11

Seasonal and longer-range forecasts

Simon J. Mason
International Research Institute for Climate and Society, Palisades, NY, USA

11.1 Introduction

As discussed in Chapter 1, verification procedures should be tailored to the specific questions that are being asked, and to the nature of the forecasts as they are presented. Procedures for verifying seasonal and longer-range forecasts need only differ from those for shorter-range forecasts to the extent that different questions about the quality and value of the forecasts are asked, that data are available to answer those questions, and that the forecasts at the different timescales are presented in dissimilar formats. In this regard, two key issues have to be considered when performing verification analyses on forecasts at these longer timescales: limited sample size and low levels of predictability.

In most parts of the world seasonal forecasts were initiated only in the 1990s or later, and are rarely issued more frequently than once per month, and so there are currently very few examples of operational forecasts with more than about two decades of history. [Forecasts for the Indian monsoon, which were started in the 1880s (Blanford, 1884), are a notable exception.] Thus, sample sizes of seasonal forecasts typically are highly limited, while at the longest timescales there may not be any verifiable earlier predictions at all. Although it may be possible to generate hindcasts, it is often difficult to do this in a way that does not introduce an element of artificial skill (as discussed in Sections 1.4.2 and 11.3.1), and so there is a danger of overestimating the quality of the forecast system. In addition, generating the hindcasts may not even be viable: in decadal forecasting, for example, potential predictability is believed to come largely from subsurface ocean conditions, but observational data for initializing the models are severely lacking prior to the 1990s. Even when there is a history of forecasts and corresponding observations available, the quality of the forecasts over this period is unlikely to have remained constant because of model revisions and changes in observation accuracy. Consequently, verification results will give an indication of the average performance of the forecasts over the period of analysis, but will not necessarily give an accurate indication of the expected quality of subsequent predictions. The net effect of this sample size problem is that uncertainty estimates on measurements of the quality (or value) of seasonal and longer-range forecasts are typically large, and so assessing 'skill', whether against a baseline or against a competing forecast system, can be difficult (Tippett *et al.*, 2005).

The second overriding consideration in the verification of seasonal and longer-range forecasts is

Forecast Verification: A Practitioner's Guide in Atmospheric Science, Second Edition. Edited by Ian T. Jolliffe and David B. Stephenson.

that levels of predictability are almost invariably much lower than those of weather forecasts. This relatively poor predictability is an inherent part of the climate system itself, but is compounded by the fact that computational demands and poor availability of observational data mean that the models used to make predictions, and the initialization of such models, are of weaker quality than for the weather forecasting problem. By far the majority of verification analyses conducted to date have sought to address the simple question of whether the forecasts have any 'skill'. As discussed in Section 11.3.1, this question is often poorly posed: careful construction of the verification question, and interpretation of the results, may be required to avoid unnecessarily pessimistic conclusions about the potential usefulness of some forecasts.

In this chapter the primary focus is on verification of seasonal forecasts for the simple reason that longer-range predictions almost invariably do not have sufficient sample sizes to perform a verification analysis. However, that is not to say that longer-range predictions cannot be evaluated at all, and some guiding principles are provided in Section 11.5 at the end of this chapter. Before discussing verification procedures for seasonal forecasts, the most common forecast formats are briefly described (Section 11.2) because the appropriate verification options are constrained by the type of information that is being communicated in the forecasts. What constitutes 'skill' and ways of measuring it are reviewed in Section 11.3. In Section 11.4 some issues regarding the verification of individual forecasts are discussed.

11.2 Forecast formats

11.2.1 Deterministic and probabilistic formats

Most seasonal forecasts fall into one of two broad categories (Section 2.2): firstly, one or more 'deterministic' predictions of a seasonally averaged or integrated meteorological variable (e.g., mean temperature or total rainfall; Figure 11.1); and, secondly, a set of probabilities for the verification to fall within each of two or more predefined ranges (e.g. Figure 11.2). The deterministic forecasts are most often statistical or dynamical model outputs,

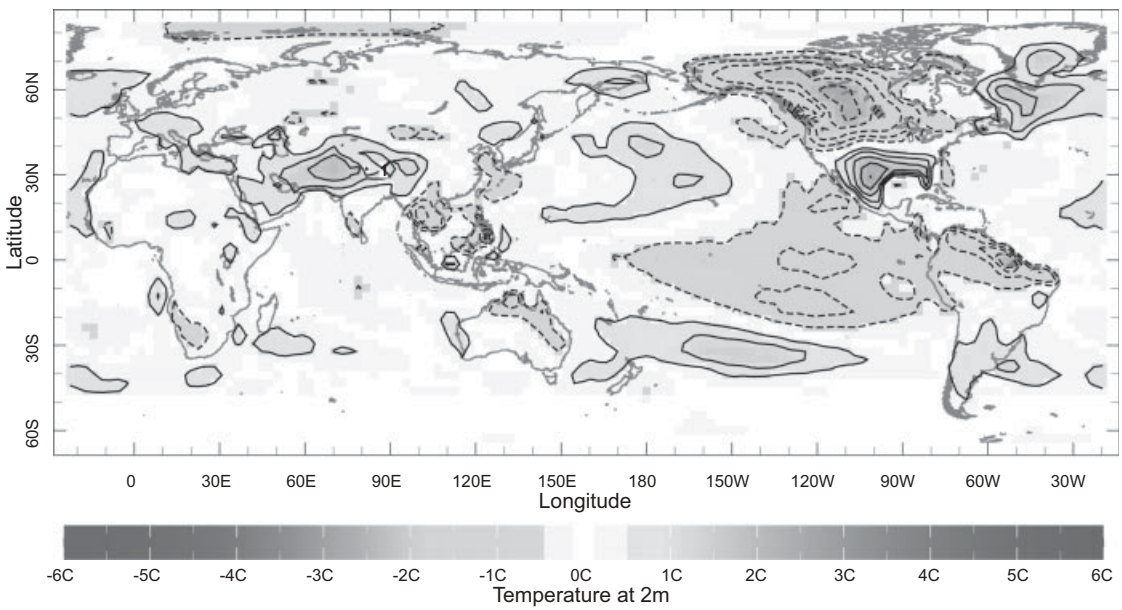

Figure 11.1 Example of a 'deterministic' seasonal prediction made in February 2011 using the ECHAM 4.5 model (Roeckner *et al.* 1996). The prediction is expressed as the ensemble mean March–May 2011 temperature anomaly with respect to the model's 1971–2000 climatology. A full colour version of this figure can be found in the colour plate section

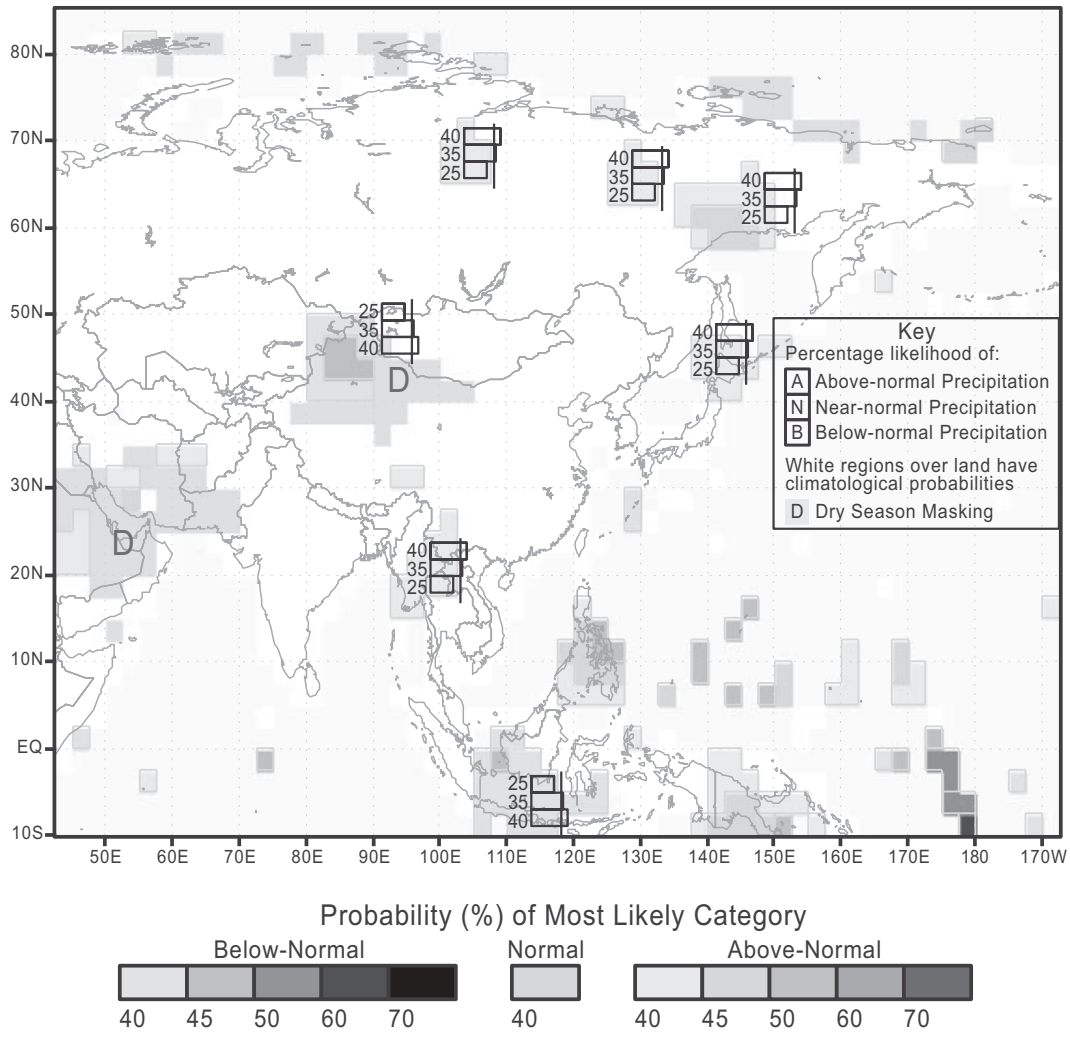

Figure 11.2 Example of a 'probabilistic' seasonal prediction issued in February 2011 by the International Research Institute for Climate and Society. The prediction shows the probabilities that the March–May 2011 precipitation total will be in one of the three categories 'above-normal', 'normal' and 'below-normal', with these categories defined using the terciles of March–May totals between 1971 and 2000. The probabilities of the most likely category are shaded, but the probability for the normal category is constrained to a maximum of 40% based on prior verification analyses. Probabilities for all three categories are only shown for large areas and for areas with relatively sharp probabilities. The three horizontal bars are scaled by the corresponding forecast probability, and the thin vertical line indicates the climatological probability of 33%. A full colour version of this figure can be found in the colour plate section

and the predictand is usually expressed as a 'best-guess' value on a continuous scale (van Oldenborgh *et al.*, 2005). This value may represent an area-average, which is typically the case for dynamical model outputs, or the forecast may be for a specific location, such as a meteorological station. Other seasonal forecast formats include counts, such as hurricane frequencies (Owens and Landsea, 2003; Vitart, 2006; Wang *et al.*, 2009) or rain-day frequencies (Moron *et al.*, 2006; Robertson *et al.*, 2009), and dates, such as the onset of a rainfall season (Moron *et al.*, 2009). When a set of deterministic forecasts is available, the ensemble mean is often represented as the 'best-guess' forecast, and the

uncertainty in the prediction is represented by some measure of the ensemble spread, or the model outputs may be presented in the probabilistic format described below. For statistical models, prediction intervals can be used to represent the uncertainty in the prediction.

The most common probabilistic format is for three climatologically equiprobable categories to be defined using a reference ('climatological') period, and the probability that the verification will fall within each of these categories is indicated (Livezey and Timofeyeva, 2008; Barnston *et al.*, 2010). Forecasts from the Regional Climate Outlook Forums (RCOFs) are typical examples of such forecasts (Ogallo *et al.*, 2008), and this format is followed closely by many national meteorological and hydrological services (NMHSs). However, examples of climatologically unequal categories, and of the use of more than three categories, are not uncommon (Ward and Folland, 1991; Tippett and Barnston, 2008). The Australian Bureau of Meteorology (BoM) uses an equiprobable two-category system, providing the probability that the forecast parameter will be above- or below-median (Fawcett, 2008). This two-category format is followed frequently in climate change projections, where the proportion of models indicating a change in one direction is indicated; usually care is taken to communicate that this proportion should not be taken as a forecast probability.

11.2.2 Defining the predictand

For some forecasts it is not always clear whether the forecast is for an area-average or is valid for all specific locations. This problem is most common where some subjective input has been introduced into a probabilistic forecast. Since different results can be realized depending upon how the forecast is interpreted, this ambiguity is an undesirable property. In such cases it may not be possible to verify the predictand precisely as it has been defined, and a new interpretation may need to be imposed. Reinterpretation of a forecast prior to verification may be quite intentional even when there is no ambiguity. For example, seasonal forecasts are often presented as coarse spatial and temporal averages, which tend to limit their usefulness because the predictand is of little relevance in most practical settings (Vogel and O'Brien, 2006; Hansen *et al.*, 2011), but it is perfectly valid to transform or reinterpret the forecast to a variable of more direct interest, and then verify the reinterpreted forecasts. While the verification results would no longer necessarily indicate whether or not the forecasts themselves are 'good' in Murphy's (1993) 'quality' sense, they would indicate whether or not the forecasts can be successfully reinterpreted to be more directly useful for some specific purpose. See Chapter 9 for a more detailed discussion of measuring the potential usefulness of forecasts.

11.2.3 Inclusion of climatological forecasts

Climatological forecasts are frequently issued in seasonal and longer-range forecasts either because of no skill or because of no signal for the current target period. Climatological probabilities are an explicit indication that each of the possible outcomes is as likely to occur as it has done over the climatological period, and they should be seen as distinct from areas of no-forecast where no statement is made about changed or unchanged probabilities. Climatological forecasts are usually included in verification analyses, but no-forecasts excluded. However, if there are a large number of climatological forecasts, these can dominate the verification analyses, and the forecasts may score poorly because of the lack of sharpness (e.g., Wilks, 2000; Wilks and Godfrey, 2002; Livezey and Timofeyeva, 2008; Barnston and Mason, 2011). While this poor scoring is appropriate because the forecasts do not contain much useful information, it can give the impression that the occasional non-climatological forecasts are not particularly useful. When comparing forecasts (perhaps for another region or season, or from another forecast system), the climatological probabilities should be included in the analysis because credit should be given for issuing sharper forecasts if those forecasts contain potentially useful information, while if they do not the forecasts should score badly. However, if the objective is to determine whether the forecasts are believable there may be justification in omitting the climatological forecasts.

11.3 Measuring attributes of forecast quality

As discussed in Section 1.1.2, the WMO's Commission for Basic Systems (CBS) established a Standardized Verification System for Long-Range Forecasts (SVSLRF; World Meteorological Organization, 1992) as part of a set of minimum requirements for qualification as a Global Producing Centre (GPC) for long-range forecasts. The SVSLRF addresses verification requirements for deterministic and probabilistic forecasts. The recommended verification scores for deterministic forecasts are based on the mean-squared error and its decomposition into terms measuring conditional and unconditional biases and Pearson's correlation (Chapters 2 and 5). The probabilistic procedures include reliability and relative operating characteristics (ROC) diagrams. All these procedures are discussed extensively in Chapters 3 to 5, 7 and 8, and so only issues related to their specific application to seasonal forecasts are discussed in this Chapter. The WMO's Commission for Climatology (CCl) guidelines for the verification of seasonal forecasts are targeted exclusively at probabilistic forecasts (Mason, 2011). The CCl guidelines include considerable overlap with the CBS guidelines for probabilistic forecasts, and so again details of the procedures are provided in Chapters 7 and 8. The aim in this section and in Section 11.4 is to highlight some of the peculiar issues in applying such verification procedures to seasonal and longer-range forecasts.

11.3.1 Skill

As discussed in Chapter 1, there are many possible reasons for verifying seasonal and longer-range forecasts. However, by far the most dominant question in the verification of seasonal and longer-range forecasts has been whether the forecasts have any 'skill'. There are particular difficulties in addressing this question with forecasts at these timescales, and so the measurement of this attribute is considered in detail here. Of course, other attributes are of interest, and modellers in particular are often interested in more detailed analyses that can reveal systematic errors in their models. Methods for identifying conditional and unconditional errors are therefore required, but one problem that often arises in verification of seasonal forecasts is that procedures are often selected that measure multiple attributes of forecast quality making the interpretation of the results difficult. It is argued in this section that procedures that measure individual attributes are to be preferred.

The underlying objective in measuring the skill of seasonal and longer-range forecasts against a naive forecast strategy such as guessing or perpetual forecasts (always forecasting the same thing) is almost invariably to answer the question of whether the forecasts are worth considering. Unfortunately, this question has often been poorly formulated, which has resulted in frequent misinterpretation. Much of the problem is that 'skill' is a vague attribute: as discussed in Sections 1.4 and 2.7, skill is a relative concept – a forecast has skill if it is better than another set of forecasts. But better in what respect? Skill requires reference to another attribute of forecast quality, and this attribute is frequently left undefined. Instead, skill has often been imprecisely interpreted as whether the forecasts have outscored climatological forecasts or some other naive forecast strategy without considering what attributes the chosen score might be measuring. In seasonal forecasting, because of the weak levels of predictability and suboptimal quality of prediction models, the diagnosis of skill can generally be reduced to the search for some potentially useful information. For deterministic forecasts this interpretation translates to a requirement that observed values should increase and decrease at least to some degree with the forecasts, while for probabilistic forecasts categories should verify more and less frequently as the probability increases and decreases. Of course, the forecasts are potentially useful if the observations vary in the opposite direction to that implied by the forecasts, and this possibility can be measured by negative skill.

Given the limited sample sizes of operational long-range forecasts, skill is often estimated using hindcasts to obtain larger samples. There are a number of problems in trying to generate a set of hindcasts that will give accurate indications of the expected skill of operational forecasts, and these problems are discussed separately below.

When assessing the skill of seasonal forecasts regardless of their format, trends in the data have to be considered. Correlations, for example, between two series that both contain trends are likely to be non-zero, even if the year-to-year variability is not successfully predicted. Similarly, probabilistic forecasts are likely to score well if the probabilities for the category in the direction of trend are consistently inflated. If trends are ignored, spurious forecasts may easily be falsely identified as being skilful, while the quality of low-skill forecasts may be exaggerated. It is often recommended that trends be removed before any skill calculations, although an argument could be made that the successful prediction of a trend should at least be acknowledged. One solution is to measure the skill of trend and interannual components separately, and to quote both. Another, related, solution is to consider presenting the forecast with reference to a shorter and more recent climatological period, which is likely to be of more relevance to many user communities anyway since it will focus the forecast on comparisons with more readily remembered climate variability.

Skill of deterministic forecasts

By far the most commonly used skill measure for deterministic forecasts is Pearson's product moment correlation coefficient. This coefficient is discussed extensively in Section 5.4.4, and so is considered only briefly here. Pearson's correlation is implicitly scaled as a skill score, with reference strategies of perpetual forecasts and of random forecasts both having expected scores of zero. Its distributional properties for random forecasts are well known and so it is possible to calculate statistical significance analytically, and to estimate its sampling uncertainty on condition that assumptions about independence of the forecasts and the observations, and of their respective distributions, are met. Pearson's correlation is a preferred measure of choice also because of its wide use and hence familiarity, and its relationship to the percentage of explained (or predicted) variance, which provides it with a reasonably intuitive interpretation.

A further feature of Pearson's correlation is often considered an advantage: it ignores conditional and unconditional biases, which can be quite large in seasonal forecasts derived from global dynamical models, but which, in principle, should be easily correctable given a sufficient sample of forecasts to estimate the biases accurately. In practice, the correlation and the biases are not typically independent (DelSole and Shukla, 2010; Lee *et al.*, 2010), but the common variability that the correlation measures seems a reasonable minimum requirement for forecasts to have some potentially useful information: if observed values do not increase and decrease with the forecasts at least to some extent then there seems little reason to consider them.

As discussed in Chapter 5, there are a number of problems with using Pearson's correlation coefficient as a skill measure. The interpretation of the coefficient's value is complicated by the fact that it is a function not only of the potential skill of the forecasts, but also of the precise distribution of the data. An example is shown in Figure 5.5, which illustrates that large Pearson correlations can result from the influence of only a few extreme values. Although this problem can be addressed to some extent by calculating bootstrapped estimates of uncertainty in the correlation, the problem remains that the results can be misleading. Consider the example shown in Figure 11.3, which compares a set of forecasts and observations of January–March seasonal rainfall totals for 1971–2000 for Kalbarri, in Western Australia. The forecasts were calculated as the mean of 85 ensemble members, each with different initial conditions, using the ECHAM4.5 model (Roeckner *et al.*, 1996), and have a correlation with the observed rainfall of about 0.39 (90% bootstrap confidence limits of 0.08 and 0.64). If the three wettest years, which are not known *a priori*, are omitted from the analysis the correlation drops to 0.08 (−0.23 to 0.40). Is it to be concluded that virtually all the skill is provided by only 10% of the cases? What is clear is that the large bootstrap confidence intervals need to be taken seriously, especially when distributional assumptions are not strictly met, and sample sizes are small. Much of the underlying difficulty is that in small samples the most extreme values can contribute much of the total variance. In Figure 11.3, for example, the three wettest years represent over 50% of the total variance (and the wettest two years over 45% of the total), and so regardless of the quality of the forecasts the score will be heavily weighted by the forecasts on only a very few of the observations.

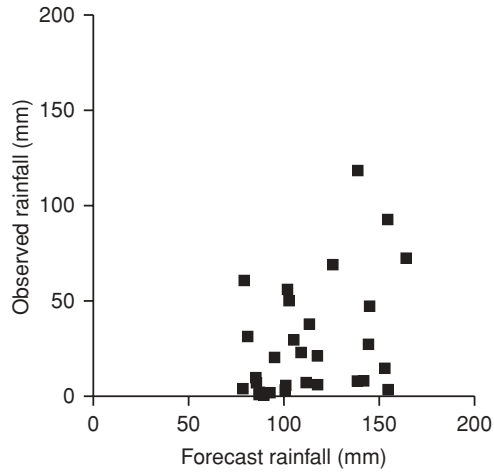

Figure 11.3 Forecasts and observations of January–March seasonal rainfall totals for 1971–2000 for Kalbarri in Western Australia (27°42′43″S, 114°09′54″E). The forecasts were calculated as the mean of 85 ensemble members, each with different initial conditions, using the ECHAM4.5 model (Roeckner *et al.*, 1996)

A difficulty with Pearson's correlation is that it imposes a stricter definition of 'skill' than the requirement defined above that observed values should increase and decrease with the forecasts. Pearson's correlation imposes the additional criterion that the observed values should increase and decrease by precisely defined amounts as the forecasts vary. Although a better set of forecasts will predict more precise increases and decreases in the observations than will a poorer set, if the objective is to identify whether forecasts are potentially useful, and if the predictability is inherently weak, the weaker skill definition is likely to be more appropriate.

As argued in Section 5.4., it would be better still not to use Pearson's correlation at all in cases where its assumptions are violated: the intuitive sense of what constitutes a 'good' correlation that experienced practitioners may have is largely irrelevant, and even misleading, when the data are not normally distributed. Instead, alternative measures of skill should be considered, specifically Spearman's and Kendall's correlations. These measures are better suited to verification of variables with non-Gaussian distributions, such as precipitation and the counts and onset dates that were mentioned in Section 11.1. While Spearman's is the more widely used of the two correlations because of its close association with Pearson's correlation, the advantages of Kendall's τ are worthy of consideration. In addition to the advantages listed in Section 5.4.6, Kendall's τ has an intuitive interpretation: Kendall's τ (depending on how ties are handled, as discussed in the following sub-section) can be transformed simply to a scale ranging from 0 to 1 to represent the probability that the forecasts successfully discriminate the larger value of any two observations (Mason and Weigel, 2009). A second, related advantage is that Kendall's τ has close affinities to other widely used verification measures such as the area beneath the ROC curve (Chapters 2 and 7), as discussed in the following subsection.

Skill of probabilistic forecasts

The most commonly used skill measure for probabilistic forecasts is the ranked probability score (RPS), and its skill score (RPSS; Sections 7.3.2 and 8.4.2), although the ignorance score (Sections 7.3.2 and 8.4.2) and similar information theory-based scores are becoming increasingly popular. In the context of objective forecasts, in which the forecast probabilities are estimated by counting the proportion of ensemble members predicting a value in each of the categories, the RPSS is biased because of reliability errors that result in turn from sampling errors in estimating the forecast probabilities given limited ensemble sizes (Section 8.4.3). Adjustments can be made to the score to remove this source of bias, but such an option is not available for subjectively derived probabilities. The need for the correction points to one of the difficulties in interpreting the RPS and its skill score: they measure multiple attributes, and so forecasts can score imperfectly if the forecasts are good in some respects, but poor in others. The RPS, Brier score, and ignorance score can each be decomposed into reliability, resolution and uncertainty terms (see Chapter 7 for further details), and in each case skill is achieved if the resolution term is larger than the reliability term. While this requirement for skill may be meaningful in some contexts, it is unnecessarily strict when trying to identify whether forecasts might be potentially useful (Mason, 2004). When verifying seasonal forecasts, which generally

Table 11.1 Three idealized sets of seasonal forecasts for above-median rainfall, and corresponding observations (1 indicates above-median and 0 indicates below-median), with Brier score calculations for the base rate (50%) and for the three sets of forecasts

		Forecasts			Brier score			
Year	Observed	I	II	III	Base rate	I	II	III
1	1	80%	60%	60%	0.25	0.04	0.16	0.16
2	1	80%	60%	60%	0.25	0.04	0.16	0.16
3	1	80%	60%	60%	0.25	0.04	0.16	0.16
4	0	80%	60%	40%	0.25	0.64	0.36	0.16
5	0	80%	60%	40%	0.25	0.64	0.36	0.16
6	0	20%	40%	40%	0.25	0.04	0.16	0.16
7	0	20%	40%	40%	0.25	0.04	0.16	0.16
8	0	20%	40%	40%	0.25	0.04	0.16	0.16
9	1	20%	40%	60%	0.25	0.64	0.36	0.16
10	1	20%	40%	60%	0.25	0.64	0.36	0.16
				Average	0.25	0.28	0.24	0.16

suffer from overconfidence (i.e. poor reliability) and weak resolution, skill scores can often be negative, and there is then a danger of rejecting potentially valuable forecasts as useless.

Consider an idealized example in which ten seasonal rainfall forecasts are to be evaluated against corresponding observations. For the sake of simplicity it will be assumed that there are only two equiprobable categories. In one set of ten forecasts (marked I) five of the ten forecasts indicate an 80% probability of above-median rainfall, while the remaining five indicate a 20% probability (Table 11.1). Above-median rainfall occurs on 60% (i.e. three out of five) of the occasions that the forecast indicated an 80% probability, and on 40% (i.e. two out of five) of the occasions that the forecast indicated a 20% probability. The 80% forecasts correctly indicated an increase in the probability of above-median rainfall, and the 20% forecasts correctly indicated a decrease, but did so (somewhat typically for seasonal forecasts) overconfidently. Brier score (Section 7.3.2) calculations are shown in the Table (similar results are obtained using the ignorance score), and the skill score is -0.12, which suggests that the forecasts are worse than climatological forecasts. If the forecasts had been perfectly reliable (marked II), the score would naturally improve despite there being no gain in resolution, and the skill becomes marginally positive (0.04). Similarly the skill can be raised by improving the resolution at the cost of reliability: for forecast set III the forecasts are under-confident, but have maximum resolution, which more than offsets the loss in reliability (the Brier skill score is 0.36). The progression in skill from set I to set III indicates that skill can increase on this measure, but with no indication of whether that is because reliability or resolution has improved, and there is no guarantee that either of these attributes has not deteriorated.

From a Brier and ignorance score perspective forecast set I is worse than information only about the base-rate, but to conclude that it would therefore be better not to have the forecasts at all is surely incorrect: the forecasts successfully indicate increased and decreased chances of above-median rainfall. The problem with set I is that the reliability errors are larger than the gain in resolution, but because both components are being measured together the resolution may be missed unless the skill is diagnosed carefully. Such difficulties in interpretation result from the skill scores imposing an arbitrarily high maximum acceptable level of reliability error for a given level of resolution. The situation is somewhat analogous to that of Pearson's correlation, which requires the observations to increase and decrease by precise amounts along with the forecasts, rather than just to increase or decrease; so also the reliability terms in the Brier and ranked

probability skill scores require the predicted events to be more and less frequent by precise amounts as the forecast probability increases and decreases.

Although the previous examples illustrate that when reliability is measured with resolution there are difficulties in interpreting the result, the reliability term cannot be completely ignored for now. On its own the resolution term is not generally considered a satisfactory indication of skill: as long as the observed relative frequency is much higher for some forecast probabilities than for others the resolution term of the Brier, RPS and ignorance scores is large. In effect, the resolution term is measuring whether the expected observation differs given different forecasts, regardless of whether or not the observations vary arbitrarily with the forecasts. Imposing the requirement that the observed relative frequency should increase as the forecast probability increases therefore seems quite reasonable.

The interpretation problems that can affect scores that measure multiple attributes, or scores such as the resolution score that have an unsatisfactorily weak definition of skill arise only if such scores are calculated in isolation: when accompanied by analyses of reliability diagrams (Section 7.6.1), for example, the scores can be valuable summaries, and their decompositions can be informative. A primary difficulty in measuring resolution and constructing reliability diagrams for seasonal and longer-range forecasts is the severe sample-size restriction. The sampling errors in constructing the graph are likely to be prohibitively large, at least for some of the points (Section 7.6.1; Bröcker and Smith, 2007a). One possible solution is to bin the forecasts into only a few bins, although there is then likely to be a deterioration in skill (Stephenson *et al.*, 2008b). Typically reliability diagrams can only be constructed meaningfully by pooling forecasts over large areas. When pooling forecasts, corrections need to be made for the effects of decreasing grid areas towards the poles, either by sampling fewer grids at higher latitudes (Wilks, 2000; Wilks and Godfrey, 2002) or by weighting each grid by its area (Barnston *et al.*, 2010; Barnston and Mason, 2011).

A more widely adopted approach is to calculate the frequency of hits only for the category with the highest probability (e.g. Livezey and Timofeyeva, 2008). If one of two categories with tied highest probabilities verifies, a half-hit is usually scored, or a third-hit if one of three categories with tied highest probabilities verifies. In some of the RCOFs a half-hit is scored if the middle category has the highest probability, and one of the outer categories verifies, but the probability for that category is higher than for the other extreme. Instead of redefining the score, and thus complicating its interpretation, a more detailed perspective of the resolution of the forecasts could be obtained by calculating the frequency of hits for the highest probability category, but also calculating how often the category with the second highest probability verifies, etc., through to how often the category with the lowest probability verifies.

A widespread criticism of seasonal forecasts is that the sharpness of the forecasts is low (or overconfident when sharpness is strong). A major problem with the frequency of hits is that it does not consider sharpness at all. However, once the probabilities themselves are considered it is difficult to avoid mixing measurement of resolution and reliability. As an alternative, measures of discrimination could be considered. Whereas resolution (in its more strict sense defined earlier) indicates whether the frequency of a category occurring increases or decreases with its forecast probability, so discrimination indicates whether the forecast probability increases and decreases as the category increases or decreases in frequency (see Chapter 2). As discussed below, although measures of discrimination are insensitive to sharpness, they do at least consider the rankings in the probabilities, but the analysis is not complicated by consideration of reliability. One other advantage of measuring discrimination instead of resolution for typical seasonal forecast formats is that it is easier to measure the conditional distribution of the forecasts on the observations than vice versa because there are usually only three possible outcomes (or very few) whereas there are many possible forecast probabilities. Given small sample sizes the sampling errors in the conditional distribution of the forecasts are therefore likely to be smaller than in the conditional distribution of the observations.

The relative operating characteristics (ROC) graph (Chapter 3), and the area beneath its curve, are widely used measures of discrimination in seasonal forecast verification, and these procedures are explicitly recommended in the SVSLRF and in the

CCl verification guidelines (Mason, 2011). As a measure only of discrimination, the area beneath the ROC curve is insensitive to some reliability errors (Kharin and Zwiers, 2003; Glahn, 2004), which may render it an inadequate summary measure of forecast quality, but is a distinct advantage when combined with measures of other attributes. Its insensitivity to the overconfidence that is commonly observed in seasonal forecasts makes the score useful for identifying skill, and the graphs can be helpful in more detailed diagnoses of forecasts at these timescales (Mason and Graham, 1999; Kharin and Zwiers, 2003).

The ROC area is calculated under the assumption of a two-category forecast system, and separate ROC areas can be calculated for each of the categories. Since most probabilistic seasonal forecasts have three or more categories, a generalized version of the ROC area may be a useful summary of the discriminatory power of the forecasts. This generalized discrimination score (Mason and Weigel, 2009) calculates the probability that given two observations the forecasts can successfully discriminate the observation in the higher category. For example, assuming that the predictand is rainfall, what is the probability of successfully discriminating the wetter of two observations? In the classical ROC, the test can be applied in a three-category system, for example, to calculate the probability that an above-normal observation could be successfully distinguished from an observation that was not above-normal, but a separate test would have to be conducted to distinguish normal and below-normal observations. A normal and below-normal observation would therefore be treated as indistinguishable when the ROC test is applied to above-normal events. In the generalized version of the test all the categories can be distinguished.

The generalized discrimination score, D, can be calculated as follows. Assume a forecast system with m mutually exclusive and exhaustive categories (i.e., each observation has to be in one and only one of the categories). As mentioned, m typically is 3, but regardless of how many, the categories are ranked from lowest values (category 1) to highest (category m). Each forecast is a vector of probabilities, $\mathbf{p}$, which consists of m probabilities, one for each category, which must total to 1.0. Next assume that category k verified n_k times, and that the i^{th} of the n_k vector of probabilities for when this category verified is given by $\mathbf{p}_{k,i}$. The generalized discrimination score can be defined as

$$D = \frac{\sum_{k=1}^{m-1} \sum_{l=k+1}^{m} \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} I\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right)}{\sum_{k=1}^{m-1} \sum_{l=k+1}^{m} n_k n_l} \tag{11.1}$$

where

$$I\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) = \begin{cases} 0.0 & \text{if } F\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) < 0.5 \\ 0.5 & \text{if } F\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) = 0.5 \\ 1.0 & \text{if } F\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) > 0.5 \end{cases} \tag{11.2}$$

and where

$$F\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) = \frac{\sum_{r=1}^{m-1} \sum_{s=r+1}^{m} p_{k,i}(r)\, p_{l,j}(s)}{1 - \sum_{r=1}^{m} p_{k,i}(r)\, p_{l,j}(r)} \tag{11.3}$$

In Equation 11.3 $p_{k,i}(r)$ is the forecast probability for the r^{th} category, and for the i^{th} observation in category k.

It can be shown that when $m = 2$, Equation 11.1 reduces to the area beneath the ROC curve (Mason and Weigel, 2009). Similarly, if m is set in Equations 11.1–11.3 to the number of observations (assuming there are no ties), and if the forecasts are deterministic (represented by x with a subscript) then D becomes

$$D = \frac{2 \sum_{k=1}^{n-1} \sum_{l=k+1}^{n} I\left(x_k, x_l\right)}{n\,(n-1)} \tag{11.4}$$

which is related to Kendall's τ by

$$\tau = 2D - 1 \tag{11.5}$$

(Mason and Weigel, 2009). These relationships are examples of how the generalized discrimination score is essentially equivalent to various so-called non-parametric statistical tests. For example, Kendall's τ was compared with Pearson's correlation in the previous section, where it was explained

Table 11.2 Example of seasonal forecasts and corresponding observations (A indicates above-normal, N indicates normal, and B indicates below-normal), with cumulative profits and losses based on an initial investment of $100. The interest earned or lost each year is shown in the last column, together with the effective rate of interest over the ten years

		Forecasts				
Year	Observed	B	N	A	Profit	Interest
					$100.00	
2001	A	20%	50%	30%	$90.00	−10.0%
2002	A	20%	55%	25%	$67.50	−25.0%
2003	A	25%	35%	40%	$81.00	20.0%
2004	B	15%	30%	55%	$36.45	−55.0%
2005	N	45%	35%	20%	$38.27	5.0%
2006	A	20%	50%	30%	$34.45	−10.0%
2007	N	35%	40%	25%	$41.33	20.0%
2008	A	20%	50%	30%	$37.20	−10.0%
2009	A	25%	35%	40%	$44.64	20.0%
2010	B	40%	35%	25%	$53.57	20.0%
					Effective	−6.1%

that Kendall's τ is a correlation based on the ranked values. Similarly, the area beneath the ROC curve is equivalent to the Mann–Whitney U-statistic (Mason and Graham, 2002), which is a non-parametric version of the more widely used Student's t-test for comparing central tendencies. When applied to forecasts, the U-test assesses whether there is any difference in the forecasts when an event occurs compared to when the event does not occur (and, thus, whether the forecasts can discriminate between events and non-events). More specifically, it indicates whether the forecast (whether probability or value) was higher, on average, when an event occurred compared to when not.

Thus, by using the generalized discrimination score, a consistent test can be applied to measure whether forecasts of virtually any format have skill in the sense defined earlier: do the observations increase, whether in value or in frequency, as the forecast value or probability increases, without specifying by how much the increases and decreases should be? This measure is useful in low-skill settings, where it may be acknowledged upfront that the forecasts themselves may be poorly calibrated, whether overconfident or biased. It is helpful, and more informative, to measure the quality of the calibration separately.

To illustrate the importance of considering calibration separately when testing whether forecasts may be potentially useful, consider a fictional set of ten probabilistic forecasts and corresponding observed categories as shown in Table 11.2. The ranked probability skill score (RPSS) for these forecasts is marginally negative (approximately −0.005); similarly the Brier skill scores for all three categories are negative. These results suggest the forecasts are effectively useless. However, given that the category with the highest probability occurs four times, while that with the lowest probability occurs only once, it seems reasonable to assume that the forecasts may have some useful information. The RPSS and Brier scores are negatively impacted by what appears to be hedging on the normal category. Acknowledging that the probabilities are poorly calibrated, but that increases and decreases in probabilities may be meaningful, the generalized discrimination score can be used to indicate whether the forecasts may be potentially useful.

Instead of comparing each forecast with its corresponding observation, as is typical of most verification scores, Equation 11.1 is calculated by comparing each year with all other years that have different observations. For example, starting with

years with below-normal rainfall ($k = 1$), the first year available ($i = 1$) is 2004. This year is compared to all the years with normal rainfall ($l = 2$), the first of which is 2005. Given that the observations differ, Equation 11.2 indicates whether the forecast for 2005 successfully indicated that 2005 was likely to be the wetter of the two years. The answer to this question is based on the probability that a value randomly drawn from the distribution represented by the forecast for 2005 will exceed one randomly drawn from that represented by the forecast for 2004, conditioned upon the two values differing (Equation 11.3). For 2004 and 2005, Equation 11.3 gives approximately 0.20. Because this value is less than 0.5, the forecasts fail to discriminate the year with the higher rainfall category (Equation 11.2). Proceeding to the next year with normal rainfall ($l = 2$), 2004 is compared with 2007. For these two years Equation 11.3 gives 0.25, and so again the discrimination is incorrect. Since there are no more years with normal rainfall ($n_2 = 2$), 2004 is then compared with all the years with above-normal rainfall ($l = 3$). This procedure is then repeated for 2010 ($i = 2$), and then the years with normal rainfall ($k = 2$) are compared with the years with above-normal rainfall ($l = 3$). For the example, $D \approx 0.68$, indicating that the forecasts discriminated the observed categories with a success rate of about 68%, and suggesting that the forecasts may be potentially useful.

Skill of hindcasts

It has long been recognized that in-sample estimates of performance provide overestimates of operational performance (Allen, 1974; Davis, 1976; Rencher and Pun, 1980; Wilkinson and Dallal, 1981). The need for out-of-sample estimates of skill is more of an issue with statistical models than it is with dynamical models (except in the context of skill-weighted multi-model combinations, which is essentially a statistical procedure anyway) because statistical models generally can be more easily tuned to compare favourably with the verification data. However, because dynamical model parameterizations are generally tuned to optimize performance over a verification period, independent verification is still required.

Cross-validation (Section 1.4.2) is the most commonly used method of trying to obtain independent estimates of predictive skill (Michaelsen, 1987). In the atmospheric sciences the most common approach to cross-validation is to predict each observation once, omitting that observation, and possibly some adjacent observations, to reconstruct the model. The re-specification of the model at each cross-validation step should involve not only recalculating the model parameters, but also reselecting the predictors to be included (Elsner and Schmertmann, 1994). In any predictor selection procedure there is a danger of selecting additional spurious predictors or the wrong predictors entirely. As the candidate pool of predictors is enlarged, the danger of choosing spurious predictors increases, and thus the probability of the hindcast skill estimates overestimating those of the operational skill also increases (Barnett *et al.*, 1981; Katz, 1988; Brown and Katz, 1991). The same is true of cross-validated skill estimates if the procedure is not implemented carefully. The problem can be reduced by leaving more than one observation out at each step, but there are few guidelines as to how many observations should be omitted. Xu and Liang (2001) recommend omitting as much as 40–60% of the observations, and even more if the candidate pool of predictors is large. This proportion may be impractical given the small sample sizes available for seasonal forecasting, but the clear message is that, given the vast pool of candidate predictors many modellers consider, the risk of overestimating operational performance is high.

An alternative to cross-validation is the verification of retroactive forecasts (Mason and Baddour, 2008). Retroactive forecasts are generated by withholding the later part of a data set, selecting and parameterizing the model on the first part of the data, and then predicting the subsequent values, possibly repeating the model construction process as observations from the second part are predicted. This process attempts to reproduce the forecasts that would have been made operationally given access to current data sets and models (Mason and Mimmack, 2002; van den Dool *et al.*, 2003). There are, however, two sources of bias. Firstly, even if implemented properly, the procedure is likely to underestimate operational performance because the model should improve gradually over time as more data

become available. A more serious source of bias, however, occurs because it is virtually impossible to avoid including predictors based on knowledge of their association with the predictand over the full sample period. Since some of these predictors may be spurious, it is essential that there is a strong theoretical base to their selection prior to producing any hindcasts.

In conclusion, all hindcasting procedures will unavoidably have some biases in their estimates of operational forecast skill. While there are sources of both positive and negative bias, the positive biases are likely to outweigh the negative given how hindcasts are most frequently calculated. One specific recommendation is that leave-one-out cross-validation should almost always be avoided even if there are no problems with temporal autocorrelation. Further research is required to make more specific recommendations about how many years to omit in a cross-validation procedure, but considerably larger numbers than those most frequently used almost certainly need to be considered, especially when the candidate pool of predictors is large. Retroactive skill estimates are normally to be preferred to cross-validated estimates because they have fewer of the problems outlined above (Jonathan *et al*., 2000). They are not calculated as often as cross-validated skill estimates because of limited sample sizes, but retroactive verification is worth attempting even if only 5 or 10 years are predicted (Landman *et al.*, 2001; Landman and Goddard, 2002; Shongwe *et al*., 2006), and even very wide uncertainty estimates on verification scores can be useful information.

11.3.2 Other attributes

If skill is to be defined in terms of a single attribute (discrimination, or possibly resolution), as proposed in Section 11.3.1, it is essential to measure additional attributes subsequent to concluding that the forecasts may be at least worth considering. The measurement of accuracy and reliability associated with the central tendency of the ensemble and of its distribution can be addressed using procedures detailed in Chapters 7 and 9. For probabilistic procedures, Chapter 8 provides extensive coverage of options for diagnosing over- and under-confidence, and unconditional biases. Attributes or reliability diagrams are particularly useful in this regard, although forecasts will inevitably have to be pooled over large areas and possibly different seasons to allow for sufficiently large sample sizes.

One aspect of seasonal forecasts that is of interest and is partly a reflection of limited sample size, and partly of longer-term variability, is the degree to which the seasonal forecasts over a limited period of perhaps a few years have indicated the extent to which the observed climate over this period has differed from that of the reference climatological period. For example, if the forecasts have successfully indicated that the verification period would be generally dry, some skill should be acknowledged, but this may not be identified using some of the procedures described above. In areas of significant decadal variability or with long-term trends, for example, the discrimination skill may have been poor because of an inability to distinguish which years are drier than others when all or most of the years are dry. Measurement of the unconditional bias in the forecasts is appropriate in this regard, and procedures for measuring the bias of deterministic forecasts are described in Chapter 5. For probabilistic forecasts, tendency diagrams provide a simple visual indication of any unconditional bias (Mason, 2011). These diagrams compare the average forecast probabilities for each category with their observed relative frequencies; if the forecasts had been reliable, one would expect the observed relative frequencies to be approximately equal to the average probabilities. In the example provided in Figure 11.4 based on the data in Table 11.2, it is evident that above-normal rainfall occurred much more frequently than the other categories, but the forecasts implied that the normal category would be most frequent (perhaps a reflection of a tendency to hedge).

One attribute that is of common interest subsequent to the demonstration of at least some skill, is the potential economic value of forecasts. Appropriate measures are discussed in Chapter 9, but one that is particularly well suited to the standard probabilistic format of seasonal forecasts is the effective interest rate (Hagedorn and Smith, 2009; Tippett and Barnston, 2008), which in turn is based on the ignorance score (Roulston and Smith, 2002; Benedetti, 2010). The ignorance score, *Ign*, can be

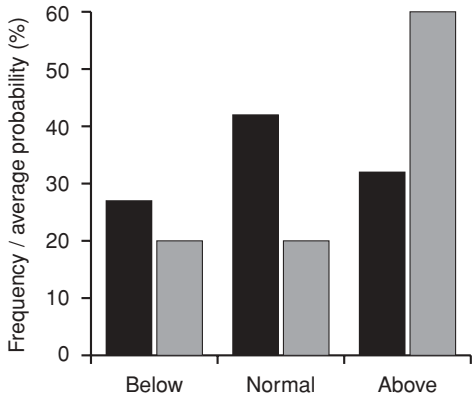

Figure 11.4 Example tendency diagram for the data from Table 11.2. The black bars show the average forecast probabilities for each category, and the grey bars show the observed relative frequencies for each category

transformed to the effective interest rate using

$$\text{effective interest rate} = \left(2^{Ign(ref)-Ign} - 1\right) \quad (11.6)$$

where $Ign(ref)$ is the ignorance score for the reference (base-rate forecasts). The effective interest rate provides an indication of the average returns an investor would make if (s)he invested on the forecasts, and received fair odds against the climatological probabilities. For example, given three equiprobable categories, the returns on the verifying category would be three times the investment. The investor will then make a profit whenever the forecast probability on the verifying outcome exceeds the base rate. For the data from Table 11.2, the effective interest rate is about -6% per year, suggesting that the forecasts are not useful.

Equation 11.6 is only valid if the forecasts are for a single location and if all the forecasts are for discrete periods (e.g. a specific 3-month season over a number of years) since it assumes that earnings (and losses) are carried over from forecast to forecast. If some of the forecasts are for different locations or for overlapping periods (or, more specifically, if any of the target periods expire after any of the release dates for subsequent forecasts), then the initial investment has to be divided between each of the s locations and periods, and the effective interest rate has to be averaged using the ignorance score for each instance:

$$\text{average effective interest rate} = \frac{1}{s}\sum_{k=1}^{s}\left(2^{Ign(ref)-Ign_k} - 1\right) \quad (11.7)$$

where Ign_k is the ignorance score for the k^{th} location/season. For the data in Table 11.2, the average interest would have been -2.5% if independent investments had been made on each forecast. However, as discussed in Section 11.3.1, the forecasts do have good discrimination and so could potentially be useful if they could be calibrated reliably.

Even with very good forecasts, the investor could occasionally make a loss because categories with probabilities lower than the base-rate should verify sometimes (otherwise they would be unreliable). However, in the long run, if the forecasts are good, the gains will exceed the losses, and the effective interest rate will be greater than zero. Given that the returns on the investments each time are a direct function of the forecast probability, in order for the effective interest rate to be positive the reliability of the forecasts is important, and the forecasts therefore must have skill higher than the minimum requirement as defined in Section 11.3.1. A plot of gains and losses over time provides a useful graphical illustration of potential forecast value. Such a graph can be constructed by plotting

$$\left(\prod_i \left(\frac{1}{s}\sum_{k=1}^{s}\frac{p_{k,i}}{c_{k,i}}\right)\right) - 1 \quad (11.8)$$

on the y-axis against time, i, on the x-axis, where s is the number of locations/seasons, $p_{k,i}$ is the forecast probability for the verifying category at location/in season k, and c_i is the corresponding base rate. An example is provided in Figure 11.5, with corresponding data in Table 11.2, using the same forecasts and observations as for the generalized discrimination score example.

11.3.3 Statistical significance and uncertainty estimates

Regardless of whether it is the skill of operational forecasts or of hindcasts that is being estimated,

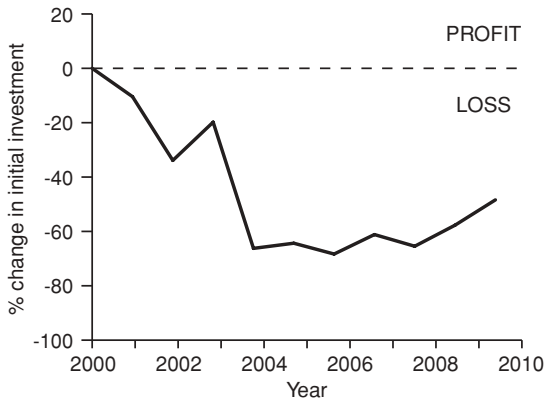

Figure 11.5 Cumulative profits and losses diagram based on data from Table 11.2

some indication of whether the measured skill provides a basis for concluding that the forecasts are good or bad is required (Section 1.4.3). The preferred approach has been to calculate statistical significance, or *p*-values, which indicate the probability that a result at least as good as that measured could have been achieved by chance. Some of the problems in interpreting *p*-values when sample sizes are large (Mason, 2008) are rarely an issue for seasonal and longer-range forecasts, but other problems of interpretation remain (Nicholls, 2001; Jolliffe, 2004, 2007). Confidence intervals remain underutilized. Statistical procedures for calculating *p*-values and confidence intervals for the measures discussed above are described in other chapters, and so only a few comments are included here that pertain specifically to seasonal forecasts.

When skill is calculated for specific seasons and locations, temporal autocorrelation in the data is often not a major problem except in the presence of trends, and so confidence intervals and *p*-values can often be calculated using distributional assumptions (Jolliffe, 2007) when available. However, when data are pooled from different locations, or if field significance is being assessed (Livezey and Chen, 1983; Wilks, 2006a), then spatial correlation has to be accounted for. Similarly, if data from overlapping or adjacent seasons are being pooled, temporal correlation can affect the results, and block bootstrapping may be required (Barnston and Mason, 2011).

Statistical significance for differences in model skill levels is rarely calculated. When considering whether a model revision improves the forecast skill compared to an earlier version it is generally impractical to demonstrate a significant improvement because the uncertainty estimates on the skill levels are so large due to limited sample sizes, and so any improvement in skill may be acceptable. However, when considering how to weight different models in some form of skill-based multi-model average the need to demonstrate robust differences in model skill levels becomes more important, otherwise the unequally weighted model average will reflect sampling errors in differences in model skill, and will therefore likely perform less well than an equally weighted average (Kang and Yoo, 2006; Weigel *et al.*, 2010).

11.4 Measuring the quality of individual forecasts

It is a perfectly reasonable question to ask whether the forecast for a specific season was good or bad even if the forecasts are probabilistic. Mathematically, most of the probabilistic verification scores discussed in this book could be calculated using forecasts representing different locations rather than different times. However, many such calculations would involve incorrect interpretations of the forecasts. For example, consider a set of forecasts for ten locations all for the same season, and all of which indicate an 80% probability of above-median rainfall. If above-median rainfall occurred at 60% of the stations rather than at 80%, one cannot necessarily conclude that the forecasts were overconfident: the forecasts were not stating that 80% of the area would be wet, only that at each station on 80% of the occasions on which an 80% probability of above-median rainfall is issued can above-median rainfall be expected to occur. Any attempt to measure the reliability of the forecast probabilities by considering the forecasts at different locations for an individual season represents an incorrect interpretation of the forecast.

The primary reason why the reliability of the forecasts cannot be measured by verifying spatially rather than temporally is that in most practical settings forecasts for different locations will not be independent. In effect, there is a sample size

problem: because of strong spatial correlation there are very few independent realizations in a forecast for any individual season.

Unfortunately, although the generalized discrimination score may be useful for a relatively short series of forecasts, there may be problems of interpretation when it is used for verifying individual forecasts, especially when the spatial domain is small. When a discrimination score is applied spatially, it indicates whether the forecast correctly discriminated wet from dry areas, not whether the forecast gave a good indication of whether the specific season would be unusually wet or dry. The forecast may successfully have indicated a high likelihood of unusually wet or dry conditions over the entire domain, but this information is ignored because of the score's insensitivity to calibration, To answer this second question, the specific contribution that the forecast in question would make to the generalized discrimination score could be calculated: the probability that the specific season of interest would have been correctly identified as wetter or drier (or warmer or colder, or whatever) than each other season for which forecasts are available, could be calculated. [To calculate this probability, set k or l in Equation 11.1 to the verifying category, and then n_k or n_l to 1.]

The frequency of hits is widely calculated as a verification measure for individual forecasts. The score indicates the proportion of the area in which the verifying categories had the highest probability, but, as discussed in Section 11.3.1, the score is more informative when scores for the second highest probability category, etc., are calculated. The frequency of hits for the various probability rankings still ignore much of the information in the probabilities, and are unable to credit sharp probabilities. The linear probability score (Wilson *et al.*, 1999) and the average interest rate (Equation 11.7) are worth considering, despite the fact that both scores lack propriety (Bröcker and Smith, 2007b). This lack of propriety is not necessarily a problem if no attempt is made to optimize these values or conclude naively that one forecast is better than another simply because the score is higher. Scores for individual years are generally calculated to tell us something about the temporal variability of predictability (Livezey and Timofeyeva, 2008), just as in Figure 11.4, for example.

Instead of calculating scores for individual seasons, much more can be discerned from a detailed diagnostic of dynamical model outputs (Jakob, 2010). By diagnosing the model's atmospheric structure, useful insights into its strengths and weaknesses can be derived (e.g., Lyon and Mason, 2009).

11.5 Decadal and longer-range forecast verification

For forecasts at decadal and longer timescales there are at best too few realizations to perform any meaningful significance testing on the kind of scores described above. Hindcasting is not a realistic option to expand the sample size of decadal forecasts because of the lack of subsurface ocean observations required to initialize the models (Smith *et al.*, 2008), and unpredictable events such as major volcanic eruptions add an important noise component to the observed climate that exacerbates the sampling problem. Traditional verification analyses that compare a set of forecasts with the corresponding observations may therefore not be a viable option. However, there are some evaluations that can be usefully performed that may not directly answer the most immediate questions of interest regarding forecast quality at these long timescales, but do at least provide some information that may help in deciding whether the forecasts are worth considering (Fildes and Kourentzes, 2011).

A common starting point in place of rigorous verification analyses is some measurement of consistency in predictions. Perhaps the simplest such measure that is widely used for climate change projections is the proportion of models agreeing upon the sign of the anomaly in the target variable (Whetton *et al.*, 2007; Hawkins and Sutton, 2009). It is usually assumed that if this proportion is close to 50% then there is little agreement between the models, and so confidence in any prediction should be low. However, this procedure rests upon the rather unreasonable assumption that the models are independent, and upon the only sometimes reasonable assumption that the underlying data are symmetrically distributed. Further, in the situation that models are closely agreed upon minimal change, the level of agreement in the sign of the anomaly may

be very low, and some measure of spread in the predictions would be more informative.

Following a similar principle of consistency in predictions, but involving more sophisticated diagnostics, so-called 'perfect model' experiments test how well the model is able to predict one of its own ensemble members (typically as measured by the root mean squared error) as additional data are assimilated into the model (Dunstone and Smith, 2010; Meehl *et al.*, 2010). If the model is able to predict its own behaviour more successfully as more data thought to be relevant to predictability are assimilated then there is some basis for suspecting that the assimilated data indicate a process of variability in the real world that the model may be able to predict. This assumption is, of course, problematic (Stainforth *et al.*, 2007), but there seems little point in verifying a model that cannot even predict itself, and in the absence of any verification results against real-world observations, such improvements in the signal-to-noise ratio provide some grounds for credibility. However, ensemble-member and inter-model consistency should, at best, be considered a very weak form of validation, and at least some attempt at comparison between model outputs and observed data should be made (Fildes and Kourentzes, 2011).

A starting point of any verification procedure should be to evaluate the accuracy with which the model's climatology matches that of the observations (Caminade and Terray, 2010; Gent *et al.*, 2010). If possible, this assessment should be performed over a number of climatological periods to test for robustness of results, especially if the skill levels of models are being compared (Macadam *et al.*, 2010). The mean squared error and its decomposition into conditional and unconditional biases (Chapter 5) can be used, although the interpretation of results given non-normally distributed data can be complicated, and the calculation of absolute errors may be more appropriate (Section 5.3). A variety of other statistics have been proposed, all generally based on mean squared or absolute errors, and differences are largely a matter of scaling, and sensitivity to extremes (Watterson, 1996). Regardless of the measure, if a model is not reproducing the observed climate realistically, there is no compelling reason to assume that simulated variability and change in its climate will match that of the real world. This assertion is certainly borne out at the seasonal timescale, where skill is a function of models' unconditional biases (DelSole and Shukla, 2010; Lee *et al.*, 2010), but it is not necessarily the case that an accurate model climatology implies forecast skill (Knutti *et al.*, 2010).

Even though the number of realizations may be trivially small, it is still worth calculating verification scores with whatever data are available. While it may be impossible to demonstrate statistically significant skill, the extent to which the models improve their simulation of the observed large-scale climate variability as improved data sets are assimilated, for example, reinforces the belief that the models may be able to make useful predictions (Doblas-Reyes *et al.*, 2006; Keenlyside *et al.*, 2008; Mochizuki *et al.*, 2010). Some account may need to be taken for the loss of skill resulting from the unpredictability of major volcanic eruptions and their effects on climate. Selecting start dates that avoid periods with major eruptions (Troccoli and Palmer, 2007) may be useful for model validation, but gives a biased estimate of operational forecast skill.

While there is no guarantee that models that produce skilful forecasts at one timescale will be skilful at other timescales, verification information for timescales for which more data are available can be informative. For example, predictability at decadal timescales is premised partly on the ability to predict sea-surface temperatures from subsurface conditions (Meehl *et al.*, 2009), and so skill at the seasonal scale, which depends largely on ocean-atmosphere coupling, may provide some indication of skill at decadal scales (Palmer *et al.*, 2008; Caminade and Terray, 2010). Of course, there may be other sources of predictability at decadal scales such as 'committed climate change' and future greenhouse gas emissions (Meehl *et al.*, 2009), the effects of recent volcanic eruptions (Troccoli and Palmer, 2007), land-surface feedbacks and cryospheric effects, and so more detailed diagnostics are therefore usually to be recommended (Scaife *et al.*, 2009), including investigations into the processes of climate variability (Giannini, 2010). Even where there is no discrimination or resolution skill at seasonal timescales, information about the reliability of the ensemble spread provides some basis for assessing the reliability of the spread at longer timescales (Palmer *et al.*, 2009).

11.6 Summary

Although many of the procedures used in seasonal forecast verification are similar to those used at shorter timescales, problems of limited sample size and low levels of predictability are invariably major factors in verification analyses at these and longer timescales. Both limitations contribute to a strong focus on measuring 'skill', although conclusions can be misleading if skill is not precisely defined. It has been argued in this chapter that widely used definitions of 'skill' for seasonal forecasts are unduly strict, and that some commonly used verification measures may therefore not be the most appropriate ones to use. For deterministic forecasts, for example, skill can be defined as increases and decreases in the observed values as the forecasts increase and decrease. This definition points to a measure of association based on the ranks of the forecasts. Similarly, for probabilistic forecasts, skill can be defined as increases and decreases in the frequency of events or a verifying category as the probability increases and decreases. This definition can be measured either by resolution or by discrimination, although the latter is usually easier to measure when sample sizes are small. In either case, it is helpful, and more informative, to consider the measurement of reliability as a separate verification question. The generalized discrimination score is proposed since it can be applied to an extensive range of forecast and verification data formats, and provides a useful indication of whether there is any potentially useful information in the forecasts. Separate tests for conditional and unconditional biases, and other reliability checks should be applied subsequently.

Partly because of the infrequency with which seasonal and longer-ranger forecasts verify, there is widespread interest in whether a specific forecast was good or bad. When forecasts are expressed probabilistically this question becomes complicated because attributes such as resolution, discrimination and reliability can change their meaning, and may become inappropriate. Much of the difficulty arises from the fact that the number of spatially independent forecasts is likely to be very low, and so these attributes cannot be measured meaningfully. However, some measures can be informative, including ones that are not strictly proper, as long as they are interpreted appropriately and their limitations recognized.

At longer timescales the sample sizes can become so small that no meaningful verification results can be realized. However, even in these cases measures of adequacy of model climate, and of forecast and/or model consistency can be helpful. It is also possible to use verification results for shorter timescales to provide some indication of credibility.