7 Probability and Ensemble Forecasts

ZOLTAN TOTH[1], OLIVIER TALAGRAND[2], GUILLEM CANDILLE[2] AND YUEJIAN ZHU[1]
[1]SAIC at National Centers for Environmental Prediction, Camp Springs, MD, USA
[2]Laboratoire de Météorologie Dynamique, Paris cedex, France

7.1 INTRODUCTION

The previous chapters have focused on verification procedures for environmental predictions given in the form of a single value (out of a continuum) or a discrete category. This chapter is devoted to the verification of probabilistic forecasts, typically issued for an interval or a category. Probabilistic forecasts differ from the previously discussed form of predictions in that, depending on the expected likelihood of forecast events, they assign a probability value between 0 and 1 to possible future states.

It is well known that all environmental forecasts are associated with uncertainty and that the amount of uncertainty can be situation dependent. Through the use of probabilities the level of uncertainty associated with a given forecast can be properly conveyed. Probabilistic forecasts can be generated through different methods. By considering a wide range of forecast information, forecasters can subjectively prepare probabilistic forecasts. Alternatively, statistical (empirical) techniques can be used either on their own, based on historical observational data (e.g., Mason and Mimmack 2002; Chatfield 2001), or in combination with a single dynamical model forecast and its past verification statistics (e.g., Atger 2001).

Probabilistic forecasts can also be based on a set of deterministic forecasts valid at the same time. Assuming the forecasts are independent realizations of the same underlying random process, an estimate of the forecast probability of an event is provided by the fraction of the forecasts predicting the event among all forecasts considered. This technique, known as *ensemble forecasting* (see Leith 1974; Ehrendorfer 1997; Stephenson and Doblas-Reyes 2000; and references therein), can produce probabilistic forecasts based on a set of deterministic forecasts, without relying on past verification

Forecast Verification: A Practitioner's Guide in Atmospheric Science. Edited by I.T. Jolliffe and D.B. Stephenson ISBN: 0–471–49759–2

statistics. In certain fields of environmental science, such as meteorology and hydrology, the ensemble forecasting technique is now becoming widely used. Therefore, this chapter will also present some of the methods that have been developed to directly evaluate a set of ensemble forecasts, before they are interpreted in probabilistic terms.

In our analysis, the expectation taken over all available realizations of a probabilistic forecast system will be denoted by the operator $E(\cdot)$, whereas the conditional expectation of a quantity B over the subset of all values of A satisfying a condition C will be denoted by $E_A(B|C)$. Note that in this chapter $\hat{p}$ will be used interchangeably to denote the forecast probability density function (p.d.f.) of a continuous variable as well as the forecast probability distribution (mass function) of a discrete variable. A more precise notation would be to use $\hat{f}(\cdot)$ for the forecast p.d.f. of a continuous variable, $\hat{F}(x)$ for the forecast cumulative distribution function (c.d.f.) of a continuous variable, and $\hat{p}(x_i)$ for the forecast probability distribution of a discrete variable.

The next section (Section 7.2) is devoted to a discussion of the two most important attributes of probabilistic forecasts referred to as 'reliability' and 'resolution'. Sections 7.3 and 7.4 will introduce a set of basic verification statistics that can be used to measure the performance of probabilistic forecasts for binary and multi-outcome events with respect to these attributes. Similar verification statistics are presented in Section 7.5 for probabilistic forecasts for continuous variables, while some measures of ensemble performance are introduced in Section 7.6. Many of these forecast verification measures will be illustrated with recent meteorological applications. Some limitations to probabilistic and ensemble verification are discussed in Section 7.7, while the concluding remarks are made in Section 7.8. Further background on the probability scores to be discussed in this chapter can be found in the reviews by Murphy and Winkler (1987), Murphy and Daan (1985), Stanski *et al.* (1989) and Wilks (1995).

7.2 MAIN ATTRIBUTES OF PROBABILISTIC FORECASTS

How can one objectively evaluate the quality of probabilistic forecasts? Let us consider the following prediction: 'There is a 40 % probability that it will rain tomorrow'. Assuming that the event 'rain' is defined unambiguously, it is clear that neither its occurrence nor its non-occurrence can be legitimately used to validate, or invalidate the prediction. This apparent lack of accountability in case of a single forecast is in contrast with categorical deterministic forecasts ('it will rain' or 'it will not rain'), which can be unambiguously validated, or invalidated for each individual event.

Whether a single forecast is valid or not tells little about the performance of a forecast system. If the goal is the evaluation of a forecast system, whether it is deterministic or probabilistic, one must use a statistical

approach, based on a sufficiently large set of cases. In the case of the probabilistic forecast example cited above, one must wait until the 40% probability forecast has been made a number of times, and then first check the proportion of occurrences when rain was observed. If that proportion is equal or close to 40%, one can legitimately claim the forecast to be statistically correct. If, on the contrary, the observed proportion is significantly different from 40%, the forecast is statistically inconsistent.

One condition for the validity of probabilistic forecasts for the occurrence of an event is therefore statistical *consistency* between *a priori* predicted probabilities and *a posteriori* observed frequencies of the occurrence of the event under consideration. Consistency of this kind is required, for instance, for users who want to make a decision on the basis of an objective quantitative risk assessment (see Chapter 8). Following Murphy (1973), this property of statistical consistency is called *reliability*. A forecast system is called reliable if it provides unbiased estimates of the observed frequencies associated with different forecast probability values. Note that the word *consistency* has several different meanings in verification (see Chapter 3, Section 3.3, for an alternative definition) and so it should be used carefully.

Reliability alone is not sufficient for a probabilistic forecast system to be useful. Consider the extreme situation where one would predict, as a form of probabilistic forecast for rain, the climatological frequency of occurrence of rain. The forecast system would be reliable in the sense that has just been defined, since the observed frequency of rain would be equal to the (unique) predicted probability of occurrence. However, the system would not provide any forecast information beyond climatology. It follows that reliability in itself says nothing about whether the forecasts are able to discriminate in advance between situations that lead to different verifying observed events. As a second condition, a forecast system must be able to distinguish among situations under which an event occurs with lower or higher than climatological frequency values. After Murphy (1973), the ability of a forecast system to *a priori* separate cases when the event under consideration occurs more or less frequently than the climatological frequency is called *resolution*. The better it separates cases when an event in the future occurs or not, and gets it right, the more resolution a forecast system has. Interestingly, it is a perfect deterministic forecast system that achieves maximum resolution, indicating that deterministic forecasts can be considered as a special case of probabilistic forecasts, with only the 0 and 1 probability values used.

What has just been described for probabilistic prediction of occurrence of events easily extends to all other forms of probabilistic forecasting. Consider a real-valued continuous predictand x (for instance, temperature at a given time and location), and the corresponding forecast p.d.f., $\hat{p}(x)$, represented by the full curve in Fig. 7.1. An example of a subsequent verifying observation value is shown by x_0 in Fig. 7.1. Note that both the forecast p.d.f. $\hat{p}(x)$ and the verifying observation x_0 are different for each individual

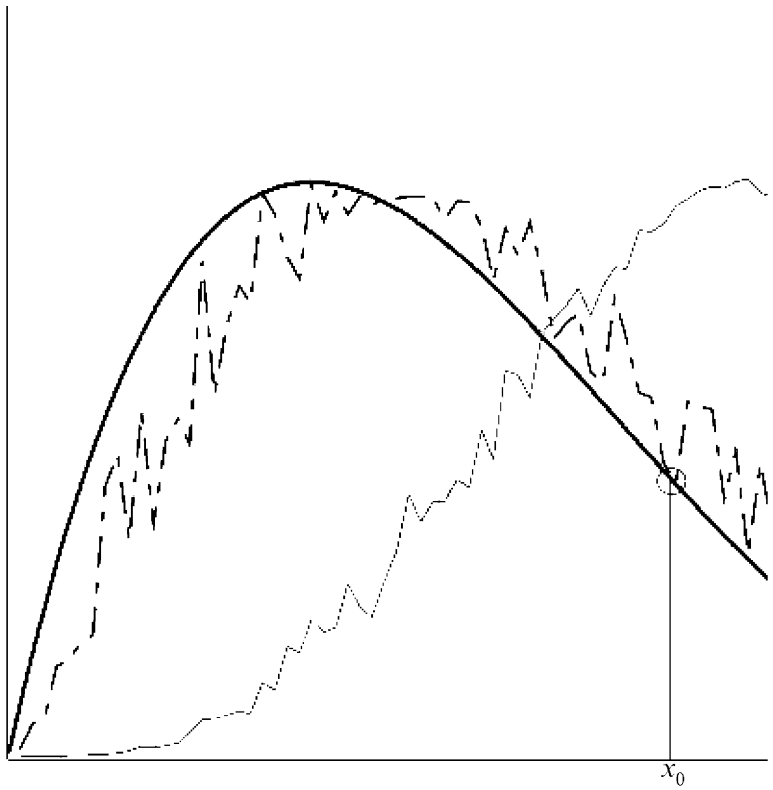

Figure 7.1 A hypothetical forecast probability density function $p(x)$ (full curve) for a one-dimensional variable x, along with a verifying observed value x_0 for a single case. The additional two curves represent possible distributions for the verifying values observed over a large number of cases when $p(x)$ was forecast by two probabilistic forecast systems. The distribution $p_1(x)$ (dash-dotted curve) is close to the forecast distribution $p(x)$, while the distribution $p_2(x)$ (dashed curve) is distinctly different from $p(x)$ (see text for discussion)

forecast time (and so implicitly include time t as a labeling index). If, as is the case in Fig. 7.1, x_0 falls within a range where the forecast probability density is non-zero, the observation can neither validate nor invalidate the forecast. The difficulty here is that, contrary to what happens with a single-value forecast (see Chapter 5), it is not obvious how to define, in a trivial way, a 'distance' score between the forecast p.d.f. $\hat{p}(x)$ and the single observed value x_0. A potential distance is provided by the $\hat{F}(x_0)$ measure used to evaluate the reliability of density forecasts in macroeconomics (see Chapter 9, Section 9.3.2).

A probabilistic (or any other) forecast system, as pointed out above, can be validated only in a statistical sense. Therefore, similarly to the case of probabilistic forecasts for a given event discussed above (There is a 40 % probability that it will rain tomorrow), one has to wait until a particular probability distribution $\hat{p}(x)$ has been predicted a number of times. Let

us denote $p_o(x)$ the frequency distribution of observations corresponding to the cases when $\hat{p}(x)$ is forecast. If $p_o(x)$ is similar to $\hat{p}(x)$ (as is shown by the dash-dotted curve in Fig. 7.1), then the prediction $\hat{p}(x)$ can be described as statistically consistent with observations. If, however, $p_o(x)$ is distinctly different from $\hat{p}(x)$ (as is the case for the distribution $p_2(x)$ shown by the dashed curve in Fig. 7.1), then the forecast $\hat{p}(x)$ is statistically *inconsistent* with observations.

This example calls for a more precise definition of reliability. A probability forecasting system is *reliable* if, and only if, the conditional probability distribution $p(x_0|\hat{p} = q)$ of the verifying observations given *any* chosen forecast probability distribution $q(x)$ is itself equal to $q(x)$ (i.e., $p(x_0|\hat{p} = q) = q(x)$ for all possible $q(x)$). In other words, the p.d.f. of the observed value, when compiled over (stratified on) the cases when the forecast probability density equalled $q(x)$, is exactly equal to $q(x)$. This definition of reliability can also be extended to multi-dimensional and any other type of probabilistic forecasts.

As noted earlier, reliability, albeit necessary, is not sufficient for the practical utility of a probabilistic forecast system. Systematic prediction of the climatological distribution of a meteorological variable is reliable yet provides no added forecast value. Probability forecasts should be able to reliably distinguish among situations for which the probability distributions of the corresponding verifying observations are distinctly different. Such a system can 'resolve' the forecast problem in a probabilistic sense, and is said to have resolution. Similarly to the case of binary events, the more distinct the observed frequency distributions for various forecast situations are from the full climatological distribution, the more resolution the forecast system has. Also maximum resolution is obtained when reliable forecast probability distributions have zero spread, i.e., they are concentrated on single points as Dirac delta functions. Again, such a probabilistic forecast system generating perfectly reliable forecasts at maximum resolution is a perfect deterministic forecast system.

Given a large enough sample of past forecasts, reliability of a forecast system can be improved by a simple statistical calibration that relabels the forecast probability values. For example, assume that the forecast distribution $\hat{p}(x) = q(x)$ is associated with a distinctly different distribution of observations $p_2(x)$ (dashed curve in Fig. 7.1), i.e. $p(x_0|\hat{p} = q) = p_2(x_0)$. The next time the system predicts $\hat{p}(x) = q(x)$, one can use previous knowledge to substitute $p_2(x)$ as the forecast, i.e., use the calibrated forecast $\hat{p}' = p(x_0|\hat{p} = q_2)$ instead of the original forecast $\hat{p}$. This *a posteriori* calibration will make a forecast system reliable. For statistically stationary forecast and observed systems, perfect reliability can always be achieved, at least in principle, by such an *a posteriori* calibration given a large enough sample of past forecasts.

The two main attributes of probabilistic forecasts, *reliability* and *resolution*, are a function of both the forecasts and the verifying observations.

Resolution was defined above as the variability of the observed frequency distributions associated with different forecast scenarios around the climatological p.d.f. Another property, *sharpness*, measures the variability of the *forecast* (and not the corresponding observed) probability distributions around the climatological p.d.f. Note that in a perfectly reliable (well calibrated) forecast system the forecast probability values, by definition, are identical to the corresponding frequency of verifying observations. For a reliable forecast system sharpness is therefore identical to resolution.

Since it is only a function of the forecast (and not the corresponding observed) distributions, sharpness is not a verification measure. It follows that in general an arbitrary increase in sharpness (e.g., an increase in the highest forecast probability values) will not lead to enhanced resolution. Resolution cannot be improved through a simple adjustment of probability values – it can only be improved by a clearer discrimination of situations where the event considered is more or less likely to occur as compared to the climatological expectation. This suggests that the intrinsic value of forecast systems lies not in their reliability (that can be improved by a calibration procedure described above) but in the resolution that cannot be improved by simply post-processing forecast probability values.

In summary, resolution and reliability together determine the usefulness of probabilistic forecast systems. Assuming they behave stationarily in time (no long-term changes in their behavior), there seems to be no desirable property of probabilstic forecast systems other than reliability and resolution. A useful forecast system must be able to *a priori* separate cases into groups with as different future outcome as possible, so as each forecast group is associated with a distinct distribution of verifying observations. This is the most important attribute of a forecast system and is called resolution. The other important attribute, reliability pertains to the proper labeling of the different groups of cases identified by the forecast system. It was pointed out that even if the forecast groups originally were designated improperly by the forecast system, they could be rendered reliable by 'renaming' them according to the observed frequency distributions associated with each forecast group, based on a long series of past forecasts. The different scores that are introduced in the rest of this chapter for the evaluation of binary, multi-categorical, and continuous variable probabilistic forecasts and ensembles will be systematically analyzed to assess which of the two main forecast attributes (resolution and reliability) they measure.

7.3. PROBABILITY FORECASTS OF BINARY EVENTS

In Sections 7.3.1–7.3.3, we will consider verification methods for the simplest conceptual case of probability forecasts of binary events. Such events, marked by A, can be defined in different ways. One can use an inequality of the form $\{A: X > u\}$, where X is a scalar variable for which a probabilistic

forecast is made, and u is a given threshold value. Examples of this type include the occurrence or not of a particular binary event A such as 'the temperature at a given location x at forecast lead-time t will be-greater than 0 °C', or 'the total amount of precipitation over a given area and a given period of time will be more than 50 mm'. Other events, like 'Tropical storm Emily will hit land', or 'Electric power distribution will be disrupted by thunderstorms', cannot be easily expressed in terms of a meteorological parameter exceeding a certain threshold, yet are equally interesting. This section is devoted to the verification of probabilistic forecasts of binary events, regardless of how they are defined.

By considering individual values/categories, probabilistic forecasts of discrete variables or categories having multiple values can also be considered as a set of probabilistic binary events. Probabilistic forecasts of binary events are, therefore, of fundamental importance in the verification of probability forecasts.

Let us introduce the binary random variable, X, that takes the value 1 when the event occurs (e.g., exceedance of a threshold value) and 0 when the event does not occur. Now consider the conditional probability $f(q) = p(X = 1|\hat{p} = q)$ for the probability of the event to occur given that the forecast probability was equal to q. In the special case of binary events, $f(q)$ is equal to the conditional expectation $E(X|\hat{p} = q)$. An (frequentist) estimate of $f(q)$ is easily obtained by counting the relative frequency of the observed event over cases when event A was forecast to occur with probability q. The condition for reliability, as defined in the previous section, is simply that $f(q) = q$ for all possible values of q.

7.3.1 The Reliability Curve

As an example, let us consider winter 1999 probabilistic forecasts produced by the National Centers for Environmental Prediction (NCEP) Ensemble Forecast System (Toth and Kalnay 1997). The event is defined here as the 850-hPa temperature being at least 4 °C below its climatological mean value. Diagnostics are accumulated over all grid-points located between longitudes 90W and 45E, and between latitudes 30N and 70N, and over 65 forecasts issued between 1 December 1998 and 28 February 1999, for a total of $n = 16,380$ realizations. Forecast probability values for the event at each grid-point are estimated by the relative frequencies, i/m, where $i = 0, 1, 2, \ldots, m$ indicate the number of members in the ensemble of $m = 16$ forecasts that predict the event to occur. The forecast probabilities are thus restricted to $m + 1 = 17$ equally spaced discrete values. For deciding whether the event occurred or not, the NCEP operational analysis will be used as the best estimate of truth.

The solid line in Fig. 7.2 shows the *reliability curve* obtained by plotting values of $f(q)$ against q, for the forecast system and event described above. Although rather close to the line $f(q) = q$ (i.e., perfect reliability), the

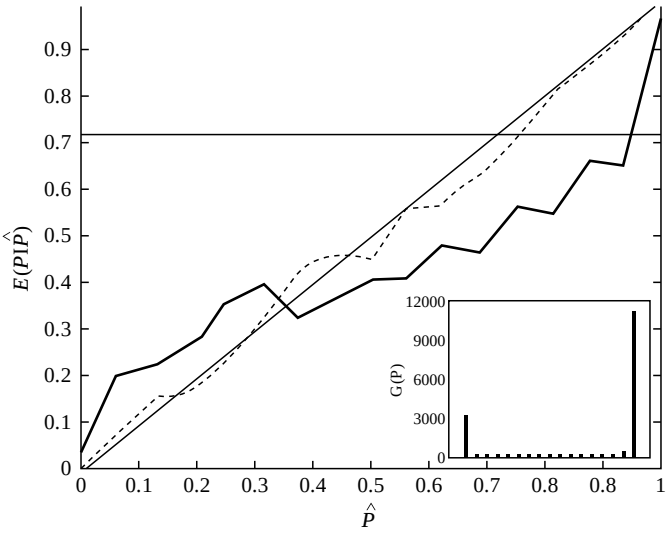

Figure 7.2 Reliability diagram for the NCEP Ensemble Forecast System (see text for the definition of the event E under consideration). Full line: reliability curve for the operational forecasts. Dash-dotted line: reliability curve for perfect ensemble forecasts where 'observation' is defined as one of the ensemble members. The horizontal line shows the climatological frequency of the event, $s = 0.714$. Insert in lower right: sharpness graph (see text for details)

reliability curve does show some significant deviations from it. In particular, the slope of the reliability curve in Fig. 7.2 is below that of the $f(q) = q$ diagonal. Note that deviations from the diagonal are not necessarily indicative of true deviations from reliability but can also be due to sampling variations. When statistics, as in our example, are based on a finite sample, the reliability curve for even a perfectly reliable forecast system is expected to exhibit sampling variations around the diagonal. The amount of sampling variability can be easily assessed by plotting reliability curves for the same forecast system except now using a randomly chosen member of the ensemble of forecasts in place of the verifying observations. By definition, the forecast system should be perfectly reliable in this case, and so deviations from the diagonal in this case are due to sampling variations. When compared to the diagonal, the difference between the perfect (dash-dotted line in Fig. 7.2) and operational ensemble curve (solid line) reflect the true lack of reliability in the forecast system, irrespective of the size of the verification sample. Bootstrap methods (see Efron and Tibshirani 1993) could easily be developed to quantify the sampling uncertainty in these estimates of reliability.

The histogram in the lower right corner of Fig. 7.2 is known as a sharpness diagram which shows the relative frequencies for the forecast probabilities, i.e., sample estimates of the marginal probability distribution of the forecast probabilities q. Since the probabilities of zero or one are used

in 90% of the forecast cases, the forecast system exhibits a considerable degree of sharpness, as defined above. This is due to the small spread in temperatures in these short-range 2-day lead-time forecasts, resulting in either none or all of the forecasts often falling 4 °C below the climatological mean value.

7.3.2 The Brier Score

Brier (1950) proposed the quadratic scoring measure $E[(\hat{p} - X)^2]$ for the quantitative evaluation of probabilistic binary forecasts. It can be estimated from a sample of past forecasts by

$$B = \frac{1}{n}\sum_{j=1}^{n}(\hat{p}_j - x_j)^2 \tag{7.1}$$

where n is the number of realizations of the forecast process over which the validation is performed. For each realization j,$\hat{p}_j$ is the forecast probability of the occurrence of the event, and x_j is a value equal to 1 or 0 depending on whether the event occurred or not. A minimum Brier score of zero is obtained for a perfect (deterministic) system in which $\hat{p}_j = x_j$ for all j. Such a system issues probability forecasts of 1(0) every time before the event is (not) observed to occur. Since such a forecast system does not use any probabilities *between* 0 and 1, it has no uncertain cases and can be considered as a deterministic binary forecast system. On the contrary, the Brier score takes the maximum value of one for a systematically erroneous (yet perfectly resolving) deterministic system that predicts with certainty the wrong event each time, i.e.,$\hat{p}_j = 1 - x_j$.

In order to compare the Brier score, B, to that for a reference forecast system, B_{ref}, it is convenient to define the Brier skill score (BSS):

$$\text{BSS} = 1 - \frac{B}{B_{\text{ref}}} \tag{7.2}$$

Unlike the Brier score in Eqn. (7.1), the BSS is *positively oriented* (i.e., higher values indicate better forecast performance). BSS is equal to 1 for a perfect deterministic system, and 0 (negative) for a system that performs like (poorer than) the reference system. The reference system is often taken to be the low-skill *climatological forecasts* in which $\hat{p}_j = s$ for all j, where $s = p(X = 1)$ is the base rate (climatological) probability for the occurrence of the event. The Brier score for such reference forecasts is equal to $B_c = s(1 - s)$ (in the large sample asymptotic limit). Climatological forecasts have perfect reliability since $E_x(x|\hat{p} = s) = s$, but have no resolution since $_{\hat{p}}(E_x(x|\hat{p} = s)) = (s) = 0$. In the rest of this chapter, the BSS will be defined using climatological forecasts as the reference, i.e., $B_{\text{ref}} = B_c = s(1 - s)$.

Because the Brier score is quadratic, it can be usefully decomposed into the sum of three individual parts related to reliability, resolution, and the underlying uncertainty of the observations (Murphy 1973). To derive this decomposition here, we will assume that the forecast probabilities can take any continuous value of q in the range 0 to 1. In other words, the predictor values q are continuous variables with a p.d.f. $p(q)$ defined such that

$$\int_0^1 p(q)\,\mathrm{d}q = 1 \tag{7.3}$$

In realistic forecast situations, only a discrete set of forecast probabilities are issued and the integral in Eq. (7.3) over all values must then be replaced by a finite sum. The climatological base rate of the event $s = p(X = 1)$ can be written as

$$s = p(X = 1) = \int_0^1 p(X = 1|q)p(q)\,\mathrm{d}q = \int_0^1 f(q)p(q)\,\mathrm{d}q \tag{7.4}$$

Alternatively, this can be expressed in terms of expectations over X and q as

$$s = E(X) = \int_0^1 E_x(X|q)p(q)\,\mathrm{d}q = E_q[E_x(X|q)] \tag{7.5}$$

and so the base rate can be written as the expectation of $f(q)$ over all possible q values: $s = E_q[\,f(q)]$. The statistical performance of the probability forecast system is entirely determined by the functions $p(q)$ and $f(q)$ – all scores can be expressed in terms of these two calibration-refinement functions. Different prediction systems will have different functions of $p(q)$ and $f(q)$, yet the base rate that is independent of the prediction system, will always be given by Eq. (7.4). By conditioning on the forecast probabilities, the Brier score can be written as

$$B = E[(\hat{p} - X)^2] = E_q[E_x[(q - X)^2|q]\,] \tag{7.6}$$

where the expectation over X is given by

$$\begin{aligned} E_x[(q - X)^2|q] &= (q - 0)^2[1 - f(q)] + (q - 1)^2 f(q) \\ &= [q - f(q)]^2 + f(q)[1 - f(q)] \end{aligned} \tag{7.7}$$

based on the definition $f(q) = p(X = 1|q)$. By taking the expectation of Eq. (7.7) over all possible q values, one then obtains the decomposition of the Brier score:

$$B = E_q[(q - f(q))^2] - E_q[(f(q) - s)^2] + s(1 - s) \tag{7.8}$$

The first term on the right-hand side of Eq. (7.8) is an overall measure of *reliability* equal to the mean squared deviation of the reliability curve from the diagonal (see Fig. 7.2). For a perfectly reliable system, $f(q) = q$ and so this term is then zero. The second term $E_q[(f(q) - s)^2]$ is an overall measure of *resolution* identical to ${}_q[f(q)]$ – systems with good resolution have $f(q)$ that differ from the climatological base rate s. The larger the overall resolution, the better the forecast system can *a priori* identify situations that lead to the occurrence or non-occurrence of the event in question in the future. Note that resolution is entirely based on the conditional probabilities $f(q)$ and so is independent of the actual marginal distribution of forecast probability values (and thus also independent of reliability). Resolution is only a measure of how the different forecast events are classified (or 'resolved') by a forecast system. The third term $s(1 - s)$ on the right-hand side of Eq. (7.8) is known as the *uncertainty* and is equal to the variance of the observations (X). This term is independent of the forecast system and cannot be reduced by improving the forecasts. The difficulty (or lack of it) in predicting events with close to 0.5 (0 or 1) climatological probability is represented by a large (small) uncertainty term in Eq. (7.8).

By comparing the terms in Eq. (7.8) with one another, it is possible to construct relative measures of reliability and resolution as follows:

$$\begin{aligned} B_{\text{rel}} &= \frac{E_q[(q - f(q))^2]}{s(1 - s)} \\ B_{\text{res}} &= 1 - \frac{E_q[(f(q) - s)^2]}{s(1 - s)} \end{aligned} \tag{7.9}$$

Both these measures are negatively oriented, and are equal to zero for a perfect deterministic forecasting system. They are related to the BSS (defined using the climatological forecast system as a reference, BSS_c) as follows:

$$BSS_c = 1 - B_{\text{rel}} - B_{\text{res}} \tag{7.10}$$

In our operational NCEP forecasting example, the Brier score for the system represented by the solid curve in Fig. 7.2 is equal to 0.066. The base rate for the event under consideration is equal to 0.714, which yields 0.677 for the Brier skill score BSS_c. The corresponding values of the components defined in Eq. (7.9) are $B_{\text{rel}} = 0.027$ and $B_{\text{res}} = 0.296$. These

values are typical of the values produced by present-day operational short- and medium-range weather forecasting systems. It is often found that the reliability term is significantly smaller (typically one order of magnitude less) than the resolution term.

Fig. 7.3 shows the BSS defined using the climatological forecast system BSS_c (full curve) and its two components B_{rel} (short-dashed curve) and B_{res} (dashed curve), as a function of forecast lead-time, for the European Centre for Medium-range Weather Forecasts (ECMWF) Ensemble Prediction System (Molteni *et al.* 1996). The event considered here is that the 850-hPa temperature falls at least 2 °C below the mean of the 1999 winter values (sample climatology T_c). Scores were computed over the same geographical area and time period as those used to construct Fig. 7.2. Since no data were missing, the total number of cases considered is now $n = 22,680$. Forecast probabilities are defined in the same way as for Fig. 7.2, as $\hat{p} = i/m$, where $i = 0, 1, 2, \ldots, m$ is the number of ensemble members forecasting the event, $m = 50$, and the verifying 'observation' is obtained from the ECMWF operational analysis. The skill score BSS_c numerically decreases (meaning the quality of the system degrades) with increasing forecast lead-time. The decrease is entirely due to the resolution component B_{res}, whereas the reliability component B_{rel} (which, as before, is significantly smaller than B_{res}) shows no significant variation with lead-time. The deg radation of resolution corresponds to the fact that, as the lead-time

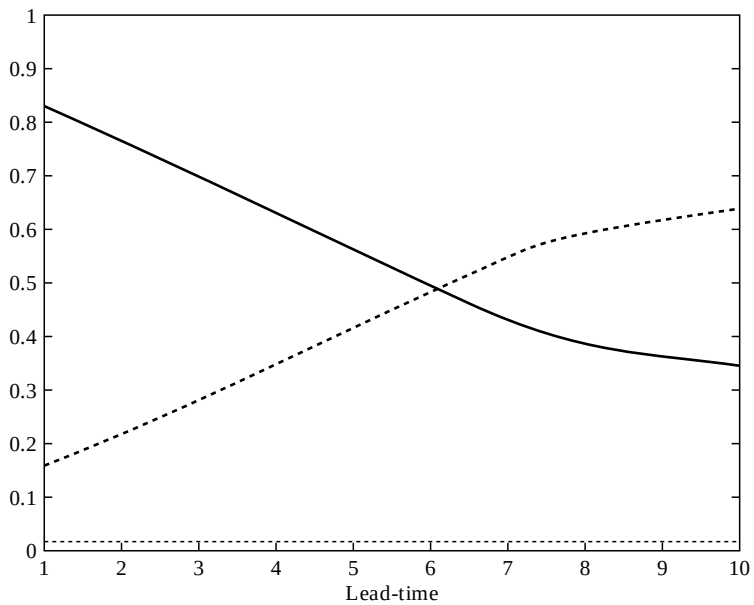

Figure 7.3 Brier skill score (BSS_c full curve, positively oriented), and its reliability (B_{rel}, short dash), and resolution (B_{res}, dashed, both negatively oriented) components, as a function of forecast lead-time (days) for the ECMWF Ensemble Prediction System (see text for the definition of the event E under consideration)

increases, the ensemble forecasts give a broader spread of temperatures, and become more similar to the climatological probability distribution. All these features are typical of what is seen in other current ensemble weather forecasting systems.

Finally, we note again that if both the forecast and observed systems are stationary in time and there is a sufficiently long record of their behavior it is possible to calibrate the forecasts to make them more reliable (see also Section 7.2). If the conditional probability of occurrence of an event $f(q)$ is different from the forecast probability q, the forecasts can be made more reliable by using the relabeled forecasts $q' = f(q)$. This, if done on all values of q, amounts to moving all points of the reliability curve horizontally to the diagonal (Fig. 7.2). As a result of this calibration, the reliability term on the right-hand side of Eq. (7.8) becomes zero, while the resolution term, that measures the variance of the *calibrated* forecasts, does not change. As pointed out earlier, resolution is invariant under calibration, hence it reflects a forecasting system's genuine ability to distinguish among situations that lead to the future occurrence or non-occurrence of an event (no matter what labels are used). We note passing by that calibration as defined above is only one type of statistical post-processing of forecasts. For example, in case of ensemble forecast systems, more complex post-processing algorithms that attempt to eliminate possible biases in the forecast values, before they are converted to probability values, can improve not only the reliability but also the resolution of the forecasts (see, e.g., Atger 2002).

7.3.3. Verification Based on Decision Probability Thresholds

A useful decision-theoretic approach to the verification of probability forecasts is to use a sequence of probability thresholds to transform a single set of probability forecasts into a continuous set of binary yes/no forecasts that can be verified using the methods presented in Chapter 3. For a given probability threshold, p_t, in the range 0 to 1, probability forecasts of a binary predictand can be converted into deterministic binary forecasts by using the following decision rule: if $\hat{p} \geq p_t$, then $\hat{X} = 1$ ('yes' forecast), otherwise $\hat{X} = 0$ ('no' forecast). This decision rule is similar to how users often make decisions based on probability information – they take protective action only when the forecast probability of the event exceeds a critical (user-specific) threshold. For an ensemble of m forecasts, there are m distinct thresholds corresponding to at least $1, 2, 3, \ldots, m$ of the forecasts predicting the chosen event. Therefore, *probability forecasts of a continuous variable* can be first converted into *probability forecasts of a binary event* (by specifying exceedance above/below a threshold for the continuous variable), and then these can be converted into a continuous set of *deterministic forecasts of a binary event* (by using a sequence of probability decision thresholds). The verification of probability forecasts therefore amounts to

verifying a continuous set of deterministic binary forecasts obtained for all the possible probability thresholds in the range 0 to 1.

As explained in detail in Section 3.4 of Chapter 3, a continuous set of deterministic binary forecasts can be verified using signal detection techniques. One of the most powerful tools is the *relative operating characteristic* obtained by plotting the hit rate versus the false alarm rate for each possible decision probability threshold: the two-dimensional locus of points $(F(p_t),\ H(p_t))$. For all probability forecasts, $H(p_t)$ and $F(p_t)$ both decrease from 1 to 0 as the decision threshold probability p_t increases from 0 to 1. The ROC curve for a climatological probability forecasting system that always forecasts the base rate probability, s, has only two points on the ROC diagram: (1,1) for $p_t > s$ and (0,0) for $p_t \leq s$. For the special case of a $(m = 1)$ deterministic binary forecast, there is only one point (F,H) on the ROC diagram in addition to the corner points (0,0) and (1,1). A probabilistic forecast system with good reliability and high resolution is similar to a perfect deterministic forecast in that it will only forecast probabilities that are close to either 0 or 1. For such a system, the majority of the points on the ROC diagram will therefore be close to the *perfect forecast* (0,1) point. It follows that in general the proximity of the ROC curve to the (0,1) point can provide an indication of overall skill of the forecasts. For example, the area under the ROC curve is one such measure that can be used to construct a skill score (see Section 3.4.4 of Chapter 3).

A more detailed interpretation of the ROC results can be obtained by noting that the probability of an event occurring is given by the threshold-dependent base rate $s(p_t) = \int_{p_t}^{1} p(q)\,\mathrm{d}q$ and that the probability of a hit for a given probability threshold p_t can be written as $\int_{p_t}^{1} f(q)\,p(q)\,\mathrm{d}q$. Therefore, the hit and false alarm rates can be written, respectively, as:

$$H(p_t) = \frac{1}{s(p_t)} \int_{p_t}^{1} f(q)\,p(q)\,\mathrm{d}q \tag{7.11a}$$

$$F(p_t) = \frac{1}{1 - s(p_t)} \int_{p_t}^{1} (1 - f(q))\,p(q)\,\mathrm{d}q \tag{7.11b}$$

The integral in Eq. (7.11a) is the average of $f(q)$ for those circumstances when $q > p_t$. When $f(q)$ is a strictly monotonically increasing function of p_t, the threshold inequality $q > p_t$ becomes equivalent to $f(q) > f(p_t)$, and the comparison with the threshold can be done on the *a posteriori* calibrated probabilities $q' = f(q)$ as well as on the directly predicted probabilities q. The same argument equally applies to the integral in (7.11b), which means that the ROC curve is invariant to *a posteriori* calibration $q' = f(q)$ (where

the points on the reliability curve are moved horizontally to the diagonal). Thus, if the reliability curve is strictly monotonically increasing, the ROC curve, just like the resolution component of the Brier score, depends only on the *a posteriori* calibrated probabilities $q' = f(q)$ (and their probability distribution). The ROC curve, in these cases, is therefore independent of reliability, and measures the resolution of the forecasting system. The resolution component of the Brier score and the ROC curve therefore often provide very similar qualitative information.

Fig. 7.4 shows the ROC curves for the same set of forecasts evaluated in terms of their Brier score in Fig. 7.3. Note that, as expected, the area under the ROC curves decreases monotonically as a function of increasing forecast lead-time (with values of 0.98, 0.95, 0.92 and 0.87 for lead-times of 2, 4, 6 and 8 days, respectively), just as the Brier score did in Fig. 7.3. While both the Brier score and ROC area indicate a loss of predictability with increasing lead-time, the corresponding values for the two scores are quantitatively different. Moreover, it can be shown that there is no one-to-one relationship between the two measures. It is not clear which measure of resolution (if any) is generally preferable for judging forecast skill. A potential advantage of skill measures such as the ROC area is that they are directly related to a decision-theoretic approach and so can be easily related to the economic value of probability forecasts for forecast users (see Chapter 8).

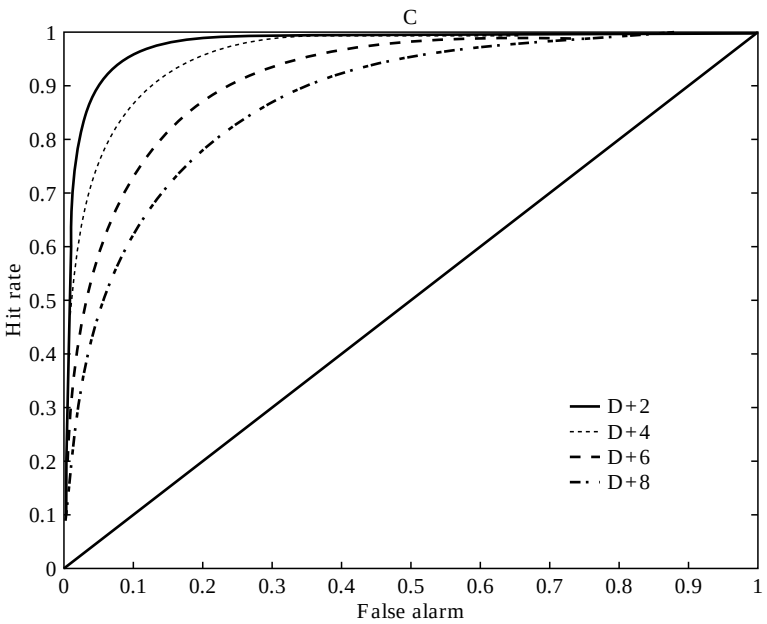

Figure 7.4 ROC curves for the same event and predictions as in Fig 7.3, for four different forecast ranges

7.4. PROBABILITY FORECASTS OF MORE THAN TWO CATEGORIES

7.4.1 Vector Generalization of the Brier Score

The Brier score was defined in Eq. (7.1) for the verification of probability forecasts of binary events. However, Brier (1950) gave a more general definition that considered multiple categories of events. Let us consider an event with K complete, mutually exclusive (and not necessarily ordered) outcomes E_k ($k = 1, \ldots, K$), of which one, and only one, is always necessarily observed (see Chapter 4 for deterministic forecasts of such predictands). A probabilistic forecast for this set of events then consists of a K-vector of probabilities $\underline{\hat{p}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K)$ such that $\sum_{k=1}^{K} \hat{p}_k = 1$. The general definition of the Brier score for probability forecasts of K categories is given by

$$B = E\left(\frac{1}{K}\sum_{k=1}^{K}(\hat{p}_k - X_k)^2\right). \tag{7.12}$$

where $X_k = 1$ if the observed outcome is E_k, and 0 otherwise. By defining the observation K-vector $\underline{x} = (x_1, x_2, \ldots, x_K)$ containing $K - 1$ zeros and a single 1, the Brier score can be written in vector notation as $E[\,\| \underline{\hat{p}} - \underline{x} \|^2 / K\,]$ where $\| \cdot \|$ denotes the Euclidean norm. The Brier score for K categories is simply the arithmetic mean of the binary Brier scores (Eq. (7.1)) for each outcome E_k. A BSS can be defined as in Eq. (7.2) by using a reference probability forecast that constantly forecasts the corresponding climatological base rate s_k for each category. Examples for the use of the multiple category Brier score can be found in Zhu *et al.* (1996) and Toth *et al.* (1998).

A reliability–resolution decomposition of the multiple-category Brier score can be obtained by averaging the components of the binary Brier scores for each individual category. For multi-event forecasts, a more discriminatory decomposition, built on the entire sequence $\underline{q}$ of predicted probabilities, seems preferable. Denoting $\mathrm{d}p(\underline{q})$ the frequency with which the sequence $\underline{q}$ is predicted by the system, and defining the sequence $\underline{f}(\underline{q}) = [\,f_k(\underline{q})]$ of the conditional frequencies of occurrence of the E_k's given that $\underline{q}$ has been predicted, a generalization of the derivation leading to Eq. (7.8) shows that

$$B_K = \frac{1}{K}\int \| \underline{q} - \underline{f}(\underline{q}) \|^2 \, \mathrm{d}p(\underline{q}) - \frac{1}{K}\int \| \underline{f}(\underline{q}) - \underline{p}_c \|^2 \, \mathrm{d}p(\underline{q}) + B_{c\,K} \tag{7.13}$$

where $\underline{p}_c$ is the sequence $(p_{c\,k})$. Similarly to Eq. (7.8), Eq. (7.15) provides a decomposition of B_K into reliability, resolution and uncertainty terms.

7.4.2. Information Content as a Measure of Resolution

It has been argued that given a suitably large sample of previous forecasts and matching observations, probabilistic forecasts can be made more reliable by calibration. Unlike reliability, the resolution of a forecasting system cannot be changed by calibration and so represents the (invariant) ability of the forecasting system to resolve future events. Various measures of resolution have been proposed for probability forecasts including ones based on information theory measures such as *information content* (entropy) (see Section 2.7; Toth *et al.* 1998; Stephenson and Doblas-Reyes 2000; Roulston and Smith 2002). One possible definition of the information content (I) of a forecast of the probabilities for K mutually exclusive, climatologically equiprobable, and exhaustive categories is given by

$$I[\hat{p}] = 1 + \sum_{i=1}^{K} \hat{p}_i \log_K \hat{p}_i \tag{7.14}$$

where $0 \leq \hat{p}_i \leq 1$ is the forecast probability for the ith category that satisfy $\sum_{i=1}^{K} \hat{p}_i = 1$. When forecasts are perfectly reliable, the mean information content over all forecasts can be considered to be another measure of resolution. Under these conditions, the information content ranges between zero for a uniform probability forecast that forecasts $\hat{p}_i = 1/K$ for all categories, and one for a deterministic forecast that forecasts $\hat{p}_i = 1$ for only one category and 0 for all others. Fig. 7.5 shows the mean information

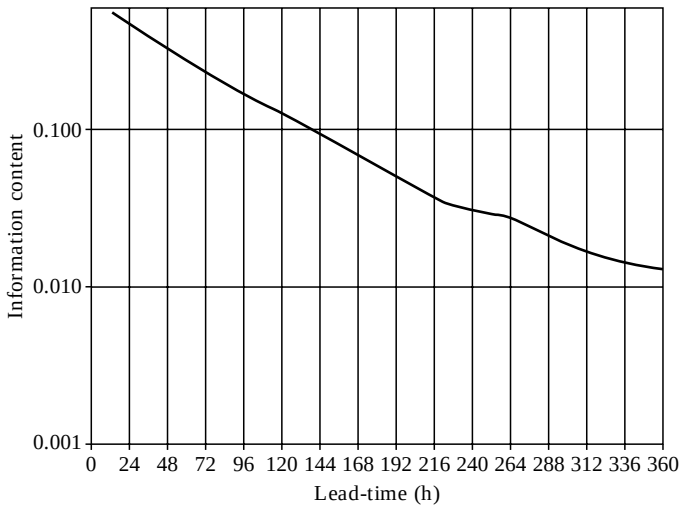

Figure 7.5 Information content as defined in text for calibrated probabilistic forecasts (with near perfect reliability) based on a 10-member subset of the NCEP ensemble. Forecasts are made for 10 climatologically equally likely intervals for 500 hPa geopotential height values over the Northern Hemisphere extratropics (20–80 N), and are evaluated over the March–May 1997 period

content of a 10-member 0000 UTC subset of the NCEP ensemble forecasts for 500 hPa geopotential height as a function of lead-time. For not perfectly reliable forecasts, a more general and invariant measure of resolution can be obtained by considering the information content of the calibrated forecasts $I[f(\hat{p})]$. It should be noted that information content defined in this way is equal to the Kullback–Leibler G^2 measure of association for $(K \times K)$ contingency tables that tends to the better known χ^2 measure of association in the limit of large cell counts (Stephenson 2000). Hence, the χ^2 measure of association for the calibrated probabilities may also provide a good overall measure of resolution for probability forecasts.

7.5 PROBABILITY FORECASTS OF CONTINUOUS VARIABLES

The previous sections in this chapter have discussed the verification of probability forecasts of nominal categories of events. Probability forecasts of continuous variables (e.g., temperature at a location) can also be treated as categorical forecasts by partitioning the range of values into a finite number of complete yet exclusive intervals (bins/classes). Categories constructed for continuous variables are ordinal categories that have a natural ordering/distance. The verification tools presented so far were developed for use with nominal categories where the order of the categories did not matter (or affect the scores). If applied with ordinal categories they can lead to loss of important verification information related to the ordering of the categories. This section will discuss two scores that have been developed specifically for accounting for the distance information implicit in categories constructed for continuous variables.

7.5.1 The Discrete Ranked Probability Score

Consider $K > 2$ thresholds $x_1 < x_2 < \cdots < x_K$ for the continuous random variable X that define the events $A_k = \{X \leq x_k\}$ for $k = 1, 2, \ldots, K$. The forecast probabilities for the events are denoted by $(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_K)$ and the binary indicator variables for the kth observed event are denoted 0_k (i.e., $0_k = 1$ if A_k occurs, and $0_k = 0$ otherwise). The discrete *ranked probability score* (RPS) is then defined as

$$\text{RPS} = E[\frac{1}{K}\sum_{k=1}^{K}(\hat{p}_k - 0_k)^2] = \frac{1}{K}\sum_{k=1}^{K} B_k \tag{7.15}$$

where B_k is the Brier score for the event $A_k = \{X \leq x_k\}$. The RPS is similar to the multiple category Brier score in Eq. (7.12), but, as its name implies, it takes into account the ordered nature of the variable X. Here the events A_k

are not mutually exclusive, and A_j implies $A_{j'>j}$. Consequently, if X is forecast for instance to fall in an interval $[x_j,\ x_j + 1]$ with probability one, but is observed to fall into another interval $[x_j',\ x_j' + 1]$, the RPS increases with the increasing absolute difference $|j - j'|$.

7.5.2 The Continuous Ranked Probability Score

A continuous extension of the RPS can be defined by considering an integral of the Brier scores over all possible thresholds x, instead of an average of Brier scores over a finite number of discrete thresholds as in Eq. (7.15). Denoting the predicted c.d.f. by $F(x) = p(X \leq x)$ and the observed value of X by x_0, the continuous ranked probability score (CRPS) can be written as

$$\text{CRPS} = E\left(\int_{-\infty}^{\infty} [F(x) - H(x - x_0)]^2 \, \mathrm{d}x\right) \tag{7.16}$$

where $H(x - x_0)$ is the Heaviside function that takes the value 0 when $x - x_0 < 0$, and 1 otherwise. Both the discrete and CRPS, just like the multiple category Brier score, can be expressed as skill scores (see Eq. (7.2)), and are amenable to reliability–resolution decompositions. For additional related information the reader is referred to Hersbach (2000).

7.6 SUMMARY STATISTICS FOR ENSEMBLE FORECASTS

Ensemble forecasting is now one of the most commonly used methods for generating probability forecasts that can take account of uncertainty in initial and final conditions. The previous sections were devoted to the verification of probabilistic forecasts in general. However, before ensemble forecasts are converted into probabilistic information, it is desirable to explore and summarize their basic statistical properties. This section will therefore present some of the statistics that are most often used to summarize ensembles of forecasts. At the initial time, an ensemble of forecasts is generally constructed to be centered on the *control analysis* – i.e., the ensemble mean at zero lead-time is the best estimate of the state of the system (obtained either directly or by averaging an ensemble of analysis fields).

Section 7.2 pointed out that the inherent value of forecast systems lies in their ability to distinguish between cases when an event has a higher or lower than climatological probability to occur in the future (resolution). As Figs. 7.3 and 7.4 demonstrate, resolution decreases rapidly with lead-time (due to the loss of information in the flow). This is because in fluid systems such as the atmosphere and oceans, naturally occurring instabilities amplify

initial and model related uncertainties. Even though skill is reduced and eventually lost, forecasts can remain (or can be calibrated to remain) statistically consistent with observations (reliable). An ensemble forecast system that is statistically consistent with observations is often called a perfect ensemble in a sense of perfect reliability. An important property of a perfectly reliable ensemble is that the verifying analysis (or observations) should be statistically indistinguishable from the forecast members. Most of the verification tools specifically developed and applied to ensemble forecasts are designed to evaluate the statistical consistency of such forecasts. These additional measures of reliability, as we will see below, can reveal considerably more detail as to the nature and causes of statistically inconsistent behavior of ensemble-based probabilistic forecasts than the reliability diagram (Section 7.3.1) or the single measure of the reliability component of the Brier score (Section 7.3.2). By revealing the weak points of ensemble forecast systems, the ensemble-based measures provide important information for the developers of such systems that can eventually lead to improved probability forecasts.

7.6.1 Ensemble Mean Error and Spread

If the verifying analysis is statistically indistinguishable from the ensemble members, then its mean distance from the mean of the ensemble members (ensemble mean error) must equal the mean distance of the individual members from their mean (ensemble standard deviation or spread) – see Buizza (1997) and Stephenson and Doblas-Reyes (2000). Fig. 7.6 compares the root mean square error of the ensemble mean forecast and the mean spread of the NCEP ensemble forecasts as a function of lead-time. Initially, the ensemble spread is larger than the ensemble mean error, indicating a larger than desired initial spread. The growth of ensemble spread, however, is less than that of the error, which then leads to insufficient spread at later lead-times. This is typical behavior in current ensemble forecast systems that often tend to underestimate ensemble spread due to not accounting for all possible sources of model related uncertainty (e.g., structural uncertainty caused by the model parameterizations being incorrect).

Since for perfectly reliable forecast systems the spread of the ensemble forecasts is equal to the error in the ensemble mean, for such systems the spread can also be considered as a measure of resolution (and therefore forecast skill in general). For example, an ensemble with a lower average ensemble spread can more efficiently separate likely and unlikely events from one another (and so has more information content). It is worth mentioning that the skill of the ensemble mean forecast is often compared to that of the single *control forecast* obtained by starting with the best initial conditions (control analysis). Once non-linearity becomes pronounced, the mean of an ensemble that properly describes the case-dependent forecast uncertainty is able to provide a better estimate of the future state of the

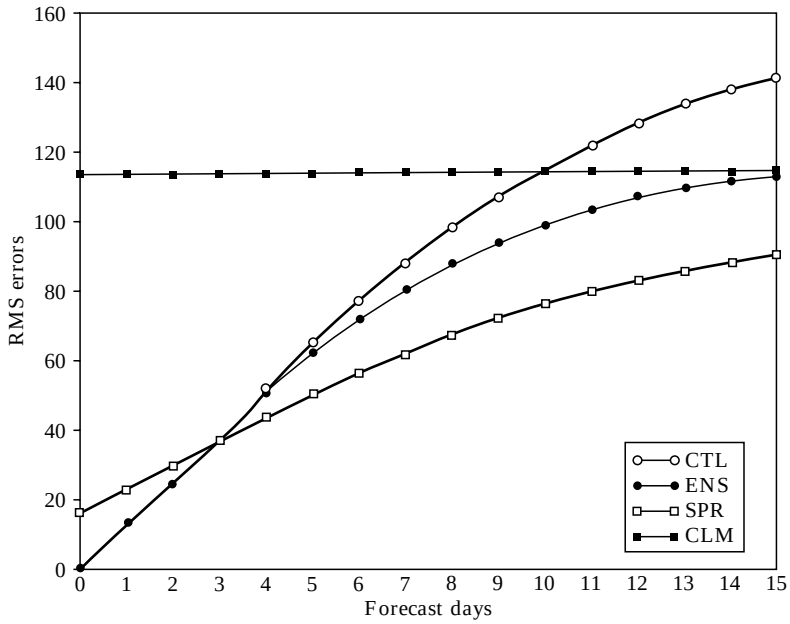

Figure 7.6 Root mean square error of 500 hPa geopotential height NCEP control (open circle), ensemble mean (full circle), and climate mean (full square) forecasts, along with ensemble spread (standard deviation of ensemble members around their mean, open square), as a function of lead-time, computed for the Northern Hemisphere extratropics, averaged over December 2001–February 2002

system than the control forecast (see Toth and Kalnay 1997). In a good ensemble forecasting system, the ensemble mean error should therefore be equal or less than the error of the control forecast (see Fig. 7.6). It follows that in a reliable ensemble the spread of the ensemble members around the mean will be less than that around the control forecast.

7.6.2 Equal Likelihood Frequency Plot

Ensemble forecast systems are designed to generate a finite set of forecast scenarios. Some ensemble forecast systems (e.g., those produced by ECMWF and NCEP) use the same technique for generating each member of the ensemble (i.e., the same numerical prediction model, and the same initial perturbation generation technique). In some other systems, each ensemble member is generated using a different model version (e.g., the ensemble forecasting system employed at the Canadian Meteorological Centre, see Houtekamer *et al.* 1996). In such systems, individual ensemble members may not all perform equally well. Similarly, if the control forecast is included in an otherwise symmetrically formed ensemble, the assumption of equal likelihood of the forecasts can become questionable.

Whether all ensemble members are equally likely or not is in itself neither a desirable nor an undesirable property of an ensemble prediction system. When ensemble forecasts are used to define forecast probabilities, however, one must know if all ensemble members can be treated in an indistinguishable fashion. This can be tested by generating a bar plot showing the number of cases (accumulated over space and time) when each member was the forecast closest to the verifying diagnostic (see Zhu *et al.* 1996). Information from such a frequency plot can be useful as to how the various ensemble members must be used in defining forecast probability values. Equal frequencies indicate that all ensemble members are equally likely and can be considered as independent realizations of the same random process, indicating that the simple procedure used in Section 7.3.1 for converting ensemble forecasts into probabilistic information is applicable.

Fig. 7.7 compares the frequency of NCEP ensemble forecasts being closest to the verifying analysis value, averaged for 10 perturbed ensemble members, with that of an equal and a higher horizontal spatial resolution unperturbed control forecast as a function of lead-time. Note first in Fig. 7.7 that the higher resolution control forecast, due to its ability to better represent nature, has an advantage against the lower resolution members of

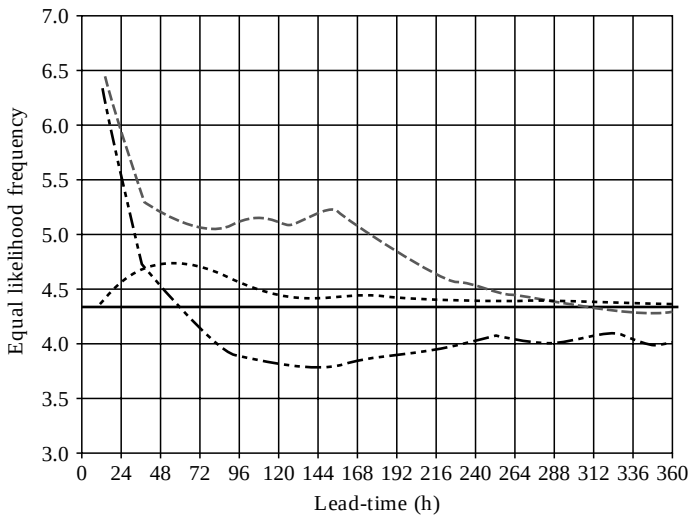

Figure 7.7 Equal likelihood diagram, showing the percentage of time when the NCEP high (dashed) and equivalent resolution control (dash-dotted), and any one of the 10 0000 UTC perturbed ensemble 500 hPa geopotential height forecasts (dotted) verify best out of a 23-member ensemble (of which the other 11 members are initialized 12 h earlier, at 1200 UTC), accumulated over grid-points in the Northern Hemisphere extratropics during December 2001–February 2002. Chance expectation is 4.35 (solid)

the ensemble. This advantage, however, is rather limited. As for the low resolution control forecast, at short lead-times, when the spread of the ensemble around the control forecast is too large (see Fig. 7.6), it is some what more likely to be closest to the verifying analysis. At longer lead-times, when the spread of the NCEP ensemble becomes underestimated due to under representation of model related uncertainty, the control forecast becomes less likely to verify best. When the spread is too low at longer lead-times, the ensemble members are clustered too densely and the verifying analysis often lies outside of the cloud of the ensemble. In this situation, since the control forecast is more likely to be near the center of the ensemble cloud than the perturbed members, a randomly chosen perturbed forecast has a higher chance of being closest to the verifying observation than the control. The opposite is true at short lead-times characterized by too large spread. The flat equal likelihood values at intermediate lead-times (i.e., the 48-h perturbed and equal resolution control forecasts have the same likelihood in Fig. 7.7) thus are indicative of proper ensemble spread (cf. Fig. 7.6), and hence good reliability.

7.6.3 Analysis Rank Histogram

If all ensemble members are equally likely and statistically indistinguishable from nature (i.e., the ensemble members and the verifying observation are mutually independent realizations of the same probability distribution), then each of the $m + 1$ intervals defined by an ordered series of m ensemble members, including the two open ended intervals, is equally likely to contain the verifying observed value. Anderson (1996) and Talagrand *et al.* (1998) suggested constructing a histogram by accumulating the number of cases over space and time when the verifying analysis falls in any of the $m + 1$ intervals. Such a graph is often referred to as the *analysis rank histogram*.

Reliable or statistically consistent ensemble forecasts lead to an analysis rank histogram that is close to flat, indicating that each interval between the ordered series of ensemble forecast values is equally likely (see the 3-day panel in Fig. 7.8). An asymmetrical distribution is usually an indication of a bias in the mean of the forecasts (see 15-day lead-time panel in Fig. 7.8) while a U (5-day panel in Fig. 7.8) or inverted U-shape (1-day panel in Fig. 7.8) distribution may be an indication of a positive or negative bias in the variance of the ensemble, respectively. Current operational ensemble weather forecasting systems in the medium lead-time range (3–10 days ahead) exhibit U-shaped analysis rank histograms, which implies the verifying analysis falls outside the cloud of ensemble forecasts more often than one can expect by chance, given the finite size of the ensemble. In other words, the ensemble forecasts underestimate the true uncertainty in the forecasts.

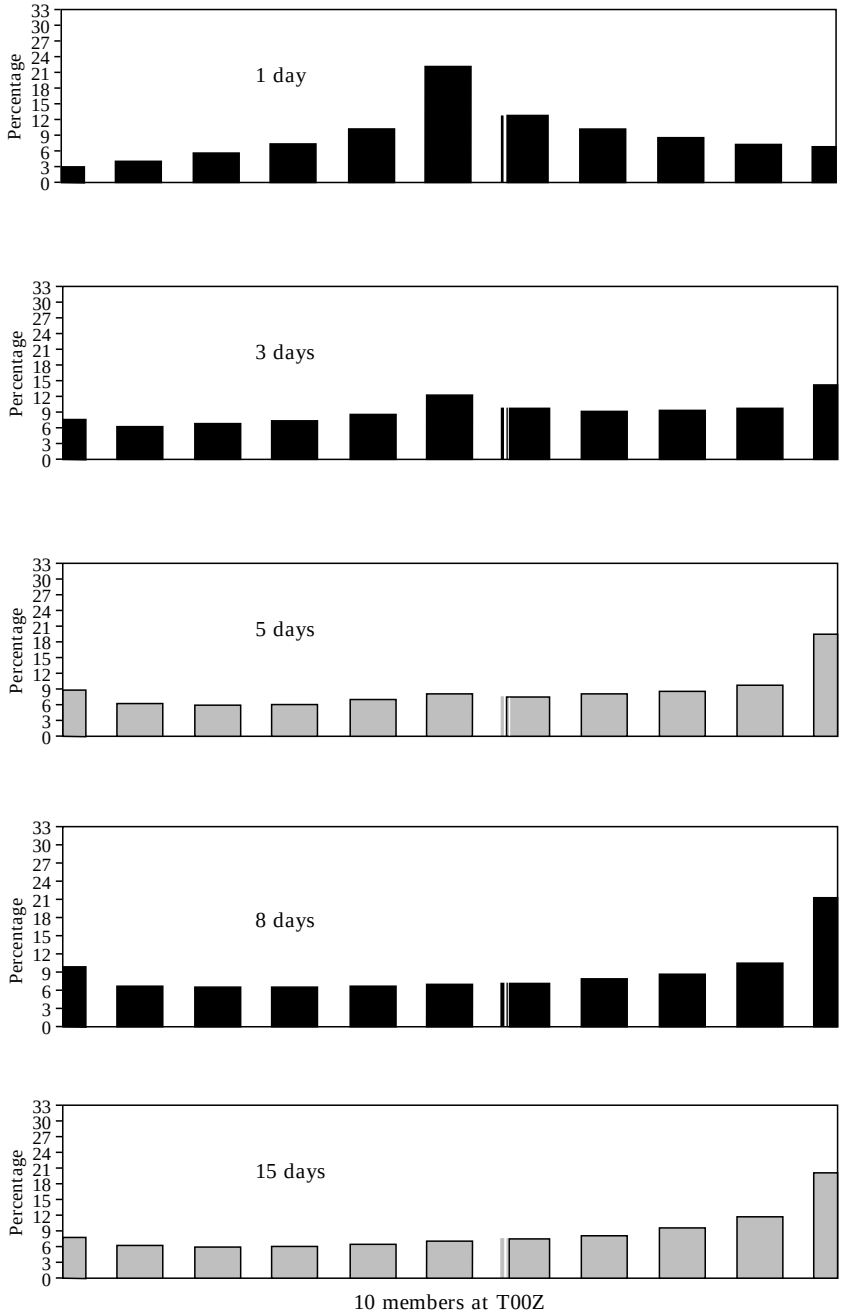

Figure 7.8 Analysis rank histogram for a 10-member 0000 UTC NCEP ensemble of 500 hPa geopotential height forecasts over the Northern Hemisphere extratropics during December 2001–February 2002

7.6.4 Multivariate Statistics

All ensemble verification measures discussed so far are based on univariate statistics (e.g., the value at one grid-point or an area-average value). However, meteorological forecasts are often issued for many variables defined at spatial grid-points and so one needs to consider multivariate statistics in order to summarize such forecasts. Recently, various multivariate approaches have been proposed to evaluate the statistical consistency of ensemble forecasts.

One approach involves the computation of various statistics (like average distance of each member from the other members) for a selected multivariate variable (e.g., 500 hPa geopotential height defined over grid-points covering a pre-selected area), separately for cases when the verifying analysis is *included in*, or *excluded from* the ensemble. A follow-up statistical comparison of the two, inclusive and exclusive sets of statistics accumulated over a spatio-temporal domain can reveal whether at a certain statistical significance level the analysis can be considered part of the ensemble in a multivariate sense (in the case when the two distributions are indistinguishable) or not. Smith (2000) suggested the use of the nearest neighbor algorithm for testing the statistical consistency of ensembles with respect to multivariate variables in this fashion.

Another approach is based on a comparison of forecast error patterns (e.g., control forecast minus verifying analysis) and corresponding ensemble perturbation patterns (control forecast minus perturbed forecasts). In a perfectly reliable ensemble, the two sets of patterns are statistically indistinguishable. The two sets of patterns can be compared either in a climatological fashion, based, on an empirical orthogonal function e.g. analysis of the two sets of patterns over a large data set (e.g., Molteni and Buizza 1999; Stephenson and Doblas-Reyes 2000), or on a case-by-case basis (e.g., Wei and Toth 2002).

7.6.5 Time Consistency Histogram

The concept of rank histograms can be used not only to test the reliability of ensemble forecasts but also to evaluate the time consistency between ensembles issued on consecutive days. Given a certain level of skill as measured by the probability scores discussed in Section 7.3, an ensemble system that exhibits less change from one issuing time to the next may be of more value to some users. When constructing an analysis rank histogram, in place of the verifying analysis one can use ensemble forecasts generated at the next initial time. The 'time consistency' histogram will then assess whether the more recent ensemble is a randomly chosen subset of the earlier ensemble set.

Ideally, one would like to see that with more information, more recently issued ensembles narrow the range of the possible, earlier indicated solutions, without shifting the new ensemble into a range that has not been included in the earlier forecast distribution. Such 'jumps' in consecutive probabilistic forecasts would result in a U-shaped time consistency histogram, indicating sub-optimal forecast performance. While control forecasts, representing a single scenario within a large range of possible solutions, can exhibit dramatic jumps from one initial time to the next, ensembles typically show much smoother variations in time.

7.7 LIMITATIONS OF PROBABILITY AND ENSEMBLE FORECAST VERIFICATION

The verification of probabilistic and ensemble forecast systems has several limitations. First, as pointed out earlier, probabilistic forecasts can be evaluated only in a statistical sense. The larger the sample size, the more stable and trustworthy the verification results become. Given a certain sample size, one often needs to, or has the option to subdivide the sample in search for more detailed information. For example, when evaluating the reliability of continuous-type probability forecasts one has to decide when two forecast distributions are considered being the same. Grouping (pooling) more diverse forecast cases into the same category will increase sample size but can potentially reduce useful forecast verification information. Another example concerns spatial aggregation of statistics. When the analysis rank or other statistics are computed over large spatial or temporal domains a flat histogram is a necessary but not sufficient condition for reliability. Large and opposite local biases in the first and/or second moments of the distribution may get cancelled out when the local statistics are aggregated over larger domains (see, e.g., Atger 2002). In a careful analysis, the conflicting demands for having a large sample to stabilize statistics, and working with more specific samples (collected for more narrowly defined cases, or over smaller areas) for gaining more insight into the true behavior of a forecast system, need to be balanced.

So far it has been implicitly assumed that observations are perfect. To some degree this assumption is always violated. When the observational error is comparable to the forecast errors, observational uncertainty needs to be explicitly dealt with in forecast evaluation statistics. A possible solution is to add noise to the ensemble forecast values with similar variance to that estimated to be present in the observations (Anderson 1996).

In case of verifying ensemble-based forecasts, one should also consider the effect of ensemble size. Clearly, a forecast based on a smaller ensemble will provide a noisier and hence poorer representation of the underlying processes, given the forecast system studied. Therefore, special care should be exercised when comparing ensembles of different sizes.

The limitations described above must be taken into account not only in probabilistic and ensemble verification studies, but also in forecast calibration where probabilistic and/or ensemble forecasts are statistically post-processed based on the previous forecast verification statistics.

7.8 CONCLUDING REMARKS

Reliability and resolution are the two main attributes of forecast systems in general. For probabilistic forecasts, reliability is defined as the statistical consistency between forecast probability values and the corresponding observed frequencies over the long run. Resolution, on the other hand, is defined as the ability of a forecast system to distinguish in advance between cases where future events are more or less likely to occur compared to the climatological frequency. A perfect forecast system uses only 0 and 1 probability values and has a perfect reliability. Note that this is a perfect deterministic forecast system.

This chapter has reviewed various methods for the evaluation of probability and ensemble forecasts. In the course of verifying probabilistic forecasts, their two main attributes: reliability and resolution are assessed. Such a verification procedure, just as that of any other type of forecasts, has its limitations. Most importantly, we recall that probabilistic forecasts can only be evaluated on a statistical (and not individual) basis using a sufficiently large sample of past forecasts and matching observations. When a stratification of all cases is required, a compromise has to be found between the desire to learn more about a forecast system and the need for maintaining large enough sub-samples to ensure good sampling of the verification statistics. Additional limiting factors include the presence of observational error, and the use of ensembles of limited size. The issue of comparative verification, where two forecast systems are inter-compared, was also raised and the need for the use of benchmark systems, against which a more sophisticated system can be compared, was stressed.

It was also pointed out that for temporally stationary forecast and observed systems, the reliability of forecasts can, in principle, be made perfect by using a calibration procedure based on (an infinite sample of) past verification statistics. In contrast, resolution cannot be improved by such a simple calibration (i.e., relabeling of forecast values). Thus, the resolution of calibrated forecasts provides an invariant measure of the performance of probabilistic forecasts.

The relationships among the different verification scores such as the ranked probability skill score, the relative operating characteristic, and the information content are not clearly understood. Which scores are best suited for certain applications is not clear either. It is important to mention in this respect that the value of forecasts can also be assessed in the context of their use by society. Some of the verification scores discussed above have

a clear link with the economic value of forecasts. For example, the resolution component of the BSS, and the ROC-area, two measures of the resolution of forecast systems, are equivalent to the economic value of forecasts under certain assumptions (Murphy 1966; Richardson 2000). The economic value of forecasts has such significance that an entire chapter in this book (Chapter 8) is devoted to this topic.