aggregate forecaster confidence.

forecast probabilities are consistently too small relative to the corresponding conditional event relative frequencies given by $p(o_1 \mid y_i)$, and so the average forecast is smaller than the average observation. Forecasts that are unconditionally biased in either of these two ways are miscalibrated, or not reliable, in the sense that the conditional event probabilities $p(o_1 \mid y_i)$ do not correspond well to the stated probabilities y_i. The vertical distances between the points and the dashed 1:1 line are nonnegligible, leading to substantial squared differences in the first summation of Equation 9.45, and thus to a large reliability term in that equation.

The deficiencies in forecast performance indicated by the calibration functions in the left and right panels of Figure 9.10a are more subtle and indicate conditional biases. That is, the sense and/or magnitudes of the biases exhibited by forecasts having these types of calibration functions depend on the forecasts themselves. In the left ("good resolution") panel, there are overforecasting biases associated with smaller forecast probabilities and underforecasting biases associated with larger forecast probabilities, and the reverse is true of the calibration function in the right ("poor resolution") panel. The calibration function in the right panel of Figure 9.10a is characteristic of forecasts showing poor resolution in the sense that the conditional outcome relative frequencies $p(o_1 \mid y_i)$ depend only weakly on the forecasts and are all near the climatological probability. (That the climatological relative frequency is somewhere near the center of the vertical locations of the points in this panel can be appreciated from the law of total probability (Equation 2.14), which expresses the unconditional climatology as a weighted average of these conditional relative frequencies.) Because the differences in this panel between the calibration probabilities $p(o_1 \mid y_i)$ (Equation 9.43) and the overall sample climatology are small, the resolution term in Equation 9.45 is small, reflecting the fact that these forecasts resolve the event o_1

poorly. Because the sign of this term in Equation 9.45 is negative, poor resolution leads to larger (worse) Brier scores.

Conversely, the calibration function in the left panel of Figure 9.10a indicates good resolution, in the sense that the weighted average of the squared vertical distances between the points and the sample climatology in the resolution term of Equation 9.45 is large. Here the forecasts are able to identify subsets of forecast occasions for which the outcomes are quite different from each other. For example, small but nonzero forecast probabilities have identified a subset of forecast occasions when the event o_1 did not occur at all. However, the forecasts are conditionally biased, and so mislabeled, and therefore not well calibrated. Their Brier score would be penalized for this miscalibration through a substantial positive value for the reliability term in Equation 9.45.

The labels underconfident and overconfident in the left and right panels of Figure 9.10a can be understood in relation to the other element of the reliability diagram, namely, the refinement distribution $p(y_i)$. The dispersion of the refinement distribution reflects the overall confidence of the forecaster, as indicated in Figure 9.10b. Forecasts that deviate rarely and quantitatively little from their average value (left panel) exhibit little confidence. Forecasts that are frequently extreme—that is, specifying probabilities close to the certainty values $y_1 = 0$ and $y_I = 1$ (right panel)—exhibit high confidence. However, the degree to which a particular level of forecaster confidence may be justified will be evident only from inspection of the calibration function for the same forecasts. The forecast probabilities in the right-hand ("overconfident") panel of Figure 9.10a are mislabeled in the sense that the extreme probabilities are too extreme. Outcome relative frequencies following probability forecasts near 1 are substantially smaller than 1, and outcome relative frequencies following forecasts near 0 are substantially larger than 0. A calibration-function slope that is shallower than the 1:1 reference line is diagnostic for overconfident forecasts, because correcting the forecasts to bring the calibration function into the correct orientation would require adjusting extreme probabilities to be less extreme, thus shrinking the dispersion of the refinement distribution, which would connote less confidence. Conversely, the underconfident forecasts in the left panel of Figure 9.10a could achieve good reliability (calibration function aligned with the 1:1 line) by adjusting the forecast probabilities to be more extreme, thus increasing the dispersion of the refinement distribution and connoting greater confidence.

A reliability diagram consists of plots of both the calibration function and the refinement distribution, and so is a full graphical representation of the joint distribution of the forecasts and observations, through its calibration-refinement factorization. Figure 9.11 shows two reliability diagrams, for seasonal (three-month) forecasts for (a) average temperatures and (b) total precipitation above the climatological terciles (outcomes in the warm and wet 1/3 of the respective local climatological distributions), for global land areas equatorward of 30° (Mason et al., 1999). The most prominent feature of Figure 9.11 is the substantial cold (underforecasting) bias evident for the temperature forecasts. The period 1997 through 2000 was evidently substantially warmer than the preceding several decades that defined the reference climate. The relative frequency of the observed warm outcome was about 0.7 (rather than the long-term climatological value of 1/3), but Figure 9.11a shows clearly that that warmth was not anticipated by these forecasts, in aggregate. There is also an indication of conditional bias in the temperature forecasts, with the overall calibration slope being slightly shallower than 45°, and so reflecting some forecast overconfidence. The precipitation forecasts (Figure 9.11b) are better calibrated, showing only a slight overforecasting (wet) bias and a more nearly correct overall slope for the calibration function. The refinement distributions (insets, with logarithmic vertical scales) show much more confidence (more frequent use of more extreme probabilities) for the temperature forecasts.

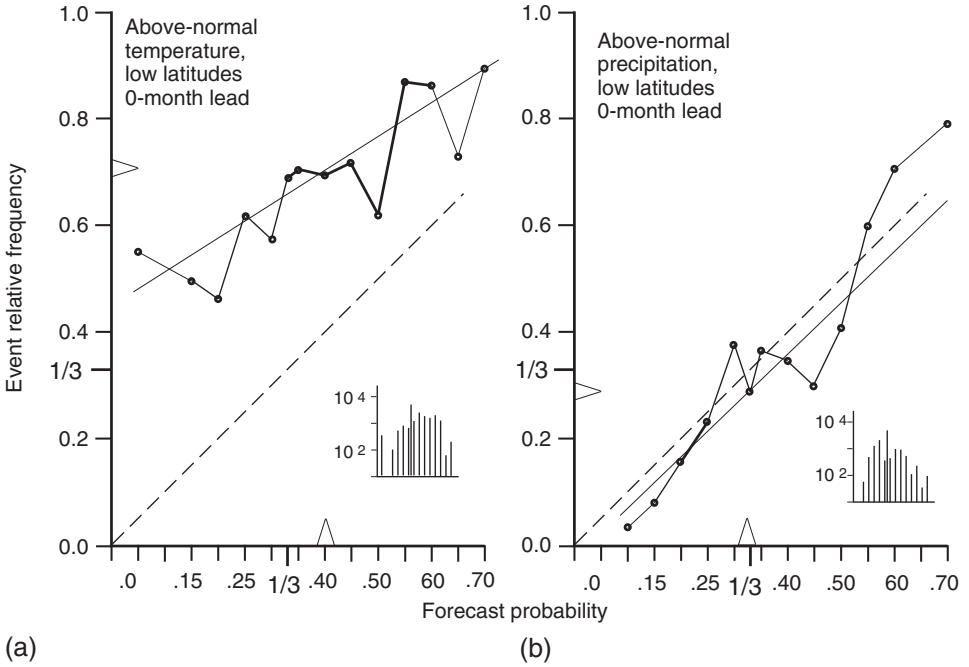

FIGURE 9.11 Reliability diagrams for seasonal (three-month) forecasts of (a) average temperature warmer than the climatological upper tercile, and (b) total precipitation wetter than the climatological upper tercile, for global land areas equatorward of 30°, during the period 1997–2000. *From Wilks and Godfrey (2002).*

The reliability diagrams in Figure 9.11 include some additional features that are not always plotted in reliability diagrams, which help interpret the results. The light lines through the calibration functions show weighted (to make points with larger subsample size N_i more influential) least-squares regressions (Murphy and Wilks, 1998), which help guide the eye through the irregularities that are due at least in part to sampling variations. In order to emphasize the better-estimated portions of the calibration function, the line segments connecting points based on larger sample sizes have been drawn more heavily. Finally, the average forecasts are indicated by the triangles on the horizontal axes, and the average observations are indicated by the triangles on the vertical axes, which emphasize the strong underforecasting of temperature in Figure 9.11a. Other modifications to the basic reliability diagram that have been suggested include plotting Brier Score contours on the diagram (Ehrendorfer, 1997) and plotting conditional boxplots in place of dots locating the conditional means (Bentzien and Friederichs, 2012). Use of nonuniformly spaced bins containing equal numbers of forecasts has also been suggested (Bröcker and Smith, 2007b), although the resulting Brier score may not be proper (Mitchell and Ferro, 2017) in the sense explained in Section 9.4.8.

An elaboration of the reliability diagram (Hsu and Murphy, 1986) includes reference lines related to the algebraic decomposition of the Brier score (Equation 9.45) and the Brier skill score (Equation 9.40), in addition to plots of the calibration function and the refinement distribution. This version of the reliability diagram is called the *attributes diagram*, an example of which (for the joint distribution in Table 9.2) is shown in Figure 9.12. The horizontal "no-resolution" line in the attributes diagram relates to the resolution term in Equation 9.45. Geometrically, the ability of a set of forecasts to identify event

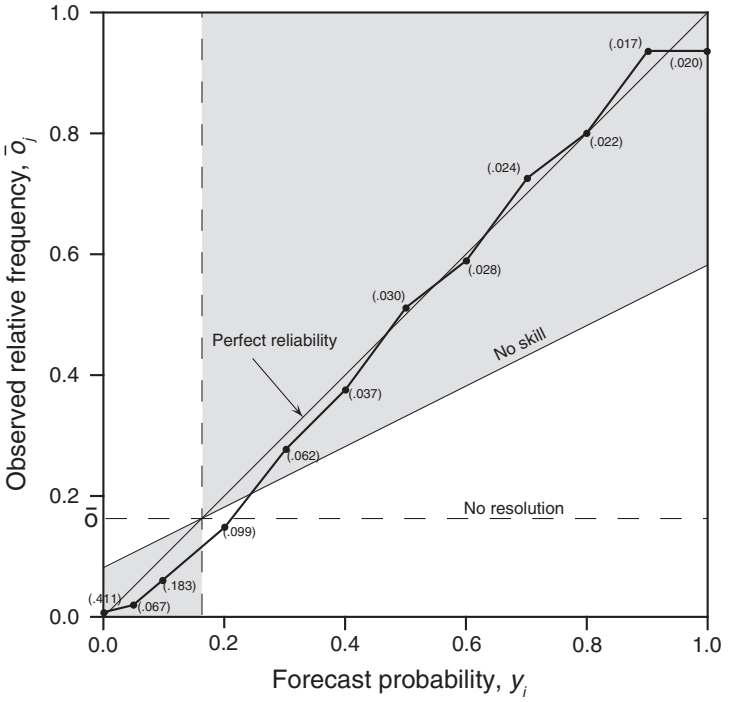

FIGURE 9.12 Attributes diagram for the $n = 12{,}402$ PoP forecasts summarized in Table 9.2. Solid dots show observed relative frequency of precipitation occurrence, conditional on each of the $I = 12$ possible probability forecasts. Forecasts not defining event subsets with different relative frequencies of the forecast event would exhibit all points on the dashed no-resolution line, which is plotted at the level of the sample climatological probability. Points in the shaded region bounded by the line labeled "no skill" contribute positively to forecast skill, according to Equation 9.40. Relative frequencies of use of each of the forecast values, $p(y_i)$, are shown parenthetically, although they could also have been indicated graphically.

subsets with different relative frequencies produces points in the attributes diagram that are well removed, vertically, from the level of the overall sample climatology, which is indicated by the no-resolution line. Points falling on the no-resolution line indicate forecasts y_i that are unable to resolve occasions where the event is more or less likely than the overall climatological probability. The weighted average making up the resolution term is of the squares of the vertical distances between the points (the subsample relative frequencies) and the no-resolution line. These distances will be large for forecasts exhibiting good resolution, in which case the resolution term will contribute to a small (i.e., good) Brier score. The forecasts summarized in Figure 9.12 exhibit a substantial degree of resolution, with forecasts that are most different from the sample climatological probability of 0.162 making the largest contributions to the resolution term.

Another interpretation of the uncertainty term in Equation 9.45 emerges from imagining the attributes diagram for climatological forecasts, that is, constant forecasts of the sample climatological relative frequency, Equation 9.44. Since only a single forecast value would ever be used in this case, there would be only $I = 1$ dot on such a diagram. The horizontal position of this dot would be at the constant forecast value, and the vertical position of the single dot would be at the same sample climatological relative frequency. This single point would be located at the intersection of the 1:1 (perfect reliability),

no-skill and no-resolution lines. Thus climatological forecasts have perfect (zero, in Equation 9.45) reliability, since the forecast and the conditional relative frequency (Equation 9.43) are both equal to the climatological probability (Equation 9.44). However, the climatological forecasts also have zero resolution since the existence of only $I = 1$ forecast category precludes discerning different subsets of forecasting occasions with differing relative frequencies of the outcomes. Since the reliability and resolution terms in Equation 9.45 are both zero, it is clear that the Brier score for climatological forecasts is exactly the uncertainty term in Equation 9.45.

This observation of the equivalence of the uncertainty term and the BS for climatological forecasts has interesting consequences for the Brier skill score in Equation 9.40. Substituting Equation 9.45 for BS into Equation 9.40, and uncertainty for $\mathrm{BS}_{\mathrm{ref}}$ yields

$$\mathrm{BSS} = \frac{\text{“Resolution”} - \text{“Reliability”}}{\text{“Uncertainty”}}. \tag{9.46}$$

Since the uncertainty term is always positive, the probability forecasts will exhibit positive skill in the sense of Equation 9.40 if the resolution term is larger in absolute value than the reliability term. This means that subsamples of the forecasts identified by the forecasts y_i will contribute positively to the overall skill when their resolution term is larger than their reliability term. Geometrically, this corresponds to points on the attributes diagram being closer to the 1:1 perfect-reliability line than to the horizontal no-resolution line. This condition defines the no-skill line, which is midway between the perfect-reliability and no-resolution lines, and delimits the shaded region, in which subsamples contribute positively to forecast skill, according to BSS. In Figure 9.12 only the subsample for $y_4 = 0.2$, which is nearly equal to the climatological probability, fails to contribute positively to the overall BSS.

Note that forecasts whose calibration functions lie outside the shaded region in an attributes diagram are not necessarily useless. Zero or negative skill according to BSS or indeed any other scalar measure may still be consistent with positive economic value for some users, since it is possible for forecasts with lower BSS to be more valuable for some users (e.g., Murphy and Ehrendorfer, 1987). At minimum, forecasts exhibiting a calibration function with positive slope different from 1 have the potential for *recalibration*, which is most often achieved by relabeling each of the forecasts y_i with the corresponding conditional relative frequencies $\overline{o}_i$ (Equation 9.43) defining the calibration function in a previously analyzed training data set. For overconfident forecasts, the recalibration process comes at the expense of sharpness, although sharpness is increased when underconfident forecasts are recalibrated. Bröcker (2008) suggests recalibration using a kernel-smoothing (Section 3.3.6) estimate of the calibration function, which is an appealing approach for continuously varying probability forecasts. Van den Dool et al. (2017) suggest that such miscalibrations might be addressed using a linear regression postprocessing step.

9.4.5. The Discrimination Diagram

The joint distribution of forecasts and observations can also be displayed graphically through the likelihood-base rate factorization (Equation 9.3). For probability forecasts of dichotomous ($J = 2$) predictands, this factorization consists of two conditional likelihood distributions $p(y_i \mid o_j)$, $j = 1, 2$; and a base rate (i.e., sample climatological) distribution $p(o_j)$ consisting of the relative frequencies for the two dichotomous events in the verification sample.

The *discrimination diagram* displays superimposed plots of the two likelihood distributions, as functions of the forecast probability y_i, together with a specification of the sample climatological probabilities $p(o_1)$ and $p(o_2)$. Together, these quantities completely represent the information in the full joint distribution. Therefore the discrimination diagram presents the same information as the reliability diagram, but in a different format.

Figure 9.13 shows an example discrimination diagram, for the probability-of-precipitation forecasts whose calibration-refinement factorization is presented in Table 9.2 and whose attributes diagram is shown in Figure 9.12. The probabilities in the two likelihood distributions calculated from their joint distribution are presented in Table 15.2. Clearly the conditional probabilities given the "no precipitation" event o_2 are greater for the smaller forecast probabilities, and the conditional probabilities given the "precipitation" event o_1 are greater for the intermediate and larger probability forecasts. Forecasts that discriminated perfectly between the two events would exhibit no overlap in their likelihoods. The two likelihood distributions in Figure 9.13 overlap somewhat, but exhibit substantial separation, indicating substantial discrimination by the forecasts of the dry and wet events.

The separation of the two likelihood distributions in a discrimination diagram can be summarized by the absolute difference between their means, called the *discrimination distance*,

$$d = \left| \mu_{y|o_1} - \mu_{y|o_2} \right|. \tag{9.47}$$

For the two conditional distributions in Figure 9.13 this difference is $d = |0.567–0.101| = 0.466$, which is also plotted in the figure. This distance is zero if the two likelihood distributions are the same (i.e., if the forecasts cannot discriminate the event at all), and increases as the two likelihood distributions become more distinct. In the limit $d = 1$ for perfect forecasts, which have all probability concentrated at $p(1\,|\,o_1) = 1$ and $p(0\,|\,o_2) = 1$.

There is a connection between the likelihood distributions in the discrimination diagram and statistical discrimination as discussed in Chapter 15. In particular, the two likelihood distributions in Figure 9.13 could be used together with the sample climatological probabilities, as in Section 15.3.3,

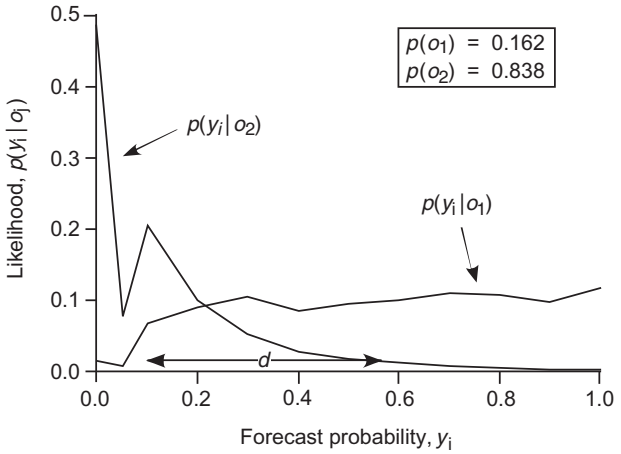

FIGURE 9.13 Discrimination diagram for the data in Table 9.2, which is shown in likelihood-base rate form in Table 15.2. The discrimination distance d (Equation 9.47) is also indicated.

to recalibrate these probability forecasts by calculating posterior probabilities for the two events given each of the possible forecast probabilities (cf. Exercise 15.5).

9.4.6. The ROC Diagram

The ROC (*relative operating characteristic*, or *receiver operating characteristic*) diagram is another discrimination-based graphical forecast verification display, although unlike the reliability diagram and discrimination diagram it does not include the full information contained in the joint distribution of forecasts and observations. The ROC diagram was first introduced into the meteorological literature by Mason (1982), although it has a longer history of use in such disciplines as psychology (Swets, 1973) and medicine (Pepe, 2003; Swets, 1979), after arising from signal detection theory in electrical engineering. The ROC diagram is a square plot, with the false alarm rate F (Equation 9.13) on the horizontal axis and the hit rate H (Equation 9.12) on the vertical.

One way to view the ROC diagram and the ideas behind it is in relation to the class of idealized decision problems outlined in Section 9.9.1. Here hypothetical decision makers must choose between two alternatives on the basis of a probability forecast for a dichotomous variable, with one of the decisions (say, action A) being preferred if the event o_1 occurs, and the other (action B) being preferable if the event does not occur. As explained in Section 9.9.1, the probability threshold determining which of the two decisions will be optimal depends on the decision problem, and in particular on the relative undesirability of having taken action B when the event occurs versus action A when the event does not occur. Therefore different probability thresholds for the choice between actions A and B will be appropriate for different decision problems.

If the forecast probabilities y_i have been issued as, or rounded to, I discrete values, there are $I - 1$ such thresholds, excluding the trivial cases of always taking action A or always taking action B. Operating on the joint distribution of forecasts and observations (e.g., Table 9.4a) consistent with each of these probability thresholds yields $I - 1$ contingency tables of dimension 2 x 2 of the kind treated in Section 9.2. A "yes" forecast is imputed if the probability y_i is above the threshold in question (sufficient probability to warrant a nonprobabilistic forecast of the event, for those decision problems appropriate to that probability threshold), and a "no" forecast is imputed if the forecast probability is below the threshold (insufficient probability for a nonprobabilistic forecast of the event). The mechanics of constructing these 2 x 2 contingency tables are exactly as illustrated in Example 9.2.

ROC diagrams are constructed by evaluating each of the $I - 1$ contingency tables using the hit rate H (Equation 9.12) and the false alarm rate F (Equation 9.13). As the hypothetical decision threshold is increased from lower to higher probabilities there are progressively more "no" forecasts and progressively fewer "yes" forecasts, yielding corresponding decreases in both H and F. In terms of the two likelihood distributions $p(o_1 \mid y_i)$ for the "yes" event and $p(o_2 \mid y_i)$ for the "no" event (e.g., Figure 9.13),

$$H = \int_{y^*}^{1} p(o_1 | y) dy \tag{9.48a}$$

and

$$F = \int_{y^*}^{1} p(o_2 | y) dy, \tag{9.48b}$$

where y^* is the decision threshold. The resulting $I - 1$ point pairs (F_i, H_i) are then plotted and connected with line segments, and also connected to the point (0, 0) corresponding to never forecasting the event (i.e., always choosing action A), and to the point (1,1) corresponding to always forecasting the event (always choosing action B).

The ability of a set of probability forecasts to discriminate a dichotomous event can be easily appreciated from its ROC diagram. Consider first the ROC diagram for perfect forecasts, which use only $I = 2$ probabilities, $y_1 = 0.00$ and $y_2 = 1.00$. For such forecasts there is only one probability threshold from which to calculate a 2 x 2 contingency table. That table for perfect forecasts exhibits $F = 0.0$ and $H = 1.0$, so its ROC curve consists of two line segments coincident with the left boundary and the upper boundary of the ROC diagram. At the other extreme of forecast performance, random forecasts consistent with the sample climatological probabilities $p(o_1)$ and $p(o_2)$ will exhibit $F_i = H_i$ regardless of how many or how few different probabilities y_i are used, and so their ROC curve will fall near (within sampling variability) the 45° diagonal connecting the points (0, 0) and (1,1). ROC curves for real forecasts usually fall between these two extremes, lying above and to the left of the 45° diagonal. Forecasts with better discrimination exhibit ROC curves approaching the upper-left corner of the ROC diagram more closely, whereas forecasts with very little ability to discriminate the event o_1 exhibit ROC curves very close to the $H = F$ diagonal.

It can be convenient to summarize a ROC diagram using a single scalar value, and the usual choice for this purpose is the area under the ROC curve, A. Since ROC curves for perfect forecasts pass through the upper-left corner, the area under a perfect ROC curve includes the entire unit square, so $A_{\text{perf}} = 1$. Similarly ROC curves for random forecasts lie along the 45° diagonal of the unit square, yielding the area $A_{\text{rand}} = 0.5$. The area A under a ROC curve of interest can therefore also be expressed in standard skill-score form (Equation 9.4), as

$$\text{SS}_{\text{ROC}} = \frac{A - A_{\text{rand}}}{A_{\text{perf}} - A_{\text{rand}}} = \frac{A - 1/2}{1 - 1/2} = 2A - 1. \tag{9.49}$$

Marzban (2004) describes some characteristics of forecasts that can be diagnosed from the shapes of their ROC curves, based on analysis of some simple idealized discrimination diagrams. Symmetrical ROC curves result when the two likelihood distributions $p(y_i|o_1)$ and $p(y_i|o_2)$ have similar dispersion, or widths, so the ranges of the forecasts y_i corresponding to each of the two outcomes are comparable. On the other hand, asymmetrical ROC curves, which might intersect either the vertical or horizontal axis at either $H \approx 0.5$ or $F \approx 0.5$, respectively, are indicative of one of the two likelihoods being substantially more concentrated than the other. Marzban (2004) also finds that A (or, equivalently, SS_{ROC}) is a reasonably good discriminator among relatively low-quality forecasts, but that relatively good forecasts tend to be characterized by quite similar (near-unit) areas under their ROC curves.

Example 9.5. Two Example ROC Curves

Example 9.2 illustrated the conversion of the probabilistic forecasts summarized by the joint distribution in Table 9.2 to nonprobabilistic yes/no forecasts, using a probability threshold y^* between $y_3 = 0.1$ and $y_4 = 0.2$. The resulting 2 x 2 contingency table consists of (cf. Figure 9.1a) $a = 1828$, $b = 2369$, $c = 181$, and $d = 8024$; yielding $F = 2369/(2369+8024) = 0.228$ and $H = 1828/(1828+181) = 0.910$. This point is indicated by the dot on the ROC curve for the Table 9.2 data in Figure 9.14. The entire ROC curve for the Table 9.2 data consists of this and all other partitions of these forecasts into yes/no forecasts using different probability thresholds. For example, the point just to the left of (0.228, 0.910) on this ROC curve is obtained by moving the threshold between $y_4 = 0.2$ and $y_5 = 0.3$. This partition produces $a = 1644$, $b = 1330$, $c = 364$, and $d = 9064$, defining the point $(F, H) = (0.128, 0.819)$.

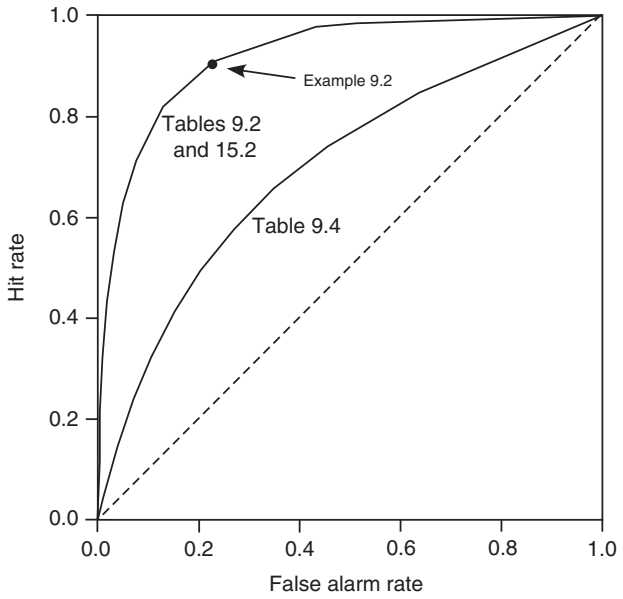

FIGURE 9.14 ROC diagrams for the PoP forecasts in Tables 9.2 and 15.2 (upper solid curve), and the hypothetical forecasts in Table 9.4 (lower solid curve). Solid dot locates the (F, H) pair corresponding to the probability threshold in Example 9.2.

Summarizing ROC curves according to the areas underneath them is usually accomplished through summation of the areas under each of the I trapezoids defined by the point pairs (F_i, H_i), $i = 1, \ldots, I-1$, together with the two endpoints (0, 0) and (1, 1). For example, the trapezoid defined by the dot in Figure 9.14 and the point just to its left has area $0.5(0.910+0.819)(0.228–0.128) = 0.08645$. This area, together with the areas of the other $I - 1 = 11$ trapezoids defined by the segments of the ROC curve for these data yield the total area $A = 0.922$.

The ROC curve, and the area under it, can also be computed directly from the joint probabilities in $p(y_i, o_j)$, that is, without knowing the sample size n. Table 9.5 summarizes the conversion of the hypothetical joint distribution in Table 9.4a to the $I - 1 = 10$ sets of 2 x 2 tables, by operating directly on the joint probabilities. Note that these data have one fewer forecast value y_i than those in Table 9.2, because in Table 9.2 the forecast $y_2 = 0.05$ has been allowed. For example, for the first probability threshold in Table 9.5, 0.05, only the forecasts $y_1 = 0.0$ are converted to "no" forecasts, so the entries of the resulting 2 x 2 joint distribution (cf. Figure 9.1b) are $a/n = 0.032+0.025+\cdots+0.013 = 0.252$, $b/n = 0.128+0.075+$ $+0.007 = 0.448$, $c/n = p(y_2, o_1) = 0.045$, and $d/n = p(y_2, o_2) = 0.255$. For the second probability threshold, 0.15, both the forecasts $y_1 = 0.0$ and $y_2 = 0.1$ are converted to "no" forecasts, so the resulting 2 x 2 joint distribution contains the four probabilities $a/n = 0.025+0.024+ \ldots +0.013 = 0.220$, $b/n = 0.075+0.056+\cdots+0.007 = 0.320$, $c/n = 0.045+0.032 = 0.077$, and $d/n = 0.255+0.128 = 0.383$.

Table 9.5 also shows the hit rate H and false alarm rate F for each of the 10 partitions of the joint distribution in Table 9.4a. These pairs define the lower ROC curve in Figure 9.14, with the points corresponding to the smaller probability thresholds occurring in the upper right portion of the ROC diagram, and points corresponding to the larger probability thresholds occurring in the lower left portion. Proceeding from left to right, the areas under the $I = 11$ trapezoids defined by these points

TABLE 9.5 The I–1 = 10 2 × 2 Tables Derived From Successive Partitions of the Joint Distribution in Table 9.4, and the Corresponding Values for H and F

Threshold	a/n	b/n	c/n	d/n	H	F
0.05	.252	.448	.045	.255	.848	.637
0.15	.220	.320	.077	.383	.741	.455
0.25	.195	.245	.102	.458	.657	.348
0.35	.171	.189	.126	.514	.576	.269
0.45	.147	.143	.150	.560	.495	.203
0.55	.123	.107	.174	.596	.414	.152
0.65	.096	.074	.201	.629	.323	.105
0.75	.071	.049	.226	.654	.239	.070
0.85	.043	.027	.254	.676	.145	.038
0.95	.013	.007	.284	.696	.044	.010

together with the points at the corners of the ROC diagram are 0.5(0.044+0.000)(0.010 – 0.000) = 0.00022, 0.5(0.145+0.044)(0.038–0.010) = 0.00265, 0.5(0.239+0.145)(0.070–0.038) = 0.00614, … , 0.5 (1.000+0.848)(1.000–0.637) = 0.33541; yielding a total area of $A = 0.698$.

Figure 9.14 shows clearly that the forecasts in Table 9.2 exhibit greater event discrimination than those in Table 9.4, because the arc of the corresponding ROC curve for the former is everywhere above that for the latter, and approaches the upper left-hand corner of the ROC diagram more closely. This difference in discrimination is summarized by the differences in the areas under the two ROC curves, that is, $A = 0.922$ versus $A = 0.698$. ◇

Because ROC diagrams display the relationship between H (Equation 9.12) and F (Equation 9.13), which are two characteristics of the 2×2 contingency table, it is not surprising that others of the verification statistics in Sections 9.2.2 and 9.2.3 can be related to these plots. For example, since PSS (Equation 9.18) can be written as PSS = H – F, and the 45° diagonal line in the ROC diagram is exactly the $H = F$ line, the PSS for any threshold is exactly the vertical distance between the ROC curve and the $H = F$ diagonal. Sometimes the partition maximizing this vertical distance is chosen as "optimal" in medical statistics. Isolines of equal bias (Equation 9.10) having slope $-p(o_2)/p(o_1) = -(b+d)/(a+c)$ can also be drawn on the ROC diagram, which intersect the vertical axis at $H = B$ (Mason, 2003). Thus the partition yielding unbiased forecasts occurs at the intersection of the ROC curve and the equation $H = 1 - F(b+d)/(a+c)$.

ROC diagrams have been used increasingly in recent years to evaluate probability forecasts for binary predictands, so it is worthwhile to reiterate that (unlike the reliability diagram and the discrimination diagram) they do *not* provide a full depiction of the joint distribution of forecasts and observations. This deficiency of the ROC diagram can be appreciated by recalling the mechanics of its construction, as outlined in Example 9.5. In particular, the calculations behind the ROC diagrams are carried out without regard to the specific values for the probability labels, $p(y_i)$. That is, the forecast probabilities are used only to sort the elements of the joint distribution into a sequence of 2 x 2 tables, but otherwise their actual numerical values are immaterial. For example, Table 9.4b shows that the forecasts defining the lower

ROC curve in Figure 9.14 are poorly calibrated, and in particular they exhibit strong conditional (overconfidence) bias. However this and other biases are not reflected in the ROC diagram, because the specific numerical values for the forecast probabilities $p(y_i)$ do not enter into the ROC calculations, and so ROC diagrams are insensitive to such conditional and unconditional biases (e.g., Glahn, 2004; Jolliffe and Stephenson, 2005; Kharin and Zwiers, 2003; Wilks, 2001). In fact, if the forecast probabilities $p(y_i)$ had corresponded exactly to the corresponding conditional event probabilities $p(o_1 | y_i)$, or even if the probability labels on the forecasts in Tables 9.2 or 9.4 had been assigned values that were allowed to range outside the [0, 1] interval (while maintaining the same ordering, and so the same groupings of event outcomes), the resulting ROC curves would be identical!

As illustrated in Example 9.5, numerical integration of the ROC curve using the trapezoidal rule is an easy and typical approach to estimating the area A under it. To the extent that the plotted points only sample an underlying smooth function, trapezoidal integration will underestimate the area because ROC curves generally exhibit negative second derivatives, although this underestimation will decrease as the number of points increases. An alternative is to represent the two likelihood distributions $p(o_1 \mid y_i)$ and $p(o_2 \mid y_i)$ as Gaussian, which is called the *bi-normal model*. This assumption is more robust that might initially be imagined, because as noted before the ROC analysis is insensitive to the numerical labels of the forecasts, so that the bi-normal model will yield good results to the extent that monotonic transformations of the two likelihood distributions yield approximately Gaussian distributions. The necessary Gaussian parameters are estimated from the data at hand, using either regression (Mason, 1982; Swets, 1979) or maximum likelihood (Dorfman and Alf, 1969). The result is a smooth continuous ROC curve that can be integrated without bias, but may yield inaccuracies to the extent that the bi-normal model is not fully adequate.

ROC-based statistics provide information that is similar to the "resolution" term in the Brier score decomposition (Equation 9.45), independently of forecast calibration or lack thereof. This insensitivity to calibration is typically not a problem for the widespread use of ROC diagrams in applications like medical statistics (Pepe, 2003), because there the "forecast" (perhaps the blood concentration of a particular protein) is incommensurate with the "observation" (the patient has disease or not): there is no expectation that the "forecasts" are calibrated to or even pertain to the same variable as the observations. In such cases what is required is evaluation of the mapping between increasing levels of the diagnostic measurement with the probability of disease, and for this purpose the ROC diagram is a natural tool.

The insensitivity of ROC diagrams and ROC areas to both conditional and unconditional forecast biases—that they are independent of calibration—is sometimes cited as an advantage. This property is an advantage only in the sense that ROC diagrams reflect potential skill (which would be actually achieved only if the forecasts were correctly calibrated), in much the same way that the correlation coefficient reflects potential skill (cf. Equation 9.38). However, this property is not an advantage for forecast users who do not have access to the historical forecast data necessary to correct miscalibrations, and who therefore have no choice but to take forecast probabilities at face value. On the other hand, when forecasts underlying ROC diagrams are correctly calibrated, dominance of one ROC curve over another (i.e., one curve lying entirely above and to the left of another) implies statistical sufficiency (DeGroot and Fienberg, 1982, Ehrendorfer and Murphy, 1988) for the dominating forecasts, so that these will be of greater economic value for all rational forecast users (Krzysztofowicz and Long, 1990).

9.4.7. The Logarithmic, or Ignorance Score

The *logarithmic score*, or *Ignorance score* (Good, 1952; Roulston and Smith, 2002; Winkler and Murphy, 1968) is an alternative to the Brier score (Equation 9.39) for probability forecasts for

dichotomous events. On a given forecasting occasion, k, it is the negative of the logarithm of the forecast probability corresponding to the event that subsequently occurs:

$$I_k = \begin{cases} -\ln(y_k), & \text{if } o_k = 1 \\ -\ln(1-y_k), & \text{if } o_k = 0, \end{cases} \tag{9.50a}$$

with the average Ignorance over n forecasting occasions being

$$\bar{I} = \frac{1}{n}\sum_{k=1}^{n} I_k. \tag{9.50b}$$

The Ignorance score ranges from zero for perfect forecasts ($y = 1$ if the binary event occurs or $y = 0$ if it does not) to infinity for certainty forecasts that are wrong ($y = 0$ if the event occurs or $y = 1$ if it does not). Another way of looking at Equation 9.50 is as the negative of the log-likelihood function (Section 4.6.1) for a sequence of n independent Bernoulli (i.e., binomial, Section 4.2.1, with $N = 1$) distributions, with probabilities y_k.

Even a single wrong certainty forecast in Equation 9.50 implies that the average Ignorance score for the entire collections of forecasts in Equation 9.50 will also be infinite, regardless of the accuracy of the other n–1 forecasts considered. Accordingly the Ignorance score is not appropriate in settings where the forecasts must be rounded before being issued. When forecast probabilities evaluated using the Ignorance score are to be estimated on the basis of a finite sample (e.g., in the context of ensemble forecasting), it would be natural to estimate the probabilities using one of the plotting position formulae from Table 3.2, for example, the implementation of the Tukey plotting position in Equation 8.5, that cannot produce either $y = 0$ or $y = 1$.

A natural skill score in this context, or relative accuracy as measured by Ignorance relative to a set of reference forecasts y_{ref}, is provided by the relative (or, reduction in) Ignorance (Bröcker and Smith, 2008; Peirolo, 2011),

$$\begin{aligned} \mathrm{RI} &= \bar{I}_{\mathrm{ref}} - \bar{I}_k \\ &= \frac{1}{n}\sum_{k=1}^{n} \ln\left(\frac{y_k}{y_{ref}}\right). \end{aligned} \tag{9.51}$$

Terms in Equation 9.51 for which $y_k > y_{ref}$ (larger probability than the reference forecast for the "yes" event) contribute positively to the sum, and those for which $y_k < y_{ref}$ contribute negatively, so that a positive value for RI corresponds to positive skill. Murphy and Epstein (1967a) pointed out that a consistent measure of forecast sharpness in this context is provided by the average negative entropy, or Shannon (1948) information

$$\bar{E} = -\frac{1}{n}\sum_{k=1}^{n} y_k \ln(y_k). \tag{9.52}$$

The Ignorance score behaves generally similarly to the Brier score, except for extreme (near-certainty) probability forecasts, for which the behavior of the two scores diverge markedly (e.g., Benedetti, 2010). Accordingly Ignorance may be better able than the Brier score to resolve the accuracy of forecasts for extreme or otherwise rare, and therefore low-probability, events. The Ignorance score

generalizes to probability forecasts for nonbinary discrete events (Section 9.4.9), and to full continuous probability distribution forecasts (Section 9.5.1).

The Ignorance score and the corresponding skill score *RI* have interesting connections to probability assessment in betting and insurance (Good, 1952; Hagedorn and Smith, 2009; Roulston and Smith, 2002). In these contexts, the base-2 logarithms are typically used in Equations 9.50–9.52. The Ignorance score can be interpreted as the expected returns to be obtained by placing bets in proportion to the forecast probabilities, and the *RI* describes the gambler's return relative to the reference forecasts of the "house."

9.4.8. Hedging and Strictly Proper Scoring Rules

It is natural for forecasters to want to achieve the best scores they can. Depending on the evaluation measure, it may be possible to improve scores by *hedging*, or "gaming," which in the context of forecasting implies reporting something other than our true beliefs about future weather events in order to produce a better score (Jolliffe, 2008; Murphy and Epstein, 1967b). For example, in the setting of a forecast contest in a college or university, if the evaluation of our performance can be improved by playing the score, then it is entirely rational to try to do so. Conversely, if we are responsible for assuring that forecasts are of the highest possible quality, evaluating those forecasts in a way that penalizes hedging is desirable.

A forecast evaluation procedure that awards a forecaster's best expected score only when his or her true beliefs are forecast is called a *strictly proper* scoring rule (Winkler and Murphy, 1968). A slightly weaker condition, that no better expected score than that earned by the forecaster's true beliefs, is called *proper*. Strictly proper scoring procedures cannot be hedged, and attempting to hedge a proper scoring rule can at best achieve only a comparable result. One very appealing attribute of both the Brier score (Equation 9.39) and the Ignorance score (Equation 9.50) is that they are strictly proper, and this is a strong motivation for using one or the other to evaluate the accuracy of probability forecasts for dichotomous predictands, although many other strictly proper scoring rules can also be formulated (Gneiting and Raftery, 2007; Merkle and Steyvers, 2013).

It is instructive to look at the strict propriety of the Brier score in more detail. Of course it is not possible to know in advance what score a given forecast will achieve, unless we can make perfect forecasts. However, it is possible on each forecasting occasion to calculate the expected, or probability-weighted, future score using our subjective probability for the forecast event. Suppose a forecaster's subjective probability for the event being forecast is y^*, and that the forecaster must publicly communicate a forecast probability, y. The expected Brier score is simply

$$E[BS] = y^*(y-1)^2 + (1-y^*)(y-0)^2, \tag{9.53}$$

where the first term is the score received if the event occurs multiplied by the subjective probability that it will occur, and the second term is the score received if the event does not occur multiplied by the subjective probability that it will not occur. Consider that the forecaster has decided on a subjective probability y^* and is weighing the problem of what forecast y to issue publicly. Regarding y^* as constant, it is easy to minimize the expected Brier score by differentiating Equation 9.53 by y, and setting the result equal to zero. Then,

$$\frac{\partial E[BS]}{\partial y} = 2y^*(y-1) + 2(1-y^*)y = 0, \tag{9.54a}$$

yielding

$$\begin{aligned} 2y\,y^* - 2y^* + 2y - 2y\,y^* &= 0 \\ 2y &= 2y^* \\ y &= y^*. \end{aligned} \tag{9.54b}$$

That is, regardless of the forecaster's subjective probability, the minimum expected Brier score is achieved only when the publicly communicated forecast corresponds exactly to the subjective probability. A similar derivation demonstrating that the Ignorance score is strictly proper can be found in Winkler and Murphy (1968). By contrast, for example, the absolute error (linear) score, $\mathrm{LS} = |y - o|$ is minimized by forecasting $y = 0$ when $y^* < 0.5$, and forecasting $y = 1$ when $y^* > 0.5$, and so is not proper.

The concept of a strictly proper scoring rule is easiest to understand and prove for a case such as the Brier score, since the probability distribution being forecast (Bernoulli) is so simple. Gneiting and Raftery (2007) show that the concept of strict propriety can be applied in more general settings, where the form of forecast distribution is not necessarily Bernoulli. It is also not necessary to invoke forecaster honesty in order to motivate the concept of strict propriety. Bröcker and Smith (2007a) and Gneiting and Raftery (2007) note that strictly proper scores are internally consistent, in the sense that a forecast probability distribution yields an optimal expected score when the verification is drawn from that same probability distribution.

Equation 9.54 proves that the Brier score is strictly proper. Often Brier scores are expressed in the skill-score format of Equation 9.40. Even though the Brier score itself is strictly proper, this standard skill score based upon it, and other skill scores as well (Winkler, 1996) are not. However, for moderately large sample sizes (perhaps $n > 100$) the BSS closely approximates a strictly proper scoring rule (Murphy, 1973a).

9.4.9. Probability Forecasts for Multiple-Category Events

Probability forecasts may be formulated for discrete events having more than two ("yes" vs. "no") possible outcomes. These events may be *nominal*, for which there is not a natural ordering; or *ordinal*, where it is clear which of the outcomes are larger or smaller than others. The approaches to verification of probability forecasts for nominal and ordinal predictands may differ, because the magnitude of the forecast error is not a meaningful quantity in the case of nominal events, but is potentially quite important for ordinal events. One approach to verifying forecasts for nominal predictands is to collapse them to a sequence of binary predictands. Having done this, Brier scores, reliability diagrams, and so on, can be used to evaluate each of the derived binary forecasting situations.

Verification of probability forecasts for multicategory ordinal predictands presents a more difficult problem. First, the dimensionality of the verification problem increases exponentially with the number of outcomes over which the forecast probability is distributed. For example, consider a $J = 3$-event situation for which the forecast probabilities are constrained to be one of the 11 values 0.0, 0.1, 0.2, … , 1.0. The dimensionality of the problem is not simply $33 - 1 = 32$, as might be expected by extension of the formula for dimensionality for the dichotomous forecast problem, because the forecasts are now vector quantities. For example, the forecast vector (0.2, 0.3, 0.5) is a different and distinct forecast from the vector (0.3, 0.2, 0.5). Since the three forecast probabilities must sum to 1.0, only two of them can vary freely. In this situation there are $I = 66$ possible three-dimensional forecast vectors, yielding a dimensionality for the forecast problem of (66 x 3)–1 = 197 (Murphy, 1991). Similarly, the dimensionality for the four-category ordinal verification situation with the same restriction on the forecast probabilities

would be (286 x 4)–1 = 1143. As a practical matter, because of their high dimensionality, probability forecasts for ordinal predictands primarily have been evaluated using scalar performance measures, even though such approaches will necessarily be incomplete.

Ranked Probability Score

Verification measures that are *sensitive to distance* reflect at least some aspects of the magnitudes of forecast errors, and for this reason are often preferred for probability forecasts of ordinal predictands. That is, such verification statistics are capable of penalizing forecasts increasingly as more probability is assigned to event categories further removed from the actual outcome. In addition, we would like the verification measure to be strictly proper (see Section 9.4.8), so that forecasters are encouraged to report their true beliefs. Several strictly proper scalar scores that are sensitive to distance exist for this type of forecast (Murphy and Daan, 1985; Staël von Holstein and Murphy, 1978), but of these the *ranked probability score* (RPS) is almost universally chosen. The RPS was originally formulated by R.J. Thompson around 1967 (Murphy, 1971), although Epstein (1969b) independently proposed a more elaborate but equivalent score.

The ranked probability score is essentially an extension of the Brier score (Equation 9.39) to the many-event situation. That is, it is a squared-error score with respect to the observation $o_j = 1$ if the forecast event j occurs, and $o_j = 0$ if the event does not occur. However, in order for the score to be sensitive to distance, the squared errors are computed with respect to the cumulative probabilities in the forecast and observation vectors. This characteristic introduces some notational complications.

As before, let J be the number of event categories, and therefore also the number of probabilities included in each forecast. For example, a common format for seasonal forecasts (pertaining to average conditions over 3-month periods) is to allocate probability among three climatologically equiprobable classes (e.g., Mason et al., 1999; O'Lenic et al., 2008). If a precipitation forecast is 20% chance of "dry," 40% chance of "near-normal," and 40% chance of "wet," then $y_1 = 0.2$, $y_2 = 0.4$, and $y_3 = 0.4$. Each of these components y_j pertains to one of the J events being forecast. That is, y_1, y_2, and y_3 are the three components of a forecast vector $\mathbf{y}$, and if all probabilities were to be rounded to tenths this forecast vector would be one of $I = 66$ possible forecasts $\mathbf{y}_i$. Similarly, in this setting the observation vector has three components. One of these components, corresponding to the event that occurs, will equal 1, and the other $J - 1$ components will equal zero. For example, if the observed precipitation outcome is in the "wet" category, then $o_1 = 0$, $o_2 = 0$, and $o_3 = 1$.

The cumulative forecasts and observations, denoted Y_m and O_m, are defined as functions of the components of the forecast vector and observation vector, respectively, according to

$$Y_m = \sum_{j=1}^{m} y_j, \quad m = 1, \ldots, J; \tag{9.55a}$$

and

$$O_m = \sum_{j=1}^{m} o_j, \quad m = 1, \ldots, J. \tag{9.55b}$$

In terms of the foregoing hypothetical example, $Y_1 = y_1 = 0.2$, $Y_2 = y_1 + y_2 = 0.6$, and $Y_3 = y_1 + y_2 + y_3 = 1.0$; and $O_1 = o_1 = 0$, $O_2 = o_1 + o_2 = 0$, and $O_3 = o_1 + o_2 + o_3 = 1$. Notice that since Y_m and O_m are both cumulative functions of probability components that must add to one, the final sums Y_J and O_J are always both equal to one by definition.

The ranked probability score is the sum of squared differences between the components of the cumulative forecast and observation vectors in Equation 9.55a and 9.55b, given by

$$\mathrm{RPS} = \sum_{m=1}^{J} (Y_m - O_m)^2, \tag{9.56a}$$

or, in terms of the forecast and observed vector components y_j and o_j,

$$\mathrm{RPS} = \sum_{m=1}^{J} \left[\left(\sum_{j=1}^{m} y_j \right) - \left(\sum_{j=1}^{m} o_j \right) \right]^2. \tag{9.56b}$$

A perfect forecast would assign all the probability to the single y_j corresponding to the event that subsequently occurs, so that the forecast and observation vectors would be the same. In this case, RPS = 0. Forecasts that are less than perfect receive scores that are positive numbers, so the RPS has a negative orientation. Notice also that the final ($m = J$) term in Equation 9.56 is always zero, because the accumulations in Equations 9.55 ensure that $Y_J = O_J = 1$. Accordingly Equation 9.56 will have only J–1 nonzero terms so that the last of these need not be computed, and the worst possible score is $J - 1$. The net effect of the calculation in Equation 9.56 is to add Brier scores for the J–1 binary probability forecasts defined by the J–1 boundaries between the J categories. For $J = 2$, the ranked probability score reduces to the Brier score, Equation 9.39.

Equation 9.56 yields the ranked probability score for a single forecast-event pair. Jointly evaluating a collection of n forecasts using the ranked probability score requires nothing more than averaging the RPS values for each forecast-event pair,

$$\overline{\mathrm{RPS}} = \frac{1}{n} \sum_{k=1}^{n} \mathrm{RPS}_k. \tag{9.57}$$

Similarly, the skill score for a collection of RPS values relative to the RPS computed from the climatological probabilities can be computed as

$$\mathrm{SS_{RPS}} = \frac{\overline{\mathrm{RPS}} - \overline{\mathrm{RPS_{clim}}}}{0 - \overline{\mathrm{RPS_{clim}}}} = 1 - \frac{\overline{\mathrm{RPS}}}{\overline{\mathrm{RPS_{clim}}}}. \tag{9.58}$$

Bradley and Schwartz (2011) point out that the RPS skill score in Equation 9.58 can also be formulated as a weighted average of the Brier skill scores for the J–1 possible binary probability forecasts defined by the boundaries of the J categories.

Example 9.6. Illustration of the Mechanics of the Ranked Probability Score

Table 9.6 demonstrates the mechanics of computing the RPS, and illustrates the property of sensitivity to distance, for two hypothetical probability forecasts for precipitation amounts. Here the continuum of precipitation has been divided into $J = 3$ categories, < 0.01 in., 0.01 – 0.24 in., and ≥ 0.25 in. Forecaster 1 has assigned the probabilities (0.2, 0.5, 0.3) to the three events, and Forecaster 2 has assigned the probabilities (0.2, 0.3, 0.5). The two forecasts are similar, except that Forecaster 2 has allocated more probability to the ≥ 0.25 in. category at the expense of the middle category. If no precipitation falls on this occasion the observation vector will be that indicated in the table. Many forecasters and forecast users would intuitively feel that Forecaster 1 should receive a better score, because this forecaster has assigned more probability closer to the observed category than did Forecaster 2. The score for Forecaster 1 is

TABLE 9.6 Comparison of Two Hypothetical Probability Forecasts for Precipitation Amount, Divided into $J = 3$ Ordinal Categories

	Forecaster 1		Forecaster 2		Observed	
Event	y_j	Y_m	y_j	Y_m	o_j	O_m
$<$ 0.01 in.	0.2	0.2	0.2	0.2	1	1
0.01–0.24 in.	0.5	0.7	0.3	0.5	0	1
$\geq$ 0.25 in.	0.3	1.0	0.5	1.0	0	1

The three components of the observation vector indicate that the observed precipitation was in the smallest category.

RPS $= (0.2–1)^2 + (0.7–1)^2 = 0.73$, and for Forecaster 2 it is RPS $= (0.2–1)^2 + (0.5–1)^2 = 0.89$. The lower RPS for Forecaster 1 indicates a more accurate forecast according to RPS.

These RPS results are equivalent to adding the $J–1 = 2$ Brier scores for the two possible binary probability forecasts based on the forecast vectors that respect the ordinal nature of the categories, that is, Pr{$<$ 0.01 in.} (*vs.* $\geq$ 0.01 in.) and Pr{$<$ 0.25 in.} (*vs.* $\geq$ 0.25 in.). For Forecaster 1, these two probability forecasts are 0.2 and 0.7, respectively, and for Forecaster 2 they are 0.2 and 0.5, which are exactly the probabilities specified by Equation 9.55a. Since the observation indicated in Table 9.6 is $<$ 0.01 in., the two Brier scores for Forecaster 1 are $(0.2–1)^2 = 0.64$ and $(0.7–1)^2 = 0.09$, which sum to 0.73, consistent with Equation 9.56. The corresponding results for Forecaster 2 are $(0.2–1)^2 = 0.64$ and $(0.5–1)^2 = 0.25$, which sum to 0.89.

Alternatively, if some amount of precipitation larger than 0.25 in. had fallen, Forecaster 2's probabilities would have been closer and would have received the better score. The score for Forecaster 1 would have been RPS $= (0.2–0)^2 + (0.7–0)^2 = 0.53$, and the score for Forecaster 2 would have been RPS $= (0.2–0)^2 + (0.5–0)^2 = 0.29$. Note that in both of these examples, only the first $J–1 = 2$ terms in Equation 9.56 were needed to compute the RPS. ◇

The Calibration Simplex

The reliability diagram (Section 9.4.4) for probability forecasts of binary events can be extended to the *Calibration Simplex* (Wilks, 2013), which is a graphical representation of the calibration-refinement factorization (Equation 9.2) for three-category probability forecasts, regardless of whether the categories are nominal or ordinal.

The reliability diagram plots the possible scalar probability forecasts along the unit interval on its horizontal axis. Each set of forecast probabilities for three categories can be plotted in two dimensions because they must sum to 1. The geometrically appropriate graph is the regular 2-simplex (e.g., Epstein and Murphy, 1965; Murphy, 1972), which takes the shape of an equilateral triangle. The area within the triangle can be partitioned to represent possible 3-dimensional forecast vectors in the same way that the reliability diagram partitions the unit interval. Figure 9.15 illustrates the plotting of discretized forecast vectors onto the simplex, which has been rendered as a tessellation of hexagons. Here the categories are labeled B (below normal), N (near-normal), and A (above normal), consistent with a typical forecast format for medium-range and seasonal forecasts. Each scalar forecast probability has been rounded to one of the ten values 0/9, 1/9, … , 9/9, yielding $I = 55$ distinct possible vector forecasts, each of which

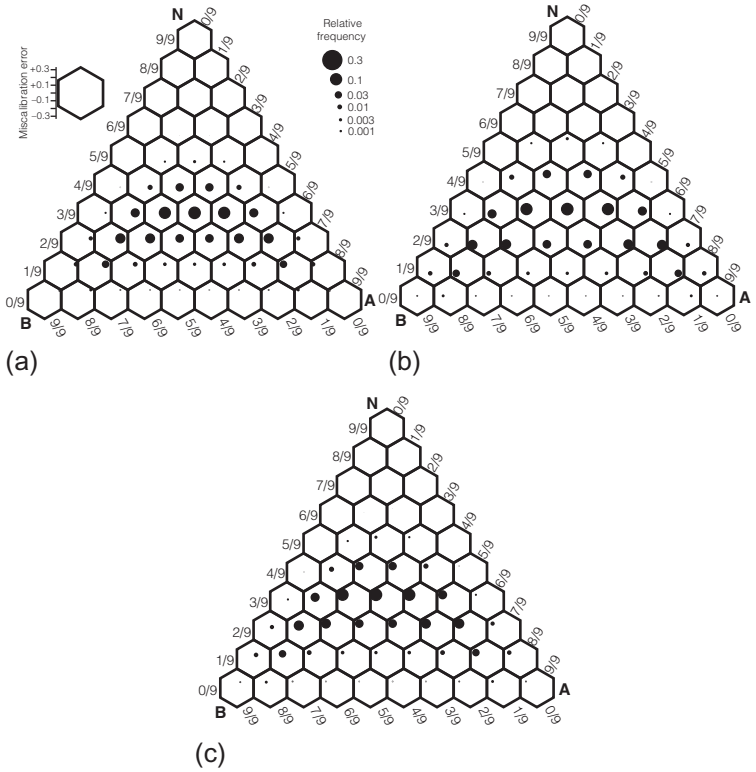

FIGURE 9.15 Calibration simplexes illustrating (a) overconfident forecasts, (b) underconfident forecasts, and (c) unconditionally biased forecasts. In each case the refinement distribution is represented by the sizes of the dots. *From Wilks (2013).*

is represented by one of the hexagons. The hexagons at the three vertices represent forecasts assigning all probability to the outcome labeled at that corner. Hexagons representing other forecast vectors are located at perpendicular distances from the respective opposite sides, which are indicated by the probability labels in the margins. The result, for example, is that the hexagon in the center of Figure 9.15 locates the climatological forecast [1/3, 1/3, 1/3].

The I conditional calibration distributions $p(o_j \mid y_i)$, for dichotomous probability forecasts are Bernoulli, each of which can be characterized by single probabilities that are indicated in the reliability diagram by the vertical positions of the I dots. In the case of three-element probability forecasts, the calibration distributions are multinomial (Equation 4.11) and are completely specified by two of the three probability parameters. The locations of dots within each hexagon of the calibration simplex indicate the magnitudes of these conditional multinomial probabilities. The calibration-refinement factorization is completed by specifying the frequencies of use of the I possible probability vectors (the refinement distribution), the elements of which are shown in proportion to the areas of the dots within each hexagon.

Figure 9.15a shows a calibration simplex reflecting overconfident forecasts, where the plotted dots have been displaced toward the center of the diagram, indicating that probabilities larger than 1/3 have been overforecast and probabilities smaller than 1/3 have been underforecast. Figure 9.15b shows a

corresponding plot for underconfident forecasts, where the probabilities larger than 1/3 have been underforecast and probabilities smaller than 1/3 have been overforecast, so that the dots have been displaced toward the margins of the diagram. Figure 9.15c shows an example plot for unconditionally biased forecasts, where the "A" category has been overforecast at the expense of the "B" and "N" categories, so that the dots are displaced away from the "A" vertex.

A similar idea, proposed independently by Murphy and Daan (1985) and Jupp et al. (2012), is to draw line segments on the 2-simplex connecting points locating each distinct forecast vector with its corresponding average outcome vector, and indicate the corresponding subsample sizes with numerals printed at the midpoints of these line segments. Jupp et al. call the resulting plot the *ternary reliability diagram*. This graphical approach also expresses the full information content of the joint distribution, but overall the diagram may be somewhat difficult to read, unless very few of the possible forecast vectors have been used.

Logarithmic (Ignorance) Score for Multiple Categories

An alternative to RPS for evaluation of probability forecasts for multicategory events is provided by an extension of the Ignorance score (Equation 9.50a), which is also strictly proper. For a single forecast k, the Ignorance score is simply the negative logarithm of that element of the forecast vector corresponding to the event that actually occurred (i.e., for which $o_j = 1$), which can be expressed as

$$I_k = -\sum_{j=1}^{J} o_j \ln\left(y_j\right), \tag{9.59}$$

where it is understood that $0 \ln(0) = 0$. As was also the case for the 2-category Ignorance score (Equation 9.50a), incorrect certainty forecasts yield infinite Ignorance, so that the Ignorance score is usually not suitable for forecasts that have been rounded to a finite set of discrete allowable probabilities. The average Ignorance score over n forecasting occasions would again be given by Equation 9.50b.

Extending Equation 9.51, a natural skill measure is the Relative Ignorance

$$\mathrm{RI} = \frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{J} o_{k,j} \ln\left(\frac{y_{k,j}}{y_{ref,j}}\right), \tag{9.60}$$

where $o_{k,j}$ is the binary observation variable for the jth category on the kth forecast occasion, $y_{k,j}$ is the corresponding forecast, and $y_{ref,j}$ is the appropriate (perhaps climatological) reference (Krakauer et al., 2013).

The Ignorance score is not sensitive to distance, and indeed exhibits a property known as *locality*, meaning that only the probability assigned to the event that occurs matters in its computation. For the forecasts in Table 9.6, $I = 1.61$ for both Forecaster 1 and Forecaster 2 because the distribution of forecast probabilities among outcome categories for which $o_j = 0$ is irrelevant. It should be clear that locality and sensitivity to distance are mutually incompatible, and some discussion of their relative merits is included in Bröcker and Smith (2007a) and Mason (2008), although as scalars both the ranked probability score and the Ignorance score are in any case incomplete measures of forecast quality. If locality is accepted as a desirable characteristic, then preference for the Ignorance score is indicated, since it is the only score that is both local and strictly proper. Use of the Ignorance score in preference to the RPS is also indicated when evaluating probability forecasts for nominal categories, in which case the concepts of ordering and distance not meaningful.

9.5. PROBABILITY DISTRIBUTION FORECASTS FOR CONTINUOUS PREDICTANDS

When the predictand is a continuous, real-valued quantity the most informative forecast format will involve communicating its predictive distribution. That is, each forecast will be a full probability distribution, which might be communicated as a PDF $f(y)$, or CDF $F(y)$. For some forecasts, such as the nonhomogeneous regressions described in Sections 8.3.2–8.3.4, these predictive distributions may be one the conventional parametric forms described in Section 4.4, in which case they might be specified in terms of a few distribution parameters. Forecasts formulated nonparametrically (e.g., Section 8.3.6), or through Bayesian methods solved using MCMC methods (Section 6.4), may be expressed as discrete approximations to continuous PDFs or CDFs. Although this latter class includes raw ensemble forecasts, and member-by-member postprocessed ensemble forecasts (Section 8.3.6), discussion of verification methods designed specifically for forecast ensembles will be deferred until Section 9.7.

9.5.1. Continuous Ranked Probability Score

Regardless of how a forecast probability distribution is expressed, providing a full forecast probability distribution is both a conceptual and a mathematical extension of multicategory probability forecasting (Section 9.4.9), to forecasts for an infinite number of predictand classes of infinitesimal width. A natural approach to evaluating this kind of forecast is to extend the ranked probability score to the continuous case, replacing the summations in Equation 9.56 with integrals. The result is the *Continuous Ranked Probability Score* (Hersbach, 2000; Matheson and Winkler, 1976; Unger, 1985),

$$\mathrm{CRPS} = \int_{-\infty}^{\infty} [F(y) - F_o(y)]^2 dy, \tag{9.61a}$$

where

$$F_o(y) = \begin{cases} 0, & y < o \\ 1, & y \geq o \end{cases} \tag{9.61b}$$

is a cumulative-probability step function that jumps from 0 to 1 at the point where the forecast variable y equals the observation o. The squared difference between continuous CDFs in Equation 9.61a is analogous to the same operation applied to the cumulative discrete variables in Equation 9.56a for the Ranked Probability Score, and accordingly CRPS can be seen as an extension of RPS for infinitely many bins of infinitesimal width. Also, just as the RPS can be viewed as the sum of Brier scores for the binary variables defined by the bin boundaries, the CRPS can be viewed as the integral of the BS over all values of the predictand (Hersbach, 2000). In common with the discrete BS and RPS, the CRPS is also strictly proper (Matheson and Winkler, 1976).

The CRPS has a negative orientation (smaller values are better), and it rewards concentration of probability around the step function located at the observed value. Figure 9.16 illustrates the CRPS with a hypothetical example. Figure 9.16a shows three Gaussian forecast PDFs $f(y)$ in relation to a single observed value of the continuous predictand y. Forecast Distribution 1 is centered on the eventual observation and strongly concentrates its probability around the observation. Distribution 2 is equally sharp (i.e., expresses the same degree of confidence in distributing probability), but is centered well away from the observation. Distribution 3 is centered on the observation but exhibits low confidence (distributes probability more diffusely than the other two forecast distributions). Figure 9.16b shows the same three

forecast distributions expressed as CDFs, $F(y)$, together with the step-function CDF $F_0(y)$ (thick line) that jumps from 0 to 1 at the observed value (Equation 9.61b). Since the CRPS is the integrated squared difference between the CDF and the step function, CDFs that approximate the step function (Distribution 1) produce relatively small integrated squared differences, and so good scores. Distribution 2 is equally sharp, but its displacement away from the observation produces large discrepancies with the step function, especially for values of the predictand slightly larger than the observation, and therefore a very large integrated squared difference. Distribution 3 is centered on the observation, but its diffuse assignment of forecast probability means that it is nevertheless a poor approximation to the step function, and so also yields a large integrated squared difference.

Alternatively, the CRPS can also be formulated as (Gneiting and Raftery, 2007)

$$\text{CRPS} = E_F|Y - o| - \frac{1}{2}E_F|Y - Y'|, \tag{9.62}$$

where E_F denotes statistical expectation with respect to the forecast distribution $F(y)$, and Y and Y' are independent realizations from $F(y)$. This alternative representation forms the basis of the Energy Score described in Section 9.5.2, and the discrete version of CRPS described in Section 9.7.3.

Equation 9.61 may be difficult to evaluate for arbitrary an forecast CDF, $F(y)$. However, if this forecast distribution is Gaussian with mean μ and variance σ^2 the CRPS when the observation o occurs is (Gneiting et al., 2005)

$$\text{CRPS}(\mu, \sigma^2, o) = \sigma\left\{\frac{o-\mu}{\sigma}\left[2\Phi\left(\frac{o-\mu}{\sigma}\right) - 1\right] + 2\phi\left(\frac{o-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right\}, \tag{9.63}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the CDF and PDF (Equation 4.25) of the standard Gaussian distribution. In Figure 9.16, $f_1(y)$ has $\mu = 0$ and $\sigma^2 = 1$, $f_2(y)$ has $\mu = 2$ and $\sigma^2 = 1$, and $f_3(y)$ has $\mu = 0$ and $\sigma^2 = 9$. Using Equation 9.63 the observation $o = 0$ yields $\text{CRPS}_1 = .23$, $\text{CRPS}_2 = 1.45$, and $\text{CRPS}_3 = .70$.

References providing closed-form expressions for CRPS when the forecast distribution takes on other known parametric forms are listed in Table 9.7.

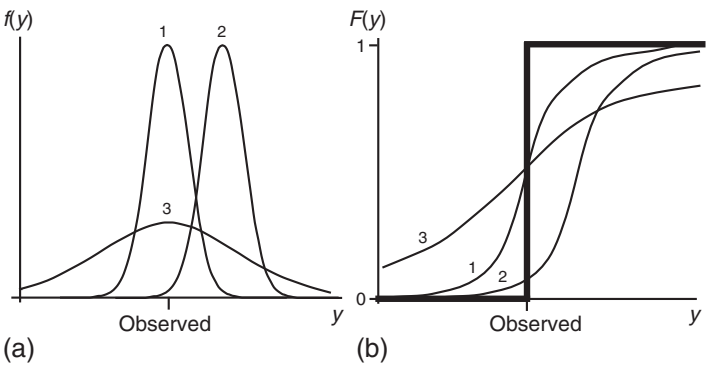

FIGURE 9.16 Schematic illustration of the continuous ranked probability score, as expressed in Equation 9.61. Three Gaussian forecast PDFs are shown in relation to the observed outcome in (a). The corresponding CDFs are shown in (b), together with the step-function CDF for the observation $F_0(y)$ (heavy line). Distribution 1 would produce a small (good) CRPS because its CDF is the closest approximation to the step function, yielding the smallest integrated squared difference. Distribution 2 concentrates probability away from the observation, and Distribution 3 is penalized for lack of sharpness even though it is centered on the observation.

Because CRPS can also be computed as the Brier score for dichotomous events integrated over all possible division points of the continuous variable y, the CRPS has an algebraic decomposition into reliability, resolution, and uncertainty components that is analogous to an integrated form of Equation 9.45 (Candille and Talagrand, 2005). Indeed, any strictly proper scoring rule can be expressed in that way (Bröcker, 2009). Hersbach (2000) also shows that for nonprobabilistic forecasts (all probability concentrated at y, with $F(y)$ also a step function in the form of Equation 9.61b), CRPS reduces to absolute error (because $0^2 = 0$ and $1^2 = 1$), in which case the average CRPS over n forecasts reduces to the MAE (Equation 9.32).

Analogously to the RPS in Equation 9.57, CRPS values for multiple occasions are typically averaged to characterize forecast quality overall in terms of a single number. Also analogously to the RPS in Equation 9.58, these average CRPS values are often expressed in the usual skill score format (Equation 9.4).

TABLE 9.7 Locations in the Literature for Analytic CRPS Formulas for Various Continuous Distributions

Distribution	References
Beta	Jordan et al. (2017), Taillardat et al. (2016)
Exponential	Jordan et al. (2017)
truncated	Jordan et al. (2017)
Gamma	Jordan et al. (2017), Scheuerer and Möller (2015), Taillardat et al. (2016)
Gaussian	
lognormal	Baran and Lerch (2015), Taillardat et al. (2016)
mixture	Gneiting et al. (2007), Grimit et al. (2006), Jordan et al. (2017)
square-root truncated	Hemri et al. (2014), Taillardat et al. (2016)
truncated	Gneiting et al. (2006), Jordan et al. (2017), Taillardat et al. (2016), Thorarinsdottir and Gneiting (2010)
Generalized Pareto	Friederichs and Thorarinsdottir (2012), Jordan et al. (2017)
GEV	Friederichs and Thorarinsdottir (2012), Jordan et al. (2017)
left-censored	Scheuerer (2014)
Laplace	Jordan et al. (2017)
log-Laplace	Jordan et al. (2017)
Logistic	Jordan et al. (2017), Taillardat et al. (2016)
censored	Taillardat et al. (2016)
log-Logistic	Jordan et al. (2017), Taillardat et al. (2016)
square-root censored	Taillardat et al. (2016)
truncated	Jordan et al. (2017), Scheuerer and Möller (2015), Taillardat et al. (2016)
t	Jordan et al. (2017)
Truncated	Jordan et al. (2017)
Uniform	Jordan et al. (2017)
Von Mises (circular)	Grimit et al. (2006)

9.5.2. Energy Score

The *Energy Score* (Gneiting and Raftery, 2007; Gneiting et al., 2008) is a multivariate extension of the CRPS, based on its representation in Equation 9.62. Defining $\boldsymbol{Y}$ and $\boldsymbol{Y}'$ as independent random vectors drawn from the multivariate forecast distribution $F(\boldsymbol{y})$, and $\mathbf{o}$ as the corresponding vector observation, the Energy Score is

$$\mathrm{ES} = E_F\|\boldsymbol{Y} - \boldsymbol{o}\| - \frac{1}{2}E_F\|\boldsymbol{Y} - \boldsymbol{Y}'\|, \tag{9.64}$$

where again E_F denotes statistical expectation with respect to F, and $||\bullet||$ indicates Euclidean distance (Equation 11.6). It is a strictly proper score (Gneiting and Raftery, 2007).

Often closed-form expressions for the expectations in Equation 9.63 will not be available, in which case Monte Carlo evaluation analogous to the scalar expression in Equation 9.83 using a large number of realizations m can be employed. If the elements of the forecast and observation vectors are measured in different physical units it may be advisable to standardize them before calculation of the score. Initial experience with the Energy Score has suggested that it may not be sufficiently sensitive to correlations among the elements of the forecast and observation vectors (Pinson and Girard, 2012; Scheuerer and Hamill, 2015b), and accordingly Scheuerer and Hamill (2015b) suggest that the Dawid-Sebastiani score (Section 9.5.3) may be preferable if the information required to compute it is available.

9.5.3. Ignorance, and Dawid-Sebastiani Scores

The strictly proper Ignorance score (Equations 9.50 and 9.59) also generalizes to probability density forecasts for a continuous predictand. When the forecast PDF is $f(y)$ and the observation is o, the Ignorance score for a single forecast is

$$I = -\ln[f(o)]. \tag{9.65}$$

Equation 9.65 is written in terms of a univariate PDF and a scalar observation, although it is equally applicable to multivariate forecast PDFs and vector observations. The Ignorance score is local, since it is simply the negative logarithm of the forecast PDF evaluated at the observation, regardless of the behavior of $f(y)$ for other values of its argument. If $f(y)$ is a Gaussian forecast PDF with mean μ and variance σ^2, the Ignorance when the observation is o is therefore

$$I = \frac{\ln(2\pi\sigma^2)}{2} + \frac{(o-\mu)^2}{2\sigma^2}, \tag{9.66}$$

where the first term penalizes lack of sharpness, independently of the observation, and the second term penalizes in proportion to the square of the standardized error in the location of the forecast distribution. Equation 9.66 is the negative of the Gaussian log-likelihood (Equation 4.82) for a single ($n = 1$) observation o.

Example 9.7. Comparison of CRPS and Ignorance for 2 Gaussian forecast PDFs

Figure 9.17b compares CRPS and the Ignorance scores for the two Gaussian forecast PDFs shown in Figure 9.17a. The forecast PDF $f_1(y)$ (solid curve) is standard Gaussian (zero mean, unit standard deviation), and $f_2(y)$ (dashed) has mean 1 and standard deviation 3. Since both forecast PDFs are Gaussian, their CRPS and Ignorance scores, as functions of the observation o, can be computed using Equations 9.63 and 9.66, respectively.

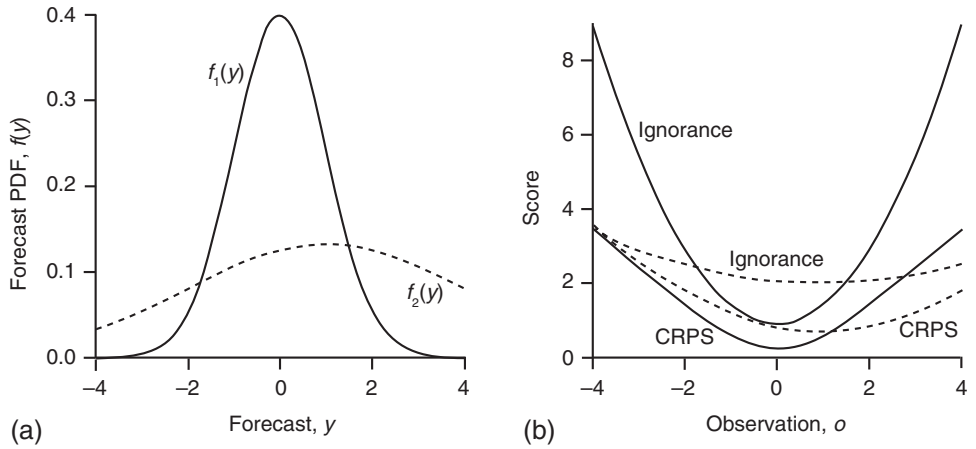

FIGURE 9.17 Comparison of the continuous ranked probability and Ignorance scores, for the two Gaussian PDFs shown in (a), as functions of the observed value (b). The solid PDF $f_1(y)$ is standard Gaussian, and the dashed $f_2(y)$ has mean 1 and standard deviation 3.

Since $f_1(y) = f_2(y)$ for $y \approx -1.7$ and $y \approx 1.5$, the Ignorance scores for the two forecasts are equal for these two values of the observation. On the other hand, CRPS yields a better score for $f_1(y)$ when $o = -1.7$, but a slightly worse score for $f_1(y)$ when $o = 1.5$. It is not immediately clear which of these two results should be preferable, and indeed the preference would generally depend on the circumstances of individual forecast users.

Even though the Ignorance score depends only on the local value $f(o)$ whereas CRPS is an integrated quantity, the two scores behave qualitatively similarly in Figure 9.17b. This similarity in behavior derives from the smoothness of the Gaussian forecast PDFs, which implicitly imparts a degree of sensitivity to distance for the Ignorance score in this example. More prominent differences for the two scores would be expected if the forecast PDFs were bi- or multimodal. The biggest differences between the two scores occurs for extreme values of the observation, 3 or 4 standard deviations away from the forecast mean, which the Ignorance score penalizes relatively more heavily than does the CRPS. ◇

The *Dawid-Sebastiani score* (Dawid and Sebastiani, 1999),

$$\mathrm{DSS} = 2\,\ln(\sigma) + \frac{(o-\mu)^2}{\sigma^2}, \tag{9.67}$$

is closely related to the Ignorance score for Gaussian predictive distributions (Equation 9.66), and indeed is simply a linear transformation of it. However, it can be applied to density forecasts in general when the mean and variance are both specified, in which case it is proper, but not strictly proper. For multivariate forecast PDFs the Dawid-Sebastiani score is

$$\mathrm{DSS} = \ln(\det[\Sigma]) + (\boldsymbol{o}-\boldsymbol{\mu})^T[\Sigma]^{-1}(\boldsymbol{o}-\boldsymbol{\mu}). \tag{9.68}$$

This is a linear transformation of the log-likelihood for the multivariate normal distribution, the PDF for which is Equation 12.1, but it can be used in connection with general multivariate forecasts PDFs and vector observations $\boldsymbol{o}$ when the covariance matrix $[\Sigma]$ is known or can be estimated well. In common with the Ignorance score for Gaussian predictive distributions (Equation 9.66) the DSS penalizes

dispersion of the forecast distribution through the determinant of [Σ] in the first term of Equation 9.68, and penalizes (Mahalanobis, Equation 11.91) distance between the forecast mean and observation vectors in the second term.

9.5.4. The PIT Histogram

It is sometimes of interest to characterize the calibration, as distinct from the accuracy, of a set of probability density forecasts. For this type of forecast, calibration implies that the forecasts and observations are consistent in the sense that, collectively, the observations can plausibly be regarded as a random draws from their respective forecast distributions. For each forecast occasion, k, the *probability integral transform* (PIT) is simply the value of the cumulative probability of the observation o_i within the forecast CDF F_k,

$$u_k = F_k(o_k). \tag{9.69}$$

The forecast distributions F_k may be different for different forecasts, but if they are calibrated then each cumulative probability in Equation 9.69 is a random variable that has been drawn from the standard uniform distribution, $f(u_k) = 1, 0 \le u_k \le 1$. Therefore if a collection of n density forecasts are calibrated, then a histogram of the corresponding collection of PIT values u_k, $k, = 1, \ldots, n$ will be uniform on the unit interval, within the limits of sampling variability. This is the *PIT histogram* (Dawid, 1984; Diebold et al., 1998). Typically between ten and twenty bins are used to discretize the unit interval.

Figure 9.18 shows PIT histograms for a collection of temperature forecasts. The unit interval has been discretized into ten bins, and the histogram bars indicate the relative frequencies of PIT values in each bin. Figure 9.18a shows the PIT histogram for Gaussian distributions fit to individual raw forecast ensembles by computing the ensemble means and ensemble standard deviations, and Figure 9.18b shows corresponding results after NGR (Section 8.3.2) postprocessing. Interpretations of characteristic PIT histogram shapes are the same as for the closely allied Verification Rank Histogram (Section 9.7.1), and the

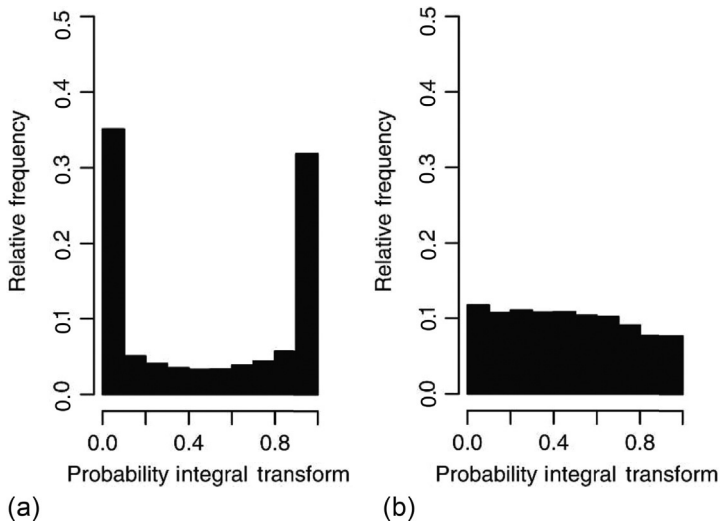

FIGURE 9.18 Example PIT histograms for 56,489 temperature forecasts made at 48h lead time, for locations in the northwestern U.S. and southwestern Canada, for (a) Gaussian distributions defined by raw ensemble means and standard deviations, and (b) NGR (Section 8.3.2) postprocessed forecasts. *From Gneiting et al. (2005). © American Meteorological Society. Used with permission.*

scalar metrics listed in Section 9.7.1 for characterizing rank histogram flatness are equally applicable to PIT histograms. The U-shaped PIT histogram in Figure 9.18a shows that the uncorrected forecast distributions are too narrow on average, so that the corresponding observations fall too frequently in the tails, corresponding to $u_k \approx 0$ or $u_k \approx 1$. The NGR-postprocessed results in Figure 9.18b are much more uniform, although they still reveal a tendency toward warm bias, so that the observations fall too frequently in the left tails of the forecast distributions.

9.6. QUANTILE FORECASTS

9.6.1. The Quantile Score

In some settings only selected quantiles of the predictive distribution are issued as forecasts. Examples include forecasts of the median, and quantile regression (Section 7.7.3) forecasts. In this setting a quantile q is forecast, with the implication that $\Pr\{o \le q\} = \alpha$, where o is the observed value and α is the probability level for the forecast quantile. Such forecasts can be evaluated using the proper *Quantile Score* (Gneiting and Raftery, 2007; Taylor, 1999),

$$\mathrm{QS} = \begin{cases} (q-o)(1-\alpha)\,, & o \le q \\ (o-q)\alpha\,, & o > q \end{cases}. \tag{9.70}$$

The Quantile Score is based on a check function (Equation 7.55). The integral of QS over all forecast quantile probability levels yields the CRPS (Gneiting and Ranjan, 2011). Bentzien and Friederichs (2014) derive a reliability-resolution-uncertainty decomposition of the average QS that is analogous to Equation 9.45 for the Brier score, and propose new verification diagrams for QS. Ben Bouallégue et al. (2015) propose computing skill scores based on QS in the form of Equation 9.4, and draw connections with economic value of forecasts.

9.6.2. Central Credible Interval Forecasts

The burden of communicating a full probability distribution is reduced considerably if the forecast distribution is merely sketched, using the central credible interval (CCI) format (Section 7.10.3). In full form, a CCI forecast consists of a range of the predictand that is centered in a probability sense, together with the probability covered by that range within the forecast distribution. Usually CCI forecasts are abbreviated in one of two ways: either the interval width is constant on every forecast occasion but the location of the interval and the probability it subtends are allowed to vary (fixed-width CCI forecasts), or the probability within the interval is constant on every forecast occasion but the interval location and width may both change (fixed-probability CCI forecasts).

The ranked probability score (Equation 9.56) is an appropriate scalar accuracy measure for fixed-width CCI forecasts (Baker, 1981; Gordon, 1982). In this case there are three categories (below, within, and above the forecast interval) among which the forecast probability is distributed. The probability p pertaining to the forecast interval is specified as part of the forecast, and because the forecast interval is located in the probability center of the distribution, probabilities for the two extreme categories are each $(1–p)/2$. The result is that $\mathrm{RPS} = (p–1)^2/2$ if the observation falls within the interval, or $\mathrm{RPS} = (p^2+1)/2$ if the observation is outside the interval. The RPS thus reflects a balance between preferring a large p if the observation is within the interval, but preferring a smaller p if it is outside, and that balance is optimized when the forecasters report their true judgment.

The RPS is not an appropriate accuracy measure for fixed-probability CCI forecasts. For this forecast format, small (i.e., better) RPS can be achieved by always forecasting extremely wide intervals, because the RPS does not penalize vague forecasts that include wide central intervals. In particular, forecasting an interval that is sufficiently wide that the observation is nearly certain to fall within it will produce a smaller RPS than a verification outside the interval if $(p–1)^2/2 < (p^2+1)/2$. A little algebra shows that this inequality is satisfied for any positive probability p.

Fixed-probability CCI forecasts are appropriately evaluated using *Winkler's score* (Dunsmore, 1968; Winkler, 1972a; Winkler and Murphy, 1979),

$$W = \begin{cases} (b-a+1)+c(a-o), & o<a \\ (b-a+1), & a \le o \le b \\ (b-a+1)+c(o-b), & b<o \end{cases}. \tag{9.71}$$

Here the forecast interval ranges from a to b, inclusive, with $a \le b$, and the value of the observed variable is o. Regardless of the actual observation, a forecast is assessed penalty points equal to the width of the forecast interval, which is $b–a+1$ to account for both endpoints when (as is usual) the interval is specified in terms of integer units of the predictand. An additional penalty is added if the observation falls outside the specified interval, and the magnitudes of these "miss" penalties are proportional to the distance from the interval. Winkler's score thus expresses a trade-off between short intervals to reduce the fixed penalty (and thus encouraging sharp forecasts), versus sufficiently wide intervals to avoid incurring the additional penalties too frequently. This trade-off is balanced by the constant c, which depends on the fixed probability to which the forecast CCI pertains, and increases as the implicit probability for the interval increases, because outcomes outside the interval should occur increasingly rarely for larger interval probabilities. In particular, $c = 4$ for 50% CCI forecasts, and $c = 8$ for 75% CCI forecasts. More generally, $c = 1/F(a)$, where $F(a) = 1 - F(b)$ is the cumulative probability associated with the lower interval boundary according to the forecast CDF. Equation 9.71 has also been called the *Interval Score* by Gneiting and Raftery (2007), who note that it is a proper score.

Winkler's score is equally applicable to fixed-width CCI forecasts, and to unabbreviated CCI forecasts for which the forecaster is free to choose both the interval width and the subtended probability. In these two cases, where the stated probability may change from forecast to forecast, the penalty function for observations falling outside the forecast interval will also change, according to $c = 1/F(a)$.

The calibration (reliability) of fixed-probability CCI forecasts can be evaluated simply, by tabulating the relative frequency over a sample of n forecasts with which the observation falls in the forecast interval, and checking numbers of observations falling above and below the interval for equality. Relative frequency of observations within the interval being less than the specified forecast probability suggests that improvements could be achieved by widening the forecast intervals, on average, and vice versa for the observed relative frequency being larger than the forecast probability. Of course good calibration does not guarantee skillful forecasts, as constant interval forecasts based on the central part of the predictand climatological distribution will also exhibit good calibration.

9.7. VERIFICATION OF ENSEMBLE FORECASTS

Chapter 8 presented the method of ensemble forecasting, in which the effects of initial-condition uncertainty on dynamical forecasts are represented by a finite collection, or ensemble, of very similar initial conditions. Ideally, an initial ensemble represents a random sample from the PDF quantifying initial-condition uncertainty, defined on the phase space of the dynamical model. Integrating the forecast model

forward in time from each of these initial conditions individually thus is a Monte Carlo approach to estimating the effects of the initial-condition uncertainty on the forecast uncertainty for the quantities being predicted. That is, if the initial ensemble members have been chosen as a random sample from the initial-condition uncertainty PDF, and if the forecast model consists of correct and accurate representations of the physical dynamics, the ensemble after being integrated forward in time represents a random sample from the PDF of forecast uncertainty. If this ideal situation could be realized, the true state of the atmosphere would be just one more member of the ensemble, at the initial time and throughout the integration period, and should be statistically indistinguishable from the forecast ensemble. This condition, that the actual future atmospheric state behaves like a random draw from the same distribution that produced the ensemble, is called *ensemble consistency* (Anderson, 1997), and is closely related to the notion of exchangeability

In light of this background, it should be clear that ensemble forecasts are probability forecasts that are expressed as a discrete approximation to a full forecast PDF. According to this approximation, ensemble relative frequency should estimate actual probability. Depending on what the predictand(s) of interest may be, the formats for these probability forecasts can vary widely. Probability forecasts can be obtained for simple predictands, such as continuous scalars (e.g., temperature or precipitation at a single location), or discrete scalars (possibly constructed by thresholding a continuous variable, e.g., zero or trace precipitation vs. nonzero precipitation, at a given location), or quite complicated multivariate predictands such as entire fields (e.g., the joint distribution of 500 mb heights at the global set of horizontal gridpoints).

When ensemble forecasts are expressed as conventional probability forecasts, perhaps after transformation to a probability for a dichotomous outcome, or after kernel density smoothing of the ensemble to yield a continuous probability distribution (e.g., Roulston and Smith, 2003), or by fitting parametric probability distributions (e.g., Hannachi and O'Neill, 2001; Stephenson and Doblas-Reyes, 2000; Wilks, 2002b), then the probabilistic forecast verification methods presented in Sections 9.4–9.6 can be applied to evaluate them. The same is true for results of ensemble postprocessing methods that yield conventional probability forecasts, such as many of the methods outlined in Section 8.3.2–8.3.4. Similarly, ensemble-mean forecasts can be evaluated using methods described in Section 9.3. However, other verification tools have been developed for evaluation of raw ensemble forecasts, or postprocessed ensembles expressed as discrete approximations to underlying probability distributions (as described in Sections 8.4.2 and 8.4.3), and these methods are presented in the following subsections. Some of these ensemble verification methods have been designed to evaluate ensemble consistency, through calibration diagnostics, and others are accuracy measures that operate on the discrete samples provided by ensembles.

9.7.1. Assessing Univariate Ensemble Calibration

To the extent that the ensemble consistency condition has been met, the observation being predicted looks statistically like just another member of the forecast ensemble. The result is that the ensemble will be calibrated, so that ensemble relative frequency can be regarded as a good estimator of probability, regardless of whether those probability estimates are sharp or not. Several methods have been proposed to evaluate ensemble calibration.

Verification Rank Histogram

Construction of a *verification rank histogram*, sometimes called simply the *rank histogram*, is the most common approach to evaluating whether a collection of ensemble forecasts for a scalar predictand satisfy

the consistency condition. That is, the rank histogram is used to evaluate whether the forecast ensembles apparently include the observations being predicted as equiprobable members. The rank histogram is a graphical approach that was devised independently by Anderson (1996), Hamill and Colucci (1998), and Talagrand et al. (1997), and is sometimes also called the *Talagrand diagram*.

Consider the evaluation of n ensemble forecasts, each of which consists of m ensemble members, in relation to the n corresponding observed values for the predictand. Within each of these n sets, if the m members and the single observation all have been drawn from the same distribution, then the rank of the observation within these $m+1$ values is equally likely to be any of the integers $i = 1, 2, 3, \ldots, m+1$. For example, if the observation is smaller than all m ensemble members, then its rank is $i = 1$. If it is larger than all the ensemble members then its rank is $i = m+1$. For each of the n forecasting occasions, the rank of the observation within this $m+1$ -member distribution is tabulated. Collectively these n verification ranks are plotted in the form of a histogram to produce the verification rank histogram. (Equality of the observation with one or more of the ensemble members requires a slightly more elaborate procedure; see Hamill and Colucci, 1998). If the ensemble consistency condition has been met this histogram of verification ranks will be uniform, reflecting equiprobability of the observations within their ensemble distributions, except for departures that are small enough to be attributable to sampling variations. In the limit of infinite ensemble size, or if the ensemble distribution has been represented as a smooth, continuous PDF, the rank histogram is identical to the PIT histogram (Section 9.5.4).

Departures from the ideal of rank uniformity can be used to diagnose aggregate deficiencies of the ensembles (Hamill, 2001). Figure 9.19 shows four problems that can be discerned from the rank histogram, together with a rank histogram (center panel) that shows only small sampling departures from a uniform distribution of ranks, or rank uniformity. The horizontal dashed lines in Figure 9.19 indicate the relative frequency $[= (m+1)^{-1}]$ attained by a uniform distribution for the ranks, which is often plotted for reference as part of the rank histogram. The hypothetical rank histograms in Figure 9.19 each have $m+1 = 9$ bars, and so would pertain to ensembles of size $m = 8$.

Overdispersed ensembles produce rank histograms with relative frequencies concentrated in the middle ranks (left-hand panel in Figure 9.19). In this situation excessive dispersion produces ensembles that range beyond the observation more frequently than would occur by chance if the ensembles exhibited consistency. The verification is accordingly an extreme member of the $m+1$ -member (ensemble + verification) collection too infrequently, so that the extreme ranks are underpopulated; and is near the center of the ensemble too frequently, producing overpopulation of the middle ranks. Conversely, a set of n underdispersed ensembles produce a U-shaped rank histogram (right-hand panel in Figure 9.19) because the ensemble members tend to be too similar to each other, and different from the verification. The result is that the verification is too frequently an outlier among the collection of $m+1$ values, so the extreme ranks are overpopulated; and occurs too rarely as a middle value, so the central ranks are underpopulated.

An appropriate degree of ensemble dispersion is a necessary condition for a set of ensemble forecasts to exhibit consistency, but it is not sufficient. It is also necessary for consistent ensembles not to exhibit unconditional biases. That is, consistent ensembles will not be centered either above or below their corresponding verifications, on average. Unconditional ensemble bias can be diagnosed from overpopulation of either the smallest ranks, or the largest ranks, in the verification rank histogram. Forecasts that are centered above the verification, on average, exhibit overpopulation of the smallest ranks (upper panel in Figure 9.19) because the tendency for overforecasting leaves the verification too frequently as the smallest or one of the smallest values of the $m+1$ -member collection. Similarly, underforecasting bias (lower panel in Figure 9.19) produces overpopulation of the higher ranks, because a consistent

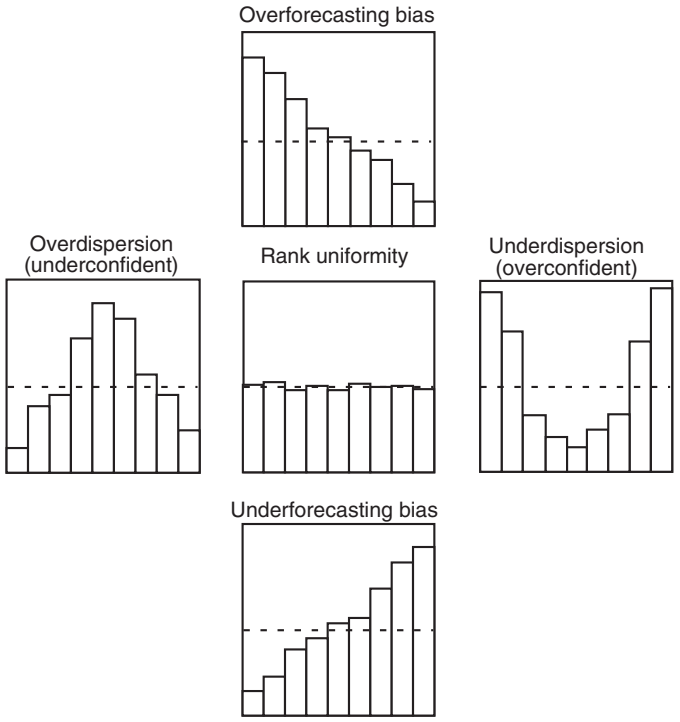

FIGURE 9.19 Example verification rank histograms for hypothetical ensembles of size $m = 8$, illustrating characteristic ensemble dispersion and bias errors. Perfect rank uniformity is indicated by the horizontal dashed lines. The arrangement of the panels corresponds to the calibration portions of the reliability diagrams in Figure 9.10a.

tendency for the ensemble to be below the verification leaves the verification too frequently as the largest or one of the largest members.

There are connections with the calibration function $p(o_j | y_i)$ that is plotted as part of the reliability diagram (Section 9.4.4), which can be appreciated by comparing Figures 9.19 and 9.10a. The five pairs of panels in these two figures bear a one-to-one correspondence for forecast ensembles yielding probabilities for a dichotomous variable defined by a fixed threshold applied to a continuous predictand. That is, the "yes" component of a dichotomous outcome occurs if the value of the continuous predictand y is at or above a threshold. For example, the event "precipitation occurs" corresponds to the value of a continuous precipitation variable being at or above a detection limit, such as 0.01 in. In this setting, forecast ensembles that would produce each of the five reliability diagrams in Figure 9.10a would exhibit rank histograms having the forms in the corresponding positions in Figure 9.19.

Correspondences between the unconditional bias signatures in these pairs of figures are easiest to understand. Ensemble overforecasting (upper panels) yields average probabilities that are larger than average outcome relative frequencies in Figure 9.10a, because ensembles that are too frequently centered above the verification will exhibit a majority of members above a given threshold more frequently than the verification is above that threshold (or, equivalently, more frequently than the corresponding probability of being above the threshold, according to the climatological distribution). Conversely, underforecasting (lower panels) simultaneously yields average probabilities for dichotomous events that are smaller than

the corresponding average outcome relative frequencies in Figure 9.10a, and overpopulation of the higher ranks in Figure 9.19.

In underdispersed ensembles, most or all ensemble members will fall too frequently on one side or the other of the threshold defining a dichotomous event. The result is that probability forecasts from underdispersed ensembles will be excessively sharp and will use extreme probabilities more frequently than justified by the ability of the ensemble to resolve the event being forecast. The probability forecasts therefore will be overconfident. That is, too little uncertainty is communicated, so that the conditional event relative frequencies are less extreme than the forecast probabilities. Reliability diagrams reflecting such conditional biases, in the form of the right-hand panel of Figure 9.10a, are the result. Conversely, overdispersed ensembles will rarely have a large majority of members on one side or the other of the event threshold, so the probability forecasts derived from them will rarely be extreme. These probability forecasts will be underconfident and produce conditional biases of the kind illustrated in the left-hand panel of Figure 9.10a, namely, that the conditional event relative frequencies tend to be more extreme than the forecast probabilities.

Several single-number summaries with which to characterize the degree of rank histogram flatness are available. They are equally applicable to characterization of the flatness of the PIT histogram (Section 9.5.4). The χ^2 statistic is the first of these summaries, which essentially addresses the goodness of the fit (Section 5.2.5) of a rank histogram to a discretized uniform distribution,

$$\chi^2 = \frac{m+1}{n} \sum_{i=1}^{m+1} \left(n_i - \frac{n}{m+1} \right)^2, \tag{9.72}$$

where n_i is the number of counts in the ith rank histogram bin. An advantage to using of the χ^2 statistic in this context is that its sampling distribution is known, and so formal testing of the null hypothesis of a flat rank histogram can be undertaken (Section 9.11.5). The *Reliability Index* (Delle Monache et al., 2006) operates similarly to the χ^2 flatness statistic, but measures absolute rather than squared discrepancies from uniformity,

$$\mathrm{RI} = \frac{1}{n} \sum_{i=1}^{m+1} \left| n_i - \frac{n}{m+1} \right|. \tag{9.73}$$

This statistic is sometimes referred to as D, or Δ by some authors, and is sometimes expressed as a percentage after multiplication by 100%. Taillardat et al. (2016) propose characterizing rank histogram flatness with the entropy statistic

$$\varphi = \frac{-1}{\ln(m+1)} \sum_{i=1}^{m+1} \frac{n_i}{n} \ln\left(\frac{n_i}{n} \right), \tag{9.74}$$

which achieves its maximum of $\varphi = 1$ for calibrated forecasts.

Lack of uniformity in a rank histogram quickly reveals the presence of conditional and/or unconditional biases in a collection of ensemble forecasts, but unlike the reliability diagram it does not provide a complete picture of forecast performance in the sense of fully expressing the joint distribution of forecasts and observations. In particular, the rank histogram does not include a representation of the refinement, or sharpness, of the ensemble forecasts. Rather, it indicates only if the forecast refinement is appropriate, relative to the degree to which the ensemble can resolve the predictand. The nature of this

incompleteness can be appreciated by imagining the rank histogram for ensemble forecasts constructed as random samples of size m from the historical climatological distribution of the predictand. Such ensembles would be consistent, by definition, because the value of the predictand on any future occasion will have been drawn from the same distribution that generated the finite sample in each ensemble. The resulting rank histogram would be accordingly flat, but would not reveal that these forecasts exhibited so little refinement as to be useless.

If these climatological ensembles were to be converted to probability forecasts for a discrete event according to a fixed threshold of the predictand, in the limit of $m -> \infty$ their reliability diagram would consist of a single point, located on the 1:1 diagonal, at the magnitude of the climatological relative frequency. This abbreviated reliability diagram immediately would communicate the fact that the forecasts underlying it exhibited no sharpness, because the same event probability would have been forecast on each of the n occasions. Of course real ensembles are of finite size, and climatological ensembles of finite size would exhibit sampling variations from forecast to forecast, yielding a refinement distribution $p(y_i)$ (Equation 9.2) with nonzero variance, but a reliability diagram exhibiting no resolution and therefore a horizontal calibration function (indicated by the "no resolution" line in Figure 9.12).

Even when a set of consistent ensemble forecasts can resolve the event better than does the climatological distribution, sampling variations in the resulting probability estimates will generally lead to reliability diagram calibration functions that suggest overconfidence. Richardson (2001) presents analytic expressions for this apparent overconfidence when the probability estimates are estimated using ensemble relative frequency (Equation 8.4), which indicate that this effect decreases with increasing ensemble size, and with forecast accuracy as measured by decreasing Brier scores. This same effect produces sampling-error-based degradations in the Brier and ranked probability scores for probability forecasts based on ensemble relative frequency (Ferro et al., 2008).

It is worth pointing out that a flat rank histogram is a necessary, but not sufficient condition for diagnosing ensemble calibration. In particular, a flat rank histogram can result from compensating effects of different types of miscalibration in subsets of the data set being examined (Hamill, 2001).

Relating Ensemble-Mean Errors and Ensemble Variances

Two consequences of the ensemble consistency condition are that forecasts from the individual ensemble members (and therefore also the ensemble-mean forecasts) are unbiased, and that the average (over multiple forecast occasions) MSE for the ensemble-mean forecasts should be equal to the average ensemble variance. If, on any given forecast occasion, the observation o is statistically indistinguishable from any of the ensemble members y_i, then clearly the bias is zero since $\mathrm{E}[y_i] = \mathrm{E}[o_i]$. Statistical equivalence of any ensemble member and the observation further implies that

$$E\left[(o-\overline{y})^2\right] = E\left[(y_i-\overline{y})^2\right], \tag{9.75}$$

where $\overline{y}$ denotes the ensemble mean. Realizing that $(o–\overline{y})^2 = (\overline{y}–o)^2$, it is easy to see that the left-hand side of Equation 9.75 is the MSE for the ensemble-mean forecasts, whereas the right-hand side expresses dispersion of the ensemble members y_i around the ensemble mean, as the average ensemble variance. It is important to realize that Equation 9.75 holds only for forecasts from a consistent ensemble, and in particular assumes unbiasedness in the forecasts. Forecast biases will inflate the ensemble-mean MSE without affecting the ensemble dispersion (because ensemble dispersion is computed relative to the sample ensemble mean), so that this diagnostic cannot distinguish forecast bias from ensemble underdispersion (Wilks, 2011). Therefore attempting to correct ensemble underdispersion by inflating

ensemble variance to match ensemble-mean MSE will yield overdispersed ensembles if the underlying forecasts are biased.

Accounting for finite ensemble size m, the sample counterpart of Equation 9.75 is (e.g., Leutbecher and Palmer, 2008)

$$\frac{m}{m+1}\left[\frac{1}{n}\sum_{t=1}^{n}(\overline{x}_t - o_t)^2\right] = \frac{m}{m-1}\left[\frac{1}{n}\sum_{t=1}^{n}s_t^2\right], \tag{9.76}$$

where the ensemble mean $\overline{x}_t$ on the left-hand side is defined by Equation 8.10, and the ensemble variance s_t^2 on the right-hand side is defined by Equation 8.9. When the ensemble size is large the corrections outside the square brackets in Equation 9.76 are often neglected. If the ensemble variances have been computed using division by m–1 rather than by m as in Equation 8.9, then the correction $m/(m–1)$ on the right-hand side is not needed.

Equation 9.76 applies both to a full n-member verification data set and also to subsets. The *binned spread-error diagram* (Van den Dool, 1989) is a graphical device that displays both the extent to which a collection of ensemble forecasts to be verified is consistent with Equation 9.75 and the degree to which variations in ensemble spread may predict variations in forecast accuracy. It is constructed by collecting subsets of the verification data having similar values for the ensemble variance into discrete bins that define points on the horizontal axis and plotting the corresponding conditional average squared errors on the vertical. Although these MSE values reflect conditional accuracy, these diagrams are sometimes inaccurately called "spread-skill" diagrams. Often it is the square roots of the two sides of Equation 9.76 that are plotted in binned spread-error diagrams, in which case it is important that the square root of the MSE of the left-hand side is used, rather than the average RMSE (Fortin et al., 2014, 2015).

Figure 9.20 shows two examples, where twenty equally populated class intervals for the ensemble spread have been chosen. The binned spread-error plot for a collection of uncorrected ensemble forecasts in Figure 9.76a shows that Equation 9.76 has not been satisfied, because the points are well away from the desired 1:1 line. The conditional ensemble-mean errors are too large relative to the binned ensemble spreads, which could indicate underdispersion or bias or both. It also indicates that there is a predictive relationship between ensemble spread and average forecast accuracy. The rank histogram for these raw ensemble forecasts (Figure 9.21) shows both underdispersion and a strong negative bias, although it does not by itself give

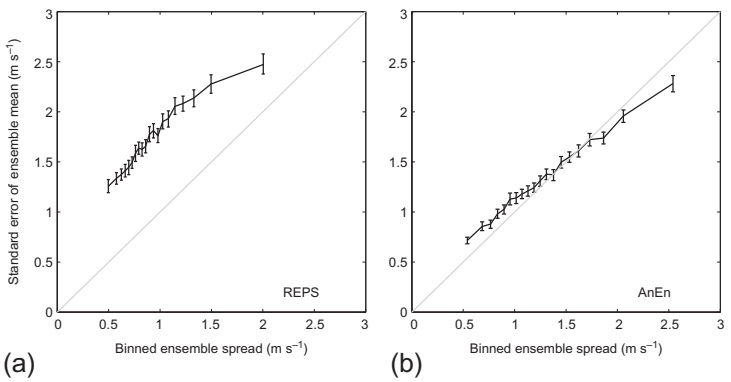

FIGURE 9.20 Binned spread-error diagrams for (a) raw ensemble forecasts and (b) corresponding postprocessed forecasts of surface wind speed at 42 h lead time. Error bars indicate 95% bootstrap confidence intervals. *From Delle Monache et al. (2013). © American Meteorological Society. Used with permission.*

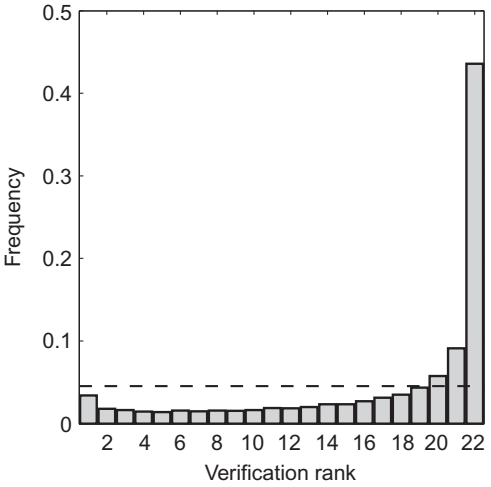

FIGURE 9.21 Rank histogram for the raw ensemble forecasts whose binned-spread error diagram is shown in Figure 9.20a. *Modified from Delle Monache et al. (2013). © American Meteorological Society. Used with permission.*

an indication of the accuracy of these forecasts. The spread-error diagram for the postprocessed ensembles in Figure 9.76b shows much better calibration, since the points falling close to the 1:1 line support Equation 9.76 having been satisfied approximately, suggesting that these postprocessed forecasts are both nearly unbiased and appropriately dispersive.

Another alternative for displaying the degree to which average ensemble spread and ensemble-mean accuracy correspond is to plot the two sides of Equation 9.76 (or their square roots) unconditionally, as functions of lead time. Figure 9.22 does this for the same forecasts underlying Figure 9.20, which shows again that the raw ensembles in Figure 9.22a are poorly calibrated whereas the postprocessed ensembles in Figure 9.22b satisfy Equation 9.76 to good approximation.

9.7.2. Assessing Multivariate Ensemble Calibration

Assessment of calibration for multivariate (i.e., vector) ensemble forecasts is more difficult, since multivariate calibration implies that both the marginal distributions of each of the elements of the forecast vectors, and their joint relationships, must be correctly represented. Several graphical approaches to assessing multivariate calibration have been proposed.

Minimum Spanning Tree (MST) Histogram

The concept behind the verification rank histogram (Section 9.7.1) can be extended to ensemble forecasts for multiple predictands, using the *minimum spanning tree* (MST) *histogram*, which allows investigation of simultaneous forecast calibration in multiple dimensions. This idea was proposed by Smith (2001) and explored more fully by Smith and Hansen (2004), Wilks (2004), and Gombos et al. (2007). The MST histogram is constructed from an ensemble of d-dimensional vector forecasts $\boldsymbol{y}_i$, $i = 1, \ldots, m$, and the corresponding vector observation $\boldsymbol{o}$. Each of these vectors defines a point in a d-dimensional space, the coordinate axes of which corresponds to the d variables in the vectors $\boldsymbol{y}$ and $\boldsymbol{o}$. In general, these vectors will not have a natural ordering in the same way that a set of $m+1$ scalars would, so

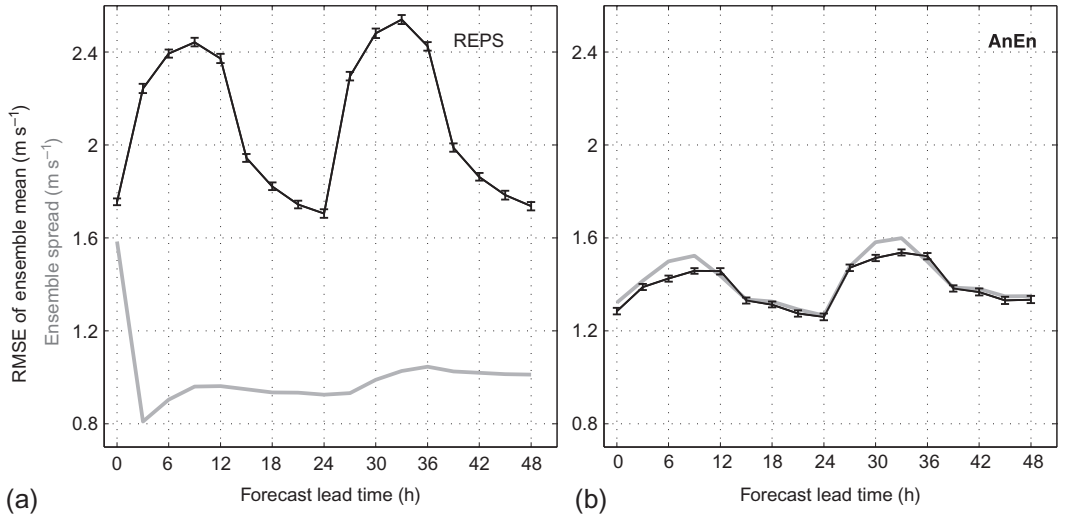

FIGURE 9.22 Square roots of the two sides of Equation 9.76 as a function of lead time, for the same forecasts underlying the 42-h lead forecasts in Figure 9.20. *From Delle Monache et al. (2013).* © *American Meteorological Society. Used with permission.*

the conventional verification rank histogram is not applicable to these multidimensional quantities. The minimum spanning tree for m the members $\boldsymbol{y}_i$ of a particular ensemble is the set of line segments (in the d-dimensional space of these vectors) that connect all the points $\boldsymbol{y}_i$ in an arrangement having no closed loops, and for which the sum of the lengths of these line segments is minimized. The solid lines in Figure 9.23 show a minimum spanning tree for a hypothetical $m = 10$-member forecast ensemble, labeled A–J.

If each (multidimensional) ensemble member is replaced in turn with the observation vector $\boldsymbol{o}$, the lengths of the minimum spanning trees for each of these substitutions make up a set of m reference MST lengths. The dashed lines in Figure 9.23 show the MST obtained when ensemble member D is replaced by the observation, O. To the extent that the ensemble consistency condition has been satisfied,

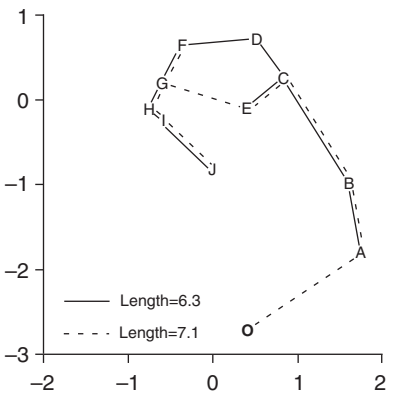

FIGURE 9.23 Hypothetical example minimum spanning trees in $d = 2$ dimensions. The $m = 10$ ensemble members are labeled A–J, and the corresponding observation is O. Solid lines indicate MST for the ensemble as forecast, and dashed lines indicate the MST that results when the observation is substituted for ensemble member D. *From Wilks (2004).*

the observation vector is statistically indistinguishable from any of the forecast vectors $\mathbf{y}_i$, implying that the length of the MST connecting only the m vectors $\mathbf{y}_i$ has been drawn from the same distribution of MST lengths as those obtained by substituting the observation for each of the ensemble members in turn. The MST histogram investigates the plausibility of this proposition, and thus the plausibility of ensemble consistency for the n d-dimensional ensemble forecasts, by tabulating the ranks of the MST lengths for the original ensembles within each group of $m+1$ MST lengths. A collection of consistent multivariate ensembles should yield a flat MST histogram, within reasonable sampling variations, and the degree of flatness can be characterized using Equations 9.72, 9.73, or 9.74.

Although the concept behind the MST histogram is similar to that of the univariate rank histogram, it is not a multidimensional generalization of the rank histogram, and the interpretations of MST histograms are different (Wilks, 2004). Some example MST histograms are shown in the bottom row of Figure 9.25. Underdispersed ensembles will yield negatively sloped (MSTs excluding the observation vector are typically shorter than those that include it) rather than U-shaped MST histograms, and overdispersed ensembles will exhibit the reverse. In raw form, the MST histogram is unable to distinguish between ensemble underdispersion and bias (the outlier observation O in Figure 9.23 could be the result of either of these problems), and deemphasizes variables in the forecast and observation vectors with small variance. However, useful diagnostics can be obtained from MST histograms of debiased and rescaled forecast and observation vectors. When the elements of the forecast vectors are measured in different physical units, or when some elements are intrinsically more variable than others, it will be advantageous to standardize each of the variables in order to avoid the result being dominated by the higher-variance elements (Wilks, 2004).

Multivariate Rank Histograms

Direct extension of the verification rank histogram for multivariate (i.e., vector) forecasts is not straightforward, because an unambiguous ranking of a collection of vectors along the real line must be defined before proceeding to assign a particular ensemble to one of the $m+1$ rank histogram bins. Gneiting et al. (2008) proposed the concept of the *prerank*, $\pi(\mathbf{y}_k)$ to define one-dimensional orderings of a collection of vectors $\mathbf{y}_i$ that allow accumulation of counts into rank histogram bins. For notational convenience in this section, define the d-dimensional vector of ensemble members $\mathbf{y}_i$, $i = 1, \ldots, m$, and define the corresponding vector observation as $\mathbf{y}_0$ rather than $\boldsymbol{o}$. Having evaluated the preranks $\pi(\mathbf{y}_i)$ for all $m + 1$ of these vectors, a forecast ensemble contributes a count to rank histogram bin b according to

$$b = 1 + \sum_{i=1}^{m} I[\pi(\mathbf{y}_i) < \pi(\mathbf{y}_0)], \tag{9.77}$$

where as before the indicator function $I(\cdot)=1$ if its argument is true, and equals zero otherwise, so that $1 \leq b \leq m+1$. That is, Equation 9.77 defines the multivariate rank histogram bin to which an ensemble contributes by counting the number of ensemble members whose prerank is smaller than the prerank of the observation and adding 1. Equation 9.77 generalizes the scalar verification rank histogram (Section 9.7.1), to which it reduces when $d = 1$ and $\pi(y_i) = y_i$.

Several prerank functions have been proposed. The original (Gneiting et al., 2008) paper used the multivariate rank histogram prerank

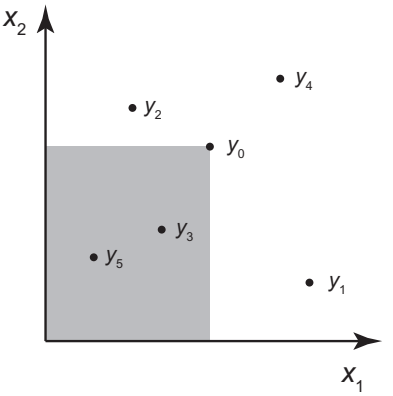

FIGURE 9.24 Illustration of the prerank function in Equation 9.78 for a $d=2$ dimensional observation vector $\boldsymbol{y}_0$ and a hypothetical $m=5$ member ensemble. Ensemble members within the gray shading contribute nonzero terms in Equation 9.78.

$$\pi_{\text{MRH}}(\boldsymbol{y}_i) = \sum_{j=0}^{m} I\left[\boldsymbol{y}_j \preceq \boldsymbol{y}_i\right], \tag{9.78}$$

where $\boldsymbol{y}_j \preceq \boldsymbol{y}_i$ if *each* of the d elements of $\boldsymbol{y}_j$ is less than or equal to its counterpart in $\boldsymbol{y}_i$. Each prerank $\pi_{\text{MRH}}(\boldsymbol{y}_i)$, $i = 0, 1, m$, is an integer between 1 and $m+1$, and ties among the preranks are resolved at random. Figure 9.24 illustrates the operation of Equation 9.78 for a hypothetical $m = 5$ member ensemble and corresponding observation, of dimension $d = 2$. Here, the prerank $\pi_{\text{MRH}}(\boldsymbol{y}_0) = 3$ for the observation vector $\boldsymbol{y}_0$ is equal to the number of ensemble members both below and to the left of $\boldsymbol{y}_0$ (i.e., within the gray shading), and including $\boldsymbol{y}_0$.

Equation 9.78 is not the only way to order a collection of vectors, and alternative prerank functions have been defined. Thorarinsdottir et al. (2016) proposed the *average-rank* histogram (ARH), computation of which begins with component-wise ranks for each of the d elements of the d-dimensional vectors $\boldsymbol{y}_k$,

$$c_l(y_i) = \sum_{j=0}^{m} I\left[y_{j,l} \le y_{i,l}\right], \quad l = 1, \ldots, d. \tag{9.79}$$

The $m+1$ preranks for the ARH are then simply the average of the d component-wise ranks in Equation 9.79,

$$\pi_{\text{ARH}}(\boldsymbol{y}_i) = \frac{1}{d}\sum_{l=1}^{d} c_l(\boldsymbol{y}_i). \tag{9.80}$$

The ARH preranks need not be integer-valued. Ties among the ARH preranks are again resolved at random. Both Equation 9.78 and 9.80 are "ascending rank" methods in that in some sense they reflect distance from the smallest values.

Thorarinsdottir et al. (2016) also introduced the band-depth rank histogram (BDH), the computations for which also begin with the component-wise ranks (Equation 9.79). The BDH prerank function is

$$\pi_{\text{BDH}}(\boldsymbol{y}_i) = \frac{1}{d}\sum_{l=1}^{d} [m+1-c_l(\boldsymbol{y}_i)][c_l(\boldsymbol{y}_i)-1]. \tag{9.81}$$

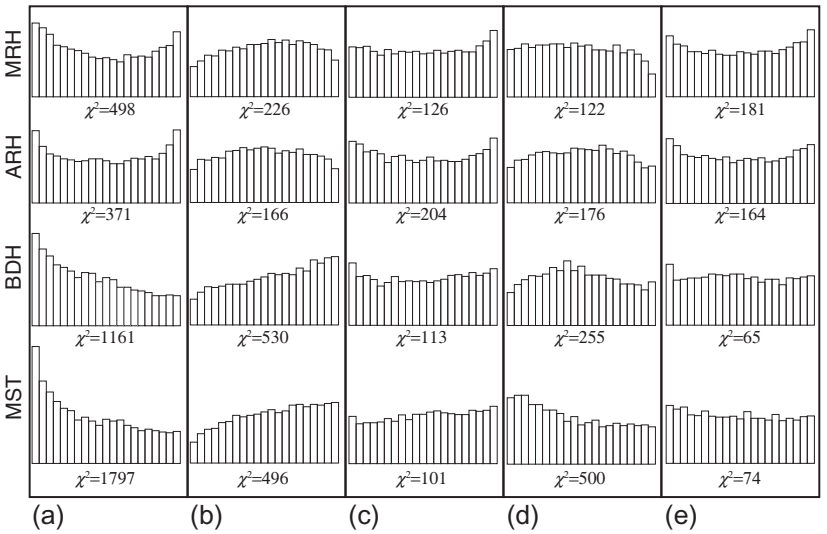

FIGURE 9.25 Representative rank histograms for unbiased ensembles according to MRH (Equation 9.78), ARH (Equation 9.80), BDH (Equation 9.81), and MST (Equation 9.82) prerank functions in an idealized situation with $m = 20$; when the variances are (a) underforecast, and (b) overforecast; when the correlations among the three variables are (c) underforecast, and (d) overforecast; and when (e) the forecast distribution is rotated 20° in the $d = 3$-dimensional forecast space. χ^2 values characterize corresponding histogram flatness according to Equation 9.72. *Modified from Wilks (2017).*

Again, prerank ties are resolved at random. Unlike the MRH and ARH prerank functions, which analogously to the familiar univariate rank histogram provide measures of "ascending rank" of the observation vector $\boldsymbol{y}_0$ relative to the ensemble, the BDH prerank function in Equation 9.81 assesses "centrality," in the sense that $\pi_{\text{BDH}}(\boldsymbol{y}_0)$ is large when $\boldsymbol{y}_0$ is near the middle of the ensemble, and is small when $\boldsymbol{y}_0$ is extreme relative to the ensemble.

Computation of the minimum spanning tree histogram can also be cast in the framework of preranks, where the MST prerank function is

$$\pi_{\text{MST}}(\boldsymbol{y}_i) = \|MST[\{\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_m\} \backslash \{\boldsymbol{y}_i\}]\|, \tag{9.82}$$

where $||\bullet||$ denotes Euclidean length (Equation 11.14). That is, the MST prerank for the vector $\boldsymbol{y}_i$ is the minimum sum of lengths of m–1 line segments connecting the m points of the ensemble plus the observation in their d-dimensional space, when (as denoted by the backslash) the vector $\boldsymbol{y}_i$ is omitted. The MST histogram also assesses centrality of the observation within the ensemble.

Figure 9.25 shows that the various prerank functions that have been proposed react differently to different types of miscalibration, although the two ascending-rank methods MRH and ARH behave similarly to each other, as do the two center-outward methods BDH and MST. Thorarinsdottir et al. (2016) and Mirzargar and Anderson (2017) provide additional comparisons. The synthetic ensembles underlying the histograms in Figure 9.25 were constructed without bias, but biases are more sensitively detected using scalar verification rank histograms separately for each of the d forecast dimensions (Scheuerer and Hamill, 2015b; Wilks, 2017). None of the four methods dominate the others in terms of detecting all types and combinations of types of miscalibration, so that it is good practice to examine the results of multiple preranking methods (Thorarinsdottir et al., 2016; Wilks, 2017).

9.7.3. Assessing Univariate Ensemble Accuracy and Skill

Ensemble CRPS

Although the continuous ranked probability score is defined in Equation 9.61 as the integral over a continuous CDF, the alternative representation in Equation 9.62 provides the basis for a discrete, or "ensemble" CRPS. In particular, straightforward substitution of sample averages within an individual forecast ensemble into Equation 9.62 yields (Van Schaeybroeck and Vannitsem, 2015)

$$\text{eCRPS} = \frac{1}{m}\sum_{i=1}^{m}|y_i - o| - \frac{1}{m(m-1)}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}|y_i - y_j|. \tag{9.83}$$

The second term in Equation 9.83 is half of the λ_2, or L-scale dispersion statistic (Equation 5.33). Terms for which $x_i = x_j$ are not included in the double summation (Ferro et al., 2008, Wilks, 2018a) because these are not independent realizations from the underlying distribution with respect to which the expectations are calculated in Equation 9.62. Alternative formulations for the eCRPS have been derived by Bröcker (2012d), Candille and Talagrand (2005), and Hersbach (2000).

Dawid–Sebastiani Score

The Dawid–Sebastiani score, which is equivalent to the Ignorance score for a continuous Gaussian predictive distribution, was expressed in Equation 9.67 as a function of the mean and variance of a continuous predictive distribution. Substitution of a sample ensemble mean (Equation 8.10) and ensemble variance (Equation 8.11) yields the discrete counterpart,

$$\text{eDSS} = \ln\left(s^2\right) + \frac{(o-\bar{y})^2}{s^2}. \tag{9.84}$$

9.7.4. Assessing Multivariate Ensemble Accuracy and Skill

Ensemble Energy Score

Gneiting et al. (2008) proposed the ensemble Energy Score, as a multivariate extension of the eCRPS (Equation 9.83) to vector forecasts, and as a discretized version of the Energy Score for continuous multivariate predictive distributions (Equation 9.64),

$$\text{eES} = \frac{1}{m}\sum_{i=1}^{m}\|\boldsymbol{y}_i - \boldsymbol{o}\| - \frac{1}{m(m-1)}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\|\boldsymbol{y}_i - \boldsymbol{y}_j\|. \tag{9.85}$$

Here the Euclidean distance in the d-dimensional space of the vector ensemble members $\boldsymbol{x}$ and observation $\boldsymbol{o}$ substitute for the absolute values in Equation 9.84, and both of these equations follow as discrete versions of the CRPS as expressed in Equation 9.62. As was also the case for calculating minimum spanning trees (Section 9.7.2), when the elements of the forecast and observed vectors are expressed in different physical units, or when their distributions differ substantially in range or spread, it may be advantageous to standardize each element before calculation of Equation 9.85, for example, by dividing by standard deviations. Although it is conceptually appealing as a direct multivariate generalization of the CRPS, experience with the ensemble Energy Score has shown it to exhibit weak sensitivity to misspecification of correlations

among the elements of the forecast vectors (Pinson and Girard, 2012; Scheuerer and Hamill, 2015b). Accordingly its use as a sole metric for ensemble forecast accuracy may be problematic.

Ensemble Variogram Score

The ensemble Variogram Score (Scheuerer and Hamill, 2015b) is an alternative to the ensemble Energy Score that is more responsive to errors in the specification of correlations among the elements of the forecast and observation vectors. The ensemble Variogram Score of order λ is defined as

$$\mathrm{eVS} = \sum_{k=1}^{d}\sum_{j=1}^{d} w_{k,j}\left(\left|o_k - o_j\right|^{\lambda} - \frac{1}{m}\sum_{i=1}^{m}\left|y_{k,i} - y_{j,i}\right|^{\lambda}\right)^2, \tag{9.86}$$

which is composed of the sum of weighted squared differences for all pairs of variogram (or *structure function*) approximations, among elements of the forecast and observed vectors. The ensemble Variogram score is proper, but not strictly proper. Both the nonnegative weights $w_{k,j}$ and the exponent order λ are adjustable parameters that must be selected. Scheuerer and Hamill (2015b) provide some guidance on criteria that could inform choices for the weights. They also found that selecting $\lambda = 0.5$ yielded good results in their example, but note that other choices are possible. Biases that are similar for all elements of the forecast vectors (e.g., in a spatial setting where gridpoint values all tend to be too wet or too warm) might not be penalized appreciably by eVS. Similarly to MST and eES calculation, standardization of the forecast and observation vector elements will be advantageous in settings where they have been expressed on incommensurate scales.

Ensemble Dawid–Sebastiani Score

In principle, the Dawid–Sebastiani score (Equation 9.68) could also be adapted for multivariate ensemble forecasts, by substituting the vector ensemble mean for $\boldsymbol{\mu}$ and an ensemble-based estimate of the covariance matrix for $[\Sigma]$. However, in practice implementation of this idea is problematic. First, unless the ensemble size m is larger than the forecast and observation vector dimension d, the estimated covariance matrix will be singular so that its inversion in Equation 9.68 will fail. Furthermore, unless $m >> d$ the covariance matrices will be estimated poorly, so that the resulting score calculation may be extremely misleading (Feldmann et al., 2015; Scheuerer and Hamill, 2015b). Since ensemble sizes are strongly limited by computational constraints, application of the Dawid–Sebastiani score to ensemble forecasts is unlikely to be feasible unless the forecast and observation vector dimension d is perhaps 2 or 3.

9.8. NONPROBABILISTIC FORECASTS FOR FIELDS

9.8.1. General Considerations for Field Forecasts

Characterization of the quality of forecasts for atmospheric fields (spatial arrays of atmospheric variables) is an important problem in forecast verification. Forecasts for such fields as surface pressures, geopotential heights, temperatures, and so on, are produced routinely by weather forecasting centers worldwide. Often these forecasts are nonprobabilistic, without expressions of uncertainty as part of the forecast format. An example of this kind of forecast is shown in Figure 9.26a, which displays 24-h forecasts of sea-level pressures and 1000-500 mb thicknesses over a portion of North America, made 4 May 1993 by the U.S. National Meteorological Center. Figure 9.26b shows the same fields

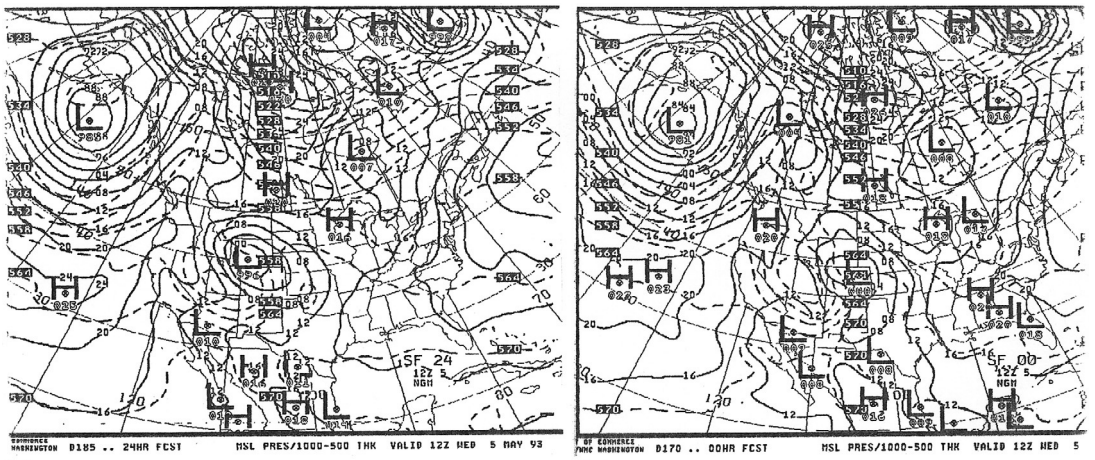

FIGURE 9.26 Forecast (a) and subsequently analyzed (b) sea-level pressures (solid) and 1000–500 mb thicknesses (dashed) over a portion of North America for 5 May 1993.

as analyzed 24 hours later. A subjective visual assessment of the two pairs of fields indicates that the main features correspond well, but that some discrepancies exist in their locations and magnitudes.

Objective, quantitative methods of verification for forecasts of atmospheric fields allow more rigorous assessments of forecast quality to be made. In practice, such methods operate on gridded fields, or collections of values of the field variable sampled at, interpolated to, or averaged over, a grid in the spatial domain. Usually this geographical grid consists of regularly spaced points either in distance, or in latitude and longitude. Figure 9.27 illustrates the gridding process for a hypothetical pair of forecast and observed fields in a small spatial domain. Each of the fields can be represented in map form as contours, or interpolated isolines, of the mapped quantity. The grid imposed on each map is a regular array of points at which the fields are represented. Here the grid consists of four rows in the north–south direction

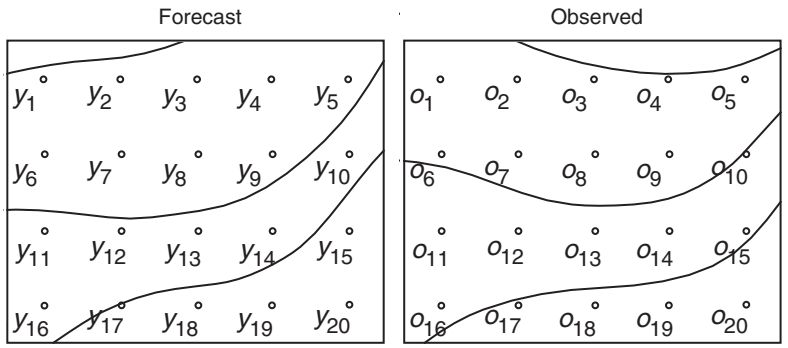

FIGURE 9.27 Hypothetical forecast (left) and observed (right) atmospheric fields represented as contour maps over a small rectangular domain. Objective assessments of the accuracy of the forecast operate on gridded versions of both the forecast and observed fields, that is, by representing them as discrete values on the same geographical grid (small circles). Here the grid has four rows in the north–south direction, and five columns in the east–west direction, so the forecast and observed fields are represented by the $M = 20$ discrete values y_m and o_m, respectively.

and five columns in the east–west direction. Thus the gridded forecast field consists of the $M = 20$ discrete values y_m, which represent the smoothly varying continuous forecast field. The gridded observed field consists of the $M = 20$ discrete values o_m, which represent the smoothly varying observed field at these same locations.

The accuracy of a field forecast usually is assessed by computing measures of the correspondence between the values y_m and o_m. If a forecast is perfect, then $y_m = o_m$ for each of the M gridpoints. Of course there are many ways that gridded forecast and observed fields can be different, even when there are only a small number of gridpoints. Put another way, the verification of field forecasts is a problem of very high dimensionality, even for small grids. Although examination of the joint distribution of forecasts and observation is in theory the preferred approach to verification of field forecasts, its large dimensionality suggests that this ideal may not be practically realizable. Rather, the correspondence between forecast and observed fields generally has been characterized using scalar summary measures. These scalar accuracy measures are necessarily incomplete but can be useful in practice.

When comparing gridded forecasts and observations, it is assumed tacitly or otherwise that the two pertain to the same spatial scale. This assumption may not be valid in all cases, for example, when the grid-scale value of a dynamical model forecast represents an area average, but the observed field is an interpolation of point observations or even the irregularly spaced point observations themselves. In such cases discrepancies deriving solely from the scale mismatch are expected (e.g., Cavanaugh and Shen, 2015; Director and Bornn, 2015), and it may be better to upscale the point values (create area averages of the smaller-scale field consistent with the larger grid scale) before comparison (e.g., Gober et al., 2008; Osborn and Hulme, 1997; Tustison et al., 2001). Even different interpolation or gridding algorithms may affect spatial verification scores to a degree (Accadia et al., 2003).

9.8.2. The S1 Score

The *S1 score* is an accuracy measure that is primarily of historical interest. It was designed to reflect the accuracy of forecasts for gradients of pressure or geopotential height, in consideration of the relationship of these gradients to the wind field at the same level (Teweles and Wobus, 1954).

Rather than operating on individual gridded values, the S1 score operates on the differences between gridded values at adjacent gridpoints. Denote the differences between the gridded values at any particular pair of adjacent gridpoints as Δy for points in the forecast field, and Δo for points in the observed field. In terms of Figure 9.27, for example, one possible value of Δy is $y_3 - y_2$, which would be compared to the corresponding gradient in the observed field, $\Delta o = o_3 - o_2$. Similarly, the difference $\Delta y = y_9 - y_4$, would be compared to the observed difference $\Delta o = o_9 - o_4$. If the forecast field reproduces the signs and magnitudes of the gradients in the observed field exactly, each Δy will equal its corresponding Δo.

The S1 score summarizes the differences between the $(\Delta y, \Delta o)$ pairs according to

$$S1 = \frac{\sum_{\text{adjacent pairs}} |\Delta y - \Delta o|}{\sum_{\text{adjacent pairs}} \max\{|\Delta y|, |\Delta o|\}} \times 100. \tag{9.87}$$

Here the numerator consists of the sum of the absolute errors in forecast gradient over all adjacent pairs of gridpoints. The denominator consists of the sum, over the same pairs of points, of the larger of the absolute value of the forecast gradient, $|\Delta y|$, or the absolute value of the observed gradient, $|\Delta o|$. The resulting ratio is multiplied by 100 as a convenience. Equation 9.87 yields the S1 score for a single pair of

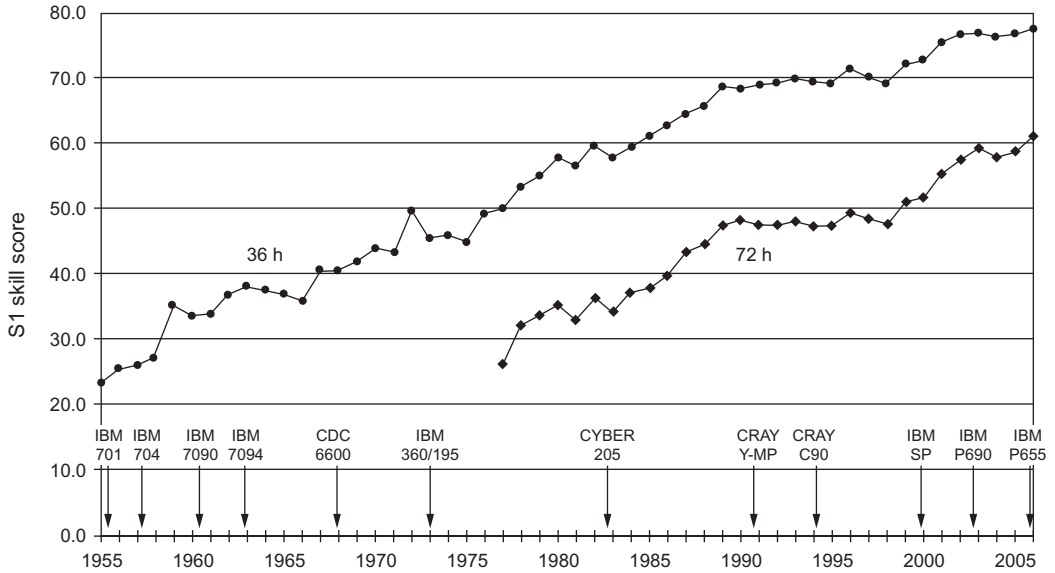

FIGURE 9.28 Average S1 score, converted to skill scores using S1 = 70 as the reference, for NCEP 36-h and 72-h hemispheric forecasts of 500 mb heights for 1955–2006. *From Harper et al. (2007). © American Meteorological Society. Used with permission.*

forecast-observed fields. When the aggregate performance of a series of field forecasts is to be assessed, the S1 scores for each forecast occasion are simply averaged. This averaging smooths sampling variations and allows trends through time of forecast performance to be assessed more easily.

Clearly, perfect forecasts will exhibit S1 = 0, with poorer gradient forecasts being characterized by increasingly larger scores. The S1 score exhibits some undesirable characteristics that have resulted in its going out of favor. The most obvious is that the actual magnitudes of the forecast pressures or heights are unimportant, since only pairwise gridpoint differences are scored. Thus the S1 score does not reflect bias. Summer scores tend to be larger (apparently worse) because of generally weaker gradients, producing a smaller denominator in Equation 9.87. The S1 score can also be improved by deliberately overforecasting the magnitudes of gradients (Thompson and Carter, 1972). Finally, the score depends on the size of the domain and the spacing of the grid, so that it is difficult to compare S1 scores not pertaining to the same domain and grid.

The S1 score has limited operational usefulness for current forecasts, but its continued tabulation has allowed forecast centers to examine very long-term trends in their field-forecast accuracy. Decades-old forecast maps may not have survived, but summaries of their accuracy in terms of the S1 score have often been retained. For example, Figure 9.28 shows S1 scores, converted to skill scores (Equation 9.4) using $S1_{ref} = 70$ (a traditional subjective rule of thumb for "useless" forecasts), for hemispheric 500 mb heights at 36- and 72-h lead times, over the period 1955-2006.

9.8.3. Mean Squared Error

The *mean squared error*, or MSE, is a much more common accuracy measure for field forecasts. The MSE operates on the gridded forecast and observed fields by spatially averaging the individual squared differences between the two at each of the M gridpoints. That is,

$$\mathrm{MSE} = \frac{1}{M}\sum_{m=1}^{M}(y_m - o_m)^2. \tag{9.88}$$

This formulation is mathematically the same as Equation 9.33, with the mechanics of both equations centered on averaging squared errors. The difference in application between the two equations is that the MSE in Equation 9.88 is computed over the gridpoints of a single pair of forecast/observation fields—that is, to $n = 1$ pair of maps—whereas Equation 9.33 pertains to the average over n pairs of scalar forecasts and observations. Clearly the MSE for a perfectly forecast field is zero, with larger MSE indicating decreasing accuracy of the forecast.

Often the MSE is expressed as its square root, the *root-mean squared error*, $\mathrm{RMSE} = \sqrt{\mathrm{MSE}}$. This transformation of the MSE has the advantage that it retains the units of the forecast variable and is thus more easily interpretable as a typical error magnitude. To illustrate, the solid line in Figure 9.29 shows RMSE in meters for 30-day forecasts of 500 mb heights initialized on 108 consecutive days during 1986–1987 (Tracton et al., 1989). There is considerable variation in forecast accuracy from day to day, with the most accurate forecast fields exhibiting RMSE near 45 m, and the least accurate forecast fields exhibiting RMSE around 90 m. Also shown in Figure 9.29 are RMSE values of 30-day forecasts of persistence, obtained by averaging observed 500 mb heights for the most recent 30 days prior to the forecast. Usually the persistence forecast exhibits slightly higher RMSE than the 30-day dynamical forecasts, but it is apparent from the figure that there are many days when the reverse is true, and that at this extended range the accuracy of these persistence forecasts was competitive with that of the dynamical forecasts.

The plot in Figure 9.29 shows accuracy of individual field forecasts, but it is also possible to express the aggregate accuracy of a collection of field forecasts by averaging the MSEs for each of a collection of paired comparisons. This average of MSE values across many forecast maps can then be converted to an average MSE as before, or expressed as a skill score in the same form as Equation 9.37. Since the MSE for perfect field forecasts is zero, the skill score following the form of Equation 9.4 is computed using

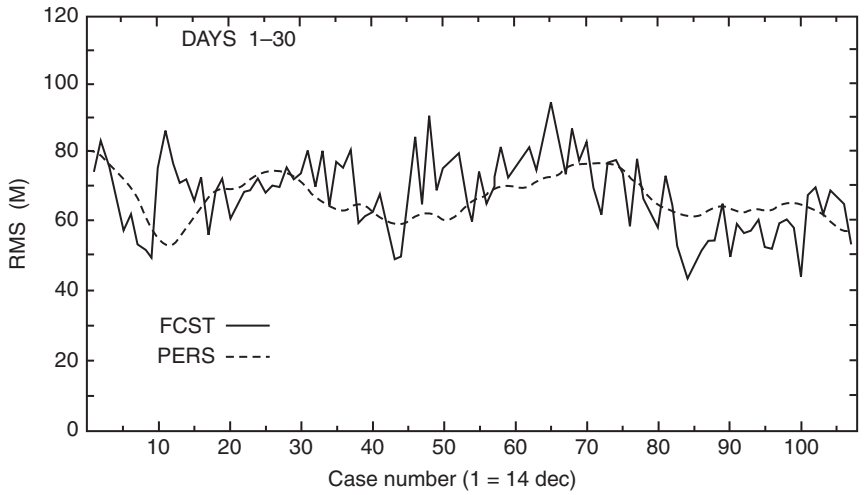

FIGURE 9.29 Root-mean squared error (RMSE) for dynamical 30-day forecasts of 500 mb heights for the northern hemisphere between 20° and 80° N (solid), and persistence of the previous 30-day average 500 mb field (dashed), for forecasts initialized 14 December 1986 through 31 March 1987. *From Tracton et al. (1989). © American Meteorological Society. Used with permission.*

$$\mathrm{SS} = \frac{\sum_{i=1}^{n} \mathrm{MSE}(i) - \sum_{i=1}^{n} \mathrm{MSE}_{\mathrm{ref}}(i)}{0 - \sum_{i=1}^{n} \mathrm{MSE}_{\mathrm{ref}}(i)} = 1 - \frac{\sum_{i=1}^{n} \mathrm{MSE}(i)}{\sum_{i=1}^{n} \mathrm{MSE}_{\mathrm{ref}}(i)}, \tag{9.89}$$

where the aggregate skill of n individual field forecasts is being summarized. When this skill score is computed, the reference field forecast is usually either the climatological average field (in which case it may be called the reduction of variance, in common with Equation 9.37) or individual persistence forecasts as shown in Figure 9.29.

The MSE skill score in Equation 9.89, when calculated with respect to climatological forecasts as the reference, allows an interesting interpretation for field forecasts when algebraically decomposed in the same way as in Equation 9.38. When applied to field forecasts, this decomposition is conventionally expressed in terms of the differences (anomalies) of forecasts and observations from the corresponding climatological values at each gridpoint (Murphy and Epstein, 1989),

$$y'_m = y_m - c_m \tag{9.90a}$$

and

$$o_m = o_m - c_m, \tag{9.90b}$$

where c_m is the climatological value at gridpoint m. The resulting MSE and skill scores are identical, because the climatological values c_m can be both added to and subtracted from the squared terms in Equation 9.88 without changing the result, that is,

$$\mathrm{MSE} = \frac{1}{M}\sum_{m=1}^{M}(y_m - o_m)^2 = \frac{1}{M}\sum_{m=1}^{M}([y_m - c_m] - [o_m - c_m])^2 = \frac{1}{M}\sum_{m=1}^{M}\left(y'_m - o'_m\right)^2. \tag{9.91}$$

When expressed in this way, the algebraic decomposition of MSE skill score in Equation 9.38 becomes

$$\mathrm{SS}_{\mathrm{clim}} = \frac{r^2_{y'o'} - \left[r_{y'o'} - \left(s_{y'}/s_{o'}\right)\right]^2 - \left[\left(\overline{y'} - \overline{o'}\right)/s_{o'}\right]^2 + \left(\overline{o'}/s_{o'}\right)^2}{1 + \left(\overline{o'}/s_{o'}\right)^2} \tag{9.92a}$$

$$\approx r^2_{y'o'} - \left[r_{y'o'} - \left(s_{y'}/s_{o'}\right)\right]^2 - \left[\left(\overline{y'} - \overline{o'}\right)/s_{o'}\right]^2. \tag{9.92b}$$

The difference between this decomposition and Equation 9.38 is the normalization factor involving the average differences between the observed and climatological gridpoint values, in both the numerator and denominator of Equation 9.92a. This factor depends only on the observed field, and Murphy and Epstein (1989) note that it is likely to be small if the skill is being evaluated over a sufficiently large spatial domain, because positive and negative differences with the gridpoint climatological values will tend to balance. Neglecting this term leads to the approximate algebraic decomposition of the skill score in Equation 9.92b, which is identical to Equation 9.38 except that it involves the differences from the gridpoint climatological values, y' and o'. In particular, it shows that the (squared) correlation between the forecast and observed fields is at best a measure of potential rather than actual skill, which can be realized only in the absence of conditional (second term in Equation 9.92b) and unconditional (third term in Equation 9.92b) biases. It is worthwhile to work with these climatological anomalies when

investigating skill of field forecasts in this way, in order to avoid ascribing spurious skill to forecasts for merely forecasting a correct climatology.

The Taylor Diagram

The joint contributions to the RMSE of the correlation between two fields, and the discrepancy between their standard deviations, can be visualized using a graphical device that is known as the *Taylor diagram* (Taylor, 2001), although a very similar diagram was proposed independently by Lambert and Boer (2001). The Taylor diagram is based on a geometrical representation of an algebraic decomposition of the debiased MSE, which is the MSE after subtraction of contributions due to overall bias errors:

$$MSE' = MSE - (\overline{y} - \overline{o})^2 \tag{9.93a}$$

$$= \frac{1}{M}\sum_{m=1}^{M}[(y_m - \overline{y}) - (o_m - \overline{o})]^2 = \sigma_y^2 + \sigma_o^2 - 2\,\sigma_y \sigma_o r_{yo}. \tag{9.93b}$$

Equation 9.93a defines the debiased MSE to be equal to MSE after subtraction of the squared bias. Clearly MSE = MSE' if the means over the M gridpoints, $\overline{y}$ and $\overline{o}$, are equal, and otherwise MSE' reflects only those contributions to MSE not deriving from the unconditional bias. The first equality in Equation 9.93b indicates that MSE' is equivalent to MSE calculated after subtracting the map means (i.e., averages over the M gridpoints) $\overline{y}$ and $\overline{o}$, so that both transformed fields have equal, zero, area averages. The second equality in Equation 9.93b suggests the geometrical representation of the relationship between the standard deviations over the M gridpoints, σ_y and σ_o, their correlation $r_{y,o}$, and RMSE', through the direct analogy to the *law of cosines*,

$$c^2 = a^2 + b^2 - 2ab\cos\theta. \tag{9.94}$$

The correspondence is between the three terms on the right-hand sides of Equation 9.93b and 9.94.

The Taylor diagram represents the two standard deviations σ_y and σ_o as the lengths of two legs of a triangle, which are separated by an angle θ whose cosine is equal to the correlation $r_{y,o}$ between the two fields. The length of the third side is then RMSE'. The diagram itself plots the vertices of these triangles in polar coordinates, where the angle from the horizontal is the cosine of the correlation, and the radial distances from the origin are defined by the standard deviations. The correlation of the observed field with itself is 1, so the corresponding vertex is at an angle $\cos^{-1}(1) = 0°$ from the horizontal, with radius σ_o. The vertices for each forecast field are represented as points, plotted at a radius σ_y and angle $\cos^{-1}(r_{y,o})$.

Taylor diagrams are most useful when multiple "y" fields are being compared simultaneously to a corresponding reference field "o." Figure 9.30 shows a superposition of three such Taylor diagrams, showing performance of 16 dynamical climate models in representing global fields of precipitation, surface temperatures, and sea-level pressures. Since the "o" fields are different in the three comparisons, all standard deviations have been divided by the appropriate σ_o, so that the "b" vertex (c.f. Equation 9.94) of each triangle is located at unit radius on the horizontal axis, at the point labeled "observed." The tight clustering near the "observed" of the points representing the "a" vertices for the simulated temperature fields indicate that these patterns have been best simulated of the three variables, with standard deviations nearly correctly simulated (all points are near unit radius from the origin), and correlations with the observed temperature field that all exceed 0.95. The distances from each of these points to the reference "observed" vertex are geometrically equal to RMSE'. In contrast, precipitation is the least well simulated variable of the three, with simulated standard deviations ranging from approximately 75% to 125% of the

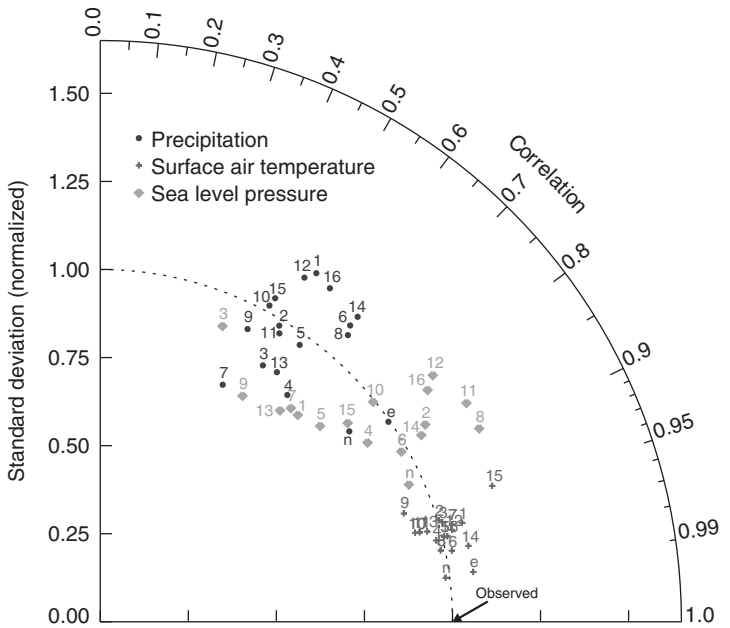

FIGURE 9.30 Taylor diagram comparing 16 climate-model generated fields of precipitation, surface air temperatures, and sea-level pressures to instrumental ("observed") values. Points labeled "n" and "e" are corresponding reanalysis values, which are model results that have been strongly constrained by actual data. All standard deviations have been divided by the appropriate σ_o in order to superimpose the Taylor diagrams for the three variables. Transformation of correlation to angle from the horizontal is indicated by the cosine scale along the curved edge. *From McAvaney et al. (2001).*

correct value, and correlations ranging from about 0.5 to 0.7, so that RMSE' (distances to "observed") is substantially larger in all cases than for the temperature simulations.

In order to emphasize that distances to the reference point indicates RMSE', Taylor diagrams are sometimes drawn with semicircles of equal RMSE', centered at radius σ_o on the horizontal axis. Negative correlations can be accommodated by extending the diagram counterclockwise to include an additional quadrant. While Taylor diagrams are most frequently used to illustrate the phase association (correlation) and amplitude (standard deviation) errors of spatial fields over M gridpoints, the mathematical decomposition and its geometrical representation are equally applicable to MSEs of nonprobabilistic forecasts for scalar predictands, over n forecasting occasions (Equation 9.33), again after removal of any contributions to the MSE from squared bias.

Recently, Koh et al. (2012) have proposed a suite if diagrams devised in a similar spirit, that portray biases, in addition to phase and amplitude errors.

9.8.4. Anomaly Correlation

The *anomaly correlation* (AC) is another commonly used measure of association that operates on pairs of gridpoint values in the forecast and observed fields. To compute the anomaly correlation, the forecast and observed values are first converted to anomalies in the sense of Equation 9.90: the climatological

average value of the observed field at each of M gridpoints is subtracted from both forecasts y_m and observations o_m.

There are actually two forms of anomaly correlation in use, and it is unfortunately not always clear which has been employed in a particular instance. The first form, called the *centered* anomaly correlation, was apparently first suggested by Glenn Brier in an unpublished 1942 U.S. Weather Bureau mimeo (Namias, 1952). It is computed according to the usual Pearson correlation (Equation 3.28), operating on the M gridpoint pairs of forecasts and observations that have been referred to the climatological averages c_m at each gridpoint,

$$\mathrm{AC}_C = \frac{\sum_{m=1}^{M} \left(y'_m - \overline{y'}\right)\left(o'_m - \overline{o'}\right)}{\left[\sum_{m=1}^{M} \left(y'_m - \overline{y'}\right)^2 \sum_{m=1}^{M} \left(o'_m - \overline{o'}\right)^2\right]^{1/2}}. \tag{9.95}$$

Here the primed quantities are the anomalies relative to the climatological averages (Equation 9.90), and the overbars refer to these anomalies averaged over a given map of M gridpoints. The square of Equation 9.95 is thus exactly $r^2_{y'o'}$ in Equation 9.92.

The other form for the anomaly correlation differs from Equation 9.95 in that the map-mean anomalies are not subtracted, yielding the *uncentered* anomaly correlation

$$\mathrm{AC}_\mathrm{U} = \frac{\sum_{m=1}^{M} (y_m - c_m)(o_m - c_m)}{\left[\sum_{m=1}^{M} (y_m - c_m)^2 \sum_{m=1}^{M} (o_m - c_m)^2\right]^{1/2}} = \frac{\sum_{m=1}^{M} y'_m o'_m}{\left[\sum_{m=1}^{M} \left(y'_m\right)^2 \sum_{m=1}^{M} \left(o'_m\right)^2\right]^{1/2}}. \tag{9.96}$$

This form was apparently first suggested by Miyakoda et al. (1972). Superficially, the AC_U in Equation 9.96 resembles the Pearson product-moment correlation coefficient (Equations 3.28 and 9.95), in that that both are bounded by ± 1, and that neither is sensitive to biases or scale errors in the forecasts. However, the centered and uncentered anomaly correlations are equivalent only if the averages over the M gridpoints of the two anomalies are zero, that is, only if $\Sigma_m(y_m - c_m) = 0$ and $\Sigma_m(o_m - c_m) = 0$. These conditions may be approximately true if the forecast and observed fields are being compared over a large (e.g., hemispheric) domain, but will almost certainly not hold if the fields are compared over a relatively small area. In this latter situation DelSole and Shukla (2006) recommend use of the uncentered anomaly correlation as conforming better to plausible subjective evaluation of forecast accuracy.

The anomaly correlation is designed to detect similarities in the patterns of departures (i.e., anomalies) from the climatological mean field and is sometimes therefore referred to as a *pattern correlation*. However, as Equation 9.92 makes clear, the anomaly correlation does not penalize either conditional or unconditional biases. Accordingly, it is reasonable to regard the anomaly correlation as reflecting potential skill (that might be achieved in the absence of conditional and unconditional biases), but it is incorrect to regard the anomaly correlation (or, indeed, any correlation) as measuring actual skill (e.g., Murphy, 1995).

The anomaly correlation often is used to evaluate extended-range (beyond a few days) forecasts. Figure 9.31 shows anomaly correlation values for the same 30-day dynamical and persistence forecasts of 500 mb height that are verified in terms of the RMSE in Figure 9.29. Since the anomaly correlation has

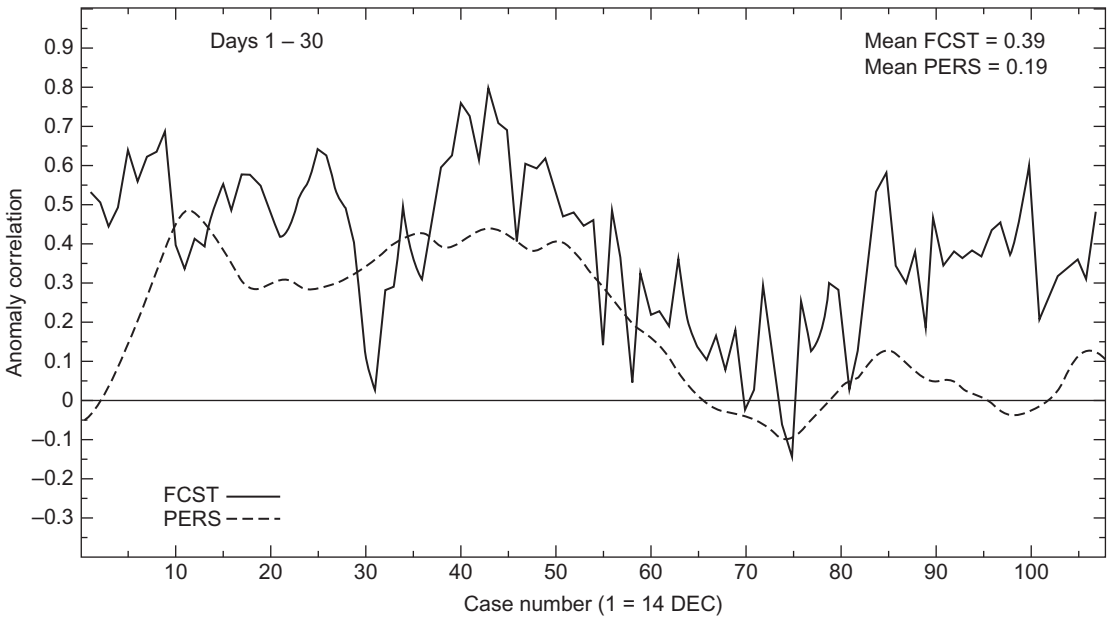

FIGURE 9.31 Anomaly correlations for dynamical 30-day forecasts of 500 mb heights for the Northern Hemisphere between 20° and 80° N (solid), and persistence of the previous 30-day average 500 mb field (dashed), for forecasts initialized 14 December 1986 through 31 March 1987. The performance of the same forecasts is characterized in Figure 9.29 using the RMSE. *From Tracton et al. (1989). © American Meteorological Society. Used with permission.*

a positive orientation (larger values indicate more accurate forecasts) and the RMSE has a negative orientation (smaller values indicate more accurate forecasts), we must mentally "flip" one of these two plots vertically in order to compare them. When this is done, it can be seen that the two measures usually rate a given forecast map similarly, although some differences are apparent. For example, in this data set the anomaly correlation values in Figure 9.31 show a more consistent separation between the performance of the dynamical and persistence forecasts than do the RMSE values in Figure 9.29, suggesting that the patterns of the fields may have been forecast better than the magnitudes.

As is also the case for the MSE, aggregate performance of a collection of field forecasts can be summarized by averaging anomaly correlations across many forecasts. However, skill scores of the form of Equation 9.4 usually are not calculated for the anomaly correlation. For the uncentered anomaly correlation, AC_U is undefined for climatological forecasts, because the denominator of Equation 9.96 is zero. Rather, AC skill generally is evaluated relative to the reference values $AC_{ref} = 0.6$ or $AC_{ref} = 0.5$. Individuals working operationally with the anomaly correlation have found, subjectively, that $AC_{ref} = 0.6$ seems to represent a reasonable lower limit for delimiting field forecasts that are synoptically useful (Hollingsworth et al., 1980). Murphy and Epstein (1989) have shown that if the average forecast and observed anomalies are zero, and if the forecast field exhibits a realistic level of variability (i.e., the two summations in the denominator of Equation 9.95 are of comparable magnitude), then $AC_C = 0.5$ corresponds to the skill score for the MSE in Equation 9.89 being zero. Under these same restrictions, $AC_C = 0.6$ corresponds to the MSE skill score being 0.20.

Figure 9.32 illustrates the use of the subjective $AC_{ref} = 0.6$ reference level. Figure 9.32a shows average AC values for 500 mb height forecasts made during the winters (December-February) of

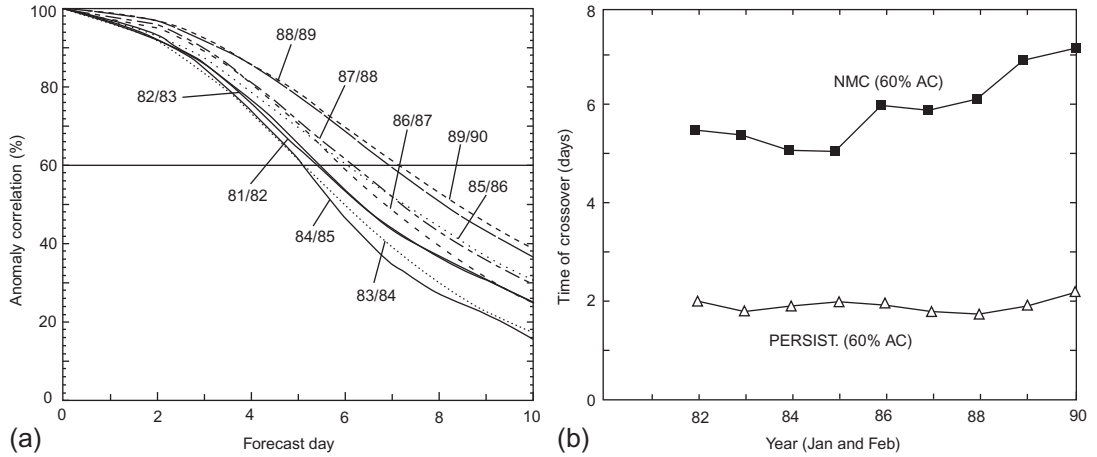

FIGURE 9.32 (a) Average anomaly correlations as a function of forecast lead time for 1981/1982 through 1989/1990 winter 500 mb heights between 20°N and 80°N. Accuracy decreases as forecast lead time increases, but there are substantial differences among winters. (b) Average lead time at which anomaly correlations for the dynamical and persistence forecasts cross the $AC_{ref} = 0.6$ level, for Januarys and Februarys of these nine winters. *From Kalnay et al. (1990). © American Meteorological Society. Used with permission.*

1981/1982 through 1989/1990. For lead time zero days into the future (i.e., initial time), AC = 1 since $y_m = o_m$ at all grid points. The average AC declines progressively for longer lead times, falling below $AC_{ref} = 0.6$ between days five and seven. The curves for the later years tend to lie above the curves for the earlier years, reflecting, at least in part, improvements made to the dynamical forecast model during the decade. One measure of this overall improvement is the increase in the average lead time at which an AC curve crosses the 0.6 line. These times are plotted in Figure 9.32b, and range from five days in the early- and mid-1980s, to seven days in the late 1980s. Also plotted in this panel are the average lead times at which anomaly correlations for persistence forecasts fall below 0.6. The crossover time at the $AC_{ref} = 0.6$ threshold for persistence forecasts is consistently about two days. Thus imagining the average correspondence between observed 500 mb maps separated by 48 hour intervals allows a qualitative appreciation of the level of forecast performance represented by the $AC_{ref} = 0.6$ threshold.

9.8.5. Field Verification Based on Spatial Structure

Because the numbers of gridpoints *M* typically used to represent meteorological fields is relatively large, and the numbers of allowable values for forecasts and observations of continuous predictands defined on these grids may also be large, the dimensionality of verification problems for field forecasts is typically huge. Using scalar scores such as GSS (Equation 9.20) or MSE (Equation 9.88) to summarize forecast performance in these settings may seem at times to be a welcome relief from the inherent complexity of the verification problem, but necessarily masks very much relevant detail. For example, forecast and observed precipitation fields are sometimes converted to binary (yes/no) fields according to whether the gridpoint values are above or below a threshold such as 0.25 in., with the resulting field forecast scored according to a 2 x 2 contingency table measure such as the GSS. But, as demonstrated by Ahijevych et al. (2009), all such forecasts exhibiting no spatial overlap with the observed "yes" area, but which have the same number of

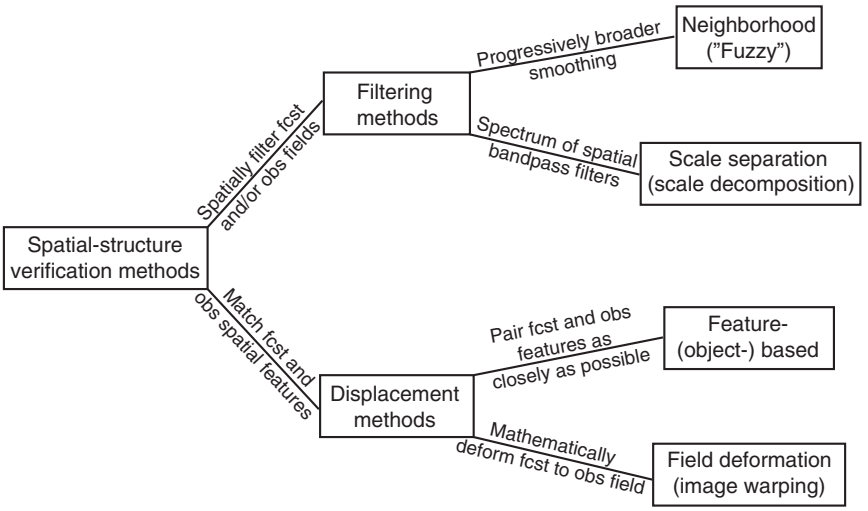

FIGURE 9.33 The taxonomy of spatial-structure verification methods proposed by Gilleland et al. (2009).

"yes" gridpoints, yield equal scores regardless of the distance between the forecast and observed features, or the similarity or dissimilarity of the shapes of the forecast and observed features. Similarly, a modest error in the advection speed of a relatively small-scale forecast feature may produce a large MSE as a consequence of the *"double penalty" problem*: the forecast will be penalized once for placing the feature where it did not occur, and penalized again for placing no feature where the actual event happened, even though the feature itself may have been well forecast with respect to its presence, shape, trajectory, and intensity.

Accordingly, forecasters and forecast evaluators often are dissatisfied with the correspondence between the traditional single-number performance summaries and their subjective perceptions about the goodness of a spatial field forecast. This dissatisfaction has stimulated work on field-verification methods that may be able to quantify aspects of forecast performance which better reflect human visual reactions to map features. Casati et al. (2008) and Gilleland et al. (2009) reviewed initial developments in this area, and Gilleland (2013) provides an extensive list of papers using these methods.

Gilleland et al. (2009) have proposed a taxonomy for these spatial verification methods. This taxonomy is shown in Figure 9.33, and Figure 9.34 illustrates the ideas behind the four types. Most of the spatial-structure verification methods that have been proposed to date can be classified either as filtering methods or displacement methods. For filtering methods, the forecast and/or observed fields (or, the difference field) are subjected to spatial filters before application of more conventional verification metrics at multiple spatial scales. In contrast, displacement methods operate on discrepancies between individual features in the forecast and observed fields, generally in terms of nontraditional metrics that describe the nature and degree of spatial manipulation necessary to achieve congruence between the manipulated forecast field and the corresponding observed field. In this context a "feature" is usually understood to be a collection of contiguous nonzero gridpoints or pixels in either the forecast or observed fields.

Software implementing many of the methods described in this section is described by Fowler et al. (2018).

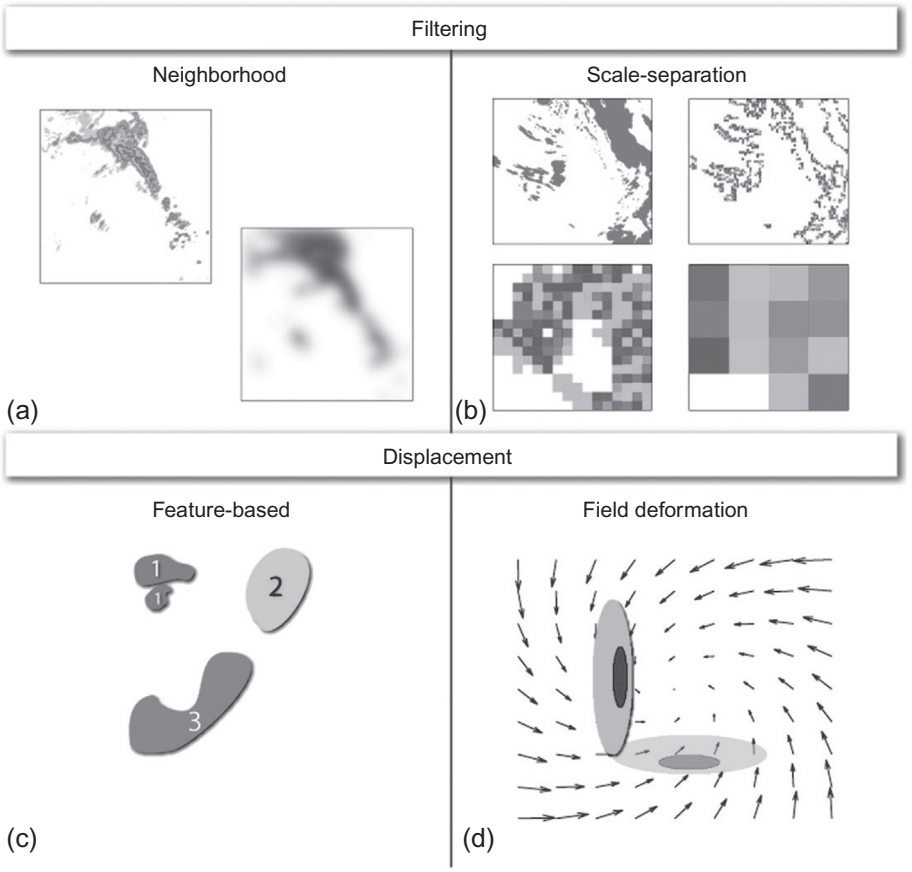

FIGURE 9.34 Graphical illustration of the four types of spatial verification methods named in Figure 9.33. *From Gilleland et al. (2009). © American Meteorological Society. Used with permission.*

Neighborhood Methods

Neighborhood (or "fuzzy") methods address the problem of excessive penalties for small displacement errors by expanding the comparisons from individual gridpoint pairs to larger spatial and/or temporal neighborhoods (Ebert, 2008). In effect these methods smooth the forecast and observation fields before computing a verification measure of interest, as suggested by Figure 9.34a. The underlying idea is particularly appealing for very high-resolution forecasts, for which exact matches between forecast and observed fields cannot be reasonably expected. In such cases forecasts that are in some sense close can be valuable, for example, in the construction of severe-weather warnings.

The *Fractions Skill Score* (Roberts and Lean, 2008) is perhaps the most widely used of the neighborhood methods. For it, and other neighborhood methods, a collection of gridpoints around each point in the domain to be verified is defined as its neighborhood. Whether these neighborhoods are defined in terms of radial distance from a base point, or as square domains centered on a base point, appears to have little effect on the results. Skok and Roberts (2016) investigated the effects of different treatments for neighborhoods that are near the edge of the domain. A binary outcome is defined for each of the M points

in the neighborhood, separately for the forecast and observed fields, which might, for example, be occurrence or not of precipitation larger than some threshold. Defining $P_{y,m}$ and $P_{o,m}$ as the fractions of above-threshold points in the forecast and observed neighborhoods around the point m, respectively, the Fractions Skill Score is computed as

$$\mathrm{FSS} = 1 - \frac{\sum_{m=1}^{M} \left(P_{y,m} - P_{o,m}\right)^2}{\sum_{m=1}^{M} P_{y,m}^2 + \sum_{m=1}^{M} P_{o.m}^2}. \tag{9.97}$$

Equation 9.97 is based on the conventional skill score (Equation 9.37) for the MSE (Roberts and Lean, 2008), so to the extent that $P_{y,m}$ and $P_{o,m}$ can be regarded as probabilities for within-neighborhood gridpoints exhibiting above-threshold values, the FSS is reminiscent of the Brier skill score (Equation 9.40), although the $P_{o,m}$ are not necessarily binary.

Because the results will be sensitive to the degree of smoothing (the sizes of the neighborhoods), the FSS and other neighborhood methods may be applied at a range of increasing scales, which may provide information on possible scale dependence of forecast performance. The score will generally be better for the larger scales (i.e., for increasing neighborhood size), as poorly- or unpredictable scales are progressively filtered. Figure 9.35 illustrates this point in a synthetic setting in which a 1-pixel-wide rain band is displaced by 1, 3, 11, and 21 pixels relative to the truth, and the neighborhood is expanded from 1 to 49 grid squares centered on each pixel. The FSS increases for increasing neighborhood size, and as expected is smaller for greater displacement between the forecast and observed features The dashed horizontal line labeled "uniform" locates the value FSS = 0.5 that is often interpreted as the threshold for a useful

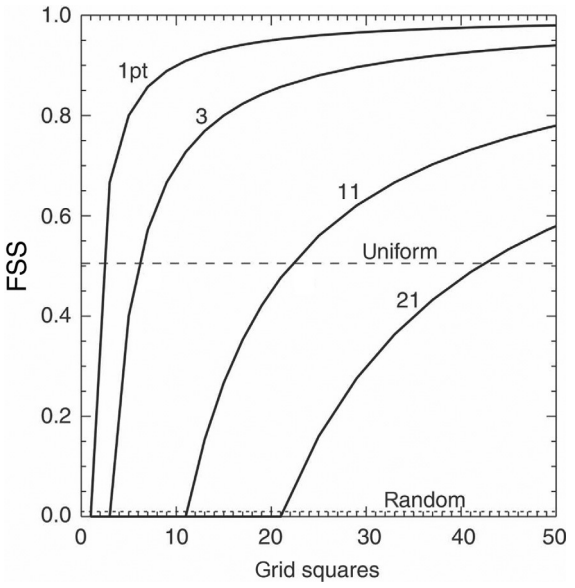

FIGURE 9.35 Illustration of the dependence of the FSS on the neighborhood size, in grid squares, for an idealized case in which a 1-pixel wide band of rain occupying 1% of the domain is displaced by 1, 3, 11, and 21 pixels. *Modified from Roberts and Lean (2008). © American Meteorological Society. Used with permission.*

forecast (e.g., Skok and Roberts, 2016), and accordingly the size of the spatial scale that might be considered as plausibly predictable with respect to a particular curve for FSS as a function of domain size, provided the domain size is large compared to the typical displacement error. Similarly, when the forecast and observation are converted to binary fields at a range of thresholds before application of neighborhood methods, information on variation of skill with event intensity (e.g., rainfall rate) can be extracted. Skok and Roberts (2018) study how a transformation of FSS can be used to diagnose overall average displacement errors, emphasizing errors for the largest objects in the field.

Scale Separation Methods

In contrast to the progressive blurring of the fields by neighborhood methods, scale separation methods apply mathematical spatial filters to the forecast and observed fields, allowing isolation and separation of the verification for features of different sizes. Unlike the results of the progressive smoothing produced by the neighborhood methods, the filtered fields generated by scale separation methods at particular spatial scales may not closely resemble the original fields, as exemplified by the maps at four spatial scales shown in Figure 9.34b.

Briggs and Levine (1997) proposed this general approach using *wavelets*, which are a particular kind of mathematical basis function (Section 10.6), applying it to forecasts of geopotential height fields. Weniger et al. (2017) review the subsequent use of wavelets to decompose forecast and observed fields into hierarchies of spatial scales for the purpose of forecast verification. Figure 9.36 illustrates the procedure, comparing a pair of example analyzed (upper left panel) and forecast (upper right panel) 500 mb fields, together with corresponding wavelet decompositions of those fields at five spatial scales. After the scale separation, conventional measures of agreement such as MSE are then applied to the pairs of forecast and observation fields at each decomposed scale. Because the wavelet basis functions are orthogonal, these scale-specific MSE values sum to the overall MSE for the unfiltered comparison, and the fractional contribution to error made by each of the scales can be easily computed

Casati (2010) extends this method to settings such as precipitation field forecasts, which include also an intensity dimension. Denis et al. (2002) and de Elia et al. (2002) consider a similar approach based on more conventional spectral basis functions. These methods allow investigation of scale dependence of forecast errors, possibly including a minimum spatial scale below which the forecasts do not exhibit useful skill.

Feature-Based Methods

The displacement methods operate by comparing the structure of specific features (contiguous gridpoints or pixels sharing a relevant property) in the forecast and observed fields, with respect to such attributes as position, shape, size, and intensity. Thus as suggested by Figure 9.34c, the methods operate by identifying discrete portions of interest of the spatial domain, which often correspond to forecast or observed precipitation amounts larger than some threshold. An early example of this class of methods was described by Ebert and McBride (2000). Differences among displacement methods relate to such issues as how features occurring in one but not the other of the forecast and observed fields are dealt with, whether two or more nearby but distinct features in one of the fields should be merged before comparison with the other field, and what summary diagnostics are used to characterize differences.

In the Method for Object-based Diagnostic Evaluation (MODE; Davis et al., 2006a, 2009), discrete objects of interest in the forecast and observation fields are identified by a combination of smoothing and thresholding operations. Figure 9.37 illustrates an example of the transformation by MODE of a

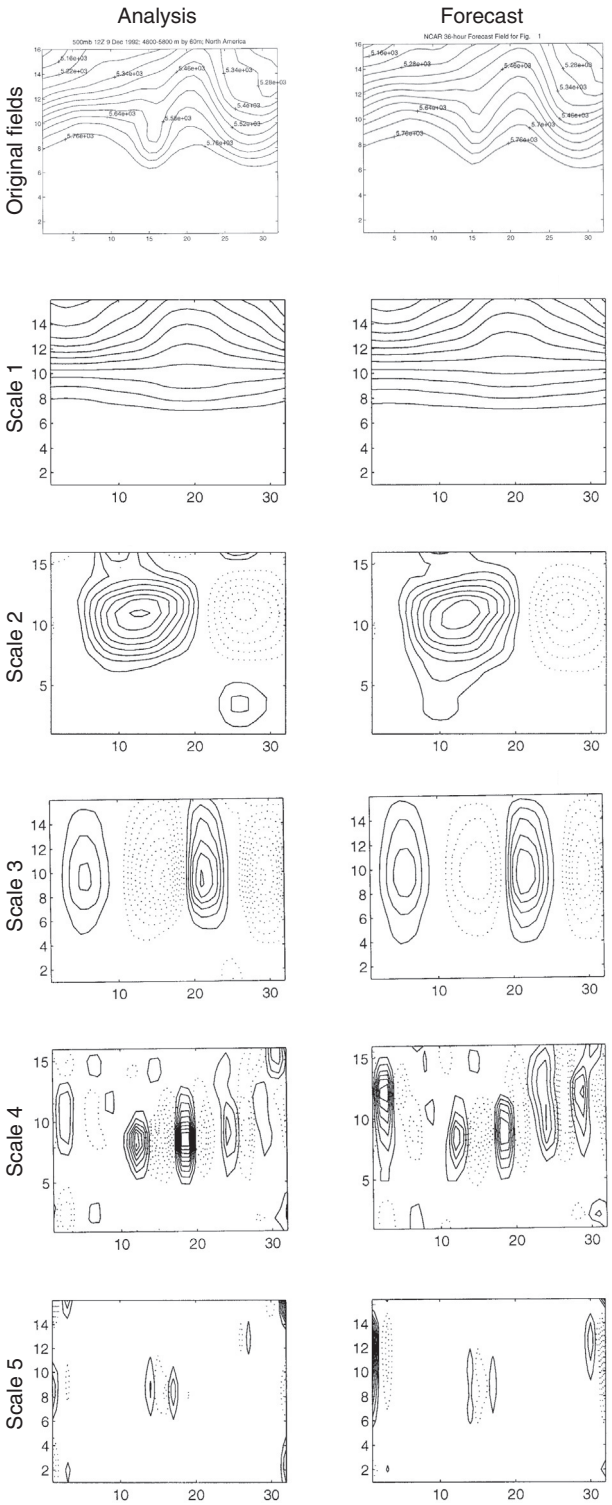

FIGURE 9.36 Example analyzed (upper left) and forecast (upper right) 500 mb height fields, and their wavelet decompositions at a sequence of five progressively finer spatial scales. *From Briggs and Levine (1997).* © *American Meteorological Society. Used with permission.*

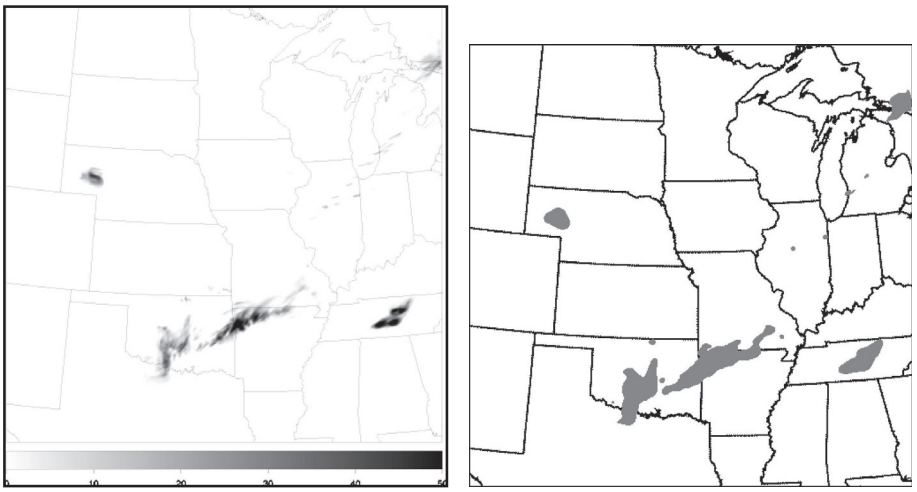

FIGURE 9.37 (a) Example field of 1-h rainfall accumulation (mm), and (b) its transformation to a collection of objects after MODE smoothing and thresholding operations. *From Davis et al. (2006b). © American Meteorological Society. Used with permission.*

radar-derived precipitation field to a collection of precipitation objects. These processed features are matched between the forecast and observed fields through a user-adjustable multiattribute agreement function that includes the separation of their centroids, minimum separation of their boundaries, orientation differences, relative sizes, and the areas of their intersections; and the degree of agreement is quantified, again with a user-adjustable multiattribute function. Mittermaier and Bullock (2013) describe extension of the 2-dimensional spatial MODE algorithm with a third dimension that represents the time evolution of the forecast and observed fields. Wolff et al. (2014) compare MODE and FSS (-Equation 9.97) for a sample of dynamical precipitation forecast fields.

The Structure-Amplitude-Location (SAL; Wernli et al., 2008, 2009) method characterizes agreement between forecast and observed fields with respect to the three characteristics that make up its name. It was designed with the verification of precipitation fields in mind, although Weniger and Friederichs (2016) apply the method to cloud fields. The SAL method begins by identifying objects that consist of contiguous gridpoints or pixels having precipitation amounts at least 1/15 of the 95th percentile of amounts larger than 0.1 mm, as evaluated over the domain. Each of these N objects in either the forecast or observed fields has center of mass s_n, and total (summed over points) precipitation amount R_n, $n = 1, \ldots, N$.

The structure error is defined as

$$S = \frac{V(\boldsymbol{y}) - V(\boldsymbol{o})}{\frac{1}{2}[V(\boldsymbol{y}) + V(\boldsymbol{o})]}, \tag{9.98}$$

where

$$V(\bullet) = \frac{\sum_{n=1}^{N} R_n^2 / R_n^{\max}}{\sum_{n=1}^{N} R_n} \tag{9.99}$$

is the mass-weighted scaled volume, and R_n^{max} is the maximum precipitation in object n. The structure error S ranges from –2 to +2, with zero reflecting perfect structure, positive values indicating forecast fields that are too smooth, and negative values indicating forecast fields that are too noisy.

The amplitude error is

$$A = \frac{D(\boldsymbol{y}) - D(\boldsymbol{o})}{\frac{1}{2}[D(\boldsymbol{y}) + D(\boldsymbol{o})]}, \tag{9.100}$$

where $D(\bullet)$ is the domain-mean value. The amplitude error ranges from –2 to +2, with zero indicating unbiasedness.

The location error is

$$L = \frac{|s(\boldsymbol{y}) - s(\boldsymbol{o})|}{d} + 2\frac{|r(\boldsymbol{y}) - r(\boldsymbol{o})|}{d}, \tag{9.101}$$

where d is the maximum distance across the domain, and

$$r(\bullet) = \frac{\sum_{n=1}^{N} R_n |\overline{s} - s_n|}{\sum_{n=1}^{N} R_n} \tag{9.102}$$

is the mass-weighted mean distance of the center of mass of each object to the overall center of mass $\overline{s}$. The location error ranges from 0 to 2, with $L = 0$ indicating perfect locations for the objects.

The three dimensions of the SAL method can be displayed compactly on a diagram in which the horizontal axis is the structure error, the vertical axis is the amplitude error, and the shading of the plotted points indicates the location error. Figure 9.38a shows an example diagram for dynamical summer precipitation forecasts, and Figure 9.38b shows corresponding results for persistence forecasts. Ideally, one would like to see points with minimal shading that are concentrated near the origin.

Field Deformation Methods

Field deformation techniques mathematically manipulate the entire forecast field (rather than only individual objects within the field) in order to match the observed field to the extent possible. This approach was first proposed by Hoffman et al. (1995), who characterized forecast errors using a decomposition into displacement, amplitude, and residual components. Location error was determined by horizontal translation of the forecast field until the best match was obtained, where "best" may be interpreted through such criteria as minimum MSE, maximal area overlap, or alignment of the forecast and observed centroids. Alternatively, the required warping of the forecast field can be characterized with a field of deformation vectors, as illustrated in Figure 9.34d, which shows idealized forecast (lighter grays) and observed (darker grays) precipitation areas, with the heavier shading indicating a region of higher precipitation. The field deformation vectors warp the forecast into congruence with the observed area. This vector field indicates that the forecast error is with respect to rotation, and not displacement or aspect ratio.

Another approach to field deformation is to iteratively displace fields though a hierarchy of progressively increasing resolution by minimizing an error statistic such as RMSE, in a method known as "optical flow" (Han and Szunyogh, 2016; Keil and Craig, 2007). Keil and Craig (2009) propose

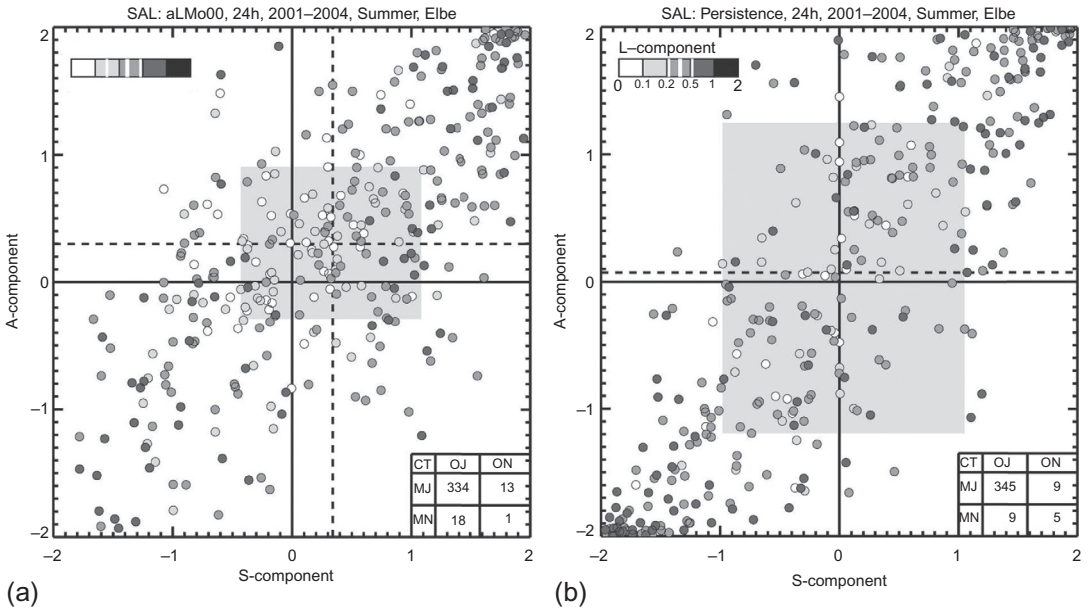

FIGURE 9.38 Example SAL diagrams for daily summer precipitation over a portion of Germany, for (a) dynamical forecasts, and (b) persistence forecasts. Median values for *S* and *A* are shown with dashed lines, and the shaded regions are defined by the quartiles of the respective distributions. *From Wernli et al. (2008). © American Meteorological Society. Used with permission.*

quantifying the agreement between the two original fields in terms of the magnitude of overall displacement required for the morphing and the amplitude errors of the morphed field.

9.9. VERIFICATION BASED ON ECONOMIC VALUE

9.9.1. Optimal Decision Making and the Cost/Loss Ratio Problem

The practical justification for effort expended in developing forecast systems and making forecasts is that these forecasts should result in better decision making in the face of uncertainty. Often such decisions have direct economic consequences, or their consequences can be mapped onto an economic (i.e., monetary) scale. There is a substantial literature in the fields of economics and statistics on the use and value of information for decision making under uncertainty (e.g., Clemen, 1996; Johnson and Holt, 1997), and the concepts and methods in this body of knowledge have been extended to the context of optimal use and economic value of weather forecasts (e.g., Katz and Murphy, 1997a; Winkler and Murphy, 1985). Forecast verification is an essential component of this extension, because it is the joint distribution of forecasts and observations (Equation 9.1) that will determine the economic value of forecasts (on average) for a particular decision problem. It is therefore natural to consider characterizing forecast goodness (i.e., computing forecast verification measures) in terms of the mathematical transformations of the joint distribution that define forecast value for particular decision problems. This notion may have first been proposed by Epstein (1962).

The reason that economic value of weather forecasts must be calculated for particular decision problems—that is, on a case-by-case basis—is that the economic value of a particular set of forecasts

will be different for different decision problems (e.g., Roebber and Bosart, 1996; Wilks, 1997a). However, a useful and convenient prototype, or "toy," decision model is available, called the *cost/loss ratio* problem (e.g., Katz and Murphy, 1997b; Murphy, 1977; Thompson, 1962). This simplified decision model apparently originated with Anders Angstrom, in a 1922 paper (Liljas and Murphy, 1994), and has been frequently used since that time. Despite its simplicity, the cost/loss problem nevertheless can reasonably approximate some simple real-world decision problems (Roebber and Bosart, 1996).

The cost/loss decision problem relates to a hypothetical decision maker for whom some kind of adverse weather may or may not occur, and who has the option of either protecting or not protecting against the economic effects of the adverse weather. That is, this decision maker must choose one of two alternatives in the face of an uncertain dichotomous weather outcome. Because there are only two possible actions and two possible outcomes, this is the simplest possible decision problem: no decision would be needed if there was only one course of action, and no uncertainty would be involved if only one weather outcome was possible. The protective action available to the decision maker is assumed to be completely effective, but requires payment of a cost C, regardless of whether or not the adverse weather subsequently occurs. If the adverse weather occurs in the absence of the protective action being taken, the decision maker suffers a loss L. The economic effect is zero loss if protection is not taken and the event does not occur. Figure 9.39a shows the loss function for the four possible combinations of decisions and outcomes in this problem.

Probability forecasts for the dichotomous weather event are assumed to be available and, depending on their quality, better decisions (in the sense of improved economic outcomes, on average) may be possible. Taking these forecasts at face value (i.e., assuming that they are calibrated, so $p(o_1 \mid y_i) = y_i$ for all forecasts y_i), the optimal decision on any particular occasion will be the one yielding the smallest expected (i.e., probability-weighted average) expense. If the decision is made to protect, the expense will be C with probability 1, and if no protective action is taken the expected loss will be $y_i\,L$ (because no loss is incurred, with probability $1-y_i$). Therefore the smaller expected expense will be associated with the protection action whenever

$$C < y_i L, \tag{9.103a}$$

or

$$C/L < y_i. \tag{9.103b}$$

That is, protection is the optimal action when the probability of the adverse event is larger than the ratio of the cost C to the loss L, which is the origin of the name cost/loss ratio. Different decision makers face problems involving different costs and losses, and so their optimal thresholds for action will be different. Clearly this analysis can be relevant only if $C < L$, because otherwise the protective action offers no potential gains, so that meaningful cost/loss ratios are confined to the unit interval, $0 < C/L < 1$.

Mathematically explicit decision problems of this kind not only prescribe optimal actions, but also provide a way to calculate expected economic outcomes associated with forecasts having particular characteristics. For the simple cost/loss ratio problem these expected economic expenses are the probability-weighted average costs and losses, according the probabilities in the joint distribution of the forecasts and observations, $p(y_i, o_j)$. If only climatological forecasts are available (i.e., if the climatological relative frequency, $\overline{o}$, is forecast on each occasion), the optimal action will be to protect if this climatological probability is larger than C/L, and not to protect otherwise. Accordingly, the expected expense associated with the climatological forecast depends on its magnitude relative to the cost/loss ratio:

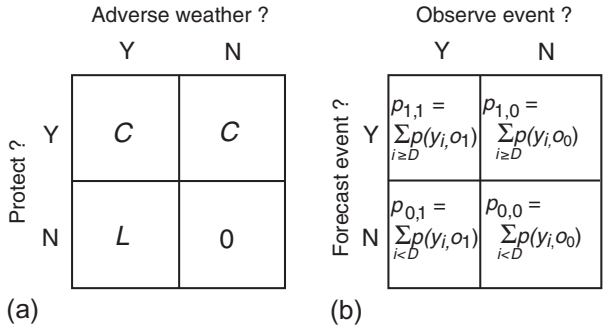

FIGURE 9.39 (a) Loss function for the 2×2 cost/loss ratio situation. (b) Corresponding 2×2 verification table resulting from probability forecasts characterized by the joint distribution $p(y_i, o_j)$ being transformed to nonprobabilistic forecasts according to a particular decision maker's cost/loss ratio. *Adapted from Wilks (2001).*

$$\mathrm{EE}_{\mathrm{clim}} = \begin{cases} C\,, & \text{if } C/L < \overline{o} \\ \overline{o}L, & \text{otherwise.} \end{cases} \tag{9.104}$$

Similarly, if perfect forecasts were available the hypothetical decision maker would incur the protection cost only on the occasions when the adverse weather was about to occur, so the corresponding expected expense would be

$$\mathrm{EE}_{\mathrm{perf}} = \overline{o}C. \tag{9.105}$$

The expressions for expected expenses in Equation 9.104 and 9.105 are simple because the joint distributions of forecasts and observations for climatological and perfect forecasts are also very simple. More generally, a set of probability forecasts for a dichotomous event would be characterized by a joint distribution of the kind shown in Table 9.4a. A cost/loss decision maker with access to probability forecasts that may range throughout the unit interval has an optimal decision threshold, D, corresponding to the cost/loss ratio, C/L. That is, the decision threshold D is that value of the index i corresponding to the smallest probability y_i that is larger than C/L. In effect, the hypothetical cost/loss decision maker transforms probability forecasts summarized by a joint distribution $p(y_i, o_j)$ into nonprobabilistic forecasts for the dichotomous event "adverse weather," in the same way that was described in Sections 9.2.5 and 9.4.6: probabilities y_i for which $i \geq D$ are transformed to "yes" forecasts and forecasts for which $i < D$ are transformed to "no" forecasts. Figure 9.39b illustrates the 2×2 joint distribution (corresponding to Figure 9.1b) for the resulting nonprobabilistic forecasts of the binary event, in terms of the joint distribution of forecasts and observations for the probability forecasts. Here $p_{1,1}$ is the joint frequency that the probability forecast y_i is above the decision threshold D and the event subsequently occurs, $p_{1,0}$ is the joint frequency that the forecast is above the probability threshold but the event does not occur, $p_{0,1}$ is the joint frequency of forecasts below the threshold and the event occurring, and $p_{0,0}$ is the joint frequency of the probability forecasts being below threshold and the event not occurring.

Because the hypothetical decision maker has constructed yes/no forecasts using the decision threshold D that is customized to a particular cost/loss ratio of interest, there is a one-to-one correspondence between the joint probabilities in Figure 9.39b and the loss function in Figure 9.39a. Combining these leads to the expected expense associated with the forecasts characterized by the joint distribution $p(y_i, o_j)$,

$$\mathrm{EE}_{\mathrm{f}} = (p_{1,1} + p_{1,0})\,C + p_{0,1}\,L \tag{9.106a}$$

$$= C \sum_{j=0}^{1} \sum_{i \geq D} p(y_i, o_j) + L \sum_{i<D} p(y_i\, o_1). \tag{9.106b}$$

This expected expense depends both on the particular nature of the decision maker's circumstances, through the cost C, the loss L, and their ratio that defines the decision threshold D; and on the quality of the probability forecasts available to the decision maker, as summarized in the joint distribution of forecasts and observations $p(y_i, o_j)$.

9.9.2. The Value Score

Economic value as calculated in the simple cost/loss ratio decision problem is, for a given cost/loss ratio, a rational and meaningful single-number summary of the quality of probabilistic forecasts for a dichotomous event. However, this measure of forecast quality is different for different decision makers (i.e., different values of C/L). Richardson (2000) proposed using economic value, plotted as a function of the cost/loss ratio, as a graphical verification device for probabilistic forecasts for dichotomous events, after a transformation that ensures retrospective calibration of the forecasts (i.e., $y_i \equiv p(o_1 \mid y_i)$). The ideas are similar to those behind the ROC diagram (see Section 9.4.6), in that forecasts are evaluated through a function that is based on reducing probability forecasts to yes/no forecasts at all possible probability thresholds y_D, and also because conditional and unconditional biases are not penalized. The result is a strictly nonnegative measure of potential (not necessarily actual) economic value in the simplified decision problem, as a function of C/L, for $0 < C/L < 1$.

This basic procedure can be extended to reflect potentially important forecast deficiencies by computing the economic expenses using the original, uncalibrated forecasts (Wilks, 2001). A forecast user without the information or sophistication necessary to recalibrate the forecasts would need to take them at face value and, to the extent that they might be miscalibrated (i.e., that the probability labels y_i might be inaccurate), make suboptimal decisions. Whether or not the forecasts are preprocessed to remove biases, the calculated expected expense (Equation 9.106) can be expressed in the form of a standard skill score (Equation 9.4), relative to the expected expenses associates with climatological (Equation 9.104) and perfect (Equation 9.105) forecasts, called the *value score*:

$$VS = \frac{EE_f - EE_{clim}}{EE_{perf} - EE_{clim}} \tag{9.107a}$$

$$= \begin{cases} \dfrac{(C/L)(p_{1,1} + p_{1,0} - 1) + p_{0,1}}{(C/L)(\overline{o} - 1)}, & \text{if } C/L < \overline{o} \\ \dfrac{(C/L)(p_{1,1} + p_{1,0}) + p_{0,1} - \overline{o}}{\overline{o}\,[(C/L) - 1]}, & \text{if } C/L > \overline{o}. \end{cases} \tag{9.107b}$$

The advantage of this rescaling of EE_f is that sensitivities to particular values of C and L are removed, so that (unlike Equations 9.104–9.106) Equation 9.107 depends only on their ratio, C/L. Perfect forecasts exhibit $VS = 1$, and climatological forecasts exhibit $VS = 0$, for all cost/loss ratios. If the forecasts are recalibrated before calculation of the value score, it will be nonnegative for all cost/loss ratios. Richardson (2001) called this score, for recalibrated forecasts, the potential value, V. However, in the more realistic case that the forecasts are scored at face value, $VS < 0$ is possible if some or all of the hypothetical decision makers would be better served on average by adopting the climatological decision rule, leading to EE_{clim} in Equation 9.104. Related ideas have been presented by Thompson and

Brier (1955), Murphy (1977), and Granger and Pesaran (2000). Mylne (2002) has extended this verification framework for 2×2 decision problems in which protection against the adverse event is only partially effective, and Diebold et al. (1998) present a similar approach for continuous density forecasts.

Figure 9.40 shows VS curves for binary frost forecasts derived from ensembles postprocessed using an MBMP (Equation 8.58) method, with (black solid and dashed) and without (gray) being constrained to maximize calibration on an independent training data set (Wilks, 2018a). The larger economic values occur for cost/loss ratios near the climatological probability of about 0.66, and the economic value is smallest for extreme C/L because in those cases the best decisions are usually obvious according to the climatological probabilities. Although the forecasts postprocessed using the MBMP method (Equation 8.58) are sufficiently well calibrated that no negative economic values occur in Figure 9.40, imposing the calibration constraint improves their economic value for nearly all C/L ratios (i.e., for nearly all forecast users), even though the improved calibration has come at the cost of a slightly reduced eCRPS (Equation 9.83). Comparing Figures 9.40a and 9.40b shows that the improvement in economic value deriving from the calibration constraint is similar to the difference in value between 24–48 h and 48–72 h lead times.

9.9.3. Connections Between VS and Other Verification Approaches

Just as ROC curves are sometimes characterized in terms of the area beneath them, value score curves also can be collapsed to scalar summary statistics. The simple unweighted integral of VS over the full unit interval of C/L is one such summary. This simple function of VS turns out to be equivalent to evaluation of the full set of forecasts using the Brier score, because the expected expense in the cost/lost ratio situation (Equation 9.106) is a linear function of BS (Equation 9.39) (Murphy, 1966). That is, ranking competing forecasts according to their Brier scores, or Brier skill scores, yields the same result as a ranking based on the unweighted integrals of their VS curves. To the extent that the forecast user community might have a nonuniform distribution of cost/loss ratios (e.g., a preponderance of forecast users for whom the protection option is relatively inexpensive), single-number weighted averages of VS also can be computed as statistical expectations of VS with respect to the probability density function for C/L across users of interest (Richardson, 2001; Wilks, 2001).

The VS curve is constructed through a series of 2×2 verification tables, and there are accordingly connections both with scores used to evaluate nonprobabilistic forecasts of binary predictands, and with the ROC curve. For correctly calibrated forecasts, maximum economic value in the cost/loss decision problem is achieved for decision makers for whom C/L is equal to the climatological event relative frequency, because for these individuals the optimal action is least clear from the climatological information alone. This explains the peaks in Figure 9.40 coinciding with C/L equal to the climatological frost probability. Gandin et al. (1992) called these *ideal users*, recognizing that such individuals will benefit most from forecasts. Interestingly, this maximum (potential, because calibrated forecasts are assumed) economic value is given by the Peirce skill score (Equation 9.18), evaluated for the 2×2 table appropriate to this "ideal" cost/loss ratio (Richardson, 2000; Wandishin and Brooks, 2002). Furthermore, the odds ratio (Equation 9.9) $\theta > 1$ for this table is a necessary condition for economic value to be imparted for at least one of the possible cost/loss ratio decision problems (Richardson, 2003; Wandishin and Brooks, 2002). The range of cost/loss ratios for which positive potential economic value can be realized for a given 2×2 verification table is given by its Clayton skill score (Equation 9.19) (Wandishin and Brooks, 2002). Semazzi and Mera (2006) show that the area between the ROC curve and a line that depends on C/L and the economic value of baseline forecasts is proportional to the

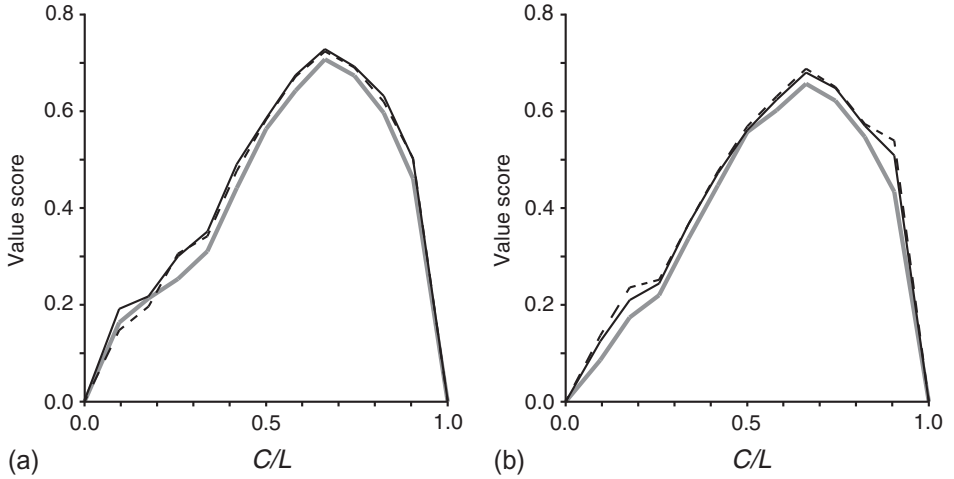

FIGURE 9.40 VS curves for binary January New York city frost forecasts derived from ensembles postprocessed without (gray) and with (light solid and dashed) calibration constraints, at (a) 24–48 h and (b) 48–72 h lead times. *From Wilks (2018a).*

potential economic value V of forecasts that are free of conditional and unconditional biases. Additional connections between VS and attributes of the ROC diagram are provided in Marzban (2012), Mylne (2002), and Richardson (2003).

9.10. VERIFICATION WHEN THE OBSERVATION IS UNCERTAIN

Forecast verification is traditionally undertaken under the tacit assumption that the verifying observation is a true and error-free representation of the predictand. This assumption may be reasonable when observation errors are small relative to forecast errors, but the true state of the predictand can never really be known with certainty because of measurement (or instrument) errors and representativeness errors, and as forecast systems improve the gap between forecast and observation error narrows (e.g., Bowler, 2008; Mittermaier and Stephenson, 2015). Measurement errors may be quite small when instruments such as ordinary thermometers and raingauges, which interact fairly directly with the process or quantity being measured, are used. *Representativeness errors* occur when the measured quantity differs from the predicted quantity. These occur in a spatial sense when there is a scale mismatch between that of the instrument (e.g., a raingage with area approximately 350 cm^2) and the predictand (which might be rainfall averaged over a 100 km^2 area), and in a temporal sense when there is a mismatch between observation time and prediction time.

Bowler (2006) considered the effects of observation error in the 2×2 contingency table setting. Given an (externally derived) characterization of the observational error characteristics, it is possible to reconstruct an expected 2×2 table for hypothetical error-free observations. The existence of observational error degrades apparent forecast skill relative to what might be achieved with error-free observations if that skill is positive, but errors in the observations tend to make negatively skillful forecasts appear less bad. Alternatively, one could consider randomly perturbing the binary observations according to misclassification probabilities for the "yes" and "no" events (Briggs et al., 2005).

Ciach and Krajewski (1999) partition the MSE for radar-derived area-averaged rainfall into two terms: MSE for instrument error (discrepancies between radar-estimated and gauge measurements at the gauge locations) and MSE for representativeness error (discrepancies between local gauge measurements and the true area average). The first of these contributions is straightforward to characterize, using prior data sets for radar-derived precipitation at raingauge locations. The second contribution is much more difficult to calculate without data from a very dense raingauge network, although a modeling-based estimate can be derived by assuming characteristics of the spatial correlation structure of the rainfall fields. Results of this study showed the representativeness error component to be most important at the shortest timescales, but that it remained a significant contribution to overall MSE even for 4-day rainfall accumulations.

The notion of uncertain observations fits well in the context of ensemble forecasting (Chapter 8). The conceptual basis for ensemble forecasting begins with the idea of uncertainty in the initial condition (i.e., the observation at initialization time) for a dynamical forecast model. Therefore a necessary condition for a forecast ensemble to be consistent (for the observation of the predictand to be statistically indistinguishable from the forecast ensemble members) is that the verification, which is the basis for initial conditions in the next forecast cycle, must also be subject to errors and uncertainty. When the magnitude of the observation error is a substantial fraction of the ensemble spread, ignoring observation errors produces overpopulation of the extreme bins of the rank histogram, which leads to an erroneous (or, at least, exaggerated) diagnosis of ensemble underdispersion. The most usual remedy in such cases is to simulate the effects of observational errors by adding random numbers with error characteristics mimicking the observational errors to each ensemble member (Anderson, 1996; Candille and Talagrand, 2008; Hamill, 2001; Saetra et al., 2004).

Replacing the observation with a probability distribution is a seemingly natural approach when uncertainty about the observation will have an important effect. Candille and Talagrand (2008) proposed treating verifying observations explicitly as probabilities (for discrete events) or probability distributions (for continuous predictands). This approach is mathematically tractable, while continuing to allow use of the reliability diagram, the Brier score (including its usual algebraic decomposition), the Brier skill score, the ranked probability score, the continuous ranked probability score, and the ROC diagram. Other studies proceeding along these lines include Pappenberger et al. (2009) and Santos and Ghelli (2012), who consider probability forecasts for binary observations evaluated with the Brier score; Friederichs and Thorarinsdottir (2012), who replace the step function in the CRPS (Equation 9.61) with a Gaussian distribution for the observational uncertainty; and Vannitsem and Hagedorn (2011) who account for observational uncertainty in the fitting of MOS regression equations. Ahrens and Jaun (2007) and Pinson and Hagedorn (2012) consider verification against Monte Carlo samples from observation distributions, Gorgas and Dorninger (2012) propose use of an "analysis ensemble," and Röpnack et al. (2013) and Weijs and van de Giesen (2011) outline Bayesian approaches. Of course, in order to implement the methods just outlined, appropriate distributions characterizing the observation errors must still be defined and estimated, either externally to the verification data being evaluated or subjectively (Briggs et al., 2005).

Ferro (2017) has introduced the notion of the unbiased and proper scoring rule for forecasts of uncertain observations. The expected value of an unbiased proper score S evaluated with respect to the uncertain observation o will be the same as the expected value of the score that would hypothetically be achieved by evaluating its counterpart proper score S_0 with respect to the true but unknown value x. He further defines two unbiased and proper scoring rules pertaining to the Dawid-Sebastiani score (Equation 9.67) for PDF forecasts of a continuous variable, which require knowledge of only the first

two moments of the conditional distribution of the observation o given the truth x. The first of these assumes that the observational uncertainty can be represented as additive white noise, so that $\mathrm{E}(o|x) = a + bx$ and $\mathrm{Var}(o|x) = c^2$, where a, b, and c are constants that are assumed known. In this case the score that is unbiased and proper for the DSS is

$$\mathrm{DSS_A} = 2\,\ln(\sigma) + \frac{(o - a - b\mu)^2 - c^2}{b^2\sigma^2}, \tag{9.108}$$

where μ and σ^2 are the mean and variance of the forecast distribution. In the limit of no observational uncertainty ($a = c = 0$ and $b = 1$), Equation 9.108 reduces to the DSS in Equation 9.67. The second unbiased and proper score for the DSS pertains to a multiplicative error distribution for o, for which $\mathrm{E}(o|x) = bx$ and $\mathrm{Var}(o|x) = c^2x^2$, yielding the score

$$\mathrm{DSS}_M = 2\,\ln(\sigma) + \frac{(o - b\mu)^2 - o^2c^2/(b^2 + c^2)}{b^2\sigma^2}. \tag{9.109}$$

Naveau and Bessac (2018) extend these ideas to encompass also multiplicative gamma-distributed errors, applicable to such quantities as precipitation, and provide results for CRPS.

9.11. SAMPLING AND INFERENCE FOR VERIFICATION STATISTICS

Practical forecast verification is necessarily concerned with finite samples of forecast-observation pairs. The verification statistics that can be computed from a particular data set are no less subject to sampling variability than are any other sort of statistics. If a different sample of the same kind of forecasts and observations were hypothetically to become available, the value of verification statistic(s) computed from it likely would be at least somewhat different. To the extent that the sampling distribution for a verification statistic is known or can be estimated, confidence intervals around it can be obtained, and formal tests (e.g., against a null hypothesis of zero skill) can be constructed.

In many cases the procedures presented in Chapter 5 can be used to construct statistical inferences for forecast verification statistics, and therefore to assess statistical significance for sample assessments of accuracy, skill, and so on. Notably, when results for a large number of forecasts and observations have been averaged to form a sample mean score, the Central Limit Theorem would suggest use of a t test (Section 5.2.1–5.2.4) to make inferences about that mean. Of course, when such averages have been computed over a sequence of autocorrelated (e.g., daily) forecasts and observations, the effects of autocorrelation on their sampling distributions must be represented. Similarly, when different forecast sources are being compared with respect to the same sequence of observations, failing to represent the paired nature of these data when assessing the possible significance of their differences, whether the data are also autocorrelated (Section 5.2.4) or not (Section 5.2.3), will likely lead to erroneous inferences (DelSole and Tippett, 2014; Gilleland et al., 2018; Jarman and Smith, 2018; Siegert et al., 2017). When multiple inferences are being assessed simultaneously the inferential standards need to be adjusted appropriately (Geer, 2016), as outlined in Section 5.4.

9.11.1. Sampling Characteristics of Contingency Table Statistics

In principle, the sampling characteristics of many 2×2 contingency table statistics follow from a fairly straightforward application of independent binomial sampling (Agresti, 1996). For example, such measures as the false alarm ratio (Equation 9.11), the hit rate (Equation 9.12), and the false alarm rate

(Equation 9.13) are all proportions that estimate (conditional) probabilities. If the contingency table counts (see Figure 9.1a) have been produced independently from stationary (i.e., constant-p) forecast and observation systems, those counts are (conditional) binomial variables, and the corresponding proportions (such as FAR, H and F) are sample estimates of the corresponding binomial probabilities (Seaman et al., 1996).

A direct approach to finding confidence intervals for sample proportions x/N that estimate the binomial parameter p is to use the binomial probability distribution function (Equation 4.1). A $(1–\alpha)\cdot 100\%$ confidence interval for the underlying probability that is consistent with the observed proportion x/N can be defined by the extreme values of x on each tail that include probabilities of at least $1–\alpha$ between them, inclusive. Unfortunately the result, called the *Clopper-Pearson exact interval*, generally will be inaccurate to a degree (and, specifically, too wide) because of the discreteness of the binomial distribution (Agresti and Coull, 1998). Another simple approach to calculation of confidence intervals for sample proportions is to invert the Gaussian approximation to the binomial distribution (Equation 5.2). Since Equation 5.2b is the standard deviation σ_x for the number of binomial successes X, the corresponding variance for the estimated proportion $\hat{p} = x/N$ is $\sigma_p^2 = \sigma_x^2/N^2 = \hat{p}(1-\hat{p})/N$ (using Equation 4.17). The resulting $(1–\alpha)\cdot 100\%$ confidence interval is then

$$p = \hat{p} \pm z_{(1-\alpha/2)}[\hat{p}(1-\hat{p})/N]^{1/2}, \tag{9.110}$$

where $z_{(1-\alpha/2)}$ is the $(1–\alpha/2)$ quantile of the standard Gaussian distribution (e.g., $z_{(1-\alpha/2)} = 1.96$ for $\alpha = 0.05$).

Equation 9.110 can be quite inaccurate, in the sense that the actual probability of including the true p is substantially smaller than $1–\alpha$, unless N is very large. However, this bias can be corrected using the modification (Agresti and Coull, 1998) to Equation 9.110,

$$p = \frac{\hat{p} + \dfrac{z_{(1-\alpha/2)}^2}{2N} \pm z_{(1-\alpha/2)}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{N} + \dfrac{z_{(1-\alpha/2)}^2}{4N^2}}}{1 + \dfrac{z_{(1-\alpha/2)}^2}{N}}. \tag{9.111}$$

The differences between Equations 9.111 and 9.110 are in the three terms involving $z_{(1-\alpha/2)}^2/N$, which approach zero in Equation 9.111 for large N. Standard errors according to Equation 9.110 are tabulated for ranges of $\hat{p}$ and N in Thornes and Stephenson (2001).

Marzban and Sandgathe (2008) derive the approximation to the standard deviation of the sampling distribution of the threat score (Equation 9.8),

$$\hat{\sigma}_{TS} \approx TS\sqrt{\frac{1}{a}\left(\frac{b}{a+b} + \frac{c}{a+c}\right)}. \tag{9.112}$$

Another relevant result from the statistics of contingency tables (Agresti, 1996), is that the sampling distribution of the logarithm of the odds ratio (Equation 8.9) is approximately Gaussian distributed for sufficiently large $n = a+b+c+d$, with estimated standard deviation

$$\hat{\sigma}_{\ln(\theta)} = \left[\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}\right]^{1/2}. \tag{9.113}$$

Thus a floor on the magnitude of the sampling uncertainty for the odds ratio is imposed by the smallest of the four counts in Table 8.1a. When the null hypothesis of independence between forecasts and observations (i.e., $\theta = 1$) is of interest, it could be rejected if the observed $\ln(\theta)$ is sufficiently far from $\ln(1) = 0$, with respect to Equation 9.113.

For large n, sampling distributions for the contingency-table statistics are expected to be approximately Gaussian, which suggests use of one-sample t tests as the basis of inferences about them, provided the underlying data are stationary and independent. Hogan and Mason (2012) tabulate expressions for the sampling variance of additional 2×2 contingency table verification measures that could be used in a similar way.

Radok (1988) shows that the sampling distribution for the multicategory Heidke skill score (Equation 9.24) is proportional to that of a chi-square variable.

Example 9.8. Inferences for Selected Contingency Table Verification Measures

The hit and false alarm rates for the Finley tornado forecasts in Table 1.1a are $H = 28/51 = 0.549$ and $F = 72/2752 = 0.026$, respectively. Because the underlying forecast situations were fairly widely separated in space and time, the underlying counts reasonably approximate independent sampling. These proportions are sample estimates of the conditional probabilities of tornados having been forecast, given either that tornados were or were not subsequently reported. Using Equation 9.111, $(1–\alpha)\cdot 100\% = 95\%$ confidence intervals for the true underlying conditional probabilities can be estimated as

$$H = \frac{.549 + \dfrac{1.96^2}{(2)(51)} \pm 1.96\sqrt{\dfrac{.549(1-.549)}{51} + \dfrac{1.96^2}{(4)(51)^2}}}{1 + \dfrac{1.96^2}{51}} = .546 \pm .132 = \{.414, .678\} \tag{9.114a}$$

and

$$F = \frac{.026 + \dfrac{1.96^2}{(2)(2752)} \pm 1.96\sqrt{\dfrac{.026(1-.026)}{2752} + \dfrac{1.96^2}{(4)(2752)^2}}}{1 + \dfrac{1.96^2}{2752}} = .0267 \pm .00598 = \{.0207, .0326\}. \tag{9.114b}$$

The precision of the estimated false alarm rate is much better (its standard error is much smaller) in part because the overwhelming majority of observations ($b+d$) were "no tornado;" but also in part because $p(1–p)$ is small for extreme values, and larger for intermediate values of p. Assuming independence of the forecasts and observations (in the sense illustrated in Equation 9.16), plausible useless-forecast benchmarks for the hit and false alarm rates might be $H_0 = F_0 = (a+b)/n = 100/2803 = 0.0357$. Neither of the 95% confidence intervals in Equation 9.114 include this value, leading to the inference that H and F for the Finley forecasts are better than would have been achieved by chance.

Stephenson (2000) notes that, because the Peirce skill score (Equation 9.18) can be calculated as the difference between H and F, confidence intervals for it can be calculated using simple binomial sampling considerations because H and F are mutually independent. In particular, since on the strength of the Central Limit Theorem the sampling distributions of both H and F are approximately Gaussian for

sufficiently large sample sizes, under these conditions the sampling distribution of the PSS will be Gaussian, with estimated standard deviation

$$\hat{\sigma}_{\mathrm{PSS}} = \sqrt{\hat{\sigma}_H^2 + \hat{\sigma}_F^2}. \tag{9.115}$$

For the Finley tornado forecasts, PSS = 0.523, so that a 95% confidence interval around this value could be constructed as $0.523 \pm 1.96\,\hat{\sigma}_{PSS}$. Estimating $\hat{\sigma}_H = .132/1.96 = 0.673$ from Equation 9.114a and $\hat{\sigma}_F = .00698/1.96 = .00305$ from Equation 9.114b, or interpolating from the table in Thornes and Stephenson (2001), this interval would be $0.523 \pm 1.96\,(0.0673^2 + 0.00305^2)^{1/2} = 0.523 \pm 0.132 = \{0.391, 0.655\}$. Since this interval does not include zero, a reasonable inference would be that these forecasts exhibited significant skill according to the Peirce skill score. Hanssen and Kuipers (1965) and Woodcock (1976) derive the alternative expression for the sampling variance of the PSS,

$$\hat{\sigma}_{\mathrm{PSS}}^2 = \frac{n^2 - 4(a+c)(b+d)\,\mathrm{PSS}^2}{4n\,(a+c)(b+d)}. \tag{9.116}$$

Again assuming a Gaussian sampling distribution, Equation 9.116 estimates the 95% confidence interval for PSS for the Finley tornado forecasts as $0.523 \pm (1.96)(.070) = \{0.386, 0.660\}$.

Finally, the odds ratio for the Finley forecasts is $\theta = (28)(2680)/(23)(72) = 45.31$, and the standard deviation of the (approximately Gaussian) sampling distribution of its logarithm (Equation 9.113) is $(1/28 + 1/72 + 1/23 + 1/2680)^{1/2} = 0.306$. The null hypothesis that the forecasts and observations are independent (i.e., $\theta_0 = 1$) produces the t-statistic $[\ln(45.31) - \ln(1)]/0.306 = 12.5$, which would lead to emphatic rejection of that null hypothesis. ◇

The calculations in this section have relied on the assumptions that the verification data are independent and, for the sampling distribution of proportions, that the underlying probability p is stationary (i.e., constant across forecasts). The independence assumption might be violated, for example, if the data set consists of a sequence of daily forecast-observation pairs. The stationarity assumption might be violated if the data set includes a range of locations with different climatologies for the forecast variable. When comparing the performance of two forecasting systems pertaining to the same set of observations, use of the equations presented in this section to formulate 2-sample t tests would be inappropriate because correlations between the paired forecasts would not be represented. In cases where these assumptions might be violated, inferences for contingency-table verification measures still can be made, by estimating their sampling distributions using appropriately constructed resampling approaches (see Section 9.11.6).

9.11.2. ROC Diagram Sampling Characteristics

Because confidence intervals around sample estimates for the hit rate H and the false alarm rate F can be calculated using Equation 9.111, confidence regions around individual (F, H) points in a ROC diagram can also be calculated and plotted. A complication is that, in order to define a joint, simultaneous $(1-\alpha)\cdot 100\%$ confidence region around a sample (F, H) point, each of the two individual confidence intervals must cover its corresponding true value with a probability that is somewhat larger than $(1-\alpha)$. Essentially, this adjustment is necessary in order to make valid simultaneous inference in a multiple testing situation (cf. Section 5.4.1). Since H and F are at least approximately independent, a reasonable approach to deciding the appropriate sizes of the two confidence intervals is to use the *Bonferroni inequality* (Equation 12.56). In the case of the ROC diagram, where the joint confidence

region is $K = 2$-dimensional, Equation 12.56 says that the rectangular region defined by two $(1-\alpha/2)\cdot 100\%$ confidence intervals for F and H will jointly enclose the true (F, H) pair with probability at least as large as $1-\alpha$. For example, a joint 95% (at least) rectangular confidence region will be defined by two 97.5% confidence intervals, calculated using $z_{1-\alpha/4} = z_{.9875} = 2.24$, in Equation 9.111.

A test for the statistical significance of the area A under the ROC curve, against the null hypothesis that the forecasts and observations are independent (i.e., that $A_0 = 1/2$), is available. The sampling distribution of the ROC area computed using trapezoidal integration, given the null hypothesis of no relationship between forecasts and observations, is proportional to the distribution of the Mann-Whitney U-statistic (Equations 5.22 and 5.23) (Bamber, 1975; Mason and Graham, 2002), so that the test for the ROC area is equivalent to the Wilcoxon-Mann-Whitney test applied to the two likelihood distributions $p(y_i|o_1)$ and $p(y_i|o_2)$ (cf. Figure 9.13). To calculate this test, the ROC area A is transformed to a Mann-Whitney U variable according to

$$U = n_1 n_2 (1 - A). \tag{9.117}$$

Here $n_1 = a+c$ is the number of "yes" observations, and $n_2 = b+d$ is the number of "no" observations. Notice that, under the null hypothesis $A_0 = 1/2$, Equation 9.117 is exactly the mean of the Gaussian approximation to the sampling distribution of U in Equation 5.24a. This null hypothesis is rejected for sufficiently small U, or equivalently for sufficiently large ROC area A.

Example 9.9. Confidence and Significance Statements about a ROC Diagram

Figure 9.41 shows the ROC diagram for the Finley tornado forecasts (Table 9.1a), together with the 97.5% confidence intervals for F and H. These are $0.020 \le F \le 0.034$ and $0.396 \le H \le 0.649$, and were calculated from Equation 9.111 using $z_{1-\alpha/4} = z_{.9875} = 2.24$. The confidence interval for F is only about as wide as the dot locating the sample (F, H) pair, both because the number of "no tornado" observations is large, and because the proportion of false alarms is quite small. These two 97.5% confidence intervals define a 95% rectangular confidence region for the true (F, H) pair according to the Bonferroni inequality

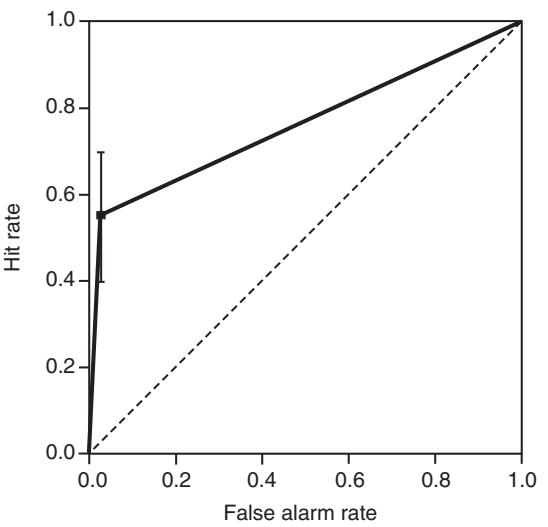

FIGURE 9.41 ROC diagram for the Finley tornado forecasts (Table 9.1a), with the 95% simultaneous Bonferroni (Equation 12.56) confidence intervals for the single (F, H) point, calculated using Equation 9.111.

(Equation 12.56). This region does not include the dashed 1:1 line, indicating that it is improbable for these forecasts to have been generated by a process that was independent of the observations.

The area under the ROC curve in Figure 9.41 is 0.761. If the true ROC curve for the process from which these forecast-observation pairs have been sampled is the dashed 1:1 diagonal line, what is the probability that a ROC area A this large or larger could have been achieved by chance, given $n_1 = 51$ "yes" observations and $n_2 = 2752$ "no" observations? Equation 9.117 yields $U = (51)(2752)(1–0.761) = 33544$, the unusualness of which can be evaluated in the context of the (null) Gaussian distribution with mean $\mu_U = (51)(2752)/2 = 70176$ (Equation 5.24a) and standard deviation $\sigma_U = [(51)(2752)(51+2752+1)/12]^{1/2} = 5727$ (Equation 5.24b). The resulting test statistic is $z = (33544–70176)/5727 = -6.4$, so that the null hypothesis of no association between the forecasts and observations would be strongly rejected. ◇

As before, when two forecast systems have been used to predict the same observations it is necessary to account for the correlation between them in order to judge the possible statistical significance of the difference in the resulting ROC areas. Hanley and McNeil (1983) and DeLong et al. (1988) provide methods for doing so when the data are serially independent. Lahiri and Yang (2018) describe computation of confidence bands around a single ROC curve, even in the presence of serial correlation.

9.11.3. Brier Score and Brier Skill Score Inference

Bradley et al. (2008) have derived an expression for the variance of the sampling distribution of the Brier score (Equation 9.39), assuming that the n forecast observation pairs (y_i, o_i) are independent random samples from a homogeneous joint distribution of forecasts and observations. Their result can be expressed as

$$\hat{\sigma}^2_{\mathrm{BS}} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^4 - 4y_i^3 o_i + 6y_i^2 o_i - 4y_i o_i + o_i\right) - \frac{\mathrm{BS}^2}{n}. \tag{9.118}$$

Similarly, Bradley et al. (2008) derive the approximate sampling variance for the Brier skill score (Equation 9.40),

$$\begin{aligned}\hat{\sigma}^2_{\mathrm{BSS}} \approx{}& \left(\frac{n}{n-1}\right)^2 \frac{\hat{\sigma}^2_{\mathrm{BS}}}{\overline{o}^2(1-\overline{o})^2} + \frac{n}{(n-1)^3}\frac{(1-\mathrm{BSS})^2}{\overline{o}(1-\overline{o})}[\overline{o}(1-\overline{o})(6-4n)+n-1] \\ &+ \left(\frac{n}{n-1}\right)^2 \frac{(2-4\overline{o})(1-\mathrm{BSS})}{\overline{o}(1-\overline{o})}\left(1 + \frac{\sum_{i=1}^{n}\left(y_i^2 o_i - 2y_i o_i\right)}{n\overline{o}} + \frac{\sum_{i=1}^{n}\left(y_i^2 o_i - y_i^2\right)}{n(1-\overline{o})}\right).\end{aligned} \tag{9.119}$$

Since Equation 9.118 and 9.119 require estimates of high (up to fourth) moments of the joint distribution of the forecasts and observations, the sample size n must be fairly large in order for these estimates to be usefully accurate. Figure 9.42 shows sample sizes necessary for Equations 9.118 and 9.119 to yield Gaussian 95% and 99% confidence intervals exhibiting approximately correct coverage. Quite large sample sizes ($n > 1000$) are required for higher-skill forecasts of relatively rare events, whereas much more modest sample sizes are adequate for low-skill forecasts of common events. Using too few samples yields estimated sampling variances, and therefore confidence intervals, that are too small. Bradley et al. (2008) also note that the sample estimate of BSS exhibits a negative bias that may be appreciable for small sample sizes and relatively rare events (small n and $\overline{o}$).

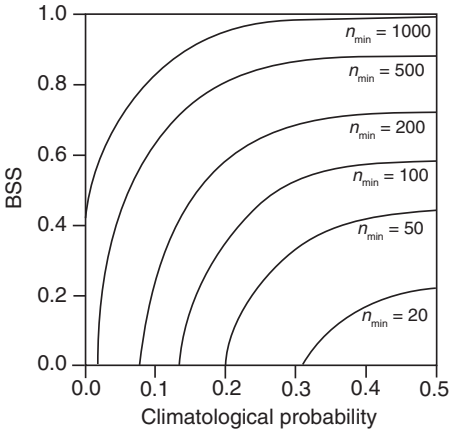

FIGURE 9.42 Sample sizes necessary for Equation 9.118 and 9.119 to yield approximately correct variance estimates, as a function of the sample climatological probability and sample Brier skill score. *From Wilks (2010).*

When the forecasts and binary events exhibit first-order autoregressive serial dependence (Section 10.3.1), Equations 9.118 and 9.119 can be used with "effective sample size" adjustments to that appropriately inflate these sampling variances (Wilks, 2010). For the sampling variance of the Brier score, the sample size n in Equation 9.118 is replaced by the effective sample size estimate

$$n' = n\frac{1-(1-\overline{o})[b(1-BS)r_1]^2}{1+(1-\overline{o})[b(1-BS)r_1]^2}, \tag{9.120}$$

and for the Brier skill score (Equation 9.119) the effective sample size can be estimated as

$$n' = n\frac{1-(1-\overline{o})[b(1-BS)r_1]^4}{1+(1-\overline{o})[b(1-BS)r_1]^4}. \tag{9.121}$$

In these two effective sample size equations r_1 is the lag-1 autocorrelation, and b is the slope of the calibration function in the reliability diagram, where $b = 1$ for calibrated forecasts, $b < 1$ for overconfident forecasts, and $b > 1$ for underconfident forecasts. Lahiri and Yang (2016) analyze inflation of the variance estimates in Equation 9.118 and 9.119 under more general time dependence conditions.

Siegert (2014) describes a method to estimate the sampling variances for each of the three components of the Brier score decomposition in Equation 9.45, assuming statistical independence among the n samples.

9.11.4. Reliability Diagram Sampling Characteristics

The calibration-function portion of the reliability diagram consists of I conditional outcome relative frequencies that estimate the conditional probabilities $p(o_1 \mid y_i)$, $i = 1, \ldots, I$. If independence and stationarity are reasonably approximated, then confidence intervals around these points can be computed using either Equation 9.110 or Equation 9.111. To the extent that these intervals include the 1:1 perfect reliability diagonal, a null hypothesis that the forecaster(s) or forecast system has produced calibrated forecasts would not be rejected. To the extent that these intervals do not include the horizontal "no resolution" line, a null hypothesis that the forecasts are no better than climatological guessing would be rejected.

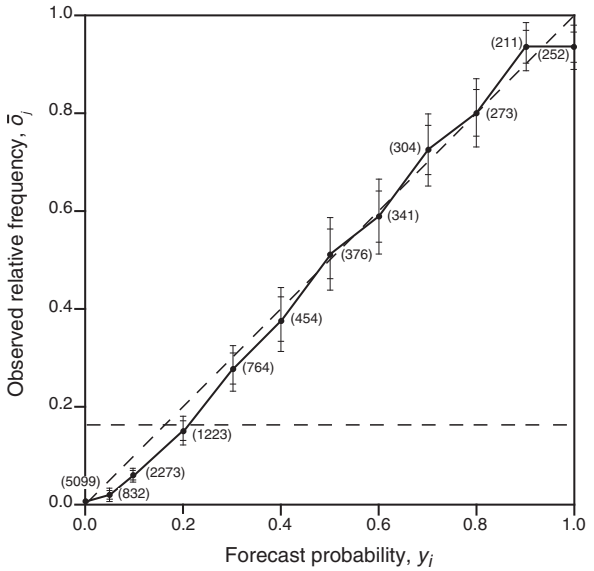

FIGURE 9.43 Reliability diagram for the probability-of-precipitation data in Table 9.2, with 95% confidence intervals on each conditional probability estimate, calculated using Equation 9.110. Inner confidence limits pertain to individual points, and outer bounds are joint Bonferroni (Equation 12.56) confidence limits. Raw subsample sizes N_i are shown parenthetically. The 1:1 perfect reliability and horizontal "no resolution" lines are dashed.

Figure 9.43 shows the reliability diagram for the forecasts summarized in Table 9.2, with 95% confidence intervals drawn around each of the $I = 12$ conditional relative frequencies. The stationarity assumption for these estimated probabilities is reasonable, because the forecasters have sorted the forecast-observation pairs according to their judgments about those probabilities. The independence assumption is less well justified, because these data are simultaneous forecast-observation pairs for about one hundred locations in the United States, so that positive spatial and temporal correlations among both the forecasts and observations would be expected. Accordingly the confidence intervals drawn in Figure 9.43 are possibly too narrow.

Because the sample sizes (shown parenthetically in Figure 9.43) are large, Equation 9.110 was used to compute the confidence intervals. For each point, two confidence intervals are shown. The inner, narrower intervals are ordinary individual confidence intervals, computed using $z_{1-\alpha/2} = 1.96$, for $\alpha = 0.05$ in Equation 9.110. An interval of this kind would be appropriate if confidence statements about a single one of these points are of interest. The outer, wider confidence intervals are joint $(1-\alpha)\cdot 100\% = 95\%$ Bonferroni (Equation 12.56) intervals, computed using $z_{1-[\alpha/12]/2} = 2.87$, again for $\alpha = 0.05$. The meaning of these outer, Bonferroni, intervals is that the probability is at least 0.95 that all $I = 12$ of the conditional probabilities being estimated are simultaneously within their respective individual confidence intervals. Thus a (joint) null hypothesis that all of the forecast probabilities are calibrated would be rejected if any one of them fails to include the diagonal 1:1 line (dashed), which in fact does occur for $y_1 = 0.0$, $y_2 = 0.05$, $y_3 = 0.1$, $y_4 = 0.2$, and $y_{12} = 1.0$. On the other hand, it is clear that these forecasts are overall much better than random climatological guessing, since the Bonferroni confidence intervals overlap the dashed horizontal no resolution line only for $y_4 = 0.2$, and are in general quite far from it.

An alternative approach to computation of confidence intervals around the points on the reliability diagram has been proposed by Bröcker and Smith (2007b). Using a bootstrap approach that accounts for

the randomness of the number of forecasts in each of the I bins, they plot "consistency bars" around the vertical projections of the reliability diagram points onto the 1:1 diagonal line, in order to evaluate the likelihoods of the observed conditional relative frequencies under a null assumption of perfect reliability. Pinson et al. (2010) extend this approach to accommodate serial correlation in the verification data.

9.11.5. Assessing Ensemble Calibration

Ensemble calibration is typically diagnosed using verification rank histograms (for scalar, discrete ensembles), PIT histograms (for ensembles postprocessed to yield continuous predictive distributions) or MST histograms (for multivariate discrete ensembles). Distinguishing between true deviations from verification rank histogram (and also PIT and MST histogram) uniformity and mere sampling variations usually is approached through the chi-square goodness-of-fit test (Section 5.2.5). Here the null hypothesis is a uniform rank (or PIT or MST) histogram, so the expected number of counts in each bin is $n/(m+1)$, and the test is evaluated using the chi-square distribution with $\nu = m$ degrees of freedom (because there are $m+1$ bins). This approach assumes independence of the n ensembles being evaluated, and so may not be appropriate in unmodified form if, for example, the ensembles pertain to consecutive days. Additive corrections to tabulated critical chi-square values appropriate to serial correlation in the forecasts are given in Table 9.8, for rank- and PIT-histogram uniformity, and in Table 9.9 for MST histogram uniformity, assuming that the time dependence can be characterized by the lag-1 autocorrelation r_1. That is, given autocorrelation of the underlying data, larger values of the χ^2 test statistic are required in order to reject a null hypothesis of rank uniformity.

A potential drawback of the ordinary chi-square test in this context is that it assesses deviations from rank uniformity independently of the ordering of the bars in the histogram. That is, deviations from the ideal uniformity level are treated equally whether they occur randomly across the histogram bars, or if (for example) the positive deviations are concentrated toward the left, and the negative deviations are concentrated toward the right. Accordingly it is not focused toward detecting coherent patterns in the

TABLE 9.8 Additive Corrections to Tabulated χ^2 Critical Values to Test Uniformity of Conventional (Scalar) Rank Histograms and PIT Histograms, as Functions of lag-1 Autocorrelation r_1

	Test level, α			
r_1	0.10	0.05	0.01	0.001
0.1	0.3	0.3	0.6	1.1
0.2	0.8	0.9	1.4	2.4
0.3	1.5	1.8	2.8	4.6
0.4	2.6	3.1	4.9	8.3
0.5	4.1	5.1	8.4	14.6
0.6	6.6	8.6	14.3	25.3
0.7	11.2	14.8	25.2	44.3
0.8	20.9	28.1	48.6	85.1
0.9	50.5	69.0	121.7	214.2

From Wilks (2004).

TABLE 9.9 Additive Corrections to Tabulated χ^2 Critical Values to Test Uniformity of MST Histograms, as Functions of lag-1 Autocorrelation r_1

	Test level, α			
r_1	0.10	0.05	0.01	0.001
0.4	0.4	0.5	0.6	1.1
0.5	0.6	0.9	1.3	2.2
0.6	1.3	1.6	2.4	4.4
0.7	2.6	3.4	5.0	8.8
0.8	5.4	7.1	11.9	22.6
0.9	15.6	21.0	37.2	68.6

Corrections for $r_1 < 0.4$ are negligible.
From Wilks (2004).

rank histogram such as those illustrated in Figure 9.19. Alternative tests designed to be sensitive to these characteristic patterns, and which therefore exhibit greater power to detect them, are described in Elmore (2005) and Jolliffe and Primo (2008).

9.11.6. Resampling Verification Statistics

Often the uncertainty characteristics of verification statistics with unknown sampling distributions are of interest. Or, sampling characteristics of verification statistics discussed previously in this section are of interest, but assumptions of independent sampling or first-order autoregressive dependence cannot be supported. In such cases, statistical inference for forecast verification statistics can be addressed through resampling tests, as described in Sections 5.3.3 through 5.3.5. These procedures are very flexible, and the resampling algorithm used in any particular case will depend on the specific setting.

For problems where the sampling distribution of the verification statistic is unknown, but independence can reasonably be assumed, implementation of conventional permutation or bootstrap tests are straightforward. Illustrative examples of the bootstrap in forecast verification can be found, for example, in Bröcker and Smith (2007b) and Roulston and Smith (2003). Bradley et al. (2003) use the bootstrap to evaluate the sampling distributions of the reliability and resolution terms in Equation 9.4, using the probability-of-precipitation data in Table 9.2. Déqué (2003) illustrates permutation tests for a variety of verification statistics.

Special problems occur when the data to be resampled exhibit spatial and/or temporal correlation. A typical cause of spatial correlation is the occurrence of simultaneous data at multiple locations, that is, maps of forecasts and observations. Hamill (1999) describes a permutation test for a paired comparison of two forecasting systems, in which problems of nonindependence of forecast errors have been obviated by spatial pooling. Alternatively, the effects of spatial correlation on resampled verification statistics can be accounted for automatically if the resampled objects are entire maps, rather than individual locations resampled independently of each other. Similarly, the effects of time correlation in the forecast verification statistics can be accounted for using the moving-blocks bootstrap (see Section 5.3.5). The moving-blocks bootstrap is equally applicable to scalar data (e.g., individual forecast-observation pairs at single locations, which are autocorrelated), or to entire autocorrelated maps of forecasts and observations

(Wilks, 1997b). Pinson et al. (2010) apply a spectrally based (i.e., based mathematically on concepts from Chapter 10) resampling procedure to account for the effects of serial correlation on the reliability diagram.

9.12. EXERCISES

9.1. For the forecast verification data in Table 9.2,
   a. Reconstruct the joint distribution, $p(y_i, o_j)$, $i = 1, \ldots, 12$, $j = 1, 2$.
   b. Compute the unconditional (sample climatological) probability $p(o_1)$.

9.2. Construct the 2×2 contingency table that would result if the probability forecasts in Table 9.2 had been converted to nonprobabilistic rain/no rain forecasts, with a threshold probability of 0.25.

9.3. Using the 2×2 contingency table from Exercise 9.2, compute
   a. The proportion correct.
   b. The threat score.
   c. The Heidke skill score.
   d. The Peirce skill score.
   e. The Gilbert skill score.

9.4. For the event o_3 (3 to 4 in. of snow) in Table 9.10 find
   a. The threat score.
   b. The hit rate.
   c. The false alarm ratio.
   d. The bias ratio.

TABLE 9.10 A 4×4 Contingency Table for Snow Amount Forecasts in the Eastern Region of the United States During the Winters 1983/1984 Through 1988/1989

	o_1	o_2	o_3	o_4
y_1	35,915	477	80	28
y_2	280	162	51	17
y_3	50	48	34	10
y_4	28	23	185	34

The event o_1 is 0–1 in., o_2 is 2–3 in., o_3 is 3–4 in., and o_4 is $\geq$6 in.
From Goldsmith (1990).

9.5. Using the 4×4 contingency table in Table 9.10, compute
   a. The joint distribution of the forecasts and the observations.
   b. The proportion correct.
   c. The Heidke skill score.
   d. The Peirce skill score.

9.6. For the persistence forecasts for the January 1987 Ithaca maximum temperatures in Table A.1 (i.e., the forecast for 2 January is the observed temperature on 1 January, etc.), compute
   a. The MAE.
   b. The RMSE.

c. The ME (bias).
d. The skill, in terms of RMSE, with respect to the sample climatology.

9.7. Hypothetical forecasts and observations for three stations on three days are shown in Table 9.11.
a. Compute the overall MSE skill score (Equation 9.37), correctly computing the skill with respect to the sample climatological means pertaining to the individual stations.
b. Compute the overall MSE skill score, incorrectly based on the overall sample climatological mean.

TABLE 9.11 Hypothetical Nonprobabilistic Forecasts and Observations of Temperature on Three Days at Three Stations Having Different Climatological Values

	Station A		Station B		Station C	
	Forecast	Observed	Forecast	Observed	Forecast	Observed
Day 1	−2	2	3	7	8	11
Day 2	2	4	7	8	12	13
Day 3	9	6	13	11	18	16

9.8. Using the collection of hypothetical PoP forecasts summarized in Table 9.12,
a. Calculate the Brier score.
b. Calculate the Brier score for (the sample) climatological forecast.
c. Calculate the skill of the forecasts with respect to the sample climatology.
d. Draw the reliability diagram.
e. Compute the average Ignorance score, assuming that $y_1 = 0.01$ and $y_{11} = 0.99$.

TABLE 9.12 Hypothetical Verification Data for 1000 Probability-of-Precipitation Forecasts

Forecast probability, y_i	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00
Number of times forecast	293	237	162	98	64	36	39	26	21	14	10
Number of precipitation occurrences	9	21	34	31	25	18	23	18	17	12	9

9.9. For the hypothetical forecast data in Table 9.12,
a. Compute the likelihood-base rate factorization of the joint distribution $p(y_i, o_j)$.
b. Draw the discrimination diagram.
c. Draw the ROC curve.
d. Test whether the area under the ROC curve is significantly greater than 1/2.

9.10. Suppose you honestly forecast the probability of precipitation for tomorrow as 20%. What is your expected Brier Score for this forecast?

9.11. For the hypothetical probabilistic three-category precipitation amount forecasts in Table 9.13, where only five distinct forecast vectors have been used,
a. Calculate the average RPS.

b. Calculate the RPS skill of the forecasts with respect to the sample climatology.
c. Calculate the average Ignorance score.

TABLE 9.13 Hypothetical Verification for 500 Probability Forecasts of Precipitation Amounts

Forecast Probabilities For			Number of Forecast Periods Verifying As		
<0.01 in.	0.01–0.24 in.	≥0.25 in.	<0.01 in.	0.01–0.24 in.	≥0.25 in.
.8	.1	.1	263	24	37
.5	.4	.1	42	37	12
.4	.4	.2	14	16	10
.2	.6	.2	4	13	6
.2	.3	.5	4	6	12

9.12. For the hypothetical five-member bivariate ensemble and corresponding forecast shown in Figure 9.24,
a. Compute the preranks π_{MRH} for the observation and the five ensemble members.
b. To which histogram bin will this observation contribute?

9.13. Suppose a particular forecaster's 75% central credible interval for the next day's maximum temperature is 27–32°F.
a. If the verifying temperature is 34°F, evaluate the accuracy of this forecast using Winkler's score.
b. Assuming that this forecaster's subjective distribution is Gaussian, what would that person's probability be for tomorrow's maximum temperature falling below 25°F?

9.14. Table 9.14 shows a set of 20 hypothetical ensemble forecasts, each with five members, and corresponding observations.
a. Plot the verification rank histogram.
b. Qualitatively diagnose the performance of this sample of forecast ensembles.

TABLE 9.14 A Set of 20 Hypothetical Ensemble Forecasts, of Ensemble Size 5, and Corresponding Observations

Case	Member 1	Member 2	Member 3	Member 4	Member 5	Observation
1	7.9	7.3	5.5	6.9	8.3	7.7
2	7.4	5.6	8.2	5.8	6.1	9.4
3	9.5	8.3	10.5	8.9	6.1	8.7
4	6.1	7.8	5.1	10.4	4.9	3.4
5	6.3	5.8	5.1	6.0	4.1	7.3
6	8.1	6.8	1.8	6.7	10.5	8.2
7	4.4	5.6	7.7	6.0	7.0	4.3
8	5.9	3.0	4.4	7.2	9.1	7.0
9	5.2	5.7	5.3	6.0	7.5	4.1

TABLE 9.14 A Set of 20 Hypothetical Ensemble Forecasts, of Ensemble Size 5, and Corresponding Observations—cont'd

Case	Member 1	Member 2	Member 3	Member 4	Member 5	Observation
10	2.7	6.6	5.8	7.5	5.1	8.3
11	6.6	5.2	5.3	5.5	3.2	4.7
12	6.7	6.0	8.6	7.7	4.8	8.7
13	8.9	1.3	5.9	7.3	6.3	8.5
14	8.5	5.0	4.6	7.6	1.4	4.8
15	9.2	4.4	8.9	5.3	6.5	9.5
16	2.7	8.7	3.4	7.6	5.1	4.3
17	4.1	7.0	7.5	7.2	7.0	5.4
18	7.7	4.7	5.7	5.7	6.8	2.1
19	6.7	7.4	6.2	5.3	5.8	3.3
20	4.4	3.3	1.9	5.4	6.6	7.4

9.15. For Case 1 in Table 9.14,
   a. Compute the eCRPS.
   b. Compute the Dawid-Sebastiani score.

9.16. For the hypothetical forecast and observed 500 mb fields in Figure 9.44,
   a. Calculate the S1 score, comparing the 24 pairs of gradients in the north-south and east-west directions.
   b. Calculate the MSE.
   c. Calculate the skill score for the MSE with respect to the climatological field.
   d. Calculate the centered AC.
   e. Calculate the uncentered AC.

9.17. Using the results from Exercise 9.1, construct the VS curve for the verification data in Table 9.2.

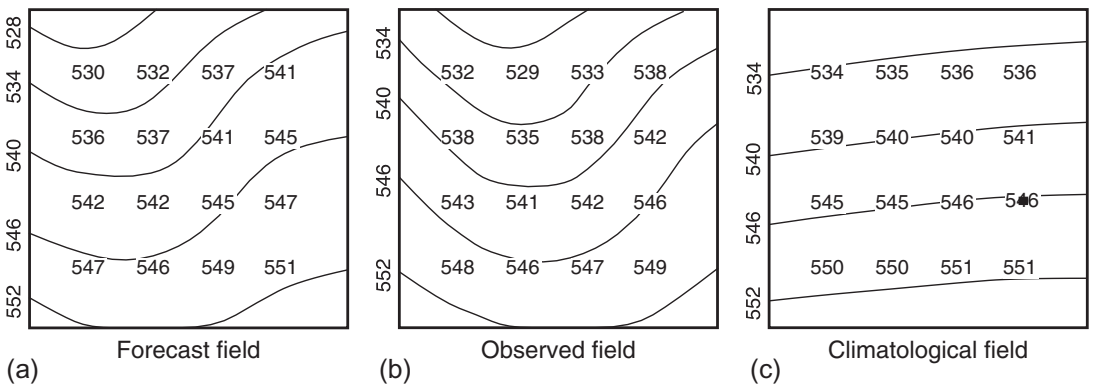

FIGURE 9.44 Hypothetical forecast (a), observed (b), and climatological average (c) fields of 500 mb heights (dam) over a small domain, and interpolations onto 16-point grids.