

CHAPTER 10

HOW DO WE KNOW WHETHER SEASONAL CLIMATE FORECASTS ARE ANY GOOD?

SIMON J. MASON

International Research Institute for Climate and Society

DAVID B. STEPHENSON

University of Exeter

When seasonal climate forecasts are expressed probabilistically, it is not possible to answer simple questions such as “how often are the forecasts correct?” The simpler attributes of forecast quality, such as “accuracy” or “correctness” are not applicable to probabilistic forecasts, and instead the main attributes of interest are: reliability, which defines whether the confidence communicated in the forecasts is appropriate; resolution, which defines whether there is any usable information in the forecasts; discrimination, which defines whether the forecasts are discernibly different given different outcomes (somewhat similar to the attribute of resolution); and sharpness, which defines the level of confidence that is communicated in the forecasts (regardless of whether that level is appropriate). How these attributes are measured depends on how the forecasts are expressed. In this chapter these attributes are explained in detail, and representation by various graphical procedures and scoring metrics is described. Partly because there is more than one desirable attribute to good probabilistic forecasts, it is argued that there is no single scoring metric that can adequately summarise forecast quality, and that in many cases graphical procedures also hide important aspects of forecast quality. The aim in this chapter is to provide some guidelines for interpreting and recognising the strengths and limitations of the most important verification tools as applied to seasonal climate forecasts.

10.1 Introduction

Forecast verification is an essential part of atmospheric science: the science of meteorology is ultimately judged by the skill of its predictions. Forecast verification is a multi-disciplinary area of research that requires careful summary and interpretation of pairs of past forecasts and observations. A comprehensive overview of forecast verification is presented in Jolliffe and Stephenson (2003); only a brief summary of issues can be presented here, and so the focus is on topics that are not discussed at length there. Specifically, because of the probabilistic nature of seasonal climate forecasts, this chapter considers only the verification of probability forecasts and of ensembles of forecasts more generally.

The chapter only considers procedures for indicating the quality of forecasts as opposed to their value; “forecast quality” is concerned with how well the forecasts match the observations, whereas “forecast value” is concerned with the benefit (whether economic, social, or otherwise) that can be realised through decisions made in response to the forecasts. In focussing on questions of quality, the *potential* for forecasts to have value is addressed; whether the forecasts can actually be used to realise that value raises questions about the impact of the climate conditions that verify, and about the options available for mitigating such impacts. Using even the simplest of decision-making models it can be demonstrated that forecasts with high quality can have negative value. For example, one such model, namely the cost-loss model, posits a specific “loss” resulting from the occurrence of adverse climate conditions, and a specific “cost” that can be incurred to mitigate these costs entirely if action is taken in advance. Given a set of forecasts and observations, it is possible to compare the costs and losses that would be incurred with and without forecasts. Despite its oversimplicity, the model is useful in demonstrating that seasonal forecasts can have value only under certain conditions: the relative costs of taking some actions compared to the losses mitigated can result in dis-benefit, even with high quality forecasts. Readers interested in procedures for estimating forecast value should consult the book by Katz and Murphy (1997).

A primary theme of the current chapter is that just as forecast quality is a necessary, but not sufficient, condition for forecasts to have value, so also individual attributes of forecast quality are necessary but not sufficient for “good” forecasts. In the following section the complex nature of forecast verification is indicated. The impossibility of summarising the quality of a set of forecasts by a single number is emphasised; because of the multifaceted nature of forecast quality, any single metric inevitably hides important information about the quality of the forecasts. Some graphical procedures are detailed (section 10.2.1) that provide a more comprehensive indication

of quality than is possible using scores. Nevertheless, for good and bad reasons, scores remain popular, and since there are large numbers of verification scores that have been proposed, the properties of such scores for probability forecasts are considered in section 10.3 so as to provide criteria for identifying which scores may be preferable to others. In section 10.4, some examples of commonly used verification scores for probability forecasts are examined.

10.2 Attributes of good probability forecasts

Perhaps the single most commonly asked verification question is “How often are the forecasts correct?” Although this question has intuitive appeal, when forecasts are presented as probabilities, questions about the “correctness” or otherwise of forecasts become unanswerable. Instead, probability forecasts are assessed on the basis of whether they reliably indicate changes in the uncertainty of the outcome: the forecasts are considered “reliable” when the forecast probability is an accurate estimation of the relative frequency of the predicted outcome (Murphy 1993).

Reliability, however, is not the only attribute of probability forecasts that is important. If the climatological probability of an outcome can be estimated accurately in advance, a set of forecasts that always indicate the climatological probability will be reliable, but will not provide any indication of the changing likelihood of the outcome from case to case. A second attribute, namely that of “resolution”, is therefore important. Probability forecasts have good resolution when they can successfully distinguish cases in which the probability of an event is high from those in which the probability is low. Forecasts with good resolution will have varying probabilities from forecast to forecast, and the more these probabilities diverge from the climatological probability, the sharper the forecasts are said to be. From an alternative perspective, if forecasts are good, the discrimination between the forecasts will be clearly defined given different outcomes.

Good probability forecasts will have good reliability as well as high resolution (and, implicitly, high sharpness), and will be well-discriminated. How these various attributes are measured depends to a large extent on the format of the probability forecasts. In the following sections the definitions of these attributes are considered in more detail. In the following discussions various scores are mentioned that aim to measure only a specific attribute of the quality of a set of forecasts. In each case, with the exception of the ROC area (section 10.2.3), these scores are distinct from scores that attempt to provide an overall summary of forecast quality. Discussion about the summary scores is reserved until section 10.4.

10.2.1 RELIABILITY

10.2.1.a Definition

As discussed in Chapter 8 (section 8.5.2), one objective in generating an ensemble of forecasts is to obtain an indication of the uncertainty in a forecast. However, it cannot automatically be assumed that the distribution of the ensemble members reliably indicates the true uncertainty: a decrease in the variance of the ensemble members does not necessarily mean that the outcome has become less uncertain. If the implied uncertainty in the forecasts is appropriate, the forecasts are said to be reliable or well-calibrated. Specifically, reliability is defined as consistency between the a priori predicted probabilities of an event and the a posteriori observed relative frequencies of this event. Reliability is measured in different ways depending on how the uncertainty in the forecast is indicated (see Chapter 9, section 9.2.3 for an introduction to the different ways in which probability forecasts can be expressed).

10.2.1.b Reliability of interval forecasts

Reliability is calculated most simply when forecast uncertainty is indicated using prediction intervals. In this case the forecast confidence is kept fixed, and so reliability can be assessed by comparing the coverage probability (sometimes called “capture rate”: the proportion of times the observed value is contained within the prediction interval) with the confidence level for the intervals. If the observed value falls too infrequently (or frequently) within the range defined by the prediction intervals then the forecasts are over-confident (under-confident).

To illustrate, two sets of forecasts of the December values of the Niño3.4 index for 1981–2000 are shown in Table 10.1. The forecasts were obtained by simple linear regression using either the June or the September values of the index as predictors. The models were trained using data for 1951–1980. Prediction intervals were calculated based on the cross-validated error variance over the training period (Chapter 7, section 7.3.3), and the widths of the intervals were set to define a 50% level of confidence (i.e. 50% of the intervals are expected to contain the observation). For both sets of forecasts, eight of the 20 years (40%) are contained within the prediction intervals. The intervals are therefore too narrow, and the forecasts are thus over-confident.

Although they have intuitive appeal, there are a number of problems with using coverage probabilities as measures of forecast quality. The first problem is that this measure of reliability does not distinguish between sets of predictions with similar coverage probabilities but different interval

Years	Obs	June	September
1981	-0.105	-0.391 (-0.807 to 0.025)	-0.191 (-0.494 to 0.112)
1982	2.590	2.019 (1.567 to 2.472)	1.998 (1.673 to 2.323)
1983	-0.464	1.343 (0.911 to 1.775)	0.044 (-0.258 to 0.347)
1984	-1.238	-0.811 (-1.231 to -0.390)	-0.017 (-0.319 to 0.286)
1985	-0.212	-0.725 (-1.144 to -0.305)	-0.253 (-0.556 to 0.050)
1986	1.261	0.283 (-0.133 to 0.699)	1.225 (0.914 to 1.536)
1987	1.167	2.218 (1.758 to 2.678)	2.421 (2.086 to 2.756)
1988	-1.892	-1.969 (-2.418 to -1.520)	-1.121 (-1.431 to -0.812)
1989	0.094	-0.812 (-1.233 to -0.391)	-0.207 (-0.510 to 0.095)
1990	0.491	0.323 (-0.093 to 0.740)	0.496 (0.192 to 0.800)
1991	1.756	1.569 (1.131 to 2.007)	0.769 (0.463 to 1.075)
1992	0.399	1.342 (0.910 to 1.775)	0.343 (0.040 to 0.647)
1993	0.371	1.341 (0.908 to 1.773)	0.742 (0.436 to 1.048)
1994	1.272	0.807 (0.386 to 1.229)	0.903 (0.595 to 1.210)
1995	-0.785	0.249 (-0.167 to 0.665)	-0.451 (-0.755 to -0.148)
1996	-0.394	-0.127 (-0.542 to 0.268)	-0.155 (-0.458 to 0.147)
1997	2.629	2.272 (1.810 to 2.734)	2.955 (2.605 to 3.305)
1998	-1.366	-0.419 (-0.836 to -0.003)	-0.644 (-0.949 to -0.339)
1999	-1.408	-1.011 (-1.435 to -0.587)	-0.838 (-1.144 to -0.532)
2000	-0.695	-0.547 (-0.965 to -0.129)	-0.309 (-0.612 to -0.006)

Table 10.1 – Observed values and forecasts of the December 1981–2000 Niño3.4 index. The upper and lower 50% prediction intervals are indicated, and intervals that capture the observed value are shaded.

widths. For example, both sets of forecasts in Table 10.1 have equal reliability, but the forecasts from September have consistently narrower intervals than those from June, and so are more informative (the narrower intervals imply less uncertainty in the forecast). A related problem is that the correct coverage probability, p say, can be achieved by unskilful forecasts simply by making the prediction interval infinitely wide $p\%$ of the time, and infinitely narrow the remaining times. These problems point to the impossibility of adequately representing forecast quality by a single score. More specifically, reliability is a necessary but not a sufficient attribute of a good set of forecasts (Murphy 1991).

10.2.1.c Reliability of probabilities for categories

When forecasts are communicated as a variable probability assigned to a predefined category, reliability is effectively defined in the same way as for the prediction intervals: forecasts are reliable if the observation falls within

the category as frequently as the forecast implies. The “observed relative frequency” (equivalent to the “coverage probability” for interval forecasts), has to be calculated for each distinct value of the forecast probability. For example, seasonal rainfall totals should be between 100 and 200 mm on 20% of the occasions in which the forecast probability for this interval is 20%, and on 40% of the occasions in which the forecast probability for this interval is 40%, etc..

The observed relative frequencies conditional upon the forecast probability can be plotted as reliability or attributes diagrams⁷⁴. Although the diagrams are designed to show the reliability of forecast probabilities for a specific event (i.e. for a two-category system), because the definition of an event does not have to remain fixed, forecasts for multiple categories can be included in the calculations⁷⁵. The interpretation of reliability diagrams may

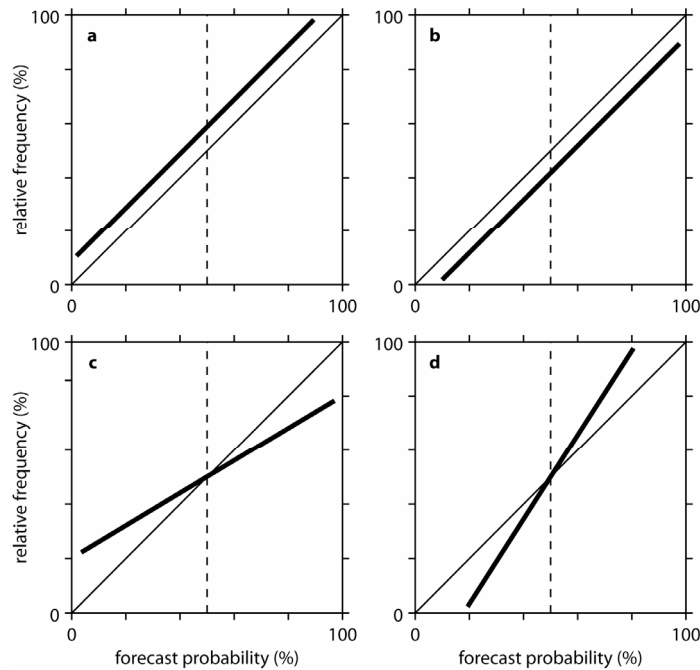


Figure 10.1 – Idealized reliability diagrams indicating cases of (a) under-forecasting, (b) over-forecasting, (c) over-confidence, (d) under-confidence. The vertical dotted line indicates the climatological probability of the event occurring, which in this case is set at 50%.

⁷⁴ The attributes diagram is the same as the reliability diagram, but includes an additional line to indicate where resolution equals reliability (see Hsu and Murphy 1986; Mason 2004).

⁷⁵ A separate multi-category reliability diagram showing the percentage of observations less than distinct percentiles of the forecast distribution has been proposed (Hamill 1997), but has not yet been widely adopted.

be facilitated by considering some idealised examples as shown in Figure 10.1. If the forecasts are perfectly reliable then the observed relative frequency will equal the forecast probability for all values of the forecast probability, and so the reliability curve will lie along the 45° diagonal. In practice, even if forecasts have excellent reliability, sampling errors result in departures from the diagonal, and so some indication of how close the curve is to the diagonal may be required to assist in the interpretation of the curve (Bröcker and Smith 2007; Kumar 2007).

More typically the forecasts are not perfectly reliable anyway, and so the curve will lie off the diagonal illustrating one or more of the characteristics shown in Figure 10.1. In Figure 10.1a the forecast probabilities are consistently lower than the observed relative frequencies, indicating that the event always occurs more frequently than anticipated. In Figure 10.1b the opposite is true, and the event occurs less frequently than anticipated. In these two cases the forecaster is under- / over-forecasting, respectively. For seasonal climate forecasts, the most common situation is indicated in Figure 10.1c. Here the event occurs more frequently than indicated when the forecast indicates a decreased probability of the event occurring compared to climatology (to the left of the dotted line), but less frequently than indicated when the forecast indicates an increased probability of the event occurring compared to climatology (to the right of the dotted line). Although the forecasts correctly indicate increases and decreases in the probabilities of the events, the changes in probability are over-stated, and the forecasts are said to be over-confident. The greater the degree of over-confidence, the shallower is the slope of the curve. If the curve becomes horizontal there is no information in the forecasts: the relative frequency of the event equals the climatological probability regardless of the forecast probability, and the forecasts are said to have no resolution (section 10.2.2). A fourth possibility is indicated in Figure 10.1d, where the changes in the forecast probabilities understate the changes in the relative frequencies of the event, and the forecasts are said to be under-confident. In this case the forecasts have high resolution, but poor reliability.

An example of a reliability diagram is shown in Figure 10.2. The diagram is based on 43 years of forecasts of Niño3.4 sea surface temperature (SST) anomalies, produced as part of the DEMETER project (Palmer et al. 2004). Forecasts from the ECMWF, Météo-France, and Met Office models for lead-times of 0–5 months and for four initialization seasons are included. For each model, nine ensemble members were available. The forecast probabilities were obtained by calculating the proportions of ensemble members forecasting temperatures in the coldest (black) and warmest (grey) 25% of years, respectively. For seasonal forecasts it is standard to bin the forecast probabilities into 11 categories, with the first bin

representing forecast probabilities of <5%, the second 5–15%, ..., and the last $\geq 95\%$ ⁷⁶. The frequencies with which forecasts in each bin occur are presented in a histogram. The reliability curves follow the 45° line closely indicating good reliability for forecasts of both anomalously warm and anomalously cold conditions. The histogram indicates that the forecast probabilities do not peak in frequency at the climatological probability of 25%, which is what would have been expected if the models had had little or no signal. Ideally, the forecasts should have high frequencies of probabilities close to 0% and 100%, whilst retaining reliability (i.e. the forecasts should be sharp), in which case the histogram would be *u*-shaped. However, the precise shape of the histogram of sharp forecasts depends on the climatological probability of the event (see section 10.2.4).

A measure of the distance between the sample reliability curve and the diagonal is an intuitive measure of forecast reliability. A commonly-used such metric is the reliability component of the Murphy (1973a) decomposition of the Brier score (section 10.4.1). Assuming that there are m points on the reliability curve, and that the forecast probability for the k th point is p_k , the reliability score is defined as:

$$\text{reliability score} = \frac{1}{n} \sum_{k=1}^m n_k (p_k - \bar{o}_k)^2, \quad (10.1)$$

where \bar{o}_k is the observed relative frequency for the k th probability bin, and n_k is the number of forecasts in this bin.

The distances defined by Eq. (10.1) represent the differences between the various forecast probabilities and the corresponding observed relative frequencies (the probabilities that should have been assigned), and thus are measures of the average “error” in the forecast probabilities. Therefore small values of Eq. (10.1) represent good reliability (see the discussion on necessarily mean that the forecasts contain useful information; perpetual forecasts of the climatological probability have perfect reliability. Therefore, as discussed in section 10.2.1.a, reliability is a necessary but not sufficient attribute of good forecasts.

Errors in calculating the observed relative frequencies for each forecast probability bin are binomially distributed with parameters n_k (the number of forecasts in bin k) and p_k (the average forecast probability for this bin).

⁷⁶ Since it is possible to tweak the binning to optimize the impression of reliability, the WMO has recommended the procedure as adopted here. These recommendations are detailed in the Standardized Verification System for Long-Range Forecasts (SVSLRF). Further details about the SVSLRF, which contains a list of recommended verification procedures, are available from the WMO Lead Centre for Verification: <http://www.bom.gov.au/wmo/lrfvs/>

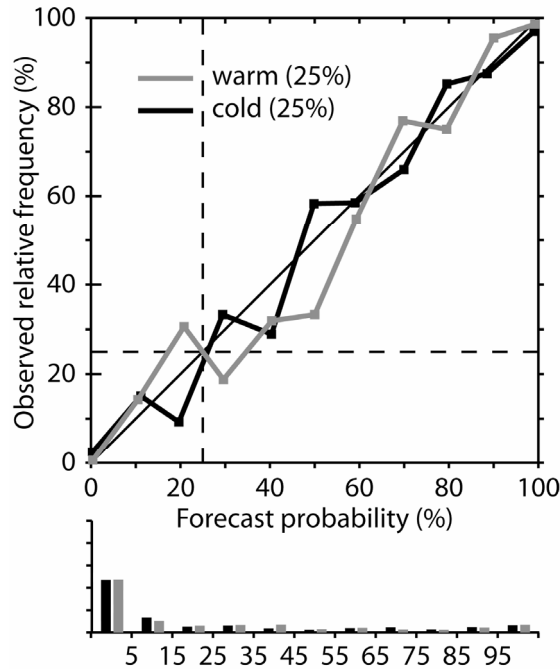


Figure 10.2 – Examples of reliability diagrams for ECMWF, Météo-France, and Met Office 0–5-month lead forecasts of Niño3.4 SST anomalies. The forecast probabilities were obtained by calculating the proportions of ensemble members forecasting temperatures in the coldest (black) and warmest (grey) 25% of years, respectively. The histogram indicates the frequency of forecasts of probabilities of <5%, 5–15%, etc..

Given the limited sample sizes of seasonal forecasts, the number of forecasts in a given bin can be too small to give a meaningful estimate of the observed relative frequency (Bröcker and Smith 2007). To increase the sample size, pooling of forecasts is necessary, whether from different lead-times or seasons (as in Figure 10.2), and or for different locations. Information about differences in the quality of the forecasts for the different lead-times, seasons, and locations is therefore masked, and there are good a priori reasons to expect the quality to differ (see Chapter 3). Pooling of forecasts for different locations is particularly problematic, not only because the forecast for proximate locations are unlikely to be independent (thus over-estimating the number of forecasts in each bin), but more specifically because it could be argued that the interpretation of the forecast probabilities is being changed. A (reliable) forecast probability of $p\%$ should imply that an event can be expected to occur on $p\%$ of the *occasions* a forecast with this probability is issued, but if forecasts are pooled for different locations the forecast is being verified with the interpretation that an event is expected to occur over $p\%$ of the *locations* at which a forecast with this

probability is issued. For all these reasons, reliability diagrams have not been used extensively for seasonal forecasts, although when sufficient forecasts are available the use of the diagrams is promoted in SVSLRF.

10.2.1.d Reliability of ensemble forecasts

Reliability diagrams are appropriate only for forecasts presented as probabilities of events (although the definition of an event does not have to remain fixed). A common method of indicating reliability when the forecast distribution is presented as percentiles is to use the ranked histogram (popularly called a Talagrand diagram). The ranked histogram is constructed by sorting the m ensemble members to form $m + 1$ bins, and then counting the numbers of times the observed value falls within each bin. If the forecast distribution reliably reproduces the distribution of possible outcomes then the observed value should be a random draw from this same distribution, and so should occur in each of the bins an equal number of times (Hamill 2001). The proportion of the total number of observations in each bin therefore should follow a uniform distribution, and a Cramér - von Mises test can be used to test for systematic errors (Elmore 2005).

Examples of ranked histograms are illustrated in Figure 10.3. The histograms were constructed using 50 years of model simulations of September–November rainfall for Brisbane (Figure 10.3a) and Kalgoorlie (Figure 10.3b), Australia⁷⁷. Nine ensemble members were considered, creating 10 bins, and so if the ensemble distribution is reliable the observed rainfall would be expected to occur in each bin 5 times. For Brisbane (Figure 10.3a), most of the observations are in the last bin, indicating that the observed rainfall is frequently more than the simulated rainfall of all nine ensemble members. The approximate upward slope of the histogram from left to right indicates that the simulated rainfall is negatively biased: the median observed rainfall over the 50-year period is about 180 mm compared to the median simulated rainfall of about 150 mm. In contrast, for Kalgoorlie (Figure 10.3b) the mean bias is minimal (41 mm observed, 44 mm simulated), but the numbers of times that the observed rainfall is either less than or more than all nine simulated values (the first and last bins) is inflated. Inflated frequencies in the outermost bins indicate that the observed rainfall falls outside the range of the ensemble distribution too frequently, and the binned histogram is said to be *u-shaped*. A *u-shaped* histogram is often interpreted to be indicative of

⁷⁷ The simulations are from the ECHAM4.5 atmospheric general circulation model (Roeckner et al. 1996), forced with observed SSTs. For this example the simulations for Brisbane and Kalgoorlie are taken simply as the model value for the grid containing 27.45°S, 153.03°E, and 30.78°S, 121.45°E, respectively.

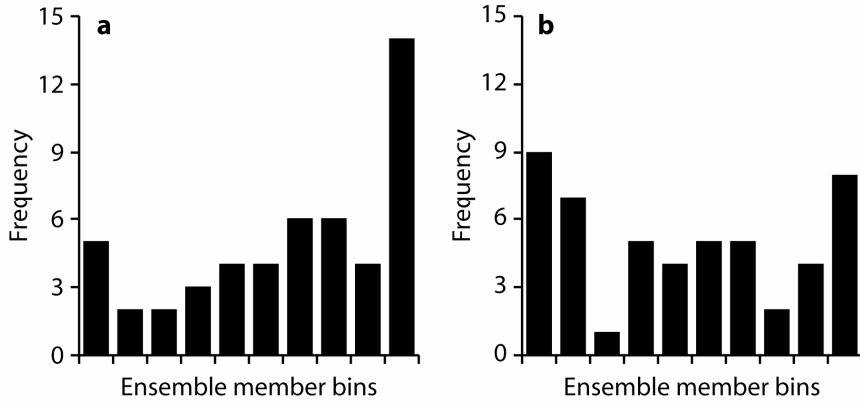


Figure 10.3 – Ranked histogram for ECHAM4.5 simulations of September–November 1951–2000 seasonal rainfall totals for (a) Brisbane and (b) Kalgoorlie, Australia.

an ensemble variance that is too small (increasing the variance of the ensemble distribution would decrease the proportion of observations outside the ensemble range), but can also be a result of conditional bias (Hamill 2001). There is no significant correlation between the ensemble mean and the observed rainfall for Kalgoorlie, and so when dry (wet) conditions are forecast, the observed rainfall is likely to be more (less) than all the ensemble members.

Since it is possible to construct a perfectly uniform ranked histogram from forecasts that do not have good resolution (Hamill 2001), a modification to the histogram has been proposed to test that the probability of each of the bins remains constant regardless of the forecast. This probability, known as the conditional exceedance probability (CEP), is defined as:

$$P(x > \hat{x}_k | \hat{x}_k) = \frac{1}{1 + \exp(-\beta_{0,k} - \beta_{1,k} \hat{x}_k)}, \quad (10.2)$$

where $\beta_{0,k}$ and $\beta_{1,k}$ are parameters to be estimated, x is the observed value, and \hat{x}_k is the k th percentile of the forecast distribution. The CEP is useful for measuring whether the probability of the observation exceeding the ensemble median, for example, increases if the ensemble forecasts are all indicating anomalously dry conditions. If this probability is conditional upon the actual forecast values then it is argued that the forecasts from the ensembles are unreliable even if, over all the forecasts, the ensemble median is exceeded 50% of the time (Mason et al. 2007).

10.2.1.e Multi-dimensional reliability

Ranked histograms and CEP diagrams both consider the prediction of a single parameter at a single point. Since dynamical models produce predictions of multiple parameters at multiple points, there may be interest in verifying the joint distributions of these predictands. For example, it is possible that a model could produce reasonable forecasts of precipitation and of temperature, but produce unrealistic simultaneous forecasts of these two parameters. Recent developments in verification methodology have begun to address the need to assess a model's ability to predict reliable joint distributions of parameters. Two such procedures are considered here: minimum spanning trees and bounding boxes.

Minimum spanning trees are a multi-dimensional adaptation of the ranked histogram (Wilks 2004). An example of a minimum spanning tree is shown in Figure 10.4a, showing the simulations (crosses) and observations (circle) of rainfall for Brisbane and Kalgoorlie for 2000. The tree is constructed by connecting each ensemble member with the nearest other member, and the total distance of all the connecting lines is then computed. This procedure is repeated replacing one of the ensemble members with the observed values, and thus treating the observation as if it were an ensemble member. The distances obtained using the outcome in place of each ensemble member are used to define the bins in the histogram. Then the distance obtained using all ensemble members without the outcome is binned. A histogram is constructed by repeating the procedure for all forecasts; as with the ranked histogram, if the outcome is indistinguishable from the ensemble members the minimum spanning distance for the tree constructed using all the ensemble members will be a random draw from the bins, and so the histogram should be level. In Figure 10.4b, the histogram shows too many distances in the first bin, indicating that the replacement of an ensemble member with the observed values typically increases the total spanning distance (i.e. the ensemble members are not a good representation of the outcome). Downward sloping histograms can result from mean biases, an ensemble spread that is too small, and/or conditional biases (cf. ranked histograms, for which only the mean bias results in a sloping histogram in either direction depending on the sign of the bias). As with ranked histograms, a uniform minimum spanning tree histogram

Bounding boxes are defined as the range of the ensemble predictions in k -dimensional space (Weisheimer et al. 2005). If the vector of observed values falls within the box then the outcome is interpreted as being consistent with the multidimensional distribution of the ensemble members, and thus is indistinguishable from an ensemble member. Bounding boxes are most commonly used with uncalibrated model output.

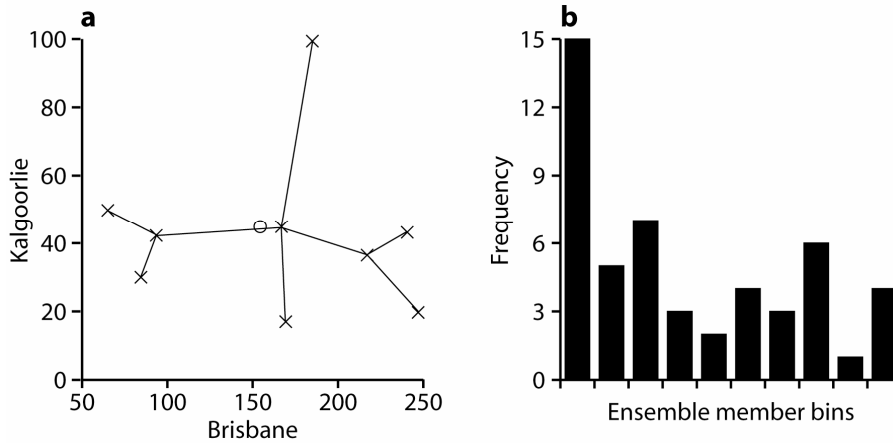


Figure 10.4 – (a) Minimum spanning tree for observed (circle) and ECHAM4.5 simulations (crosses) of September–November 2000 seasonal rainfall totals for Brisbane and Kalgoorlie, Australia. The tree is constructed for the case without the observation. (b) Minimum spanning tree histogram for 1951–2000.

10.2.2 RESOLUTION AND DISCRIMINATION

10.2.2.a Definition

It has been argued in the previous section that reliability is a necessary but not sufficient attribute of a good set of forecasts. Specifically, if the climatological probability of an event is known then a number of simple strategies can be devised to ensure that the forecasts are reliable, but which are otherwise uninformative. What is required in addition to reliability is an ability to distinguish between cases when the probability of an event is inflated from cases when the probability is deflated. More precisely, forecasts have good resolution when the outcome is strongly conditioned upon the forecast.

10.2.2.b Resolution given probabilities for categories

Resolution is most clearly defined in the context of forecasts expressed as probabilities for categories: if the forecasts have good resolution and good reliability then the probability of an event occurring should increase (or decrease) when the forecast probability increases (or decreases). In this context the most commonly used measure of forecast resolution is, like that for reliability, based upon the Murphy (1973a) decomposition of the Brier score, and is closely related to the reliability diagram (section 10.4.1). It is defined as:

$$\text{resolution score} = \frac{1}{n} \sum_{k=1}^m n_k (\bar{o}_k - \bar{o})^2, \quad (10.2)$$

where \bar{o} is the climatological probability of the event. Unlike the reliability score, the resolution score is not an error score, but can be interpreted as the weighted variance of the observed relative frequencies, with large variance representing good resolution (see section 10.3.1 on score orientation). Note that because of the squaring in Eq. (10.2) forecasts that have an increase in the observed relative frequency with a decrease in the forecast probability will score equally well on resolution as forecasts that indicate the correct direction of change in the probability of an event. As with reliability, therefore, resolution is not a sufficient attribute of good forecasts.

10.2.3 DISCRIMINATION

10.2.3.a Definition

The attribute of discrimination is similar to that of resolution, but considers the conditional distribution of the forecasts given the outcomes rather than of the outcomes given the forecasts. Whereas resolution is concerned with whether the expected outcome differs as the forecast changes, discrimination is concerned with whether the forecast differs given different outcomes. In the general framework for forecast verification introduced by Murphy and Winkler (1987), the first perspective is known as a calibration-refinement factorization, whereas the latter is called a likelihood-base rate factorization.

10.2.3.b Discrimination given probabilities for categories – the Relative Operating Characteristics (ROC)

The most commonly used method of identifying whether a set of forecasts is well-discriminated given different outcomes is the relative operating characteristics (ROC; sometimes called receiver operating characteristics) graph (Mason 2003). This procedure requires the outcome to be binary, just as in the case of a reliability diagram, and so separate results are usually calculated for each category if there are more than two categories. The rather horribly named ROC is actually equivalent to a non-parametric test commonly used for testing for differences in central tendency, namely the Mann-Whitney U -test⁷⁸ (Mason and Graham 2002). In the current context,

⁷⁸ More strictly, the Mann-Whitney U -test is used to test whether the probability that a sample from one population (e.g. a forecast for when rainfall is observed to be above-normal) has a value larger than that from another (e.g. a forecast rainfall is observed not to be above-normal) is 50%, but if assumptions are made about the distributions of the two

the *U*-test can be applied to assess whether there is any difference in (i.e. discrimination between) the forecasts when an event occurs compared to when the event does not occur. The ROC is most commonly applied to probabilistic forecasts, in which case it indicates whether the forecast probability was higher when an event occurred compared to when not, but it can as easily be applied to deterministic forecasts (in which case it indicates whether the forecast rainfall, for example, is higher when rainfall is above-normal than when not).

As a simple example, Table 10.2a contains 30 years of retroactive probabilistic forecasts of above-normal (defined as observed rainfall being above the upper-tercile) December–February total rainfall for Lusaka, Zambia, obtained using a simple statistical model⁷⁹. These forecasts can be ranked from most confident to least confident, and the ranks of the forecasts are shown in the third column. It seems reasonable to select the seasons with the highest probability (1973/74 and 1975/76, with forecasts of 65%) as the seasons in which one would be most confident about the observed rainfall being above-normal. Similarly, the season with the second highest probability (1975/76, with a forecast of 60%) would be the season in which one would be next most confident that observed rainfall was above-normal.

The table can be re-ordered so that the seasons are listed in order of the ranks of the forecasts rather than chronologically, as shown in Table 10.2b. The seasons for which one would be most confident observed rainfall was above-normal are then at the top of the table. If the forecasts are good, then the actual seasons in which rainfall was above-normal should be towards the top of the table. If the forecasts are effectively useless, the above-normal seasons will be randomly distributed through the table, and if they are bad these seasons will be towards the bottom of the table. The actual seasons of above-normal rainfall are marked by grey shading, and do appear preferentially to be towards the top of the table.

To construct the ROC, start at the top of the table and, treating each forecast as a prediction of an event (i.e. above-normal rainfall), count the proportion of correct forecasts (the hit-rate) and incorrect forecasts (the false-alarm rate). These scores are shown in Table 10.2b: for the highest

populations (specifically, that they have similar shapes and variances) then the test can be used to compare the central tendencies (the medians) of the distributions (Sheskin 2007). These assumptions generally are irrelevant in the context of forecast verification.

⁷⁹ The data are based on the example in Chapter 7, section 7.3.3. The retroactive procedure used an initial 10-year training period (1961/62 – 1970/71) to forecast the next year, and was updated each year so that the last training period for forecasting 2000/01 was 39 years long. Forecast probabilities were obtained from the error variance of the cross-validated predictions (see Chapter 7, section 7.3.3), and then rounded to the nearest 5%.

a)

Year	Forecast (%)	Rank
1971/72	45	4
1972/73	10	29
1973/74	65	1
1974/75	40	9
1975/76	65	1
1976/77	30	19
1977/78	30	19
1978/79	45	4
1979/80	35	14
1980/81	35	14
1981/82	35	14
1982/83	20	23
1983/84	40	9
1984/85	40	9
1985/86	35	14
1986/87	20	23
1987/88	15	27
1988/89	55	3
1989/90	40	9
1990/91	25	22
1991/92	20	23
1992/93	30	19
1993/94	20	23
1994/95	15	27
1995/96	45	4
1996/97	35	14
1997/98	5	30
1998/99	45	4
1999/00	45	4
2000/01	40	9

b)

Rank	Year	Hit Rate	False-Alarm Rate
1	1973/74		
1	1975/76	1 of 10	1 of 20
3	1988/89	2 of 10	1 of 20
4	1971/72		
4	1978/79		
4	1995/96		
4	1998/99		
4	1999/00	4 of 10	4 of 20
9	1974/75		
9	1983/84		
9	1984/85		
9	1989/90		
9	2000/01	7 of 10	6 of 20
14	1979/80		
14	1980/81		
14	1981/82		
14	1985/86		
14	1996/97	9 of 10	9 of 20
19	1976/77		
19	1977/78		
19	1992/93	10 of 10	11 of 20
22	1990/91	10 of 10	12 of 20
23	1982/83		
23	1986/87		
23	1991/92		
23	1993/94	10 of 10	16 of 20
27	1987/88		
27	1994/95	10 of 10	18 of 20
29	1972/73	10 of 10	19 of 20
30	1997/98	10 of 10	20 of 20

Table 10.2 – (a) Forecast probabilities and ranks (in descending order) for above-normal December–February 1981/82–2000/01 seasonal rainfall totals for Lusaka, Zambia, made using a simple statistical model. (b) Forecasts shown in order of descending rank, with corresponding hit and false-alarm rates. The ‘events’ (observed rainfall above the upper tercile) are indicated by grey shading.

ranking forecasts, rainfall was above-normal in only one of the years and so one of the ten above-normal events were correctly identified, and one of the twenty non-events. The next highest ranking forecast (1988/89) was for a season that was above-normal, and so now two of the ten events have been correctly selected. Effectively Table 10.2b involves constructing a series of contingency tables in which forecasts of an event are issued using successively lower warning thresholds, t . Initially a warning is issued when the forecast probability is at least 65%, then the threshold is lowered to 55%, etc.. So, at each point on the ROC curve, the hit and false-alarm rates, $HR(t)$ and $FR(t)$ respectively, are calculated as:

$$HR(t) = P(p \geq t | x = 1), \quad (10.3)$$

$$FR(t) = P(p \geq t | x = 0), \quad (10.4)$$

where p is the forecast probability, and $x = 1$ if the event occurs, and $x = 0$ otherwise.

The ROC graph is constructed by plotting the hit rates against the false-alarm rates. The graph for the example is shown in Figure 10.5. The diagonal line on the graph indicates the line of no-skill. If the events were uniformly distributed through the table, the hit and false-alarm rates would accumulate at approximately the same rate, and so the ROC curve would follow the diagonal line. However, if the forecasts are good, the hit rate will accumulate faster, and so the graph will curve towards the upper left. In the extreme case of perfect discrimination, the curve will reach the top left corner. The example shows that the forecasts are well-discriminated.

Noting that the procedure for constructing the ROC graph is based only on the ranks of the forecasts, it should be evident that any monotonic transformation of the forecasts will not affect the graph at all. For example, if the forecast probabilities for all forecasts were increased (or decreased) by 10%, the graph would be unaffected. Alternatively if the forecast probabilities for all forecasts above 50% were increased by a fixed amount, and those below were decreased, the graph would again be unaffected. This insensitivity has been cited as a criticism of this verification procedure since the reliability of the forecast probabilities is ignored. The message is that a good ROC graph does not necessarily imply that the forecasts are well-calibrated.

The area beneath the ROC graph is increasingly used as a measure of discrimination, partly because of the inclusion of the ROC as a recommended verification procedure in the SVSLRF. For forecasts of no skill, for which the ROC curve lies along the diagonal, the area would be 0.5, and the maximum area of perfect discrimination is 1.0. The area is related to the

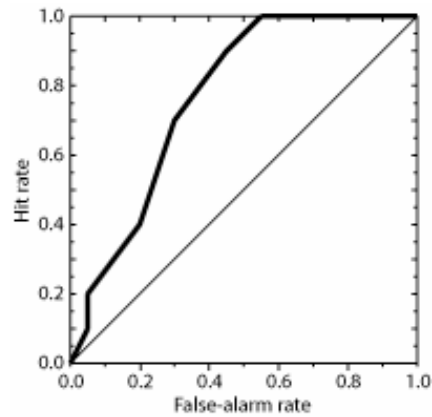


Figure 10.5 – ROC diagram for the retroactive forecasts of December–February 1981/82–2000/01 seasonal rainfall totals for Lusaka, Zambia.

U -statistic by a factor of the numbers of events and non-events, and can be interpreted as the probability of successfully distinguishing an event from a non-event given a forecast of each (Mason and Graham 2002). For the example in Figure 10.5, the area is 0.7675, implying that there is a greater than 75% probability that the forecasts can successfully discriminate an above-normal season from other seasons. The area is sometimes criticised as a summary measure of forecast performance because of its insensitivity to reliability. However, the ROC graph has an advantage over the reliability diagram in being less sensitive to sampling errors, and so can be more meaningfully constructed given the small sample sizes typical of seasonal forecasting.

10.2.4 SHARPNESS

10.2.4.a Definition

Resolution, in the sense defined above, together with reliability, incorporate the idea of “sharpness”. Although there is no formally recognised mathematical definition of sharpness⁸⁰, the general concept is usually clear: sharpness refers to the degree to which the forecasts depart from the climatology. If forecasts are expressed as intervals, sharp forecasts are indicated by narrow intervals; if as probabilities of categories, sharp forecasts are expressed as probabilities that differ from the climatological probability, and

⁸⁰ The variance of the forecasts around the climatological probability is sometimes used to define sharpness, although arguably this definition makes sense only if the climatological probability is 0.5.

are close to 0% or 100%; if as a forecast distribution, sharp forecasts are indicated as narrow distributions. Sharp forecasts imply high confidence (see Chapter 8, section 8.5.1), but do not necessarily imply good forecasts; as with reliability, sharpness is a necessary but not sufficient condition for high forecast quality.

Unfortunately, after appropriate recalibration (see Chapters 8 and 9) the sharpness of seasonal forecasts is typically much weaker than that of weather forecasts because of the large inherent uncertainty in predicting seasonal climate. In the extreme case of no predictability, the forecast probability should always be equal to the climatological probability.

A specific question of interest that is concerned with the sharpness of forecasts pertains to the ability of an ensemble to indicate changes in the uncertainty in the forecast. More specifically, does a sharper forecast mean that uncertainty is reduced? This question has received considerable attention in the context of forecast ensembles (including multi-model ensembles) where there is interest in the case-to-case variability in the ensemble spread: does this variability in the ensemble distribution contain any useful information? In other words, can one be more confident that the observed value will be close to the ensemble mean when the ensemble spread is small compared to when it is large? Two general approaches to the question have been adopted, both of which can be fraught with difficulties. One approach involves seeking a relationship between some measure of the spread in the ensemble, and some measure of accuracy in the central tendency of the ensemble distribution. These procedures are discussed in next section. The second approach attempts to measure the quality of the forecasts when the ensemble distribution is considered explicitly and to compare this with the quality when the variability in the ensemble distribution is ignored (section 10.2.4.c).

10.2.4.b Accuracy⁸¹–spread relationships

A common approach to the question of determining the information content of an ensemble distribution is to identify whether there is any relationship between the accuracy of the forecast, as measured by the “error” in the ensemble mean, and the uncertainty in the forecast, as measured by the ensemble variance (the “spread”). The theory behind this approach is that a larger ensemble spread implies greater uncertainty in the forecast, and hence larger errors in the ensemble mean can be expected. However, this theory is often based on a misconception of any accuracy-spread relation-

⁸¹ The term “skill” is often confusingly used instead of accuracy. Although “skill” is often used in a more generic sense than the definition provided in the glossary, in the current context it invariably refers to “accuracy”, and so the latter term is preferred here.

ship that may exist. Assuming that the ensemble distribution is a reliable indicator of the true distribution of possible outcomes, the expected error is zero regardless of the uncertainty; it is the variance of the errors that should increase with increasing ensemble spread, not the expected error, as indicated in Figure 10.6. This misconception is not adequately resolved by defining the error in absolute terms (or by squaring the errors), and any standard form of regression between forecast error and some measure of forecast spread is poorly designed to identify any relationship that may exist.

10.2.4.c Skill of the ensemble spread

Given the form of the relationship between accuracy and spread as indicated in Figure 10.6, it is more helpful to reconstitute the problem as identifying whether there is any useful information in the case-to-case variability in the ensemble spread (or, more generally, the ensemble distribution⁸²). An approach that offers more promise than seeking accuracy-spread relationships when sample sizes are small is to calculate whether the performance metric of the forecasts improves if the information in the spread of the ensemble distribution is considered compared to if the ensemble distribution is kept fixed. It is inferred that the variability in the ensemble spread does provide meaningful indications of changes in uncertainty if the measure of forecast quality is highest for the forecasts with varying ensemble spread.

There are numerous ways of implementing such a procedure. Perhaps the simplest is to assess the quality of the ensemble forecasts when using a counting procedure to obtain forecast probabilities (Chapter 8, section 8.5.2), and then to reassess the quality after reducing the ensemble to the ensemble mean so that the forecast probabilities are either 0% or 100%. Although such an approach is unfair because most of the information in the ensemble mean is lost by converting it to a binary forecast, it has been used occasionally, most commonly when the ROC area is the verification metric of choice. Because of the unfair treatment of the ensemble mean as a single-member ensemble, such results are heavily biased in favour of finding useful information in the ensemble spread. A fairer approach using the ROC is to calculate the area based on the ranks of the ensemble means. However, the results can then be biased against finding information in the ensemble distribution because of the unsatisfactory calculation of probabilities by counting for the ensemble.

⁸² For the sake of economy of phrase, in the rest of this section the term “spread” is assumed to incorporate changes in shape as well as variance.

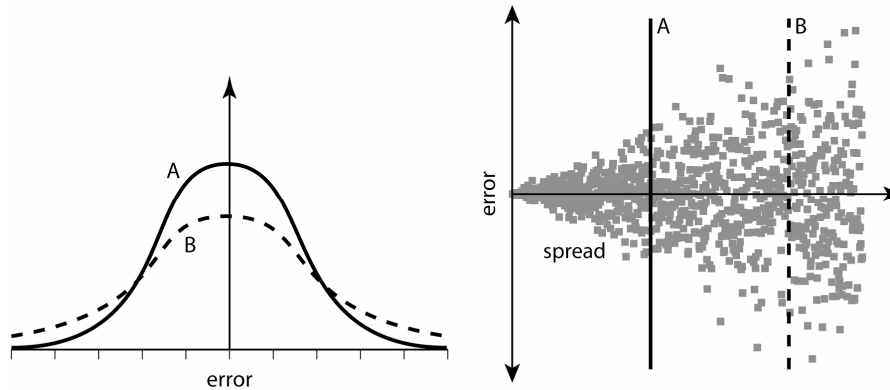


Figure 10.6 – (a) Distributions of forecast errors given high confidence forecasts with small ensemble spread (A) and low confidence forecasts with large ensemble (B); and (b) a hypothetical sample of ensemble mean error and ensemble spread measurements for a continuous range of spread values. The distributions in (a) are drawn from the points on the x-axis marked in (b).

A more satisfactory procedure is to use a distribution-fitting approach to obtain probability forecasts from the ensemble (Chapter 8, section 8.5.2). The question at hand can then be reformulated to: is there any information in the ensemble beyond the first moment of its distribution? The forecast probabilities are first obtained by fitting a distribution firstly with a fixed variance (or a variance that is a function only of the mean in the case of meteorological parameters with skewed distributions, such as precipitation), and secondly with a variance defined by the ensemble variance. If the forecast distribution does contain useful information the score should improve by allowing the variance to vary, and it can then be inferred that the ensemble provides useful information about the uncertainty in the forecast. It is thus possible to decompose the skill of the ensemble into signal and indications of uncertainty.

As an example, consider simulations of Brisbane September–November total rainfall for the 50-year period 1951–2000. The simulations are based on output from the ECHAM4.5 atmospheric general circulation model forced with observed SSTs. Fitting a gamma distribution to the 85 ensemble members, probabilities for the rainfall less than the lower tercile were calculated. The gamma distribution was fitted using a fixed and a varying shape parameter. Compared to climatology, a Brier skill score (see section 10.4.1) of 0.081 was achieved given a fixed shape parameter, and of 0.060 given a varying shape parameter, indicating positive skill in both cases. However, the probabilities given the varying shape parameter scored slightly worse, and the negative skill score of -0.022 when measured against the fixed

shape parameter indicates that there is no useful information in the ensemble shape.

10.3 Properties of summary measures for probability forecasts

Although it has been argued throughout section 10.2 that forecast quality cannot be represented adequately by a single metric because there is more than one important attribute of good probability forecasts, sometimes it is desirable to quantify the forecast performance in a summary measure. Metrics that measure the individual attributes have been discussed briefly, but in all cases were found to be problematic as summary measures of overall quality: specifically, while a bad score does indicate bad forecasts, a good score does not necessarily indicate good forecasts. Before considering examples of summary measures of forecast quality it is therefore of value to consider the desirable properties such scores should have.

There is a wide range of performance scores, partly reflecting the need for different verification methods depending on how the probability forecasts are issued. Scores for cases in which there are only two categories are discussed in detail in section 10.4.1, while scores for multiple categories are considered in section 10.4.2. There are a few measures that apply to forecast distributions on a continuous scale, and these are covered in section 10.4.3. Very few studies have addressed verification methods for probability forecasts of count data, interval and quantile forecasts, and probability forecasts of spatial distributions; these are areas of probability forecasting that merit more attention.

10.3.1 SCORE ORIENTATION AND SKILL SCORES

Before discussing the desirable properties of verification scores (section 10.3.2), it is helpful to distinguish between positively and negatively oriented scores. Good forecasts achieve a high score if the score is positively oriented, but a low score if it is negatively oriented. Negatively oriented scores frequently are some measure of the error in the score: if the forecasts are good the errors will be small, and so the score will be low. Positively oriented scores, however, give credit to good forecasts, and so a high score is desirable. In this chapter all scores are presented in their negatively oriented versions unless indicated otherwise.

Skill scores are positively oriented scores with specific characteristics: they compare the quality of one set of forecasts with that of a second set, known as the reference forecasts (sometimes the second set is implied), and equal zero if the quality of the two sets of forecasts is identical. A com-

only used formula for deriving a skill score, SS , from a positively oriented score is:

$$SS = \frac{S - S_{ref}}{S_{per} - S_{ref}}, \quad (10.5a)$$

where S is the score for the forecasts in question, S_{ref} is the score for the reference forecasts, and S_{per} is the score for a perfect set of forecasts. For a negatively oriented score, $S_{per} = 0$, and so Eq. (10.5a) reduces to:

$$SS = 1 - \frac{S}{S_{ref}}, \quad (10.5b)$$

(Murphy 1973b). Skill scores usually have a maximum value of one (or 100%), when the first set of forecasts perfectly outperforms the reference set, but their lower limit depends on the score for the reference forecasts, which can make the interpretation and comparison of negative skill scores complicated. Using Eq. (10.5), a positive skill score can be interpreted as the fractional improvement in the score for the forecasts against the score for the reference forecasts. An alternative formulation for a skill score is

$$SS = S - S_{ref}, \quad (10.6a)$$

for positively oriented scores, and

$$SS = S_{ref} - S, \quad (10.6b)$$

for negatively oriented scores. Using Eq. (10.6), a positive skill score can be interpreted as the amount of improvement in the score for the forecasts against the score for the reference forecasts.

Because a skill score is a relative score, the interpretation of the score depends upon the choice of the reference forecasts. Commonly used reference strategies include climatology and persistence, but these strategies are not necessarily equally unskilful. For example, when forecasting SSTs, the slowly evolving nature of the oceans makes persistence a much harder standard to beat for short lead-time forecasts than climatology (sections 3.4, in Chapter 3, and 5.4, in Chapter 5). In addition, because scores necessarily over-simplify the complex nature of forecast verification, a negative skill score against climatology and / or persistence does not automatically imply that the forecasts do not contain any useful information. This information may have been lost by the score (Mason 2004).

10.3.2 DESIRABLE PROPERTIES OF PROBABILISTIC FORECAST VERIFICATION SCORES

Given the wide range of scores available for assessing the quality of probabilistic forecasts, it is helpful to define a set of criteria that can be used to identify which scores may be the most appropriate. Three criteria are considered: propriety, equitability, and locality (Murphy 1993).

10.3.2.a Propriety

An important concept in verification is whether or not a score can be improved by hedging of the forecasts. The Oxford English Dictionary defines hedging as “the securing of, or limiting the possible loss on, a debt, bet, or the like”. In the context of forecast verification, “hedging” occurs when a forecaster issues a forecast different to what (s)he truly believes. Certain non-mathematical targets of performance can encourage a forecaster to issue a forecast that is inconsistent with his/her true belief: for example, not wishing to cause excessive alarm. However, some verification scores can also be hedged: the forecaster is encouraged to modify the forecast in order to improve the expected value of the score. Hedging is undesirable because it encourages the forecaster to issue a forecast that may be inconsistent with his/her true beliefs simply in order to achieve a better score, and so it is best to choose scores that cannot be improved by forecasting anything other than the forecaster’s true beliefs.

A *strictly proper score* is a probability score, S , for which the forecaster uniquely optimises the expected score by forecasting his/her true beliefs. So if the forecaster believes an event occurs with probability q then the expected score should be minimised when the forecast probability actually issued, p , equals q . The score will be minimised if there is a unique stationary point at which

$$\left. \frac{\partial}{\partial p} S(p, q) \right|_q = 0 \text{ at } p = q. \quad (10.7)$$

A score is proper, but not strictly proper, if Eq. (10.7) is true for more than one value of q . Unfortunately, most skill scores defined using Eq. (10.5) are not strictly proper unless the categories are equiprobable, and/or unless the forecaster has absolute certainty about the outcome (Murphy 1973b; Gneiting and Raftery 2007). Skill scores defined using Eq. (10.6) may therefore be preferable, although further research is required to investigate their properties.

10.3.2.b *Equitability*

Another property of scores that sometimes is considered desirable is *equitability*; a score is *equitable* if it takes the same value for all unskilful forecasts that have no association with the observations (i.e. forecasts that have no resolution). In the context of probabilistic forecasts there are a variety of forecast strategies that can be adopted that have no resolution (e.g. random forecasts, and perpetual forecasts of constant probabilities, including of climatological probabilities). Although it may seem desirable that these various naïve strategies should score equally badly, it is impossible for a probabilistic score to be equitable and strictly proper, and the latter property is to be preferred (Jolliffe and Stephenson 2007). The differences in the scores of differing no-resolution forecast strategies can be attributed to differences in their reliability.

10.3.2.c *Locality*

A skill score is local if it depends only on the probability assigned to the outcome. The desirability of this property of verification scores is disputed: two main arguments are presented against locality, but both arguments are inconclusive. The first argument is that non-local scores can be less sensitive to the categorization of the observed values than local scores; the more categories that are used the lower the score tends to be (Daan 1985). However, one could argue that a forecast system with many categories attempts to communicate more information than a system with only a few, and so a greater degree of precision is required. Another argument against locality is that it seems reasonable to account for “distance” (i.e. to credit forecasts that assign high probability close to the observed value as well as to the outcome itself). For, example, given two forecasts for three ordinal categories, A {0.2, 0.3, 0.5} and B {0.1, 0.4, 0.5}, forecast B would seem a better score if the third category verified since, while both forecasts assign the same probability to this category, forecast B assigns a higher probability to the adjacent category. Forecast B has more probability close to the verification, and so seems intuitively better than forecast A. Implicit in such reasoning is the assumption that because category 3 verified, category 1 was less likely to have occurred than category 2. However, we know only that category 3 occurred, and do not know what the relative probabilities of the other categories were. In fact, we do not even know that category 3 was the most likely. The reliability of the probabilities of all three categories can only be assessed by considering the categories individually. From one perspective, then, locality seems to be a desirable property, although if it is accepted as such, there are a number of non-local scores that are widely used, including the ranked probability score (RPS; see section 10.4.2).

10.4 Summary measures for probability forecasts

10.4.1 SOME SCORES FOR BINARY EVENTS

By far the most commonly used summary measure of the quality of probability forecasts of binary events is the Brier score⁸³. The Brier score is analogous to the mean-squared error, but is defined in terms of the “error” in the probabilities rather than in the actual units of the observations, and is calculated using:

$$\text{Brier score} = \frac{1}{n} \sum_{i=1}^n (p_i - o_i)^2, \quad (10.8)$$

where o_i is 1 if the event occurs in the i th case, or 0 otherwise. Murphy (1973a) defined a well-used algebraic decomposition of the Brier score consisting of reliability, resolution, and uncertainty. The first two terms have been discussed in sections 10.2.1 and 10.2.2, respectively, where their relationships to the reliability diagram were explained. The uncertainty term, defined as

$$\text{uncertainty} = \bar{o}(1 - \bar{o}), \quad (10.9)$$

is independent of the forecasts, but because it depends on the relative frequencies of the observed events, it can complicate comparison of scores for different sets of verification data if these relative frequencies are not constant.

The skill score version of the Brier score [using Eq. (10.5)] is negatively biased when climatology is used as a reference strategy, so that forecast quality looks worse than it really is. For ensemble forecasts this bias is related to sampling errors in calculating the forecast probabilities (Müller et al. 2005), and can be corrected by adding an additional uncertainty term that accounts for these sampling errors. Without the additional uncertainty term the imperfectly estimated probabilities for the forecasts are compared to perfectly estimated (and therefore perfectly reliable) climatological probabilities, making for an unfair comparison. The correction term, d , is calculated as

⁸³ Strictly, the Brier score, being a special case of the probability score (section 10.4.2) for when there are only two categories, is defined both for the event and for the non-event categories. Because the definition in Eq. (10.8) applies only to the event, it is more correctly called the half-Brier score. However, since the contribution from the non-event category is the same [$(p_i - o_i)^2 = ((1 - p_i) - (1 - o_i))^2$], it is redundant to score them both, and for the sake of simplicity the Brier score is widely calculated using Eq. (10.8). Throughout this chapter, the phrase “Brier score” refers to the half-Brier score, unless specified otherwise.

$$d = \frac{\bar{o}(1-\bar{o})}{m}, \tag{10.10}$$

where m is the number of ensemble members (Weigel et al. 2007). The debiased skill score is calculate using

$$SS = 1 - \frac{S}{S_{ref} + d}, \tag{10.11}$$

and since S_{ref} reduces to Eq. (10.9) for climatological forecasts, Eq. (10.11) simplifies to

$$SS = 1 - \frac{S}{d(m+1)}. \tag{10.12}$$

As an alternative score, just as the mean absolute error is sometimes used in place of the mean squared error, a mean absolute score for probability forecasts has been proposed:

$$\text{mean absolute score} = \frac{1}{n} \sum_{i=1}^n |p_i - o_i|. \tag{10.13}$$

A third score is the logarithmic score, defined as:

$$\text{logarithmic score} = \frac{1}{n} \sum_{i=1}^n [-(1-o_i) \log(1-p_i) - o_i \log p_i], \tag{10.14}$$

which is the average of the negative of the logarithm of the probability assigned to the verifying category. Since the logarithm of 1 is 0, Eq. (10.14) is an “error” score, like Eqs. (10.8) and (10.13).

The Brier, mean absolute, and logarithmic scores can be generalised as the mean error over all cases:

$$\bar{S} = \frac{1}{n} \sum_{i=1}^n S(p_i, o_i). \tag{10.13}$$

The Brier score and mean absolute scores are quadratic and linear “error” or loss functions, whereas the logarithmic score is the negative log likelihood for a set of n independent Bernoulli events. As loss functions, the scores are negatively oriented in that smaller scores imply more skilful forecasts. The loss function for the above three scores is symmetric since $S(p,0) = S(1-p,1)$ and so they are each completely defined by just the single loss function $S(p,0)$. This loss function is given by p^2 , p , and $-\log(1-p)$ for the Brier, absolute, and logarithmic scores, respectively.

The expected value of the score for an event that has probability q of occurrence (sometimes confusingly referred to as the “true probability”, and hereafter referred to as the “event probability”) is given by:

$$E(S(p, X)) = (1 - q)S(p, 0) + qS(p, 1) = S(p, q). \quad (10.14)$$

Consistent with the scores being minimized when the forecast probability equals the event probability, the minimum values occur where $p = q$ (Jolliffe and Stephenson 2007). Note, however, that the score is a function of the event probability, q , and is smallest when $q = 0$ and 1 (i.e. when the observed event is least uncertain)⁸⁴. The Brier and logarithmic scores have a saddle point (maxima of a valley and minima of a ridge) at $p = q = 0.5$ whereas the mean absolute score has a flat ridge at $q = 0.5$. This saddle point defines the unique stationary point required for a score to be strictly proper [Eq. (10.7)], thus implying that the Brier and logarithmic scores are proper scores, but the absolute score is not (if the forecaster thinks that the event probability is 0.5, it does not matter what probability (s)he issues since the expected score will be 0.5 regardless of the probability issued).

The logarithmic score penalises much more heavily poor forecasts different from $p = q$ than does either the Brier or the absolute score. The penalty becomes infinitely large when a probability of 0% is assigned to an event that does happen, or when a probability of 100% is assigned to an event that does not happen. This apparently excessive penalty can be justified on the basis that the implied odds of the actual outcome were infinitely small.

10.4.2 SCORES FOR MULTI-CATEGORY FORECASTS

The probability score is defined as the average over n forecasts of the sum of squared probability “errors” for each category, j , of m categories:

$$\text{probability score} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m (p_{ij} - o_{ij})^2. \quad (10.15)$$

The full Brier score is a special case of the probability score for when there are two categories. The probability score considers the probability assigned to all categories, and so it does not have the property of locality. Despite being non-local, since there is no implicit ordering in the categories, the probability score does not account for distance, and for this reason it is not widely used. Because of the failure to account for distance, coupled with the lack of locality, the score has some undesirable properties. Con-

⁸⁴ In the case of the Brier score, this dependence on the observed probability is reflected by the uncertainty term in Murphy’s (1973a) decomposition of the score.

sider, for example, the two forecasts $A=\{0.45, 0.55, 0.00\}$ and $B=\{0.40, 0.30, 0.30\}$. If the first category verifies, the score for B (0.5400) is less than (and hence better than) for A (0.6050), despite the fact that A issues a higher probability to the outcome, and a higher probability to the adjacent category. A simple modification to Eq. (10.15) reduces it to the half-Brier score [Eq. (10.8)] and thus would resolve these problems:

$$\text{quadratic score} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m o_{ij} (p_{ij} - o_{ij})^2. \quad (10.16)$$

A commonly used alternative to the probability score is the ranked probability score (RPS), which addresses the problem of lack of effectiveness by considering distance. Instead of comparing the probabilities assigned to each category with that of a perfect deterministic forecast, the RPS compares the cumulative probabilities:

$$\text{ranked probability score} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m (P_{ij} - O_{ij})^2. \quad (10.17)$$

where P_{ij} is the cumulative probability assigned to category j in the i th forecast, and O_{ij} is 1 if the i th observation is in any of the categories less than or equal to j , and is 0 otherwise [cf. Eq. (10.8)]. So, the example forecasts A and B above, would be expressed as $A=\{0.45, 1.00, 1.00\}$ and $B=\{0.40, 0.70, 1.00\}$. Similarly, for the observations, the cumulative probabilities given a perfectly deterministic forecast are used (i.e. $\{1.00, 1.00, 1.00\}$, $\{0.00, 1.00, 1.00\}$, and $\{0.00, 0.00, 1.00\}$ if the first, second, and third categories were to verify, respectively). The RPS for A (0.3025), given that category 1 verifies, is now less than for forecast B (0.4500), and so the RPS correctly identifies A as the better forecast. Despite being strictly proper, the RPS, by definition, does not have the property of locality.

The skill score version [Eq. (10.5)] of RPS is biased when climatology is used as the reference strategy for the same reasons as with the Brier score. The skill score can be debiased using Eq. (10.11), but d is defined as

$$d = \frac{k^2 - 1}{6mk}, \quad (10.18)$$

where k is the number of categories, as long as the categories are equiprobable. See Weigel et al. (2007) for corrections to unequal categories.

In the previous section, the linear probability score was rejected because it is not a strictly proper score. The score can be generalized for cases when there are more than two categories [in a similar way to Eq. (10.15)], but it still suffers from the same problem of lack of propriety as its two-category version.

The logarithmic score is defined as the average of the negative of the logarithm of the probability assigned to the verifying category, and so can be interpreted as a measure of probability “error”:

$$\text{logarithmic score} = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m -o_{ij} \log [p_{ij}]. \quad (10.19)$$

Just as with the binary version of the score, it penalizes large forecast errors much more severely than any of the other scores, and in the extreme case of a zero probability being assigned to a verifying event, the logarithmic score is infinite. Although not yet widely used, the logarithmic score has all the desirable properties of verification scores (ignoring equitability), and can be generalized to cases of continuous forecasts and still retain all properties (see section 10.4.3).

10.4.3 SCORES FOR CONTINUOUS FORECASTS

Probabilistic scores for continuous forecasts are not widely used for seasonal climate forecasts, partly because full forecast distributions are rarely specified, and also partly because options for scoring such forecasts have not been discussed much in the climate literature. There is a version of the linear error in probability space (LEPS) suitable for probabilistic forecasts (Ward and Folland 1991). The LEPS score was derived to measure the error in a forecast in terms of distance measured by the climatological cumulative distribution rather than in terms of the original units of the forecast (so, for example, a forecast error of 1°C for a normally distributed variable would be penalised much more heavily if the forecast and observed value are close to the mean than if near the tails of the distribution). The version of the score for continuous probabilistic forecasts lacks propriety, and so its use should be discouraged (Mason and Mimmack 2002).

A preferable option is the continuous version of the ranked probability score (CRPS):

$$\text{continuous ranked probability score} = \frac{1}{n} \sum_{i=1}^n \int_{-\infty}^{\infty} [F_i(z) - O_i(z)]^2 dz, \quad (10.20)$$

where $F_i(z)$ is the cumulative forecast probability for the i th forecast, and $O_i(z)$ is 0 if the i th observation is less than z , and is 1 otherwise [cf. Eq. (10.17)]. The score describes the average of the squared difference between the forecast and observed cumulative distributions, where the observed cumulative distribution is a step function represented by the cumulative distribution of a perfectly accurate deterministic forecast; it is calculated by

integrating the squares of the vertical distances between the two curves⁸⁵, as represented by the grey shaded areas in see Figure 10.7. Note that the squaring in Eq. (10.20) is along the probability axis not along the x -axis, and so the score reduces to the mean absolute error if the forecasts are deterministic (Hersbach 2000).

The linear error score defined in Eq. (10.13) has been generalised for ensemble forecasts (Wilson et al. 1999), but the score lacks propriety. However, it can be adjusted so that it is strictly proper:

$$\text{proper linear score} = \frac{1}{n} \sum_{i=1}^n \left[\int_{-\infty}^{\infty} f_i^2(z) dz - 2f_i(x) \right], \quad (10.21)$$

where $f_i(x)$ is the forecast probability density at the observed value, x (Bröcker and Smith 2007). The integral in Eq. (10.21) renders the score proper, and is a representation of the sharpness of the forecasts, but does make the score lack locality. The only score that has the propriety, and the locality properties is the continuous version of the logarithmic score (Roulston and Smith 2002):

$$\text{logarithmic score} = \frac{1}{n} \sum_{i=1}^n -\log[f_i]. \quad (10.22)$$

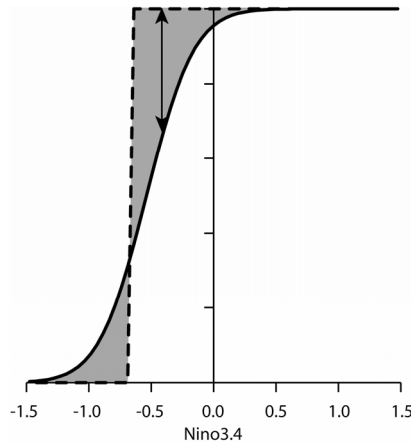


Figure 10.7 – Example of a forecast for the December 2000 Niño3.4 index (using the June predictor from Table 10.1) showing the “error” (grey shading) in the cumulative forecast probability, as measured by the squared distance between the forecast distribution (solid line) and the distribution for a perfectly accurate deterministic forecast (dashed line). The integral of the squared vertical distances between the two curves (shown by the arrow at an index value of -0.5) is the contribution to the continuous ranked probability score.

⁸⁵ Compare the Kolmogorov-Smirnov test, discussed in Chapter 8, which is based on the largest vertical distance between two such cumulative distributions.

10.5 Summary

Forecast verification, or the evaluation of the quality of a set of forecasts, is a multifaceted problem, and so there is no single metric that can comprehensively represent the quality of the forecasts. The problem is complicated in the case of probability forecasts because naïve attributes of forecast quality, such as “accuracy”, are inappropriate – the question of whether a specific probability forecast is correct or incorrect is unanswerable. Specially designed verification procedures are therefore required for probability forecasts, but there are several different types of probability forecast, and each requires its own methods for verification. The main attributes of interest are:

- reliability, whether the confidence communicated in the forecasts is appropriate;
- resolution, whether there is any usable information in the forecasts;
- discrimination, whether the forecasts are discernibly different given different outcomes (somewhat similar to the attribute of resolution);
- sharpness, the level of confidence that is communicated in the forecasts (regardless of whether that level is appropriate).

The most commonly used graphical procedures for indicating forecast quality are the reliability diagram and accompanying frequency histogram (which together indicate reliability, resolution, and sharpness), and the relative operating characteristics graph (which indicates discrimination). Numerous summary measures of forecast quality have been defined, and so in choosing between them it is helpful to define a set of properties that verification scores should have. Perhaps the most important property is that only those verification scores should be used that cannot be improved by hedging the forecasts. Scores for probability forecasts that have this property are called proper scores. Unfortunately, the score that is arguably the easiest to interpret, namely the mean absolute probability error, is not strictly proper. Squared and logarithmic scoring rules are therefore generally preferred, although the logarithmic score is the only one that can be generalised to forecast of continuous probability distributions.