Guidance on Verification of Operational Seasonal Climate Forecasts

Simon J. Mason

International Research Institute for Climate and Society

Draft: 19 November 2008

Revision: 10 February 2009

Final revision: 13 August 2013

Prepared under the auspices of

World Meteorological Organization, Commission for Climatology XIV

Expert Team on CLIPS Operations, Verification, and Application Service

Executive Summary

The purpose of this document is to describe and recommend procedures for the verification of operational probabilistic seasonal forecasts, including those from the Regional Climate Outlook Forums (RCOFs), National Meteorological and Hydrological Services (NMHSs) and other forecasting centres. The recommendations are meant to complement the WMO's Commission for Basic Systems (CBS) Standardized Verification System for Long-Range Forecasts (SVSLRF), which has the specific objective of providing verification information for Global Producing Centre (GPC) products that are used as inputs to seasonal forecasting processes, including RCOFs. The SVSLRF procedures are targeted at providing information about the quality of ensemble prediction systems. In contrast, the procedures described in this document are targeted at verification of probabilistic forecasts, and have been selected to inform end-users of the forecasts as well as the forecasters themselves. The recommended procedures range in complexity from simple measures for communicating forecast quality to non-specialists, to detailed diagnostic procedures to provide in-depth analyses of the various strengths and weaknesses of forecasts. While the focus of the recommended procedures is on how well the forecasts correspond with the observations (forecast quality), care is taken to measure those attributes that can make forecasts potentially useful (forecast value). Interpretive guides for each of the recommended procedures are included to assist the user in understanding the verification results, and worked examples are included in an appendix. A glossary of technical terms is also provided.

When measuring the quality of a series of forecasts, the most important attributes are resolution (Does the outcome differ given different forecasts?) and discrimination (Does the forecast differ given different outcomes?) because they indicate whether the forecasts contain any potentially useful information. It is generally easier to measure discrimination than it is to measure resolution because discrimination can be measured more accurately with smaller samples than can resolution. The generalized discrimination score is therefore recommended as an initial score for assessing forecast quality. This score measures the ability of the forecasts to discriminate between the wetter, or warmer, of two observations. Since forecast quality can be conditional upon the outcome, it is also recommended that the score be calculated for the individual categories. The score then generalizes to the area beneath the relative operating characteristics (ROC) curve. Construction of the ROC curve is recommended for more detailed diagnostics of the ability of the forecasts to discriminate observations in each category. These scores and procedures can be mapped, or calculated for all locations together. It may be desirable to calculate the scores for subsets of the forecasts so that changes in forecast quality can be tracked.

The reliability of the forecast probabilities is an additional important attribute. Unfortunately, measuring reliability requires large samples, and so it is only viable to measure it by pooling the forecasts from different locations. The reliability component of the Brier score is recommended, which measures the squared "errors" for discrete forecast probabilities. For scores at individual locations it is recommended that summary scores (those that measure more than one attribute) be used instead. The effective interest rate (which is based on the ignorance score) is recommended in place of the commonly used skill score versions of the Brier score and the RPS because of widespread misinterpretation of the values of these latter scores when measured against forecasts of climatological probabilities.

For more detailed diagnostics of forecast quality, reliability diagrams are recommended. Unlike in the SVSLRF, it is recommended that the diagrams be drawn with points at 5% increments in the forecast probabilities because of the general practice of issuing seasonal forecasts with probabilities rounded to the nearest 5% (except for climatological probabilities of 33% in three-category systems, which should be pooled with the 35% forecasts). A number of suggestions are made for simplifying the communication of reliability diagrams to non-specialists, including: presenting the information in tables rather than graphs; fitting regression lines to the

reliability curves and describing their slopes in terms of the change in frequency of occurrence given 10% increments in the forecast probability; and so-called “tendency” diagrams, which show the average forecast probabilities for each category against the relative frequency of occurrence, and thus provide simple visual indications of unconditional biases.

The measurement of the quality of a specific forecast map requires a different set of procedures to those used for measuring the quality of a series of forecasts. The important attributes of good probabilistic forecasts cannot be measured given a single map without changing the meaning of the forecasts. Instead it is proposed that the forecasts be measured in terms of their “accuracy” (Do they assign high probabilities to the verifying outcomes?). The hit score (sometimes called the Heidke score, and which measures whether the forecasts assign high*est*, as opposed to high, probabilities to the verifying outcomes) is commonly used to measure the accuracy of seasonal forecasts, and has a simple intuitive interpretation. The calculation of the hit scores for the categories with the second and third highest probabilities is suggested, and can be informative in cases when the forecasters have been hedging. However, instead of recommending the hit score as a primary measure of verification, the linear probability score, which is a simple average of the forecast probability for the verifying outcomes, and the average interest rate, which indicates the average rate of return of an investment or betting strategy on the forecasts, are suggested. These scores are considered suitable for communication to non-specialists, but can be hedged, and so should not be used for any formal monitoring purposes. The ignorance score is recommended for more formal measurement. The Brier and ranked probability scores are not recommended because of their association with the attributes of reliability and resolution, which have a problematic interpretation given only a single forecast map.

Given the small sample sizes typical of seasonal forecasts many of the procedures recommended are likely to have large sampling errors. There therefore remains some uncertainty as to the true quality of the forecasts even after conducting detailed verification diagnostics. The calculation of *p*-values to assess the statistical significance of any of the recommended scores is discouraged because, when interpreted strictly correctly, they do not provide direct answers to the question of interest: What are the uncertainties in the measurements of forecast quality? In addition, *p*-values are considered unnecessarily complicated for non-specialists. Instead, the calculation of bootstrapped confidence intervals are recommended, which give a more direct indication of the uncertainty in the values of the scores.

Table of Contents

List of Abbreviations

A	Above-normal
B	Below-normal
BS	Brier score
CBS	Commission for Basic Systems
CCl	Commission for Climatology
CLIPS	Climate Information and Prediction Services
FAR	False-alarm rate
GPC	Global Producing Centre
HR	Hit rate
Ign	Ignorance
N	Normal
NMHS	National Meteorological and Hydrological Services
POD	Probability of detection
PRESAO	Prévision Saisonnière en Afrique de l'Ouest
RCOF	Regional Climate Outlook Forum
ROC	Relative (or receiver) operating characteristics
RPS	Ranked probability score
RPSS	Ranked probability skill score
SVSLRF	Standardized Verification System for Long-Range Forecasts
WMO	World Meteorological Organization

List of Recommended Procedures

a. Series of forecasts

The table below lists the recommended verification scores and procedures for series of forecasts, together with the attributes (see section 4c) that they measure. The procedures marked with an asterisk are considered a minimal set that all operational forecasting centres should strive to calculate. The other procedures provide useful, more detailed diagnostics and / or additional information that may be useful to non-specialists. The third column in Table 1 indicates whether the score has to be calculated on each forecast category separately, or can be calculated over all categories. The table also indicates whether or not it is viable to calculate a score or apply the procedure for each station / gridbox / region, given realistic sample sizes of seasonal forecasts. If it is viable to calculate the score for each location then the score can be mapped. In the final column some key references are provided. Further details on most of the scores and procedures are available from Jolliffe and Stephenson (2012) and Wilks (2011). Additional scores that are suggested, but not specifically recommended are provided in section 4b.

Table 1. List of recommended scores and procedures for series of forecasts, together with a list of the attributes the procedures measure, and an indication of whether the procedure should be applied on each category individually, and at each location. Some key references are provided for further information. The procedures marked with an asterisk are considered a minimal set that all operational forecasting centres should strive to calculate.

Score or procedure	**Attributes**	**By category?**	**By location?**	**Part of SVSLRF?**	**References**
Generalized discrimination *	Discrimination, skill	No	Yes	No	Mason and Weigel (2009)
ROC graph *	Discrimination, skill	Yes	Yes	Yes	Mason (1982); Harvey et al. (1992)
ROC area *	Discrimination, skill	Yes	Yes	Yes	Hogan and Mason (2012)
Resolution score	Resolution	Yes	No	No	Murphy (1973)
Reliability score	Reliability	Yes	No	No	Murphy (1973)
Effective interest rate *	Accuracy, skill	No	Yes	No	Hagedorn and Smith (2008)
Accumulated profit graphs	Accuracy, skill	No	Yes	No	Hagedorn and Smith (2008)
Reliability diagrams *	Reliability, resolution, sharpness, skill	Yes and no	No	Yes	Hsu and Murphy (1986)
Tendency diagrams	Unconditional bias	Yes	Yes and no	No	Mason (2012)
Slope of reliability curve	Resolution, conditional bias	Yes and no	No	No	Wilks and Murphy (1998)

b. Individual forecast maps

The table below lists the recommended verification scores and procedures for forecasts of an individual season. Some key references are provided. Additional scores that are suggested, but not specifically recommended are described in section 4c.

Table 2. List of recommended scores and procedures for individual forecast maps. Some key references are provided for further information.

Score	**References**
Verification maps as percentiles	
Model diagnostics	
Hit scores for categories with highest probabilities	Mason (2012)
Hit scores for categories with second and third highest probabilities	Mason (2012)
Linear probability score	Wilson et al. (1999)
Average interest rate	Hagedorn and Smith (2008)
Ignorance score	Roulston and Smith (2002)

1. Introduction

Forecast verification is an essential component of seasonal climate forecasting: without information about the quality of the forecasts how is anybody to know whether to believe them? It is very easy to make a forecast, but it is much harder to make a good forecast, and so the onus is on the forecaster to demonstrate that his/her forecasts are worth listening to. However, the question: "Are these forecasts good?" does not usually have a simple "yes" or "no" answer for a number of reasons. Firstly, forecast quality is not a simple binary quantity: it is perfectly reasonable to ask the question "how good (or bad) are these forecasts?" Secondly, forecast quality is not a simple univariate quantity: forecasts can be "good" in a number of different ways, and so forecasts A may be better than forecasts B in some ways but not in others. Hence fairly detailed information about the quality of forecasts can be determined, and it is possible to go beyond asking "Can these forecasts be believed?" to address questions such as "How can these forecasts best be used?" and "How can these forecasts be improved?"

The purpose of this document is to describe and recommend procedures for the verification of forecasts from the Regional Climate Outlook Forums (RCOFs), and similar forecast products. More generally, the recommendations are relevant to the verification of forecasts presented as probabilities of three ordinal, mutually exhaustive categories, which is true of most real-time seasonal climate forecasts whose target audience is the general public. In the RCOFs the categories typically are climatologically equiprobable, with the category thresholds defined using the upper and lower terciles, but, with few exceptions (which are mentioned in the text), the recommendations are relevant for forecasts presented as probabilities of categories that do not have to be climatologically equiprobable, and for forecasts with any finite number of categories. In this document, these forecasts are described as "probabilistic forecasts", and are distinguished from "deterministic forecasts", which are forecasts of specific values with no indication of uncertainty. The aims are to describe the most important qualities (formally known as "attributes") of good probabilistic forecasts, and to describe recommended procedures for measuring these attributes. The recommended procedures range in complexity from simple measures for communicating forecast quality to non-specialists, to detailed diagnostic procedures to provide in-depth analyses of the various strengths and weaknesses of forecasts. Interpretive guides for each of the recommended procedures are included to assist the user in understanding the verification results.

The recommendations for verification procedures made in this document were prepared under the auspices of the World Meteorological Organization's (WMO) Commission for Climatology (CCl) XIV Expert Team 3.2 on CLIPS Operations, Verification, and Application Service. These recommendations build upon the WMO's Commission for Basic Systems (CBS) Standardized Verification System for Long-Range Forecasts (SVSLRF), which has the specific objective of providing verification information for Global Producing Centre (GPC) products that are used as inputs to seasonal forecasting processes, including RCOFs. The SVSLRF is targeted at providing information about the quality of ensemble prediction systems, and the target audiences of the verification information are model developers, and the immediate users of these products (typically forecasters). In contrast, the procedures defined in this document are targeted partly at end-users of the forecasts, and partly at the forecasters themselves. Another difference from the SVSLRF is that here the specific products to be verified are series of forecasts released operationally, and so some of the complicated questions addressed by the SVSLRF pertaining to the generation of hindcasts and issues of cross-validation are avoided. It should be noted that in most cases the available history of operational seasonal forecasts is very short, and so there may be large uncertainties in the verification results. Procedures similar to those in the SVSLRF are therefore recommended to indicate the estimated uncertainty in the results.

The most important qualities of good probabilistic forecasts are described (section 3). These descriptions form the basis for identifying which attributes it is most useful to measure in order to assess the quality of the forecasts. Using the principles established in section 3, a set of

recommended verification procedures is then defined, distinguishing methods used to measure a set of forecasts (section 4b), and those for measuring the quality of a specific forecast (i.e., the forecasts for one target period; section 4c). Separate procedures are recommended for use by forecasters, and for communication to non-specialists. In section 5 some procedures for indicating uncertainty in the estimates of forecast quality are considered. First, however, some pragmatic issues regarding the definition and calculation of the target variable, and some potential problems with the observational data are discussed in section 2.

2. Verification Data

Before discussing some of the mathematical attributes of good forecasts in section 3, it is necessary to consider an attribute that is more qualitative in nature: is what is being forecast clear? In order to verify a forecast, it is essential to be unequivocal about exactly what is being forecast. Defining exactly what is this target variable is surprisingly non-trivial, especially with regard to the way in which forecasts are often constructed in consensus-building approaches such as at RCOFs. In this section, some problems in defining the target variable are discussed, and some possible problems arising from the nature of the forecasts and of the observational data are addressed.

a. Defining the target variable

Seasonal forecasts typically are presented as maps showing probabilities of seasonal accumulations (in the case of precipitation), or averages (in the case of temperature) falling within predefined categories. However, it is not always clear whether these accumulations or averages relate to areal averages, and if they do, it is not always clear what the area is over which the target variable is to be averaged. For example, consider the idealized example shown in Figure 1, in which there are forecasts of seasonal rainfall totals for three regions. The forecasts for regions I and II were constructed by calculating a regional average rainfall index, and then forecasting the index. For region III, however, the forecast was constructed by calculating two separate regional indices, whose areas are delimited by the dashed line, and then combining the two regions because the forecasts were identical, or at least very similar. The problem now, however, is that the three forecasts no longer mean the same thing: the forecasts for regions I and II define probabilities for seasonal rainfall totals averaged over the respective regions, but the forecast for region III does not mean that the seasonal rainfall total averaged over this region has a 25% chance of being above-normal. Instead, for region III, the probability of the seasonal rainfall total averaged over sub-region *a* has a 25% chance of being above-normal, and the same is true of sub-region *b*.

Why is this difference in interpretation important? Imagine a situation in which observed rainfall over sub-region *a* is above-normal and over region *b* is below-normal, and that the spatial

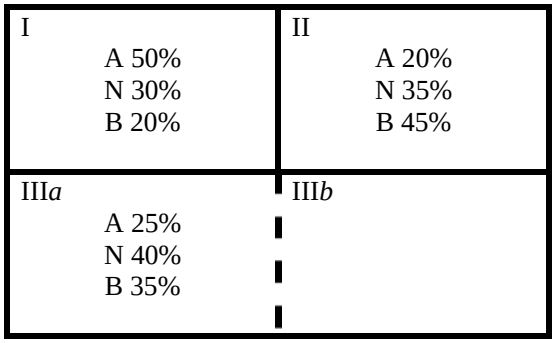

Figure 1: Idealized example of seasonal rainfall forecasts for three regions. A indicates the probability of above-normal rainfall, N of normal, and B of below-normal.

average over the whole of region III is then normal; this forecast would be scored based on the 40% for the verifying category (normal). If the sub-regions had been left as separate forecasts it would be scored based on the 25% for the above-normal category over sub-region *a*, and the 35% for the below-normal category over sub-region *b*, and so would score worse, and appropriately so. The point is that the spatial distribution of observed rainfall over region III (wet on the one half, and dry on the other) should result in a forecast that scores poorly, but because the forecasts for the two sub-regions were similar the forecast scores well, which seems inappropriate. The net effect is that by re-interpreting the forecast over region III to refer to the spatial average over the entire region, the forecasts are going to verify as better than they really are.

This situation of the forecast verification being sensitive to the degree to which regions are combined because of similar predictions has to be considered undesirable. If the forecasts for sub-regions *a* and *b* had been slightly different such that their areas had not been combined, surely it would then be unfair that the forecast map verifies so much more poorly. It seems to make sense to verify the sub-regions separately even if the forecasts are the same. Unfortunately, in most cases, these sub-regions are not indicated on the map, and so it may be impossible to identify where they are, and how many there are. In drawing up the consensus, the boundaries for the original sub-regions may have been modified anyway.

A more serious problem is that it is no longer possible to identify exactly what the forecast for any region means: given a forecast like that shown in Figure 1, but without the dotted line, how is one to know that the forecast for region III does not mean that the rainfall averaged over this entire region has a 25% probability of being above-normal? And without any information about the sub-regions, what can one conclude that the forecast means anyway? This problem of the interpretation of forecasts for regions is not only a problem for verification analyses, it is also a problem for the users of the forecasts, and needs to be addressed as an issue in the construction of the forecast. Possible solutions include providing forecasts on a gridded basis [as followed by Global Producing Centres (GPCs), for example], or indicating the sub-regions on the map as thin lines, or forecasting for stations rather than for regional indices. Forecasting for pre-defined homogeneous zones is also an attractive option. However, the solution to this problem is beyond the scope of this document. More generally, clear guidance needs to be provided on how to define and represent regions when constructing seasonal forecasts. These questions should be posed to the CCl Expert Team 3.1 on Research Needs for Intraseasonal, Seasonal and Interannual Prediction, Including the Application of these Predictions.

b. Sizes of forecast regions

A further problem arises when verifying forecasts of regional indices if the regions are not the same size. Supposing that the problem of the sub-regions can be resolved, if the regions are of differing size, the verification results should reflect this fact. For example, imagine a simple example where there are only two regions, one three times as large as the other. If forecaster A issued a 50% probability on the verifying category for the larger region, and a 20% probability for the smaller, surely (s)he should get greater credit than forecaster B, who issued a 50% probability on the the verifying category for the smaller region and a 20% for the larger. The verification results should be weighted by area to resolve this problem. Further details on weighting the results by area are provided in appendix A.

c. Gridding of forecasts

One way to address the problems of the sub-regions and of unequally sized regions, highlighted in the previous two sub-sections, is to grid the forecasts. Although gridding involves defining a set of sub-regions and assuming (quite possibly incorrectly) that the forecasts are valid at this new scale, the solution is an attractive one, especially if the verification data themselves are gridded. Either way, if any form of gridding or interpolation is required, the forecasts should be

gridded / interpolated to the resolution / locations of the verification data rather than vice versa as long as the resolution of the verification data is not substantially different from the spatial scale at which the forecasts were meant to be interpreted. If there is a marked mismatch in the spatial scales the verification data may need to be aggregated to a more compatible resolution first. (If the verification data are on a coarse resolution than the forecast then producing meaningful verification results will be difficult.) There are a few problems in gridding the forecasts, which are addressed in this sub-section.

In gridding the forecasts, inevitably some of the grids will span two or more forecast regions, and the question of what value to put down for the forecasts at such gridboxes has to be addressed. The problem can be minimized by using a high resolution grid, but is not eliminated. It is not clear that it would make sense to average the forecasts because of possible differences in variance in the different sections. There are no simple, perfectly adequate solutions, and so it is recommended that the forecast that represents the largest part of the gridbox be used. In most cases this should be the forecast at the centre of the gridbox, which simplifies the gridding procedure considerably.

For domains that span a large number of latitudes, the most poleward gridboxes may be considerably smaller than those closest to the equator. A cosine weighting scheme should be applied in these instances. Each gridbox should be weighted by the cosine of the latitude at the centre of the gridbox. Further details on weighting the results by latitude are provided in appendix A.

d. Verification using station data

If the forecast maps were constructed from forecasts for individual stations, and the regions simply delimit stations with similar forecasts, then the verification should be performed station-by-station. It is preferable to weight the results so that clustered stations receive less weight than those in more sparse areas. This weighting can be performed by constructing Thyssen polygons[1], although the weighting may be unnecessarily complicated if the station distribution is reasonably uniform. On the other hand, if there are notable clusters of stations, those stations on the edge of the clusters can receive disproportionately large weight. In these cases it is preferable to grid the verification data, or to divide the region into homogeneous zones, and to weight these by area. Another option, which is recommended if the station density is sparse, is to define a threshold distance beyond which observations are considered missing, and to not include areas beyond these thresholds in the verification analysis. No absolute value for the threshold can be recommended because of regional differences in decorrelation (the decrease in correlation between the climate at two locations as the distance between the locations increases), and because of differences in decorrelation with different variables. It is suggested that the decorrelation be calculated, or that knowledge of homogeneous climate zones be applied. Either way, the solution must draw on local expert knowledge of the climate. Further details on weighting the results by representative area are provided in appendix A.

e. Data inconsistencies

It may be the case that the observational dataset used for the climatological period (the climatological dataset) is different from that used for the verification of the actual forecasts (the verification dataset). If so inconsistencies in the two datasets may result in inaccurate assignment of the verifying observations to the categories. If it is necessary to use different datasets, they

[1] Thyssen polygons contain one station, and are constructed so that all locations within each polygon are closer to the station it contains than to any of the other stations. There areas can be approximated by superimposing a fine-resolution grid over the entire domain, and counting how many of the gridpoints are closest to each of the stations.

should be chosen so that there is a period of overlap that is as long as possible. For temperature data the following procedure is recommended to reduce the effects of inconsistencies:

1) Calculate the mean and standard deviation of the data from the climatological dataset for the period of overlap.
2) For the data in the climatological dataset prior to the period of overlap, subtract the mean and divide by the standard deviation obtained from step 1.
3) Calculate the mean and standard deviation of the data from the verification dataset for the period of overlap.
4) Multiply the results of step 2 by the standard deviation and add the mean calculated in step 3.
5) Append the verification data onto these transformed data. The first part of the climatological period should now consist of transformed data from the original climatological dataset, and the second part should consist of data from the verification dataset that has not been transformed.
6) Calculate the terciles using these merged data, and assign the verification data to the corresponding category.

For precipitation data, the transformation procedure above can give inadequate results if the data have a skewed distribution. In this case the following procedure is recommended:

1) Calculate the parameters of a gamma distribution fitted to the climatological data for the period of overlap.
2) For the data in the climatological dataset prior to the period of overlap, transform the data to quantiles of the corresponding gamma distribution using the parameters obtained from step 1.
3) Calculate the parameters of a gamma distribution fitted to the verification data for the period of overlap.
4) Transform the quantiles from step 2 to deviates using the gamma distribution parameters for the verification data obtained from step 3.
5) Append the verification data onto these transformed data. The first part of the climatological period should now consist of transformed data from the original climatological dataset, and the second part should consist of data from the verification dataset that has not been transformed.
6) Calculate the terciles using these merged data, and assign the verification data to the corresponding category.

If the climatological period is modified at any time over the verification period, only step 6 in the two algorithms above need be repeated. The aims of steps 1–5 are to obtain a consistent dataset; once this dataset is obtained the climatologies can be calculated using whichever period was referenced for each of the forecasts that are to be verified.

Having addressed various data problems and problems of interpreting the forecasts, the attributes of good forecasts can now be considered.

3. Attributes of "Good" Forecasts

a. Types of forecast "goodness"

Forecasts can be described as "good" in three different senses (Murphy 1993). Firstly, forecasts are "consistent" if the forecast is a true indication of what the forecaster thinks is going to happen. Forecasts may not be consistent with the forecaster's beliefs if the forecaster is hedging in order to avoid what may be perceived as a bad forecast, for example. It is quite easy to show

that some ways of verifying forecasts encourage the forecaster to hedge; for example, in a three-category system, the forecaster may be encouraged to predict the middle category simply because the forecast can then never be more than one category away from what happens. Hedging by the forecaster is undesirable since the user of the forecasts is not provided with the information that the forecaster would consider correct. For example, if the forecaster thinks that there is a 50% probability of below-normal rainfall, but modifies this probability to 40% in the forecast to avoid causing too much alarm (or for any other reason), the user is provided with an under-estimate of the risk of below-normal rainfall. Verification scores that encourage the forecaster to issue a forecast that is consistent with his/her true beliefs are known as "strictly proper" scores. Some scores are not uniquely optimized when the forecaster forecasts his/her true beliefs, but cannot be optimized by hedging; these scores are known as "proper" scores. It is generally accepted that scores that encourage forecasters to hedge should be avoided, although as long as they are not used as a metric for measuring improvements (or deteriorations) in forecast quality they may be of some interest.

A second sense in which forecasts can be described as "good" is how well what was forecast corresponds with what happened. This aspect is known as forecast "quality" and is typically measured by some form of mathematical relationship between the forecasts and the respective observations; forecasts that correspond well with the observations will be scored better than those that do not correspond well. There are numerous ways in which this correspondence can be described (Murphy 1991), which is one of the reasons why there are so many different ways of measuring forecast quality. For example, the first question that almost invariably gets asked about forecast quality is "How often are the forecasts correct?" Correctness seems an intuitively appealing attribute for deterministic forecasts, but is generally considered an inappropriate attribute when considering probabilistic forecasts. If, for example, the probability for above-normal rainfall is 50%, it does not seem to matter whether or not above-normal rainfall occurs, the forecast cannot be described as incorrect. Nevertheless, interest in measuring the "correctness" of probabilistic forecasts remains widespread, both amongst end-users, and amongst many of the forecasters themselves. Because of the impossibility of educating all potential users of seasonal forecasts on the complexities of forecast verification, some recommendations are made for procedures that can, in a loose sense, be interpreted as measuring the correctness of probabilistic forecasts. In recommending these procedures it is emphasized that it remains imperative, especially for the forecasters, to focus on attributes that are appropriate for describing the quality of probabilistic forecasts. These attributes are discussed in detail in the following sub-sections, and procedures for measuring them are described in section 4.

The third way in which forecasts can be described as "good" is in terms of the benefit, or "value", they can help realize, whether that be economic, or social, or otherwise. Forecast quality is a prerequisite but not a guarantee of forecast value: forecasts that have good quality have the potential for being of value, but whether they actually are depends on the impacts of the observed climate, and on the options available for mitigating (or taking advantage of) such impacts (Katz and Murphy 1997). For example, if excellent forecasts are made, but are released too late to take any useful action, they have no value. More importantly, however, is the fact that there is usually an imbalance between the losses and gains realized from the use of forecasts. Imagine another set of excellent forecasts that are released in a timely manner, and which enable a user to reap some benefit when the forecasts correspond well with the observed outcomes: it is quite possible that the costs of the occasional "bad" forecast may more than offset the benefits. Very good forecasts can therefore have no, or even negative, value, but it is not possible to realize any benefit from forecasts that have no correspondence with the observations.[2]

[2] Exceptions have to be made for cases where markets may respond to forecasts regardless of the skill of the forecast. It may then be possible to benefit from anticipating changes in the market that are a response to the forecasts rather than as a response to the climate, in which case forecasts that are otherwise useless may have some value.

In this document verification procedures for measuring forecast quality, rather than value, are proposed. In selecting some of the recommended procedures an attempt has been made to answer the question "Is it possible that the forecasts have value?" If the forecasts contain any information about the outcomes then it may be possible to make beneficial decisions in response to the forecasts. What those beneficial decisions might be lies well beyond the scope of this document. Instead, the verification procedures can offer only guidance as to whether it may be worth trying to identify such decisions.

b. Probabilistic forecasts and forecast quality

Measuring the quality of probabilistic forecasts is much more complicated than for deterministic forecasts. Consider the simple example of forecaster A, who says it is going to rain tomorrow, and forecaster B, who says there is a 60% chance of rain tomorrow. If it rains, forecaster A clearly issued a correct forecast, but what about forecaster B? And is forecaster B correct or incorrect if it does not rain? To forecaster B, it does not seem to matter whether it rains or not, (s)he has not made an incorrect forecast. The temptation is to conclude that probabilistic forecasts cannot be "wrong" (as long as probabilities of 0% are never issued on any of the outcomes), and that therefore these forecasts are always correct. While this conclusion is logically valid, it is also distinctly unhelpful since any probabilistic forecast that does not have a zero probability on the outcome is as equally "correct" as any other. Probabilistic forecasts can only be described as "correct" in the sense that they indicated that the observed outcomes could have happened, in which case the probabilities themselves are completely irrelevant (as long as they exceed zero). The question of correctness of probabilistic forecasts is then so uninformative as to be useless, and nothing is learned about whether the forecasts have successfully indicated whether or not the observed outcomes were likely or unlikely to have happened. Therefore more meaningful questions about the quality of probabilistic forecasts need to be asked.

One reasonably common practice is to define probabilistic forecasts as "correct" if the category with the highest probability verified. Hit (Heidke) scores are then calculated, and have a reasonably intuitive interpretation as long as the user has a good understanding of the base rate.[3] While it is not unreasonable to ask how often the category with the highest probability verifies, there are some inter-related and important problems with such approaches. Firstly, there is a danger that the verification procedure will be seen as implicitly condoning the interpretation of the forecast in a deterministic manner, which is a problem both for the user (who loses information about the uncertainty in the forecast), and for the forecaster (who typically becomes tempted to hedge towards issuing higher probabilities on the normal category to avoid a two-category "error"). Secondly, if the probabilities are to be considered as at all meaningful, one should not actually want to achieve a high hit score because that would indicate that the forecasts are unreliable. If the highest probability is 40% one should want a "hit" only 40% (i.e. less than half) of the time. Thirdly, the scoring system does not give any credit for issuing sharp probabilities. Thus two forecasters who always issue the same tendencies in their forecasts will score exactly the same regardless of whether one of the forecasters is more confident than the other. Finally, although this scoring system does not explicitly penalize for two-category errors[4], there is clear evidence from some of the RCOFs that scoring the forecasts in this way has encouraged the forecasters to hedge towards increasing the probability on the normal category (Chidzambwa and Mason 2008). The normal category typically has the highest probability far more frequently than the outer two categories in RCOF forecasts, and although there are many reasons for this bias, the scoring strategy is one contributor.

[3] Knowledge of the base rate is necessary because the naïve expectation is that at least 50% of the forecasts should be correct. However, with three or more categories scores of less than 50% correct may be good.

[4] Although at some of the RCOFs the numbers of one- and two-category errors are reported separately, which is likely to encourage hedging.

Rather than trying to transform the forecasts so that individual forecasts can be counted as "correct" or "incorrect" in some way, it is recommended that verification procedures be used that are suitable for the forecasts in the format in which they are presented. However, although there are numerous scores suitable for probabilistic forecasts, many of these are too technical to be suitable for communication to non-specialists, and it is not always clear what specific attributes of forecast quality they are measuring. In the following sections, arguments are made for the most important attributes.

c. Attributes of "good" probabilistic forecasts

i. Resolution

One of the most basic attributes of a good set of probabilistic forecasts is that the outcome must be different if the forecast is different. If, on average, the same thing happens regardless of what the forecast is then the forecasts are completely useless[5]. For probabilistic forecasts, above-normal rainfall, for example, should occur more frequently when the probability is high, compared to when it is low. As the probability increases (decreases) so above-normal rainfall should occur more (less) frequently. If the forecaster says there is an 80% chance of above-normal rainfall (s)he is communicating much more confident that above-normal rainfall will occur than when (s)he says there is a 20% chance, and if there is any basis to this difference in confidence, above-normal rainfall should occur more frequently given forecasts of 80% than given forecasts of 20%. Good forecasters should be able to distinguish between times when the probability of above-normal rainfall (or of any other outcome) is inflated from times when the probability is deflated. If they can make this distinction, then their forecasts will have good "resolution" (assuming consistency, as defined in section 2a). Resolution can be determined by measuring how strongly the outcome is conditioned upon the forecast. If the outcome is independent of the forecast, the forecast has no resolution and is useless – it can provide no indication of what is more or less likely to happen. It is quite possible for forecasts to have good resolution in the wrong sense: if the above-normal rainfall occurs less frequently as the forecast probability is increased the outcome may be still strongly conditioned on the forecast, but the forecast is pointing to changes in probability in the wrong direction. In this case the forecasts have good resolution, but would otherwise be considered "bad". Forecasts with no resolution are neither good nor bad, but are useless. Metrics of resolution distinguish between potentially useful and useless forecasts, but not all these metrics distinguish between good and bad forecasts.

ii. Discrimination

A similar perspective to resolution is to ask "Do the forecasts differ given different outcomes?" rather than "Do the outcomes differ given different forecasts?" Discrimination, along with resolution, can be considered one of the most basic attributes of a good set of probabilistic forecasts: If, on average, a forecaster issues the same forecast when rainfall is above-normal compared to when rainfall is below-normal the forecasts cannot "discriminate" between these different outcomes[6]. Whereas resolution is concerned with whether the expected outcome differs as the forecast changes, "discrimination" is concerned with whether the forecast differs given

[5] One can imagine a pathological case in which the variance of forecasts is very large when rainfall is above-normal, and virtually zero when rainfall is below-normal, but the mean forecast is the same in each case. These forecasts could be considered useful in that an extreme forecast (wet or dry) would point to high probabilities of above-normal rainfall. This kind of situation is likely to be rare, but does point to the fact that forecast quality cannot be adequately summarized by a single metric.

[6] As with resolution, one can imagine a pathological case in which the variance of the observations is very large when the probability for above-normal rainfall is high, and virtually zero when the probability is low, but the mean of the observations is the same in each case. These forecasts could be considered useful in that a forecast with low probability would enable one to be very confident about the outcome being close to the conditional mean.

different outcomes. Just as for resolution, it is not necessarily the case that forecasts with good discrimination are good forecasts: a forecaster may issue lower probabilities of above-normal rainfall when above-normal rainfall occurs compared to when below-normal rainfall occurs. Again measures of discrimination distinguish between potentially useful and useless forecasts, but not all these metrics distinguish between good and bad forecasts.

Resolution and discrimination cannot be improved by statistical recalibration of forecasts[7] whereas the subsequent attributes can be. This inability is a result of the fact that statistical procedures cannot increase the information in the forecasts, and can only communicate the information more effectively.

iii. Reliability

The purpose of issuing probabilistic forecasts is to provide an indication of the uncertainty in the forecast. The forecast probabilities are supposed to provide an indication of how confident the forecaster is that the outcome will be within each category, and so they could be interpreted as the probability that a deterministic forecast of each category will be "correct". For example, given probabilities of 40% for below-normal, 35% for normal, and 25% for above-normal, if someone were to issue a deterministic forecast of below-normal, the probabilistic forecaster thinks there is a 40% chance that the deterministic forecast will be correct. Similarly, if someone were to forecast above-normal, there is thought to be a 25% probability (s)he will be correct. Forecasts are reliable, or well-calibrated, if the observation falls within the category as frequently as the forecast implies (Murphy 1993).

More often than not seasonal forecasts are unreliable. The commonest situation is that the forecasts are over-confident – increases and decreases in probability are too large. Over-confidence occurs when the forecaster thinks that the probability of a specific category is increased (or decreased), but over-estimates that increase (or decrease), and thus issues a probability that is too high (or too low). For example, if a forecaster thinks that the chances of above-normal rainfall have increased (decreased) from their climatological value of 33%, and indicates a probability of 50% (20%), but above-normal rainfall occurs on only 40% (as much as 25%) of these occasions, (s)he has correctly indicated increases and decreases in the probability, but over-stated these changes, and thus was over-confident. In relatively rare cases the forecasts may be under-confident, in which cases the increases and decreases in the probabilities are too small. Over- and under-confidence are examples of conditional biases: the errors in the forecasts depend upon whether the forecasts indicate increased or decreased probabilities. For over-confident forecasts the increased (decreased) probabilities are too high (low); for under-confident forecasts the increased (decreased) probabilities are too low (high). Sometimes the forecasts are unconditionally biased: the probabilities are too high (or too low) regardless of whether the forecasts indicate increased or decreased probabilities. If the probabilities are generally too high the category occurs less frequently than implied by the forecasts, and the category is over-forecast. Similarly, if the probabilities are generally too low the category is under-forecast.

Although reliability is widely recognized as being an important attribute of probabilistic forecasts, it cannot be considered the only attribute that is important: if the climatological

[7] It is possible to improve resolution and discrimination if the forecasts are recalibrated with the data at a higher measurement scale than is used to evaluate the forecasts. Imagine a set of forecasts of temperatures that are unconditionally biased (perhaps consistently too warm), and which are expressed as binary forecasts of temperatures exceeding a predefined threshold. Because of the bias the forecast temperatures will exceed the threshold more frequently than the observations, and so any resolution or discriminatory skill in the forecasts will be weakened. If the forecasts are available in °C (or on some other continuous scale), the bias could be removed, and the verification results re-calculated. Some of the forecasts that previously exceeded the threshold will no longer do so, and the resolution / discriminatory skill is likely (although not necessarily) to improve. However, this recalibration is impossible if the forecasts are available only as binary values because there would be no basis for selecting which forecasts to reclassify.

probability of an outcome can be estimated accurately in advance, a set of forecasts that always indicate the climatological probability will be reliable, but will not provide any indication of the changing likelihood of the outcome from case to case.

iv. Sharpness

Complaints are often expressed that the probability shifts indicated in most seasonal forecasts are small – the forecast probabilities are nearly always close to the climatological probabilities even though they are frequently over-confident. In many of the RCOF forecasts, for example, the highest probability on any of the categories often is only 40%, and only occasionally exceeds 45%. In these cases, if the forecasts are reliable, the category with the highest probability is still more likely not to occur than it is to occur. (If the highest probability is for the above-normal category and is 40%, there is a 60% chance that rainfall will be normal or below-normal). These forecasts express low confidence, and because the uncertainty is large, end-users of the forecasts can have correspondingly low confidence in realizing benefit from any decisions they might make in response to the forecasts. The problem is exacerbated by the fact that the forecasts are often over-confident, so even the fairly small shifts in probability indicated by the forecasts are often larger than justified. Assuming they are reliable, forecasts that express high confidence are more useful than forecasts with low confidence because they enable the end-user to be more confident in making decisions. It was explained in the previous section that climatological forecasts may have good reliability, but are not otherwise very useful because they provide no indication from case to case as to whether the probabilities for any of the categories have increased or decreased – the forecast communicates no reduction in the uncertainty in the outcome. Forecasts for which the probabilities differ markedly from the climatological values communicate high confidence in the outcome, and are said to be "sharp". If the probability is very high, there is high confidence that category will verify; if the probability is very low, there is high confidence that category will not verify. Of course, sharp forecasts are not necessarily reliable, nor do they necessarily provide any resolution. Sharpness is defined only in terms of the forecasts, and makes no reference to whether the forecasts correspond well with the observations. Nevertheless, forecasts that have good resolution and reliability, will also have good sharpness.

Because sharpness is defined only in terms of the forecasts it is not generally measured separately. Instead, some of the scores that are proposed implicitly consider sharpness with at least one of the other attributes of interest.

v. Skill

A set of forecasts is "skilful" if it is better than another set, known as the "reference" forecasts. Skill is therefore a comparative quantity, rather than an absolute quantity: it is quite possible for both sets of forecasts to be good, or for both to be bad. Because skill is relative, a set of forecasts may be considered skilful compared to a second set of forecasts, but unskilful compared to a third set. In order to measure whether forecast set A is better than set B, some measure of forecast quality is required, and so the forecasts have first to be scored on one or more of the other attributes mentioned in the previous sections (or on other attributes not mentioned). The scores on this (these) attribute(s) can then be compared to determine the level of skill. With probabilistic forecasts it is standard practice to use climatological probabilities as the reference forecast since these would be the best information available in the absence of any forecasts (and for most practical purposes would be considered better than randomly assigned probabilities, or perpetual forecasts of non-climatological probabilities). Of course it is not necessarily the case that the user knows what the climatological probabilities are, but it seems reasonable to ask whether the provision of forecasts is an improvement upon the information that could be gleaned from providing only historical information about past climate variability. Another useful reference strategy is to use "persistence" – to assume that the most recently observed climate anomalies will continue into the target period. However, it is not obvious how to express a persistence forecast in

probabilistic terms; some kind of statistical model is required, which immediately involves communicating a wealth of climate information within the forecast beyond some simple summary statistics that are implied by the climatological probabilities. Regardless of which reference strategy is used, it is important to communicate carefully what exactly the comparison reveals about the quality of the forecasts.

d. Attributes of individual probabilistic forecast maps

There is often interest in knowing whether the forecast for a specific season was good or bad. At many of the RCOFs, for example, the observed rainfall and forecast for the previous season are reviewed, and attempts are made to assess whether or not the forecast gave a good indication of what transpired. If the forecast is to be scored, it is important to recognize that many of the attributes described above are no longer applicable because an incorrect interpretation of the forecasts can otherwise be invoked. For example, consider a set of forecasts for 10 locations all for the same season, and suppose that these forecasts form a map and that the intention is to obtain a score for these forecasts. Imagine further that the forecasts at all 10 stations indicated a 10% probability of above-normal rainfall, and that above-normal rainfall occurred at two of the ten (20%) stations. The temptation may be to conclude that these forecasts were over-confident because above-normal rainfall occurred over more than 10% of the area, but the forecasts were not stating that 10% of the area would be dry, only that at each station 10% of the occasions on which a 10% probability of above-normal rainfall is issued can above-normal rainfall be expected to occur. To try and measure the reliability of the forecasts for an individual season therefore represents an incorrect interpretation of the forecast.

Part of the problem in this faulty interpretation is that the climate over a specific forecast region is expected to be homogeneous, and so unless a large number of sub-regions have been combined (see section 2a), the forecaster would normally expect the same category to occur over the entire region, or at least over a large part of it. So for example, if the forecast indicates a 50% probability of above-normal rainfall over a region, the forecaster may think there is a 50% chance that almost all or almost none of the region will be above-normal.[8] In effect, therefore, we are only expecting one realization of the forecast, and so trying to calculate the reliability of the forecast measured over space is like trying to calculate the reliability of a probabilistic forecast for a single location and a single target period. The problem of trying to measure the forecast quality of a single map therefore is partly a problem of severely small sample size, even if there are large numbers of stations or gridboxes.

To resolve this problem of identifying what attributes might be of interest when scoring an individual forecast map, it is helpful to consider what the objective might be in performing such a verification analysis. There appear to be two main possible sets of objectives. The first set of objectives may be to score the forecast in some way so that changes in forecast quality with time can be tracked and communicated to forecast users. The tracking of the forecast quality could be used to monitor improvements in forecast quality over time, or to see how forecast quality changes with ENSO phase, for example. Rather than measuring the changes in forecast quality by calculating a score for each target period, it is recommended that subsets of forecasts be used (for example, by considering all of the forecasts for a calendar year, or by grouping separately forecasts before and after a modification in the operational forecast procedure). However, there may still be occasion for scoring each target period: for example, users may be interested in how the quality of last season's forecast compares with that of the previous year, or to those of the last few years. If a user of last year's forecast made some particularly beneficial (or detrimental) decisions in response to the forecast they may find it helpful to know whether to expect similar

[8] We cannot go so far as to say that the forecaster thinks there is an approximately 50% chance of virtually all of the region being above-normal, and another 50% chance that virtually all the region will not be above-normal (and therefore a virtually 0% probability that about half of the region will be above-normal) since the forecaster may be making a statement about spatially averaged conditions.

levels of benefit (or loss) given subsequent forecasts. Users of seasonal forecasts must not expect that even well-formulated decision-making will result in beneficial outcomes all the time, but must be prepared for occasional losses. A forecast system itself can only be properly evaluated on the basis of a long series of forecasts, but knowing how individual forecasts score against an expected level is useful information. Most verification scores are simply an average of scores on individual forecasts anyway, and it is quite reasonable to consider the distribution of the scores for the individual years rather than looking only at the average.

A second set of objectives may be to diagnose whether the forecast could have been improved given better information (more accurate sea-surface temperature forecasts, for example, or by giving greater consideration to other sources of forecast information such as from the Global Producing Centres). For this second set, various sensitivity analyses would be required (for example, running models with improved sea-surface temperature forecasts, or even with observed temperatures), and the problem is more one of model diagnostics than one of forecast verification. In both cases, however, there is an interest in knowing how sharp the forecasts were on the verifying categories. The forecast will be considered an unusually good (or bad) one if these probabilities were uncharacteristically high (or low), and it may be concluded that the forecast could have been improved if the additional information results in a re-forecast with higher probabilities on the verifying categories. This attribute is called "accuracy' in this document; an accurate forecast has high probabilities on the verifying outcomes.

It has to be emphasized that the "accuracy" (defined as high probabilities on verifying outcomes) of an individual forecast map is a very incomplete perspective of the quality of forecasts. It is quite possible for a highly unreliable forecast system to occasionally issue highly accurate forecasts because of a tendency to over-confidence. The multifaceted quality of a set of forecasts (i.e., as described by the attributes defined above) should therefore be considered before concluding whether a forecaster or forecast model is to be believed on a consistent basis.

4. Measuring Forecast Quality

A set of good probability forecasts will have good reliability as well as high resolution (and, implicitly, high sharpness), will be well-discriminated, and have high skill, resulting in consistently accurate forecasts. Because there are a number of attributes of interest, forecast quality cannot be adequately measured by a single metric, and so it is imperative that a multifaceted approach to forecast verification be taken, and that it be made clear upfront that more than one score is required. Attempts to devise scores that measure all attributes at once typically result in an abstract number that has little intuitive appeal, and is difficult to interpret. While such scores have their uses, they are not generally recommended in this document.

How the various attributes of good forecasts can best be measured depends to a large extent on the format of the forecasts. In section 4b procedures are recommended for measuring these attributes given forecasts expressed as probabilities of categories. It is assumed that the forecasts are presented in map form, and that a set of maps is available forming a history of forecasts. Procedures for measuring the accuracy of an individual forecast map are considered in section 4c, with the forecasts again being expressed as probabilities of categories. First, however, it is necessary to consider whether to include climatological forecasts in the verification.

a. The verification of climatological forecasts

When calculating any verification score or constructing a graph, a decision has to be made as to whether or not to include climatological forecasts that were made because the forecaster had no reason for expecting any one outcome over any other. This ignorance could either be because of a lack of skill or because of a lack of signal for that particular instance. If there are a large number of climatological forecasts, these can dominate the verification analyses, and many of the

verification procedures will score rather poorly because of the lack of sharpness. On the one hand this poor scoring is appropriate because the forecaster is not providing much information, but on the other hand it can give the impression that the non-climatological forecasts are not particularly useful. So the decision as to whether to include climatological forecasts in the analysis depends on why the verification analysis is being conducted. If the interest is in comparing the forecasts with another set (perhaps for another region, or season, or from another forecast system), then the climatological probabilities should be included because the forecaster should be credited for issuing sharper forecasts if those forecasts contain potentially useful information, and should be penalized if they do not. If the interest is in whether to believe the forecasts, then it may be better not to include the climatological probabilities because then the forecasts will be assessed only on the basis of when the forecaster has something to say. However, this recommendation is only valid if climatological forecasts are meant as a statement of lack of knowledge about the outcome. The possible meanings of climatological forecasts, and the implications of these differences, should therefore be considered.

When climatological probabilities are indicated is the forecaster saying (s)he does not know what is going to happen, or is (s)he saying that there is a 33% (assuming three equiprobable categories) probability of each of the possible outcomes? The difference is subtle, but in the latter case the forecaster is stating that (s)he thinks that the climatology of the verification period is likely to be similar to that of the training period, whereas in the former case the forecaster would prefer simply to state that (s)he does not have any useful information at all. If the forecasts include a statement about the climatology of the verification period then they should be assessed accordingly. Forecasters are encouraged to be clear in future about making this distinction, and to indicate "no forecast" rather than issue climatological probabilities if they do not even want to make a statement about climate trends. In most cases, the forecaster is likely to be expected to make a forecast, and will thus default to issuing climatological probabilities. However, the forecaster should at least be encouraged to think about recent climate trends if other methods of forecasting are indicating no signal.

If climatological forecasts are to be included in the analyses, and if any of the climatological probabilities are not a multiple of 5%, it is recommended that these values be rounded to the nearest 5% only for drawing the reliability diagrams, but that even here the corresponding point on the reliability curve should be drawn at the average of the forecast probabilities for this bin on the *x*-axis. For all other procedures the climatological probabilities should be analyzed as a separate probability value where appropriate.

b. Measuring the quality of series of forecasts

Before recommending any verification scores, it is important to consider what information is required from the verification analysis. The first question to be addressed is: Do the forecasts contain any potentially useful information, and if so, how much? It was argued in section 2 that resolution and discrimination are the key attributes for addressing this question (Murphy 1966), and so the next subsections describe some recommended procedures for measuring these attributes. The scores are intended to be used both to provide an easy-to-understand indication of the quality of all the forecasts, as well as to be suitable for producing a map to indicate where the forecasts may be most potentially useful. Scores suitable for the communication to the general public are described first, followed by scores that may be more informative to forecasters and users with some expertise in the field of forecast verification. The second question to be addressed is: How can the forecasts be improved? The objective is no longer to attempt to summarize forecast quality in a single number, but to provide diagnostic information on systematic errors in the forecasts that can point to the need for correcting the forecast system. These diagnostics are also useful for forecast users who, having accepted that the forecasts contain useful information, may then be interested in whether to take the forecasts at face value. They may wish to know whether the forecaster is over-confident or issues biased forecasts, or even to diagnose whether

the forecaster is hedging. Procedures for providing detailed diagnostics of forecast quality are described, and some guidelines on how to summarise this information for non-experts are provided.

i. Measuring discrimination

It was argued in section 2 that resolution and discrimination are the most fundamental attributes of good probabilistic forecasts. For a number of reasons, it is easier to measure discrimination than resolution given the standard three-category probabilistic systems that are typical of seasonal climate forecasting. One simple reason is that there are only three possible outcomes whereas there are many different possible forecast probabilities. Therefore we have to ask "Do the forecasts differ given different outcomes?" only three times (once for each category), rather than having to ask "Do the outcomes differ given different forecasts?" for each of the multitude of different forecast probabilities that may have been issued. Both resolution and discrimination are useful attributes, and procedures for measuring both are recommended. However, as a starting point for assessing the quality of forecasts it is a measure of discrimination that is recommended.

Accepting that discrimination is the most logical attribute to measure as a starting point, the relative operating characteristics (ROC; sometimes called receiver operating characteristics) graph, and the area beneath the curve (Hogan and Mason 2012), seem logical recommendations. Indeed these procedures are recommended, but, because they require the outcome to be binary, separate results have to be calculated for each category if there are more than two categories, and so the areas beneath the curves cannot be used as a single summary metric. Instead the generalized discrimination score (Mason and Weigel 2009) is recommended, which can be viewed as a multi-category version of the area beneath the ROC graph. The generalized discrimination score, *D*, is defined as

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

where m is the number of categories, n_k is the number of times the observation was in category k, $\mathbf{p}_{k,i}$ is the vector of forecast probabilities for the i^{th} observation in category k, and

$$I\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) = \begin{cases} 0.0 & \text{if } F\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) < 0.5 \\ 0.5 & \text{if } F\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) = 0.5 \\ 1.0 & \text{if } F\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) > 0.5 \end{cases}, \tag{1b}$$

and where

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

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

Although Eq. (1) may seem complicated, its interpretation is fairly simple: What is the probability of successfully discriminating the wetter (or warmer) of two observations? Equation (1a) compares each of the observations in the normal and above normal categories with each of those in the below-normal category in turn, and calculates the probability that the forecasts

correctly point to the observation in the normal or above-normal category as the wetter (warmer). This procedure is then repeated comparing each of the observations in the normal category with each of those in the above-normal category. The selection of the wettest observation is based on Eq. (1c), which defines the probability that a value randomly drawn from $\mathbf{p}_{l,j}$ will exceed one randomly drawn from $\mathbf{p}_{k,i}$. If this probability is greater than 0.5 [Eq. (1b)] the forecasts suggest that it is more likely that the second observation (that corresponding to $\mathbf{p}_{l,j}$) is wetter (or warmer) than the first (that corresponding to $\mathbf{p}_{k,i}$).

The generalized discrimination score can be calculated for each location, and then a map of the scores can be drawn to indicate areas in which the forecasts have some discrimination. The score has an intuitive scaling that is appealing to many non-specialists: it has an expected value of 50% for useless forecast strategies (guessing, or always forecasting the same probabilities), and good forecasts will have a score greater than 50%, reaching 100% given perfect discrimination. Scores of less than 50% indicate bad forecasts (forecasts that can discriminate, but which indicate the wrong tendency – for example, high forecast probabilities on below-normal indicate a low probability that below-normal rainfall will actually occur), and can reach a lower limit of 0% given perfectly bad forecasts. The score represents the most meaningful answer to the naïve question: "How often are the forecasts correct?" without having to reduce the forecasts to a deterministic format. One major problem with the score is that it is insensitive to the reliability of the forecasts, and so is unaffected by any monotonic transformation of the forecasts. This problem is considered less severe than for the insensitivity of the area beneath the ROC graph (discussed below) to reliability of the forecasts because of the way in which the probabilities are compared across the categories in the generalized discrimination score. With the ROC area, for example, all the probabilities could be multiplied by a factor k, where $0 < k < 1$, and the area is unaffected, but since the probabilities across the categories have to add to 1.0, simple rescalings cannot be applied without affecting the probabilities in the other categories, and thus also affecting the probability of successfully discriminating observations.

For a number of reasons the generalized discrimination score is recommended in place of the more commonly used ranked probability skill score (RPSS) as a summary measure of skill to indicate where the forecasts are good. Firstly, the RPSS does not have any intuitive interpretation, which essentially renders it an abstract number to most non-specialists. Secondly, the RPSS's scaling can be confusing: while positive values indicating skill should be simple enough to understand, the fact that it does not have a lower bound of -100% means that the score is asymmetric, so that forecasts with a score of -50% are not as bad as forecasts with a score of 50% are good. But even the idea of having 0% as no skill rather than 50% seems much more logical to verification experts than it does to users with only weak mathematical backgrounds.

The main reason for not recommending the RPSS is that it is frequently misinterpreted even by forecasters, and typically results in a more pessimistic view of the quality of the forecasts than is warranted. There is a widespread belief that if the score is less than zero the forecasts are worse than climatological forecasts, and the user would therefore have been better off with the latter, but this is not necessarily the case. Consider a set of 10 forecasts, one per year, five of which indicate a probability of 60% for above-normal rainfall, and the other five only a 10% probability. The climatological probability of above-normal rainfall is 33%. For the sake of simplicity, the below-normal and normal categories can be combined. Now imagine that for two of the years for which a 60% probability was issued rainfall was above-normal; 40% of these years were thus above-normal, and the forecasts have successfully, but over-confidently, indicated an increase in the probability of above-normal rainfall. For the years in which a 10% probability was issued, only one of these was above-normal; thus 20% of these years were above-normal, and the forecasts have successfully, but over-confidently, indicated a decrease in the probability of above-normal rainfall. These forecasts appear to contain useful information because they have correctly indicated increases and decreases in the probability of above-normal rainfall. However, the RPSS (which in this case is equivalent to the Brier skill score, because the normal and below-normal categories are combined) is about -7%, indicating negative skill. The score is slightly worse still if

it is objected that above-normal rainfall was observed only 30% of the time over the 10-year period rather than 33%. The problem is that the RPSS has an arbitrary requirement that the resolution of the forecasts must be greater than the errors in the reliability, and since these forecasts show marked over-confidence they score badly[9]. By trying to measure resolution and reliability at the same time, the score becomes difficult to interpret, whereas the generalized discrimination score measures only the one attribute, and thus provides a simpler indication of whether the forecasts might be useful.

In addition to being useful for mapping discrimination, the score may also be used for providing an indication of the overall discriminatory power of all the forecasts by pooling all the forecasts for all locations and seasons. Note that this pooling does not normally result in a simple average of the score for the individual locations because probabilities now have to be compared for different locations, and so some calibration problems at individual locations that were not reflected in the score are more likely to be detected when the forecasts are pooled. It is not even possible to state whether the pooled score will be less than or greater than the average of the individual scores.

From section 3d it may be concluded that the pooling of forecasts for different locations is invalid because the implied interpretation of the forecasts was shown to be invalid – an area with a 10% probability on above-normal does not mean that 10% of the area is expected to be above-normal. However, it was argued that this problem of interpretation is partly a sampling issue, and so the pooling of the forecasts helps to increase the sample size. For this reason, spatial pooling of forecasts is widely practised, and is recommended for some of the verification procedures in the CBS SVSLRF.

An example of the calculation of the generalized discrimination score is provided in appendix B, section a.*i*.

ii. Diagnosing discrimination

The generalized discrimination score provides an indication of the ability of the forecasts to discriminate wetter (or warmer) observations from drier (or cooler) ones. In most cases it is likely that seasonal forecasts are good at discriminating observations in the outer two categories, but are not as successful at discriminating observations in the normal category. In addition, in some cases it may be the case that the forecasts are good at discriminating observations in either the above- or below-normal categories, but not both. In all these examples, the implication is that the forecasts contain useful information only sometimes. It will not be possible to discern these subtleties from a single score, and the single score may even hide the fact that there is some useful information in some of the forecasts since it looks at all the forecasts together. It is therefore recommended that the ability of the forecasts to discriminate observations in each of the categories be calculated. The generalized discrimination score [Eq. (1)] can be used for these purposes, but reduces to a statistic more widely known as the area beneath the ROC graph (assuming that the area is calculated using the trapezoidal rule). The calculation of these areas for each of the forecast categories is recommended. The use of the trapezoidal rule is recommended rather than using the normal distribution assumption, which gives a smoother fit, for the following reasons. Firstly, if the trapezoidal area is calculated, the area has an intuitive interpretation: it defines the probability of successfully discriminating observations in the respective category. Secondly, the smoothed fit is suitable for convex or concave ROC curves, but these are much more common with shorter

[9] Some attempts have been made to correct these so-called biases in the ranked probability skill score by introducing an adjustment for the uncertainty in the climatological probabilities, but the corrections are applicable only for ensemble prediction systems, and so it is not clear how they could be applied for consensus forecasts. These debiased scores are considered useful in the context of the CBS SVSLRF, which targets GPC products, but cannot be applied in the current context. Besides, the criticism remains that these scores are still abstract numbers, and so are difficult to understand by all but specialists in forecast verification.

range forecasts than with seasonal forecasts, where the curve can be quite irregular. Thirdly, the smoothed fit is most suitable for ensemble prediction systems where the resolution of the ROC graph is constrained by the number of ensemble members (assuming a counting system for the probabilities), but seasonal forecasts are currently issued purposely as discrete probabilities in integrals of 5% (except for the climatological probability), and it is recommended that the graph be constructed at this resolution.

In constructing an ROC graph, one of the categories is selected as the current category of interest, and an occurrence of this category is known as an event. An observation in any of the other categories is defined as a non-event. Separate ROC graphs are completed for each category. Setting $p_{1,j}$ as the forecast probability for the j^{th} observed event, and $p_{0,i}$ as the probability of an event for the i^{th} non-event, the ROC area, A, can be calculated without constructing the graph using

$$A = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I\left(p_{0,i}, p_{1,j}\right), \tag{2a}$$

where n_0 is the number of non-events, and n_1 the number of events, and the scoring rule[10] $I\left(p_{0,i}, p_{1,j}\right)$ is defined as

$$I\left(p_{0,i}, p_{1,j}\right) = \begin{cases} 0.0 & \text{if } p_{1,j} < p_{0,i} \\ 0.5 & \text{if } p_{1,j} = p_{0,i} \\ 1.0 & \text{if } p_{1,j} > p_{0,i} \end{cases}. \tag{2b}$$

However, given that the ROC graph itself is a useful diagnostic, it may be more convenient to calculate the ROC area directly from the graph, as explained below.

The interpretation of the ROC area is similar to that of the generalized discrimination score. If the category of interest is above-normal, the score indicates the probability of successfully discriminating above-normal observations from normal and below-normal observations. The scaling is identical to that for the generalized discrimination score, with a score of 50% representing no skill, 100% indicating perfect discrimination, and 0% indicating perfectly bad discrimination. Given this simple interpretation, the scores for the separate categories are considered suitable for communication to non-specialists, although the name of the score is unnecessarily intimidating, and it is recommended that if a name for the score needs to be communicated it be described as the discrimination score.

The ROC area is equivalent to the Mann-Whitney *U*-statistic after some simple rescaling, and this equivalency may help in understanding the score (Mason and Graham 2002). The Mann-Whitney *U*-test is often used to compare the central tendencies of two sets of data, and is a non-parametric version of the more widely used Student's *t*-test. 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 probability was higher, on average, when an event occurred compared to when not.

The ROC graph is constructed by calculating the ability of the forecasts to successfully identify the events. Starting with the forecasts with highest probabilities, the observations that are most confidently indicated as events are highlighted. The events that are selected are called "hits". The proportion of all events thus selected is calculated, and is known as the hit rate (HR), or probability of detection (POD):

[10] A scoring rule is the form of the score for a single forecast-observation pair.

$$\mathrm{HR} = \frac{\text{number of hits}}{\text{number of events}}. \tag{3}$$

It is possible that some non-events have been selected incorrectly, and these are known as "false alarms". The proportion of non-events incorrectly selected [the false-alarm rate (FAR)] is calculated also:

$$\mathrm{FAR} = \frac{\text{number of false-alarms}}{\text{number of non-events}}. \tag{4}$$

The hit and false-alarm rates are commonly tabled (see example in appendix B section a.*ii*), and given the general practice in seasonal forecasting that probabilities are rounded to the nearest 5% (except for forecasts of the climatological probability), it is recommended that the table be constructed for each discrete probability value.

If the forecasts have no useful information, the hit and false-alarm rates will be identical, but if the forecasts can discriminate the events, the hit rate will be larger than the false-alarm rate. Since it is unlikely that all the events were correctly selected using only the forecasts with highest probabilities, additional selections are made using the next highest probability, and the hit and false-alarm rates are updated. The difference in the increments of the hit and the false-alarm rate is expected to be a little less than at the first step, since we are less confident about having correctly selected the events. These steps are continued until all the events have been selected. The hit rates are then plotted against the false-alarm rates. See the example in appendix B section a.*ii* for more details.

To calculate the area beneath the curve by the trapezoidal rule, the following equation can be used:

$$A = 0.5 \times \left[1 + \sum_{k=0}^{d} \left(y_k x_{k+1} - y_{k+1} x_k \right) \right]. \tag{5}$$

where d is the number of discrete probability values, and y_1 and x_1 are the hit and false-alarm rates for the highest probability value only, y_2 and x_2 are the rates for the highest and second highest probabilities, etc.. For $i = 0$ the hit rate and false-alarm rates are defined as 0.0, and for $i = d + 1$ they are defined as 1.0 to ensure that the curve starts in the bottom-left, and ends in the top-right corners, respectively.

The ROC curves can provide some useful diagnostics about the quality of the forecasts, and some guidelines for interpreting the curves are provided below by considering some idealized curves as indicated in Figure 2. For good forecasts, it has just been argued that the hit rate will be initially much larger than the false-alarm rate, and so the graph should be fairly steep near the left-hand corner. As the forecasts with progressively lower probabilities are used, hits are likely to accumulate at a progressively slower rate, while false-alarms will be accumulated at a progressively faster rate, and so the curve is likely to be convex (Figure 2*a*). Conversely, if the forecasts are bad only relatively few events will be selected initially, and the curve will be concave (Figure 2*b*). The more successfully the forecasts can discriminate the events, the steeper the curve will be near the left, and the shallower the curve will be near the right, and will thus embrace more area (Figure 2*c*). If all the events are selected before any of the non-events then the hit rate will reach 1.0 while the false-alarm rate is still zero, and so the area under the curve will be 1.0 (Figure 2*d*). Forecasts that used a guessing strategy, or one that is just as naïve as a guessing strategy, will score similar hit and false-alarm rates, and so the curve will follow the 45° diagonal, and enclose an area of 0.5 (Figure 2*e*). This area should be compared with the 50% success rate for discriminating events and non-events that is characteristic of useless forecasts, and to which it is equivalent. Perpetual forecasts of the same probabilities (including climatological forecasts) will select all or none of the observations at the same time since there is no basis for selecting some over others. The hit and false-alarm rates will then both be 1.0 if all

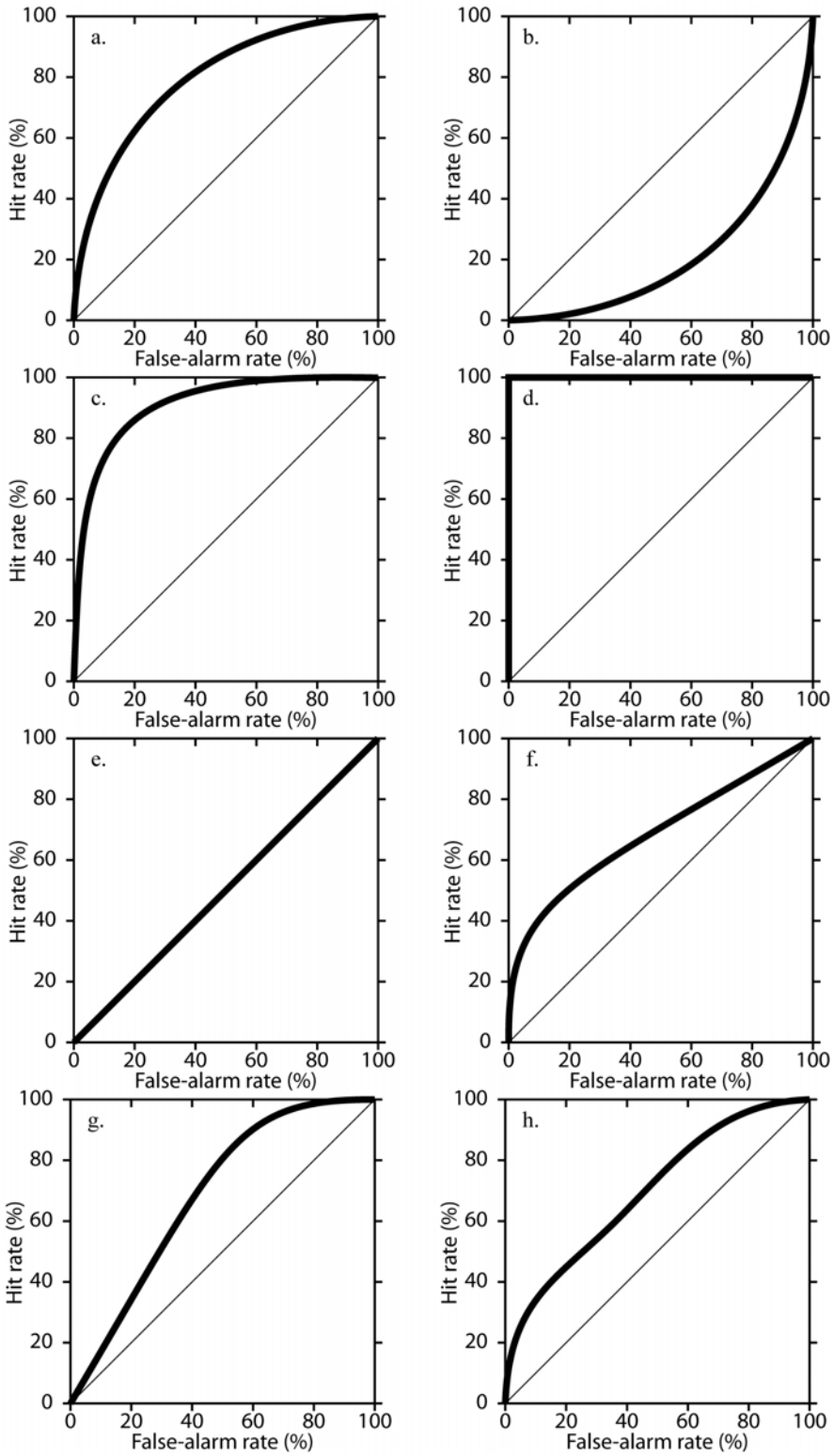

Figure 2. Idealized examples of ROC curves showing forecast systems with (*a*) good discrimination and good skill, (*b*) good discrimination but bad skill, (*c*) excellent discrimination, (*d*) good discrimination, (*e*) no discrimination, (*f*) good discrimination for high probability forecasts, (*g*) good discrimination for low probability forecasts, and (*h*) good discrimination for confident (high and low probability) forecasts.

the observations are selected, and both 0.0 when none of the events are selected, and the curve joining these points will again follow the 45° diagonal, and enclose an area of 0.5. It is possible that only the forecasts with the highest probabilities are useful in discriminating events, in which case the curve will be initially steep, but then flatten out (Figure 2*f*). These forecasts successfully indicate when the probability of an event is greatly inflated, but are otherwise useless. Sometimes the forecasts are useful when they are equally confident, but confident that an event is unlikely to happen. These forecasts contain useful information only when the probabilities are low, and so the curve will be initially close to 45°, but will flatten out towards the top right (Figure 2*g*). More commonly, the seasonal forecasts are useful when the probabilities are either high or low, but are not useful when confidence is low and the probabilities take more intermediate values. In this situation the curve is close to 45° in the middle, but is steep near the left and shallow near the right (Figure 2*h*).

As with the generalized discrimination score the ROC areas can be mapped, and calculated using forecasts pooled from all, or a regional selection of, locations. The pooled scores are unlikely to be a simple average of the scores for the individual locations. If, for example, all the forecasts at location A are biased high, but those at location B are biased low, the scores at the individual stations will be unaffected because of the insensitivity of the ROC area to monotonic transformations of those probabilities. It is possible that the discrimination at both stations is good, despite the poor reliability (over-forecasting at station A, and under-forecasting at station B). When these forecasts are pooled, the events at location B are likely to be poorly discriminated from the non-events at location A because the probabilities for the former are too low, and those for the latter are too high. For example, imagine that we are measuring the ability to discriminate below-normal rainfall over a domain that includes station A and B. At station A, when the probability for below-normal rainfall is relatively high (by station A's standards; e.g., 70%) below-normal rainfall occurs much more frequently than when the probability is relatively low (e.g., 50%). Similarly, at station B, when the probability for below-normal rainfall is relatively high (by station B's standards; e.g., 40%) below-normal rainfall occurs much more frequently than when the probability is relatively low (e.g., 10%). Although below-normal rainfall is discriminated well at the individual stations, when forecasts at station B are compared with those at station A, the inflated probabilities at station A (50% for when below-normal occurs infrequently) and deflated probabilities at station B (40% for when below-normal occurs frequently) result in poor discrimination.

iii. Measuring resolution

As defined in section 3c.*i*, resolution measures how different the outcomes are given different forecasts. A commonly used measure of resolution is the resolution component of the Murphy (1973) decomposition:

$$\text{resolution} = \frac{1}{n}\sum_{k=1}^{d} n_k \left(\overline{y}_k - \overline{y}\right)^2 , \tag{6}$$

where n is the total number of forecasts, d is the number of discrete probability values (or number of probability bins), n_k is the number of forecasts for the k^{th} probability value, $\overline{y}_k$ is the observed relative frequency for that value, and $\overline{y}$ is the observed relative frequency for all forecasts. The observed relative frequency for the k^{th} forecast probability, $\overline{y}_k$, is the number of times an event occurred divided by the number of times the respective probability value was forecast:

$$\overline{y}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} y_{k,i} . \tag{7}$$

where n_k is the number of forecasts of the k^{th} forecast probability, and $y_{k,i}$ is 1 if the i^{th} observation was an event, and is 0 otherwise. The individual values of Eq. (7) are often plotted on a reliability diagram (see section 4b.*vi*).

Equation (6) defines the variance of the observed relative frequencies for different forecast probabilities. The larger the variance, the more the outcome is conditioned upon the forecast, and so a resolution score of 0.0 is undesirable: an event occurs with the same relative frequency regardless of the forecasts and so the forecasts are useless. The resolution score has a maximum value of $\bar{y}(1-\bar{y})$, which indicates that the occurrence of an event or non-event can be determined with complete success by the forecasts, but which does not necessarily indicate that the forecasts are good because, as Eq. (6) indicates, the forecast probabilities are not explicitly considered, and so the score does not specifically measure whether the event occurs more frequently as the forecast probability increases. Thus, the resolution score distinguishes useless from potentially useful forecasts, but is unable to distinguish between forecasts that are useful because they are good from those that are useful because they are so bad that one can be confident they are giving completely the wrong message, or even from those for which the observed relative frequency increases and decreases randomly with changing forecast values![11] For this reason it is not recommended that a skill score version of the resolution score be calculated: positive skill could indicate either that the forecasts are better than the reference or that the forecasts are worse than the reference forecasts are good.

One other problem with Eq. (6) is that it has large sampling errors given a small number of forecasts because $\bar{y}_k$ needs to be estimated accurately for each value of the forecast probability. As a result, it is usually necessary to pool forecasts from a number of different locations. Alternatively, if spatial diagnostics of resolution are required, Eq. (6) could be applied by binning the forecast probabilities into a small number of bins[12]. The bins will likely have to be very crude given realistic sample sizes, and bins of <30%, 30-35% inclusive, and >35% are suggested, representing forecasts of decreased, near-climatological, and increased probabilities, respectively. However, experiments will have to be conducted to test how accurately the observed relative frequency is calculated for each bin. Errors in calculating these values are binomially distributed (Bröcker and Smith 2007a), so the user can get an indication of the expected size of errors given their sample of forecasts.

A variation on the idea of binning the forecasts and then calculating the resolution score is simply to calculate Eq. (7) for the bin with the highest probabilities. This procedure amounts to calculating the hit (Heidke) score for the highest probability forecasts. The hit score is derived from Eq. (7):

$$\bar{y}_d = \frac{1}{n}\sum_{i=1}^{n} y_{d,i} \ . \tag{8a}$$

where the largest probability bin, *d*, is defined flexibly to include the highest probability for each forecasts (hence there are *n* forecasts in this bin). The hit score is usually expressed as a percentage, and can be more easily interpreted as

$$\text{hit score} = \frac{\text{number of hits}}{\text{number of forecasts}} \times 100\% , \tag{8b}$$

where a "hit" is the occurrence of the category with the highest probability, (i.e., $y_{d,i} = 1$). Adjustments to the score can be made if two or more categories tie for the highest probability. In this case, if one of two categories with tied highest probabilities verifies, score a half-hit, or a third-hit if one of three categories with tied highest probabilities verifies. The hit score ranges from 0% for the worst possible forecasts (the category with the highest probability never verifies) to 100% for the best possible forecasts (the category with the highest probability always verifies).

[11] In the limiting case of $d = n$ (the number of probability bins is equal to the number of forecasts), the resolution term reduces to its maximum value of $\bar{y}(1-\bar{y})$, which is the same as the uncertainty term in the Brier score decomposition, and, like the uncertainty term, contains no information about the quality of the forecasts.

[12] Note that the binning of probabilities results in a deterioration in skill through the inclusion of additional terms into the Murphy (1973) decomposition of the Brier score (Stephenson *et al.* 2008).

In both cases resolution is strong. For forecasts with no resolution the score depends upon the climatological probabilities, as discussed below.

In some of the RCOFs a half-hit has been scored if the forecasts have the "correct tendency" – specifically when the below-normal (or above-normal) category is observed, and the probability for that category is higher than the climatological probability, but the highest probability is on the normal category. This practice should be discouraged because it is unclear what the expected score of no-resolution forecasts would be, and because it becomes unclear what the score actually means. Mixing different meanings of a "hit" in a single score, can be very confusing, and generally gives the impression that the forecasts are better than they really are. At the least, if the interest is in whether the forecasts are indicating the correct "tendency", "hits" should not be scored for correct forecasts of the normal category, but only on the basis of whether the tendency in the forecast correctly indicated whether the observation was above- or below-average (or, better still, above- or below-median). If the forecaster wishes only to forecast the tendency (whether rainfall will be above- or below-median) then two-category, instead of three-category, forecast probabilities should be issued.

It was argued in section 3b that the practice of measuring the quality of seasonal forecasts by counting the number of times the category with the highest probability verified is not specifically recommended because, as should be evident from a comparison of Eq. (8a) with Eqs (6) and (7), the hit score gives an extremely limited perspective on the resolution of the forecasts. Nevertheless, the hit score does have a simple interpretation, and the question "How often does the category with the highest probability verify?" is perfectly reasonable, as long as it is acknowledged that much of the important information in the forecast is being ignored. Therefore, rather than explicitly recommending against calculating hit scores, it is recommended that the score not be used as a primary verification measure, and that additional information be provided about resolution (and about the other attributes of good probabilistic forecasts). It is proposed that a more complete view of resolution be obtained using the hit score by asking the additional questions "How often does the category with the second highest probability verify?" and "How often does the category with the lowest probability verify?" These questions can be answered by generalizing the hit score to

$$\overline{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{j,i}\,, \tag{8c}$$

where n is the number of forecasts, and $y_{j,i}$ is 1 if the i^{th} observation was in the category with the j^{th} highest probability, and is 0 otherwise. Note that the subscript j has a distinct meaning from the subscript k in Eq. (7): in Eq. (8c) $\overline{y}_j$ for $j = 1$ is the hit score for the category with the highest probability, whereas in Eq. (7) $\overline{y}_j$ for $j = 1$ is the hit score for the lowest probability bin. An example is provided in appendix B section a.*iii*.

Calculating hit scores for the categories with the second and third highest probabilities is preferable to calculating the number of "bad misses" (two-category errors in a three category system) because it does not encourage hedging towards normal. The additional information about the resolution also can be informative. For example, if the hit score for the highest probability categories is only marginally larger than for climatological forecasts, there may still be reasonably good resolution, and therefore useful information, if the category with the lowest probability verifies only very rarely. In the context of some of the RCOFs, for example, where there is a tendency to hedge towards highest probabilities on the normal category (Chidzambwa and Mason 2008), the hit score may be fairly low because of an inability to score a hit when above- or below-normal verify, but these outer categories are rarely given identical probabilities, and it is possible that the forecasts successfully point to the category least likely to verify even with the hedging. Such questions help to indicate how the outcome verifies given different forecasts rather than just how the outcome differs from climatology given high probabilities. Of course, if sample sizes

permit, it is preferable to calculate the hit score for all values of the forecast probability, and to analyse these in the context of the reliability diagram, as discussed in section 4b.*vi*.

Because the expected value of the hit score for forecasts with no resolution depends upon the base rate (and upon the forecast strategy if the categories are not equiprobable), the score can be a little difficult for non-specialists to interpret. A skill score version of Eq. (8) is therefore frequently calculated of the form

$$\text{hit skill score} = \frac{\text{number of hits} - \text{expected number of hits}}{\text{number of forecasts} - \text{expected number of hits}} \times 100\% \,. \qquad (9)$$

where the expected number of hits is calculated for the reference (no-resolution) forecasts. However, Eq. (8) generally has a simpler meaning than Eq. (9), and so the former may be preferable for communication to non-specialists; if so, the expected number of hits should be indicated. The expected number of hits is usually calculated from either a strategy with no resolution, (e.g., of random guessing or of perpetual forecasts of one category). It should be noted that the score for perpetual forecasts of one category may not be the same as that of random forecasts if the verification period has a different relative frequency of each category compared to the climatological period. The scores for the perpetual forecasts will differ depending upon which category is selected. Which strategy to use as the reference depends on whether it seems reasonable to have predicted the trend. For example, if the forecasts are for temperature, then it may seem a more reasonable strategy to predict perpetual above-normal temperatures than to randomly guess because of the known global warming signals. For precipitation, however, it may not seem quite so obvious whether any trend could have been predicted so easily.

As an alternative skill score to Eq. (9) the difference between the hit scores for the categories with the highest and lowest probabilities would provide a simple indication of resolution. This skill score would range from 100% for forecasts that always identify the correct category, to 0% for forecasts with no resolution (the category with the lowest probability occurs just as often as the category with the highest probability), to -100% for forecasts with perfect resolution, but which consistently point to the wrong category. One limitation of this skill score is that it considers the scores only for the outer categories, and so is a rather incomplete measure of resolution. It is quite possible, for example, for the category with the second highest probability to verify most (or least) frequently, in which case the forecasts do have some resolution even if the scores for the outer categories are identical. However, as discussed in section 3c.*i*, considering resolution of this type as good seems questionable, one would have to have a very good reason for believing that the forecasts have a genuine systematic error before considering them to give reliable indications of the most likely outcome.

Use and interpretation of the hit scores in any of the forms above are likely to be complicated if the categories are not equiprobable. In this case, a recommended alternative would be to calculate separate hit scores: one for when probabilities for any of the categories are increased, and one for when they are decreased. These results are likely to be most informative if they are performed for each category separately.

iv. Measuring reliability

The generalized discrimination score, the ROC areas, and the resolution score were all recommended because they measure the attributes of discrimination or resolution, which were considered essential measures of whether the forecasts may be potentially useful. However, in focusing only on these attribute, the reliability of the forecasts has been ignored, and so additional scores are recommended. Ideally it would be useful to measure reliability in isolation from other attributes, but the only option is the reliability component of Murphy's (1973) decomposition of the Brier score, which is subject to large sampling errors (large numbers of forecasts for each discrete value of the forecast probability are required to calculate the observed relative frequency

of an event accurately). The score is therefore very noisy when calculated at individual locations. However, it is a recommended procedure for calculating the reliability of forecasts when pooled from a number of locations. Given *d* discrete probability values

$$\text{reliability} = \frac{1}{n}\sum_{k=1}^{d} n_k \left(\overline{p}_k - \overline{y}_k \right)^2 , \tag{10}$$

where, n_k is the number of forecasts for the k^{th} probability value ($\overline{p}_k$), and $\overline{y}_k$ is the observed relative frequency for that value. As with the resolution score, if spatial diagnostics of reliability and resolution are required, Eq. (10) could be applied by binning the forecast probabilities into a small number of bins. The reliability score measures "errors" in the reliability of the forecasts, and so ranges from 0.0 for perfectly reliable forecasts to a maximum value of 1.0, which is only possible if the forecasts were perfectly bad (the forecast probability was 100% every time an event did not occur, and 0% every time an event did occur.

A skill score version of the reliability score could be calculated of the form

$$\text{reliability skill score} = \left(1 - \frac{\text{reliability of forecasts}}{\text{reliability of reference forecasts}} \right) \times 100\% , \tag{11}$$

This score ranges from 0% for no improvement in reliability, to 100% if the forecasts are perfectly reliable (and the reference forecasts are not). Negative values of the score indicate deterioration in reliability. It should be noted that although the reliability of climatological forecasts is unlikely to be perfect (because the climatology is defined using separate data from that over the verification period) it is likely to be very good unless there is a significant change in climate between the climatological and verification periods, and / or if there are large sampling errors. Reliability skill may therefore be low or negative.

As discussed in section 3c.*iii*, errors in reliability can be conditional or unconditional. The reliability score does not distinguish between these different errors. Scores for measuring the two components separately are recommended in section 4b.*vi*.

v. Measuring other attributes

In the absence of scores that are ideally suited for measuring reliability and resolution at specific locations, the next best option is to use a score that measures all the important attributes of good probabilistic forecasts. The Brier and ranked probability skill scores are logical alternatives, but for the reasons mentioned in section 4a.*i* their interpretation needs to be considered carefully. Specifically, their skill score versions may need to be avoided, but the scores themselves can be usefully mapped, although the ignorance score transformed to an effective interest rate is recommended in preference.

The (half-) Brier score[13] is calculated on each category, *j*, separately, and is defined as

$$BS_j = \frac{1}{n}\sum_{i=1}^{n} \left(y_{j,i} - p_{j,i} \right)^2 . \tag{12}$$

where *n* is the number of forecasts $y_{j,i}$ is 1 if the i^{th} observation was in category *j*, and is 0 otherwise, and $p_{j,i}$ is the i^{th} forecast probability for category *j* (Brier 1950). The score is the average squared "error" in the forecasts, and it ranges between 0% for perfect forecasts (a probability of 100% was assigned to the observed category on each forecast) to 100% for perfectly bad forecasts (a probability of 0% was assigned to the observed category on each forecast). An example is provided in appendix B section a.*iv*.

[13] Strictly speaking, the Brier score is calculated over all categories, but when there are only two categories the score is identical for each, and so it is standard to calculate the score for only the category defined as "events" (Wilks 2006).

The ranked probability score (RPS) is defined as

$$RPS = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{k=1}^{m-1} \left(\sum_{j=1}^{k} \left(y_{j,i} - p_{j,i} \right) \right)^2 , \tag{13}$$

where n is the number of forecasts, m is the number of categories, $y_{j,i}$ is 1 if the i^{th} observation was in category j, and is 0 otherwise, and $p_{j,i}$ is the i^{th} forecast probability for category j (Epstein 1969; Murphy 1969, 1970, 1971). The score is the average squared "error" in the cumulative probabilistic forecasts, and it ranges between 0% for perfect forecasts (a probability of 100% was assigned to the observed category on each forecast) to a maximum of 100% that can only be achieved if all the observations are in the outermost categories, and if the forecasts are perfectly bad forecasts (a probability of 100% was assigned to the opposite outermost category to that observed). The summation over j is to m-1 rather than to m because the cumulative forecasts and observations over all categories are both 100%. An example is provided in appendix B section a.*v*.

Given the problems in interpreting the skill score versions of the Brier score and the RPS, the use of the ignorance score (Roulston and Smith 2002) is recommended, which is defined as

$$\text{ignorance score} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{j,i} \log_2 \left[p_{j,i} \right], \tag{14}$$

where n is the total number of forecasts, m is the number of categories, $y_{j,i}$ is 1 if the i^{th} observation is in category j, and is 0 otherwise, and $p_{j,i}$ is the corresponding forecast probability. The score ranges from 0.0 for a perfect set of forecasts, to infinity for a perfectly bad set of forecasts[14]. This score for the perfectly bad forecasts may seem unduly harsh, but could be considered appropriate given that a 0% probability implies that the category in question is an absolute impossibility.

It is recommended that the ignorance score be transformed to an effective interest rate, for ease of interpretation (Hagedorn and Smith 2008). This transformation involves comparing the ignorance score of the forecasts against that for forecasts of climatological probabilities (Tippett and Barnston 2008):

$$\text{effective interest rate} = \left(2^{Ign(ref) - Ign} - 1 \right), \tag{15}$$

where $Ign(ref)$ is the ignorance score for the reference (climatological forecasts), and Ign is the score for the forecasts. The effective interest rate provides an indication of the average returns a gambler would make if (s)he betted on the forecasts, and was paid out against odds based on the climatological probabilities. For example, given three equi-probable categories, a gambler would be paid out three times the amount bet on the verifying category. Imagine a gambler who has \$100 to bet on the three categories, and (s)he chooses to divide the bet between the three categories based on the forecast probabilities[15]. For example, if the forecast indicates a 50% chance of category-1, the gambler bets \$50 on this category, and the rest on the other categories, and then if category-1 occurs (s)he would be paid out \$150, making a profit of \$50 (or 50%). If the gambler had bet an equal amount on each category (\$33.33), then no matter which category verified (s)he would be paid out \$100, and would thus break even. The ignorance score for a 50% probability on the verifying category is 1.0, and for 33% on the verifying category is about 1.58; applying Eq. (15) yields an effective interest rate of 50%.

[14] In fact, if any of the forecasts has a probability of 0% for the verifying category the score will be infinity even if all the other forecasts have 100% on the verifying categories.

[15] Betting proportionally on the forecast probabilities is a strategy that maximizes the growth in the initial \$100. This strategy is distinct from maximizing the expected profit, which would involve betting all the money on the category with the highest probability, and thus creating a chanc of winning a very large amount of money over the long run, but at the very high risk of going bankrupt.

Given a series of forecasts, the gambler may choose to bet all takings from the previous round, dividing it according to the forecast probabilities, as before. So, for example, after the first round the gambler now has \$150 to bet on the second round. Assuming that the probabilities for the three possible outcomes are now 40%, 35%, and 25%, the gambler would then bet \$60 (40% of \$150) on the first category, \$52.50 on the second, and \$37.50 on the third. If the first category verifies again, the gambler would now have \$180. The initial \$100 has grown by 80%, but over two periods: 50% was made after the first round (from \$100 to \$150), and 20% in the second (from \$150 to \$180). The gain over the two rounds represents an average of about 34% per round[16]. The ignorance score for the forecasts after two rounds is about 1.16, and for the climatological forecasts is about 1.58. Applying Eq. (15) yields the average rate of return of about 34%. A more detailed example is provided in appendix B section a.*vi*.

If at any round a category verifies that was assigned a zero probability, the gambler will lose all their money, and will have nothing to invest in further rounds. In this case, the ignorance score is infinity, and the effective interest rate becomes -100%, which is its lower bound. The upper bound depends on the climatological probabilities of the events, which set the returns the gambler can make. If the forecasts are perfect (they always indicate 100% on the verifying category) then the ignorance will be zero, and so the effective interest rate reduces to $2^{Ign(ref)}-1$, which in the case of three equi-probable categories is 200%. An effective interest rate of zero indicate that in the long-term the gambler will neither win nor lose betting on the forecasts, and they thus contain no useful information. Although the effective interest rate does not have an upper bound of 100%, it can be considered a skill score because it compares one set of forecasts to a reference set, and any positive values indicate an improvement over the reference.

When calculating the effective interest rate using Eq. (15) it is assumed that all the forecasts are for a single location and that the *n* forecasts are for discrete periods (typically, a specific three-month season over a number of years) so that all the winnings from the previous year's forecast can be bet on the subsequent year. If some of the forecasts are for different locations, then the initial investment has to be divided between each of the *s* locations, and the effective interest rate has to be averaged using the ignorance score for each location:

$$\text{average effective interest rate} = \frac{1}{s}\sum_{k=1}^{s}\left(2^{Ign\,(ref)-Ign_k}-1\right), \tag{16}$$

where Ign_k is the ignorance score for the k^{th} location. Similarly, if some of the forecasts are for separate, but overlapping target periods (or, more specifically, if any of the target periods expire after any of the forecast dates for subsequent forecasts) the bet has to be divided between the different periods since the returns on at least some of the initial bets are not yet available for reinvestment. The effective interest rate again has to be calculated using Eq. (16).

From Eqs (14) – (16), it can be shown that when the forecast probability on the verifying outcome exceeds the climatological probability $Ign < Ign(ref)$ and so the effective interest rate for that step is positive (the gambler makes a profit). However, when the forecast probability on the verifying outcome is less than the climatological probability $Ign > Ign(ref)$ and the gambler makes a loss. Even with very good forecasts, a loss should be expected sometimes of course (if categories with low probabilities never verified these low probabilities would be unreliable), but in the long run the wins should exceed the profits, and the effective interest rate will be greater than zero. To illustrate the fact that there are likely to be both gains and losses over a period of betting on forecasts, it is recommended that accumulated profits graphs be drawn (Hagedorn and Smith 2008). These graphs show a plot of

$$\left(\prod_i \frac{p_i}{c_i}\right)-1 \tag{17}$$

[16] The geometric mean of 50% and 20% is about 34%.

on the *y*-axis against time, *i*, on the *x*-axis, where p_i is the forecast probability for the verifying category, and c_i is its climatological probability. An example is provided in appendix B section a.*vii*. If Eq. (16) had to be used to calculate the effective interest rate because forecasts for different locations and / or overlapping forecasts were pooled in the calculation, Eq. (17) has to be modified to

$$\left(\prod_i \left(\frac{1}{s} \sum_{k=1}^{s} \frac{p_{k,i}}{c_{k,i}} \right) \right) - 1, \tag{18}$$

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 climatological probability.

vi. Detailed diagnostics

The procedures recommended thus far have provided only minimal diagnostic information. For more detailed information on forecast quality, reliability diagrams are recommended. These diagrams provide useful indications of most of the important attributes of forecast quality that were described in section 3c. Reliability diagrams are based on a diagnosis of probabilistic forecasts for a predefined set of events, and so can be constructed for each of the categories separately. However, it is not required that the definition of an event remain fixed, and so a single diagram can be constructed for all the categories combined. Both approaches are recommended; the diagrams for the individual categories are useful for indicating whether the quality of the forecasts depends on the outcome, while the combined diagram is useful for examining whether the probabilities can be interpreted consistently across the categories.

The basic idea of the reliability diagram is simple, but the diagram contains a wealth of information about the quality of the forecasts. For each discrete value of the forecast probability, the reliability diagram indicates whether the forecast event occurred as frequently as implied. The different forecast probabilities are plotted on the *x*-axis, and on the *y*-axis the "observed relative frequency" of the event is shown. The observed relative frequency for the k^{th} forecast probability, $\overline{y}_k$, is the number of times an event occurred divided by the number of times the respective probability value was forecast [Eq. (7)]. An example is provided in appendix B section a.*viii*.

The interpretation of reliability diagrams may be facilitated by considering some idealized examples as shown in Figure 3. If the forecasts are perfectly reliable then the observed relative frequency will equal the forecast probability for all values of the forecast probability, and so the reliability curve will lie along the 45° diagonal (Figure 3*a*). In Figure 3*b* the event occurs with the same relative frequency regardless of the forecasts, and so the forecasts have no resolution. They are useless. More typically the forecasts have some resolution, but are not perfectly reliable, and will frequently show over-confidence (Figure 3*c*); the event occurs more frequently than indicated when the forecast indicates a decreased probability of the event occurring compared to climatology (to the left of the dotted line), but less frequently than indicated when the forecast indicates an increased probability of the event occurring compared to climatology (to the right of the dotted line). The greater the degree of over-confidence, the shallower is the slope of the curve. If the forecasts are under-confident, the curve is steeper than the diagonal (Figure 3*d*). In Figure 3*e* the forecast probabilities are consistently lower than the observed relative frequencies, indicating that the event always occurs more frequently than anticipated, and so the event is under-forecast. In Figure 3*f* the opposite is true, and the event occurs less frequently than anticipated, and is over-forecast. Note that under- and over-forecasting will not occur on the diagram for all categories because the probabilities have to add to 1.0, and so under- or over-forecasting in one category has to be compensated for in the others.

The horizontal dashed line not only represents a line of no resolution, but also indicates how frequently an event occurred over the verification period. The line is also drawn vertically at the

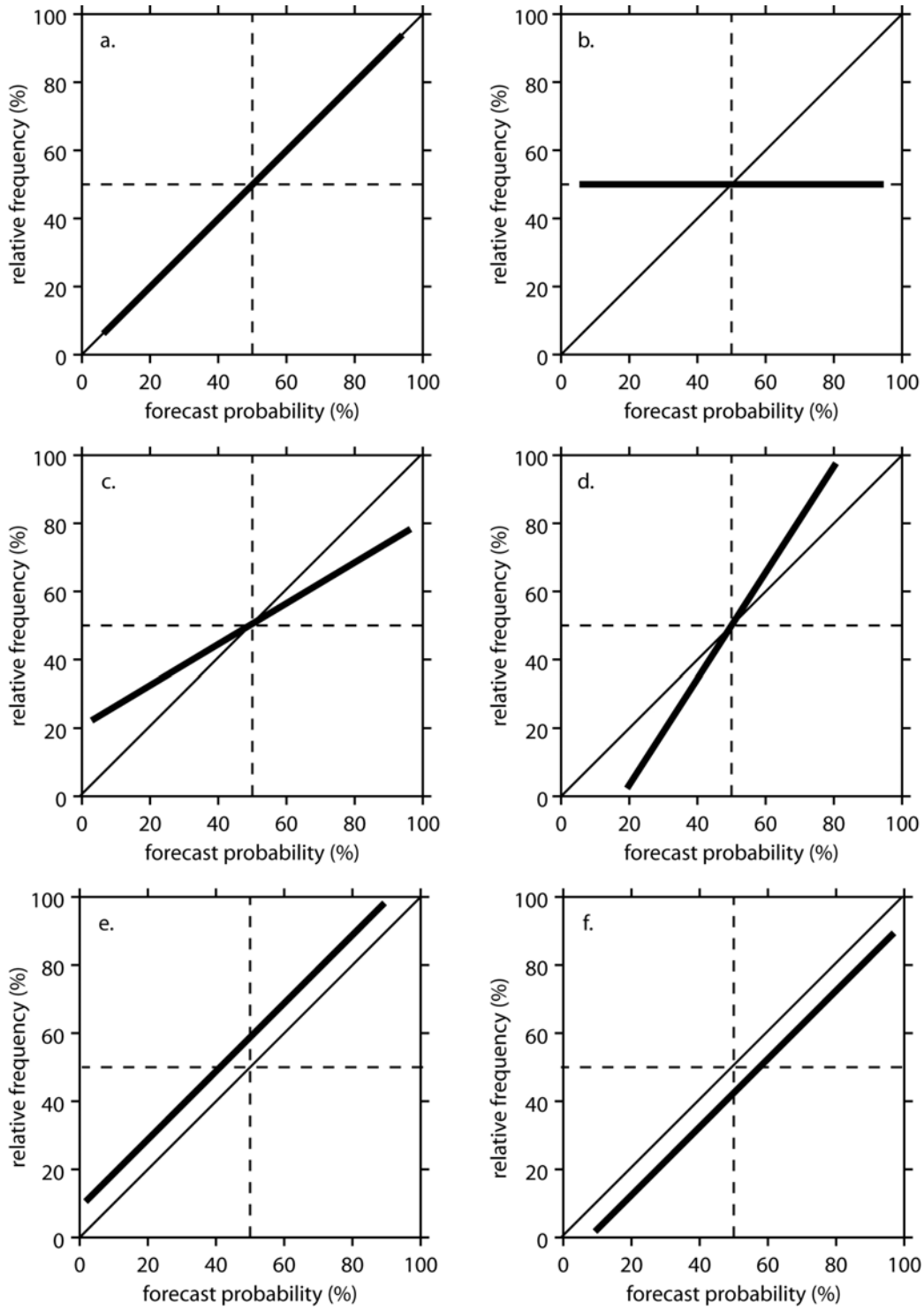

Figure 3. Idealized examples of reliability curves showing forecast systems with (*a*) perfect reliability, (*b*) no resolution, (*c*) over-confidence, (*d*) under-confidence, (*e*) under-forecasting, and (*f*) over-forecasting. The horizontal dashed line indicates the observed relative frequency of the event for all forecasts, which is shown also as a dashed vertical line.

same value rather than at the average forecast probability so that the bottom-left and top-right quadrants are divided by the diagonal, and areas of over- and under-confidence can then be seen more easily. The areas of over-confidence are represented by the two-triangles in which the reliability curve lies in Figure 3*c*, whereas the areas of under-confidence are represented by the two-triangles in which the reliability curve lies in Figure 3*d*. Because the frequency of an event is not guaranteed to be the same for the different categories, the reliability curves should be plotted on separate diagrams. Separate diagrams also help to avoid making the diagrams too cluttered.

Another reason for showing the vertical dashed line at the point marking the frequency of an event rather than at the average forecast probability is that unconditional biases in the forecast can be visualized more easily. Although the reliability curve itself can indicate unconditional biases (under- and over-forecasting; Figures 3*e* and *f*), the extent of the bias can be difficult to discern because the points on the reliability curve are not usually represented by equal numbers of forecasts. It is common practice to include a histogram showing the frequency of forecasts for each point on the curve. Examples are shown in Figure 4, which are for the first ten years of the PRESAO (Prévision Saisonnière en Afrique de l'Ouest) seasonal rainfall forecasts for the July to September season (Chidzambwa and Mason 2008). Over-forecasting in the normal-category is clearly evident from the displacement of the reliability curve, but is even more clearly evident from the histogram, which shows that all the forecasts had probabilities for this category that were higher than the observed relative frequency of normal rainfall. If there were no unconditional bias

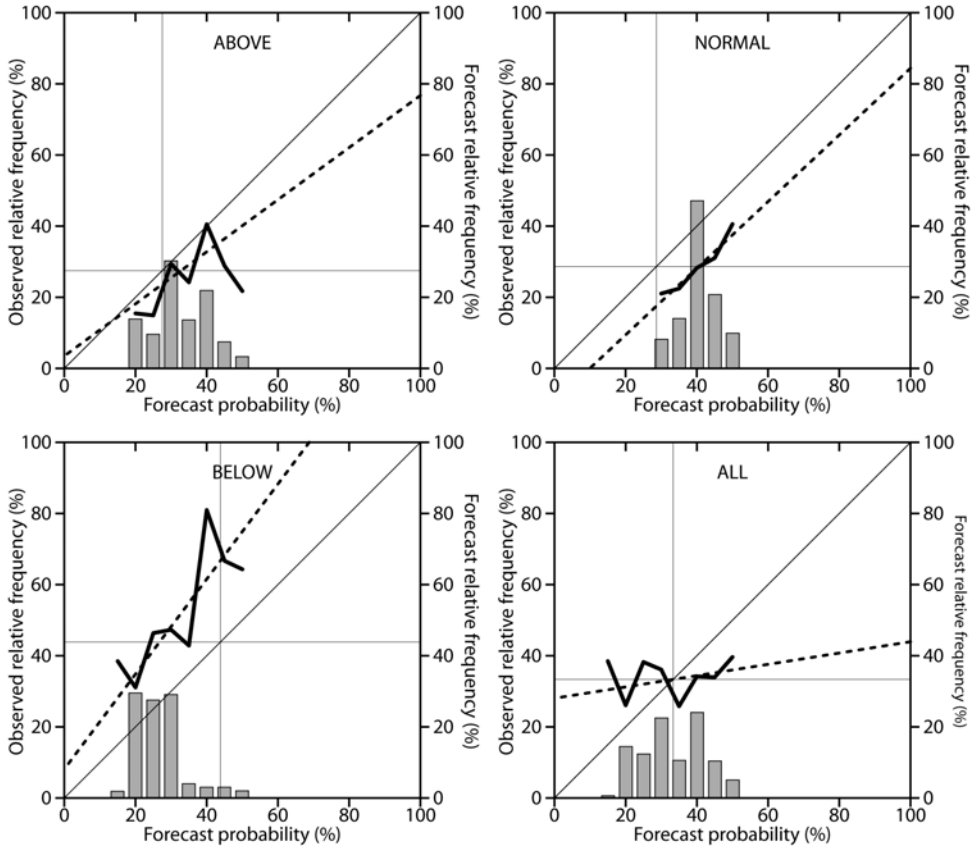

Figure 4. Example reliability diagrams for the first ten years of the PRESAO seasonal rainfall forecasts for the July to September season. The thick black line shows the reliability curve, and the thick dashed line is the least squares weighted regression fit to the reliability curve. The weights are shown by the grey bars, which indicate the relative frequency of forecasts in each 5% bin. The thin horizontal and vertical lines indicate the relative frequency of occurrence of rainfall in the respective category, while the thin diagonal represents the line of perfect reliability.

the vertical line would divide the histogram into two equal halves. Similarly, the under-forecasting of the below-normal category is evident. It is recommended that unconditional biases in the forecasts be calculated using

$$b_k = \frac{1}{n}\sum_{i=1}^{n}\left(p_{k,i} - y_{k,i}\right) = \overline{p}_k - \overline{y}_k \,. \tag{19}$$

where n is the number of forecasts, $p_{k,i}$ is the i^{th} forecast probability for the k^{th} category, and $y_{k,i}$ is 1 if the i^{th} observation was in category k, and is 0 otherwise.

In addition to showing any unconditional biases in the forecasts the histograms show the sharpness. Forecasts with weak sharpness have histograms that have high frequencies on a narrow range of probabilities close to the climatological probability, as for the normal forecasts in Figure 4, for example. Sharp forecasts have histograms showing high frequencies of forecasts near 0 and 100% – the histograms are *u*-shaped. For seasonal forecasts, *u*-shaped histograms are exceptionally rare because of an inability to be so confident, but relatively sharp forecasts have more dispersed histograms than those shown for the normal category in Figure 4. The PRESAO forecasts do not show particularly marked sharpness for any of the categories. No specific scores are recommended for measuring sharpness.

The reliability curve itself can be deceptively difficult to interpret because it does not represent the frequency of forecasts on each probability value. Sampling errors can therefore vary quite markedly along the curve. It is recommended that least squares regression fits to the curves be calculated, weighted by the frequency of forecasts on each probability, and added to the diagrams (Wilks and Murphy 1998). The parameters of the regression fit can be estimated using

$$\beta_1 = \frac{\sum_{k=1}^{d} n_k \left(p_k - \overline{p}\right)\left(\overline{y}_k - \overline{y}\right)}{\sum_{k=1}^{d} n_k \left(p_k - \overline{p}\right)^2} \,. \tag{20a}$$

and

$$\beta_0 = \overline{y} - \beta_1 \overline{x} \,. \tag{20b}$$

where β_1 is the slope and β_0 the intercept of the fitted regression line, d is the number of discrete probability values, n_k is the number of forecasts for the k^{th} probability value, $\overline{p}_k$ is the k^{th} probability value, $\overline{p}$ is the average probability for all forecasts, $\overline{y}_k$ is the observed relative frequency for the k^{th} probability value [Eq. (7)], and $\overline{y}$ is the observed relative frequency over the verification period. It is recommended that the slope of the regression line, which can be viewed as a measure of resolution, be communicated as a percentage change in the observed relative frequency given a 10% increase in the forecast probability. If the forecasts have good resolution an event should increase in frequency by 10% as the forecast probability is incremented by each 10% (e.g., from 30% to 40%, or from 40% to 50%), and the slope will be 1.0, but if they have no resolution the slope will be zero. Over-confidence will be indicated by a slope of between 0.0 and 1.0 (the increase in frequency will be between 0% and 10%), while under-confidence will be indicated by slopes of greater than 1.0 (increases in frequency of more than 10%). An example is provided in appendix B section a.*viii*.

For many non-specialists, reliability diagrams are likely to be intimidating, and so alternative ways of presenting the information should be considered. The slope of the regression line and the unconditional biases [Eq. (20)] have been suggested as two possibilities. The actual reliability curve itself could be communicated in tabular form rather than as a graph. The biases can be usefully graphed in a "tendency diagram" by showing the average forecast probability and the observed relative frequency of each category over the verification period. An example tendency

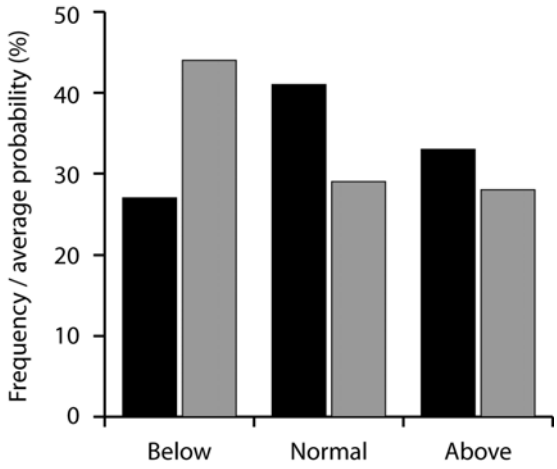

Figure 5. Example tendency diagram for the first ten years of the PRESAO seasonal rainfall forecasts for the July to September season. The black bars show the average forecast probabilities for each category, and the grey bars show the observed relative frequencies for each category.

diagram is shown in Figure 5, which indicates a shift towards drier conditions compared to the training period, which was not forecast.

One major limitation of reliability diagrams is that they require a large number of forecasts because of the need to calculate the observed relative frequencies for each forecast value. The diagrams can therefore only be constructed by pooling forecasts for different years and locations. Typically, one set of reliable diagrams will be drawn for all the forecasts available.

c. Measuring the quality of individual forecast maps

i. Scoring of attributes

It may seem attractive to propose the same verification scores as were used for verifying series of forecasts for use with individual maps, especially for scores to be communicated to the general public. However, it has been argued that the evaluation of forecasts over space is fundamentally different from evaluating them over time, and that the attributes of interest are different. The use of different procedures may help to emphasize this distinction, but more importantly provides the freedom to address the questions of direct interest rather than simply trying to calculate what would amount to an arbitrary score. Procedures for verifying the forecasts on an individual map are therefore discussed in this section, with the objective of measuring forecast accuracy (as defined in section 3d).

Measuring the quality of a map of forecasts for a specific season by counting a hit if the observed category was the one with the highest probability has become a popular practice amongst some of the RCOFs, and at some National Meteorological Services[17]. Although this

[17] Scoring a "half-hit" if the observed category had an increased probability above the climatological value, regardless of whether or not the verifying category had the highest probability is not recommended because the meaning of the score becomes complicated. Given that there is a general practice not to indicate increased probabilities of the normal category without showing one of the outer two categories as more likely than the other, this procedure seems to be motivated by the desire to indicate whether the forecasts indicate the correct "tendency" – i.e., to simplify the forecast to a two-category system. Although procedures can be proposed to verify the forecasts from such a

practice seems to have intuitive appeal, it is problematic, as discussed in section 3, and so it is recommended that the score not be used as the procedure of first choice. However, as discussed in section 4b.*iii*, the procedure is not completely without merit, and so it is included here as a suggestion with the recommendation that the hit scores for the categories that do not have the highest probabilities also be calculated. The calculation of the scores is essentially the same as for a series of forecasts, and so the reader is referred to section 4b.*iii* for details.

If the hit score is not recommended as a starting point for verifying individual forecast maps, an alternative has to be offered. However, replacing the hit score with a probabilistic verification score does not represent a simple solution, since most verification procedures are designed to evaluate forecasts issued over time rather than over a spatial field, and many of the probabilistic scores are difficult to communicate to non-specialists anyway. As argued in section 3d, interest in the accuracy of an individual forecast is concerned with having high probabilities on the verifying categories, and is not specifically concerned about the reliability of these probabilities. The linear probability score (Wilson *et al.* 1999) may be an attractive option for communicating forecast quality of individual maps to non-specialists. The linear probability score is defined as:

$$\text{linear probability score} = 100\% \times \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{j,i} p_{j,i} \,, \tag{21}$$

where n is the number of points on the map at which the forecasts are to be verified, m is the number of categories, $y_{i,j}$ is 1 if the observation at location i was in category j, and is 0 otherwise, and $p_{j,i}$ is the forecast probability for category j at location i. The score has a very simple interpretation: it is the average forecast probability assigned to the verifying categories, and it ranges between 100% for perfect forecasts (a probability of 100% was assigned to the observed category at each of the locations) to 0% for perfectly bad forecasts (a probability of 0% was assigned to the observed category at each of the locations). A "good" forecast will score more than a strategy of using the climatological forecasts, and will beat the expected score from guessing. An example is provided in appendix B section b.*i*.

What the value of the score for a useless of forecasts would be depends on the number and definition of the categories, and must be considered a disadvantage of the score. The naïve expectation will be that the score should exceed 50% for the forecasts to be good, which will only be the case for two equiprobable categories, and so it may be helpful to compare the score with the reference score to indicate whether the forecasts are skilful. However, for communication to the general public it is not recommended that the score be converted into a skill score of the commonly-used form

$$\text{skill score} = \left(1 - \frac{\text{score of forecasts}}{\text{score of reference forecasts}} \right) \times 100\% \,. \tag{22}$$

The skill score defined in Eq. (22) is measured in percent, but has a different meaning to the forecast probability, also measured in percent, and thus can get confusing to some non-specialists. Instead, it is generally simpler to calculate the skill score as:

$$\text{skill score} = \left(\text{score of forecasts} - \text{score of reference forecasts} \right) \times 100\% \,, \tag{23}$$

which simply indicates how much higher (or lower) the average probability for the verifying categories was compared to probabilities derived from the climatology. For example, assume that the average probability for the verifying categories is 45%, and that there are three equiprobable categories. These forecasts would have a skill score of approximately 12%, using Eq. (23), and the results could be communicated as: "The observed categories were forecast with an average probability of 45%, which is 12% more than for guessing." A downside of Eq. (23) is that it does not have an upper bound of 100%, and so perfect forecasts will not achieve an intuitive score

---

perspective, it is preferable to evaluate the information provided by the forecast as it stands, rather than to try and re-interpret the forecast and then verify what is inferred. If a two-category forecast is desired then make one!

value, but in practice most skill levels of seasonal forecasts are low anyway, and so interpretation problems when the skill does reach the upper limit are not likely to be experienced very often.

The average interest rate is recommended as an alternative score to the linear probability score especially if understanding the base rate is expected to be problematic (Hagedorn and Smith 2008). The average interest rate is recommended in place of the more widely used Brier score or ranked probability score primarily because of the association of these scores with attributes such as reliability and resolution, which are inappropriate in the current context. The average interest rate was described in section 4b.*v*, and an example is provided in appendix B section b.*ii*. Because the forecasts are for different locations, Eqs (14) – (16) cannot be used. Instead, the average interest rate can be calculated using

$$\text{average interest rate} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{p_i}{c_i} \right) - 1, \tag{24}$$

where p_i is the forecast probability for the verifying category at the ith of n locations, and c_i is the corresponding climatological probability.

One major problem with the linear probability score and the average interest rate is that they are not proper (they encourage the forecaster to issue 100% probability on the most likely category; Bröcker and Smith 2007b), and so they should not be used for anything but the simplest form of communication to non-specialists. Instead, if a proper score is required, the ignorance score is recommended, although it cannot be converted into the more intuitive interest rate for the reasons discussed above. The ignorance score is defined by Eq. (14), and is described in section 4b.*v*. An example is provided in appendix B section b.*iii*.

In addition to calculating a score for the map, it is recommended that the forecast be accompanied by a corresponding map of the observed rainfall or temperature, but with the observations presented in such a way as to make them correspond to the information communicated in the forecast. Maps of anomalies or, in the case of rainfall, of percentage departures from average should not be used, because it is not clear from either of these which category the observation is in. These maps can be quite misleading to non-specialists who may not have much knowledge of the climatology of the region. Instead the map of observations should be contoured by the quantile to which the rainfall or temperature corresponds. The most logical way to calculate the quantiles for each location would be from the cumulative distribution function of a distribution fitted to the climatological data, but if the terciles for defining the categories when making the forecasts are not calculated from this distribution, it would be advisable to use a method consistent with the way in which the forecast is defined. Linear interpolation of the empirical cumulative distribution function should therefore most likely be used. It is recommended that contours be shown for the 33rd and 67th percentiles, the 20th and 80th, the 10th and 90th, and for record-breaking values. An example of such a map for rainfall is shown in Figure 6 (Tall et al. 2012).

Percentiles are, of course, rather complicated concepts for non-specialists, and so the way in which the map is presented is important to avoid making it unnecessarily incomprehensible. It is suggested that the map be shaded in appropriate colours for areas that are above-normal and below-normal, and that these be labelled as such in a colour-bar, and that more intense shading be used for the more extreme outcomes. It may be helpful to try communicating the more extreme outcomes in terms of their corresponding return periods, but different user communities may have their own preferred ways of understanding these concepts.

ii. Model diagnostics

It is strongly recommended that any forecast based on dynamical model output be verified by considering its full set of fields, to address the question of whether the model got the specific field

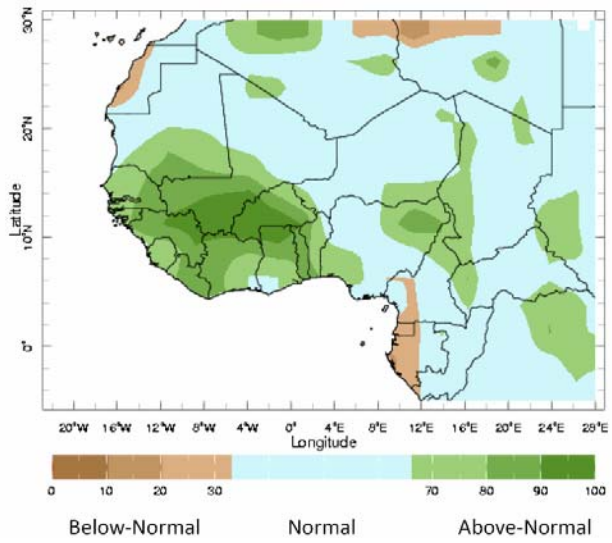

Figure 6. Example rainfall anomaly map for July – August 2008 with units in percentiles using a 1971 – 2000 climatology.

of interest "right" for the right reasons, or if not why not. For example, in considering the quality of a specific GCM's rainfall predictions during the 1997/98 El Niño, analysis of the model's atmospheric structure, may provide useful insights into its strengths and weaknesses.

5. Uncertainty of the Results

Given the small sample sizes typical of seasonal forecasts many of the procedures recommended are likely to have large sampling errors. There therefore remains some uncertainty as to the true quality of the forecasts even after conducting detailed verification diagnostics. To assess whether the results do indeed indicate that the forecasts are good it is common practice to calculate the statistical significance of verification scores, as measured by the so-called *p*-value. The objective in calculating the significance of the scores is to estimate the possibility that a result at least as good as that obtained could have been achieved by chance; if this probability is very low, the argument is that we can be very confident about the forecasts being good. However, calculating statistical significance is not recommended for a number of reasons. Firstly, the *p*-value tells us as much about the sample size as it does about the goodness of the forecasts. So for example, given a very large set of forecasts we could achieve an exceptionally low *p*-value even with very low verification scores. All the *p*-value would then tell us is that we can be highly confident that the forecasts are marginally good. A related reason is that the *p*-value does not tell us what we want to know – we want to know how good our forecasts are not how likely a set of forecasts from an inherently useless system could have outscored ours simply by luck. Thirdly, *p*-values can be difficult to understand, and so are not very suitable for communication to non-specialists.

A preferred approach is to calculate confidence intervals for the scores. Confidence intervals can provide an indication of the uncertainty in verification scores by defining a range of values that have a pre-specified probability of containing the true score. [For a detailed interpretation of confidence intervals, see Jolliffe (2007).] It is recommended that a 90% confidence interval be used. For most of the scores the confidence intervals will have to be obtained by bootstrapping techniques. The most commonly used bootstrapping procedures involve resampling with replacement, and are most easily implemented for forecasts at individual locations. In this instance if there are *n* forecast-observation pairs, then select *n* forecast-observation pairs with

replacement (i.e. so that each forecast observation pair may be selected more than once), but keep each observation matched with its corresponding forecast. The bootstrap sample should be similar to the original sample except that some forecast observation pairs will be missing, being replaced by others that are repeated, possibly more than once. The verification scores can then be recalculated using this bootstrap sample. This process should be repeated, a few hundred times at least. It is recommended that at least 1000 bootstrap samples be drawn if computing power permits. If n_b is the number of bootstrap samples, then there should be n_b bootstrapped values of the verification score. These bootstrapped scores should then be ranked in ascending order, and the 5^{th} and 95^{th} percentiles identified. [The 5^{th} percentile will be the $(n_b \times 5 \div 100)^{th}$ ranked value, and the 95^{th} percentile will be the $(n_b \times 95 \div 100)^{th}$ ranked value.] Each score can then be communicated with the confidence interval included in parentheses afterwards. [For example: "the probability of correctly discriminating the above-normal category (the ROC area) is 0.80 (0.73 – 0.85)."] For the graphs, the confidence intervals can be indicated as error bars.

APPENDIX A:

Weighted versions of the verification scores

In this appendix, weighted versions of the equations for the various verification scores are presented. Each location (i.e., gridbox, station, or region) is assigned a weight, w_i, based on its representative area, as discussed in section 2b–d. In each case the verification scores are adjusted to represent the appropriate area weighting of each location.

In most of the following equations no explicit term is included to indicate the current location. Instead it is assumed that the forecasts for all the locations and time steps are pooled to avoid unnecessary inclusion of additional summation signs and subscripts. The pooling requires a weight to be set for each forecast rather than just each location. The total number of forecasts is typically (but not necessarily) the number of locations × the number of time steps. So, for example, if there are 10 forecasts at each of 5 location, there will be a total of 50 forecasts (10 × 5), and the first 10 (or the first and every 5th thereafter) weights will be for the first location, the second 10 (or the second and every 5th thereafter) weights will be for the second location, etc.

a. Series of forecasts

i. Generalized discrimination score

The weights are applied only to Eq. (1a), which becomes:

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

where m is the number of categories, n_k is the number of times the observation was in category k, $\mathbf{p}_{k,i}$ is the vector of forecast probabilities for the i^{th} observation in category k.

ii. Relative operating characteristics (ROC)

If the ROC graph is to be constructed, the hit and false-alarm rates [Eqs (3) and (4)] can be calculated by counting each hit (or false alarm) weighted by the current location, and dividing by the total number of weighted events (non-events). The area beneath the curve can then be calculated using Eq. (5) without having to worry about weights, since the weights are already built into the hit and false-alarm rates.

Alternatively, if the ROC area is to be calculated without constructing the graph, Eq. (2a) must be modified to

$$A = \frac{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} w_i w_j I\left(p_{0,i}, p_{1,j}\right)}{\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} w_i w_j}, \tag{A.2}$$

where n_0 is the number of non-events, n_1 the number of events, $p_{1,j}$ the forecast probability for the j^{th} observed event, and $p_{0,i}$ the probability of an event for the i^{th} non-event. Equation (2b) is unaffected.

iii. Resolution score

The weighted version of the resolution score [Eq. (6)] is

$$\text{resolution} = \frac{\sum_{k=1}^{d}\left(\sum_{i=1}^{n_k} w_{k,i}\right)\left(\overline{y}_k - \overline{y}\right)^2}{\sum_{j=1}^{n} w_j}, \tag{A.3}$$

where d is the number of discrete forecast values or forecast probability bins, n_k is the number of forecasts for the k^{th} probability bin, $w_{k,i}$ is the weight for the i^{th} forecast in this bin, $\overline{y}_k$ is the observed relative frequency, and $\overline{y}$ is the observed relative frequency for all forecasts. The denominator defines the sum of the weights for all the forecasts.

iv. Hit (Heidke) score

The hit score [Eq. (8c)] can be weighted by scoring the weight of the current location, rather than scoring 1.0, for each hit, and then dividing by the accumulated weights for all forecasts:

$$\overline{y}_j = \frac{\sum_{i=1}^{n} w_i y_{j,i}}{\sum_{i=1}^{n} w_i}. \tag{A.4}$$

where n is the number of forecasts, and $y_{j,i}$ is 1 if the i^{th} observation was in the category with the j^{th} highest probability, and is 0 otherwise.

v. Reliability score

The weighted version of the reliability score [Eq. (10)] is

$$\text{reliability} = \frac{\sum_{k=1}^{d}\sum_{i=1}^{n_k} w_{k,i}\left(p_{k,i} - \overline{y}_k\right)^2}{\sum_{j=1}^{n} w_j}, \tag{A.5}$$

where d is the number of discrete forecast values or forecast probabilit bins, where, n_k is the number of forecasts for the k^{th} probability bin, $w_{k,i}$ is the weight for the i^{th} forecast in this bin, $p_{k,i}$ is the *c*orresponding forecast probability, and $\overline{y}_k$ is the observed relative frequency. The denominator defines the sum of the weights for all the forecasts (locations and time steps).

vi. Brier score

The weighted version of the Brier score [Eq. (12)] is

$$BS_j = \frac{\sum_{i=1}^{n} w_i\left(y_{j,i} - p_{j,i}\right)^2}{\sum_{i=1}^{n} w_i}, \tag{A.6}$$

where n is the number of forecasts $y_{j,i}$ is 1 if the i^{th} observation was in category j, and is 0 otherwise, and $p_{j,i}$ is the i^{th} forecast probability for category j.

vii. Ranked probability score

The weighted ranked probability score [Eq. (13)] is defined as

$$RPS = \frac{\sum_{i=1}^{n} w_i \sum_{k=1}^{m-1} \left(\sum_{j=1}^{k} \left(y_{j,i} - p_{j,i} \right) \right)^2}{(m-1) \sum_{i=1}^{n} w_i}, \tag{A.7}$$

where n is the number of forecasts, m is the number of categories, $y_{j,i}$ is 1 if the i^{th} observation was in category j, and is 0 otherwise, and $p_{j,i}$ is the i^{th} forecast probability for category j.

viii. Effective interest rate

When calculating the weighted effective interest rate it is implicit that the score is being averaged over different locations, and so the weighting is applied to Eq. (16) rather than to Eq. (15). To incorporate the weights, Eq. (16) is modified to

$$\text{average effective interest rate} = \frac{\sum_{k=1}^{s} w_i \left(2^{Ign\ (ref) - Ign_k} - 1 \right)}{\sum_{k=1}^{s} w_i}, \tag{A.8}$$

where $Ign(ref)$ is the ignorance score for the reference (climatological forecasts), and Ign_k is the ignorance score for the k^{th} location.

ix. Accumulated profits

The values on the y-axis for the accumulated profits graph [Eq. (18)] can be obtained using

$$\left(\prod_i \left(\frac{\sum_{k=1}^{s} \frac{w_k p_{k,i}}{c_{k,i}}}{\sum_{k=1}^{s} w_k} \right) \right) - 1 \tag{A.9}$$

where, s, is the number of stations / seasons, where $p_{k,i}$ is the forecast probability for the verifying category at location / in season k, and c_i is the corresponding climatological probability.

x. Reliability diagram

The weights must be applied when calculating the observed relative frequencies [Eq. (7)]:

$$\overline{y}_k = \frac{\sum_{i=1}^{n_k} w_i y_{k,i}}{\sum_{i=1}^{n_k} w_i}. \tag{A.10}$$

where n_k is the number of forecasts of the k^{th} forecast probability, and $y_{j,i}$ is 1 if the i^{th} observation was an event, and is 0 otherwise. The weights should also be applied when calculating the average forecast probability in each bin, although in most cases, because the issued forecast probabilities are discrete, the forecast probabilities are the same for every forecast in each bin. However, if weighted averages are required they can be calculated using Eq. (A.10), and replacing $y_{k,i}$ with $p_{k,i}$,

where $p_{k,i}$ is the ith forecast probability in bin k. Similarly, the bin frequencies for the frequency should be replaced by the denominator in Eq. (A.10).

The weighted unconditional biases can be calculated using

$$b_k = \frac{\sum_{i=1}^{n} w_i \left(p_{k,i} - y_{k,i} \right)}{\sum_{i=1}^{n} w_i}, \tag{A.11}$$

where n is the number of forecasts, $p_{k,i}$ is the i^{th} forecast probability for the k^{th} category, and $y_{k,i}$ is 1 if the i^{th} observation was in category k, and is 0 otherwise.

The regression fit can be calculated using Eq. (20) since the weights are already considered in the construction of the regression curve. The only modification required is to replace n_k in Eq. (20a) with the sum of the weights for all the forecasts in the current bin [i.e., with the denominator in Eq. (A.10)].

b. Individual forecast maps

i. Linear probability score

The weights are included in Eq. (21) thus:

$$\text{linear probability score} = 100\% \times \frac{\sum_{i=1}^{n} w_i \sum_{j=1}^{m} y_{j,i} p_{j,i}}{\sum_{i=1}^{n} w_i}, \tag{A.12}$$

where n is the number of points on the map at which the forecasts are to be verified, m is the number of categories, $y_{i,j}$ is 1 if the observation at location i was in category j, and is 0 otherwise, and $p_{j,i}$ is the forecast probability for category j at location i.

ii. Average interest rate

The average interest rate [Eq. (24)] is modified in a similar way to Eq. (A.8):

$$\text{average interest rate} = \left(\frac{\sum_{i=1}^{n} \frac{w_i p_i}{c_i}}{\sum_{i=1}^{n} w_i} \right) - 1, \tag{A.13}$$

where p_i is the forecast probability for the verifying category at the ith of n locations, and c_i is the corresponding climatological probability.

iii. Ignorance score

The weighted version of the ignorance score can be calculated using

$$\text{ignorance score} = -\frac{\sum_{i=1}^{n} w_i \sum_{j=1}^{m} y_{j,i} \log_2 \left[p_{j,i} \right]}{\sum_{i=1}^{n} w_i}, \tag{A.14}$$

where n is the total number of locations, m is the number of categories, $y_{j,i}$ is 1 if the i^{th} observation is in category j, and is 0 otherwise, and $p_{j,i}$ is the corresponding forecast probability.

APPENDIX B:

Calculation of the recommended scores and graphs

In this appendix, a simple set of forecasts is used to illustrate the steps in calculating each of the recommended scores, and to construct the graphs. The appendix forms a quick-reference guide to the recommendations by including the relevant equations, simple descriptions of the scaling of the scores and of the forms of the graphs, and indications of conditions under which the scores and procedures can be calculated. The example data for all the procedures for verifying series of forecasts are shown in Table B.1, except for the reliability diagram, which ideally requires a larger sample, too large to illustrate the other procedures simply. For ease of reading, the data are ordered by the observation, so that all the below-normal years are listed first, and the above-normal years last, and the years are listed from 2001 to 2008. Data for verifying forecasts for a single season are presented in Table B.2.

Table B.1. Example forecasts and observations for three equi-probable categories [below-normal (B), normal (N), and above-normal (A)].

Year	Observation	Below	Normal	Above
2001	B	0.45	0.35	0.20
2002	B	0.50	0.30	0.20
2003	B	0.35	0.40	0.25
2004	B	0.33	0.33	0.33
2005	N	0.25	0.35	0.40
2006	N	0.20	0.35	0.45
2007	A	0.20	0.35	0.45
2008	A	0.25	0.40	0.35

Table B.2. Example forecasts and observations for three equi-probable categories [below-normal (B), normal (N), and above-normal (A)].

Location	Observation	Below	Normal	Above
I	B	0.45	0.35	0.20
II	B	0.50	0.30	0.20
III	B	0.35	0.40	0.25
IV	B	0.33	0.33	0.33
V	N	0.25	0.35	0.40
VI	N	0.20	0.35	0.45
VII	A	0.20	0.35	0.45
VIII	A	0.25	0.40	0.35

a. Series of forecasts

i. Generalized discrimination score

As its name suggests, the generalized discrimination score (Mason and Weigel 2009) measures discrimination (do forecasts differ given different outcomes?). The score is a generalization of the trapezoidal area beneath the ROC curve for forecasts with more than two categories. The score, *D*, is defined as

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

where m is the number of categories, n_k is the number of times the observation was in category k, $\mathbf{p}_{k,i}$ is the vector of forecast probabilities for the i^{th} observation in category k, and

$$I\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) = \begin{cases} 0.0 & \text{if } F\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) < 0.5 \\ 0.5 & \text{if } F\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) = 0.5 \\ 1.0 & \text{if } F\left(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}\right) > 0.5 \end{cases}, \tag{B.1b}$$

and where

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

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

Setting category 1 as the below-normal category, using Eq. (B.1a) $m = 3$, $n_1 = 4$, $n_2 = 2$, and $n_3 = 2$, which means that the denominator is $n_1 \times (n_2 + n_3) + n_2 \times n_3 = 20$. The summation over k therefore considers the below-normal years firs

t, and then the normal years. The summation over l indicates that the below-normal years are compared with the normal and above-normal years, and then the normal years are compared with the above-normal years. The summation over i takes each of the three below-normal years (when $k = 1$), and compares them with each of the four normal years (when $l = 1$) using the summation over j. So taking $k = i = j = 1$, and $l = 2$, 1991 is compared with 1995 (the first normal year). Using Eq. (B.1c) the probability that a value from the forecast for 1995 is greater than one from the forecast for 1991 is calculated. A value from 1995, will be greater than from 1991 if a below-normal value is taken from 1991 ($r = 1$), and a normal or above-normal value is taken from 1995 ($s = 2$ or 3), or if a normal value is taken from 1991 (rv2), and an above-normal value from 1995 ($s = 3$). The probability is conditioned upon the two values being different, and so the probability that the two values are the same is calculated in the denominator and subtracted from 1. The calculation of Eq. (B.1c) is shown in Table B.3, and the result of Eq. (B.1b) is shown in the final column. The sum of these scores is 17.5, as shown at the foot of the table, and so the score is $17.5 \div 20 = 87.5\%$.

Eq. (1) defines the probability of successfully discriminating the wetter (or warmer) of two observations, and has an intuitive scaling that is appealing to many non-specialists: it has an expected value of 50% for useless forecast strategies (guessing, or always forecasting the same probabilities), and good forecasts will have a score greater than 50%, reaching 100% given perfect discrimination. Scores of less than 50% indicate bad forecasts (forecasts that can discriminate, but which indicate the wrong tendency – for example, high forecast probabilities on below-normal indicate a low probability that below-normal rainfall will actually occur), and can reach a lower limit of 0% given perfectly bad forecasts. The score can be calculated for each location, and then a map of the scores can be drawn to indicate areas in which the forecasts have some discrimination, or it can be calculated by pooling sets of locations as long as each location has a reasonable number of forecasts.

Table B.3. Example calculation of the generalized discrimination score.

k	l	i	j	$\text{Year}_{k,i}$	$\text{Year}_{l,j}$	Eq. (1c)	$I\left(\mathbf{p}_{k,i},\mathbf{p}_{l,j}\right)$
1	2	1	1	2001	2005	$\frac{0.45\times(0.35+0.40)+0.35\times 0.40}{1-(0.45\times 0.25+0.35\times 0.35+0.20\times 0.40)}\approx 0.70$	1.0
1	2	1	2	2001	2006	$\frac{0.45\times(0.35+0.45)+0.35\times 0.45}{1-(0.45\times 0.20+0.35\times 0.35+0.20\times 0.45)}\approx 0.74$	1.0
1	2	2	1	2002	2005	$\frac{0.50\times(0.35+0.40)+0.30\times 0.40}{1-(0.50\times 0.25+0.30\times 0.35+0.20\times 0.40)}\approx 0.72$	1.0
1	2	2	2	2002	2006	$\frac{0.50\times(0.35+0.45)+0.30\times 0.45}{1-(0.50\times 0.20+0.30\times 0.35+0.20\times 0.45)}\approx 0.76$	1.0
1	2	3	1	2003	2005	$\frac{0.35\times(0.35+0.40)+0.40\times 0.40}{1-(0.35\times 0.25+0.40\times 0.35+0.25\times 0.40)}\approx 0.63$	1.0
1	2	3	2	2003	2006	$\frac{0.35\times(0.35+0.45)+0.40\times 0.45}{1-(0.35\times 0.20+0.40\times 0.35+0.25\times 0.45)}\approx 0.68$	1.0
1	2	4	1	2003	2005	$\frac{0.33\times(0.35+0.40)+0.33\times 0.40}{1-(0.33\times 0.25+0.33\times 0.35+0.33\times 0.40)}\approx 0.58$	1.0
1	2	4	2	2003	2006	$\frac{0.33\times(0.35+0.45)+0.33\times 0.45}{1-(0.33\times 0.20+0.33\times 0.35+0.33\times 0.45)}\approx 0.63$	1.0
1	3	1	1	2001	2007	$\frac{0.45\times(0.35+0.45)+0.35\times 0.45}{1-(0.45\times 0.20+0.35\times 0.33+0.20\times 0.45)}\approx 0.74$	1.0
1	3	1	2	2001	2008	$\frac{0.45\times(0.40+0.35)+0.35\times 0.35}{1-(0.45\times 0.25+0.35\times 0.40+0.20\times 0.35)}\approx 0.68$	1.0
1	3	2	1	2002	2007	$\frac{0.50\times(0.35+0.45)+0.30\times 0.45}{1-(0.50\times 0.20+0.30\times 0.35+0.20\times 0.45)}\approx 0.76$	1.0
1	3	2	2	2002	2008	$\frac{0.50\times(0.40+0.35)+0.30\times 0.35}{1-(0.50\times 0.25+0.30\times 0.40+0.20\times 0.35)}\approx 0.70$	1.0
1	3	3	1	2003	2007	$\frac{0.35\times(0.35+0.45)+0.40\times 0.45}{1-(0.35\times 0.20+0.40\times 0.35+0.25\times 0.45)}\approx 0.68$	1.0
1	3	3	2	2003	2008	$\frac{0.35\times(0.40+0.35)+0.40\times 0.35}{1-(0.35\times 0.25+0.40\times 0.40+0.25\times 0.35)}\approx 0.61$	1.0
1	3	4	1	2004	2007	$\frac{0.33\times(0.35+0.45)+0.33\times 0.45}{1-(0.33\times 0.20+0.33\times 0.35+0.33\times 0.45)}\approx 0.63$	1.0
1	3	4	2	2004	2008	$\frac{0.33\times(0.40+0.35)+0.33\times 0.35}{1-(0.33\times 0.25+0.33\times 0.40+0.33\times 0.35)}\approx 0.55$	1.0
2	3	1	1	2005	2007	$\frac{0.25\times(0.35+0.45)+0.35\times 0.45}{1-(0.25\times 0.20+0.35\times 0.35+0.40\times 0.45)}\approx 0.55$	1.0
2	3	1	2	2005	2008	$\frac{0.25\times(0.40+0.35)+0.35\times 0.35}{1-(0.25\times 0.25+0.35\times 0.40+0.40\times 0.35)}\approx 0.47$	0.0
2	3	2	1	2006	2007	$\frac{0.25\times(0.35+0.40)+0.35\times 0.40}{1-(0.25\times 0.25+0.35\times 0.35+0.40\times 0.40)}=0.50$	0.5
2	3	2	2	2006	2008	$\frac{0.20\times(0.40+0.35)+0.35\times 0.35}{1-(0.20\times 0.25+0.35\times 0.40+0.45\times 0.35)}\approx 0.42$	0.0
						$\sum I\left(\mathbf{p}_{k,i},\mathbf{p}_{l,j}\right)$	17.5

ii. Relative operating characteristics (ROC)

To measure the discrimination for individual categories, the area beneath the relative operating characteristics curve is recommended. The area, *A*, can be calculated using

$$A = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I\left(p_{0,i}, p_{1,j}\right), \tag{B.2a}$$

where n_0 is the number of non-events, n_1 the number of events, $p_{1,j}$ is the forecast probability for the j^{th} observed event, $p_{0,i}$ is the probability of an event for the i^{th} non-event, and $I\left(p_{0,i}, p_{1,j}\right)$ is defined as

$$I\left(p_{0,i}, p_{1,j}\right) = \begin{cases} 0.0 & \text{if } p_{1,j} < p_{0,i} \\ 0.5 & \text{if } p_{1,j} = p_{0,i} \\ 1.0 & \text{if } p_{1,j} > p_{0,i} \end{cases}. \tag{B.2b}$$

Using the data in Table B.1, and setting the above-normal category as an event, from Eq. (B.2a) $n_1 = 2$, and $n_0 = 6$. The summation over *i* takes each of the years that are below-normal or normal, and compares them with each of the above-normal years. If the probability for above-normal is higher on the year that is above-normal the two years are successfully discriminated, and the forecaster scores 1.0 [Eq. (B.2b)]. A total of $n_0 \times n_1 = 12$ comparisons are made. These comparisons are shown in Table B.4. There are 9 out of 12 correct selections, and one tie, yielding a score of about 79%.

The interpretation of the ROC area is similar to that of the generalized discrimination score: if the category of interest is above-normal, the score indicates the probability of successfully discriminating above-normal from normal and below-normal observations. The scaling is identical to that for the generalized discrimination score, with a score of 50% representing no skill, 100% indicating perfect discrimination, and 0% indicating perfectly bad discrimination.

The ROC curve is constructed by calculating hit and false-alarm rates for decreasing probability thresholds. The hit and false alarm rates are calculated using (respectively)

$$y_k = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{k,i} \ . \tag{B.3}$$

$$x_k = \frac{1}{n_0} \sum_{i=1}^{n_0} y_{k,i} \ . \tag{B.4}$$

where n_1 is the number of number of events, n_0 is the number of number of non-events, y_k is the number of times an event occurred given a forecast probability exceeding the k^{th} probability threshold, and x_k is the number of times a non-event occurred given a forecast probability exceeding the k^{th} probability threshold. The calculation is indicated in Table B.5a. The table indicates a 1 if the forecast probability for above-normal is greater than or equal to the threshold, and indicates a 0 otherwise. The graph is constructed by plotting the hit rates on the *x*-axis against the false-alarm rates on the *y*-axis (Figure B.1). The area beneath the curve can then be calculated from the trapezoidal rule, as shown in Table B.5b, and is in agreement with the area obtained from Table B.5a. The trapezoidal rule is given by

$$A = 0.5 \times \left[1 + \sum_{k=0}^{d} \left(y_k x_{k+1} - y_{k+1} x_k\right)\right]. \tag{B.3}$$

where *d* is the number of discrete probability values, and y_1 and x_1 are the hit and false-alarm rates for the highest probability value only, y_2 and x_2 are the rates for the highest and second highest

probabilities, etc.. [Note slight change of notation from Eq. (5)]. For $i = 0$ the hit rate and false-alarm rates are defined as 0.0, and for $i = d + 1$ they are defined as 1.0 to ensure that the curve starts in the bottom-left, and ends in the top-right corners, respectively.

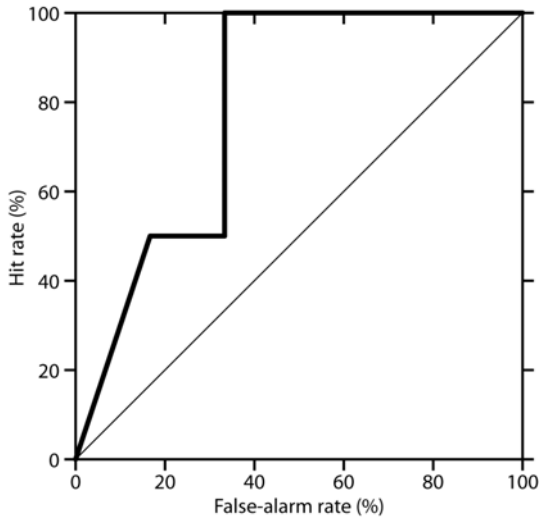

Figure B.1. Example ROC graph for the above-normal category using the hit and false-alarm rates from Table B.5.

Table B.4. Example calculation of the ROC area for the above-normal category using Eq. (2).

i	j	$\text{Year}_{0,i}$	$\text{Year}_{1,j}$	$p_{0,i}$	$p_{1,j}$	$I\left(p_{0,i}, p_{1,j}\right)$
1	1	2001	2007	0.20	0.45	1.0
1	2	2001	2008	0.20	0.35	1.0
2	1	2002	2007	0.20	0.45	1.0
2	2	2002	2008	0.20	0.35	1.0
3	1	2003	2007	0.25	0.45	1.0
3	2	2003	2008	0.25	0.35	1.0
4	1	2004	2007	0.33	0.45	1.0
4	2	2004	2008	0.33	0.35	1.0
5	1	2005	2007	0.40	0.45	1.0
5	2	2005	2008	0.40	0.35	0.0
6	1	2006	2007	0.45	0.45	0.5
6	2	2006	2008	0.45	0.35	0.0
					$\sum I\left(p_{0,i}, p_{1,j}\right)$	9.5

Table B.5a. Example calculation of the hit and false-alarm rates for the ROC graph.

Year	Event	*p*	Thresholds						
			0.45	0.40	0.35	0.33	0.30	0.25	0.20
2001	No	0.20	0	0	0	0	0	0	1
2002	No	0.20	0	0	0	0	0	0	1
2003	No	0.25	0	0	0	0	0	1	1
2004	No	0.33	0	0	0	1	1	1	1
2005	No	0.40	0	1	1	1	1	1	1
2006	No	0.45	1	1	1	1	1	1	1
False-alarm rate			0.17	0.33	0.33	0.50	0.50	0.67	1.00
2007	Yes	0.45	1	1	1	1	1	1	1
2008	Yes	0.35	0	0	1	1	1	1	1
Hit rate			0.50	0.50	1.00	1.00	1.00	1.00	1.00

Table B.5b. Example calculation of the ROC area for the above-normal category using Eq. (5).

k	Threshold	HR (y_k)	FAR (x_k)	$y_k x_{k+1}$	$y_{k+1} x_k$	$y_k x_{k+1} - y_{k+1} x_k$
0		0.00	0.00	0.00	0.00	0.00
1	0.45	0.50	0.17	0.17	0.08	0.08
2	0.40	0.50	0.33	0.17	0.33	-0.17
3	0.35	1.00	0.33	0.50	0.33	0.17
4	0.33	1.00	0.50	0.50	0.50	0.00
5	0.30	1.00	0.50	0.67	0.50	0.17
6	0.25	1.00	0.67	1.00	0.67	0.33
7	0.20	1.00	1.00	1.00	1.00	0.00
8		1.00	1.00			
$\sum_{k=0}^{d} \left(y_k x_{k+1} - y_{k+1} x_k \right)$						0.58
$0.5 \times \left[1 + \sum_{k=0}^{d} \left(y_k x_{k+1} - y_{k+1} x_k \right) \right]$						0.79

iii. Hit (Heidke) scores

The hit score is calculated using

$$\overline{y}_k = 100\% \times \frac{1}{n} \sum_{i=1}^{n} y_{k,i} \, . \tag{B.3}$$

where n is the number of number of forecasts, and $y_{k,i}$ is the number of times the observation occurred in the category with the k^{th} highest probability. If two or more categories tie for the highest probability and one of those categories is observed then $y_{k,i}$ is adjusted to one divided by the number of categories with tied highest probability. For example, if one of two categories with tied highest probabilities verifies, then $y_{k,i} = 0.5$ (a "half-hit"), and $y_{k,i} = 0.33$ if one of three categories with tied highest probabilities verifies. The hit score ranges from 0% for the worst possible forecasts (the category with the highest probability never verifies) to 100% for the best possible forecasts (the category with the highest probability always verifies). In both cases resolution is strong, but the forecasts would be considered "good" only if the score exceeded the value expected from forecasts with no resolution. For forecasts with no resolution (i.e., no skill) the score depends on the number of categories and the climatological probabilities of those

categories, and so can be difficult to understand my non-specialists. The interpretation is simplified considerably if the categories are climatologically equiprobable, in which case the score of no-skill forecasts will be one divided by the number of categories (33% in the case of the standard three-category tercile-based seasonal forecasts). If the categories are not climatologically equiprobable then the expected score of no-skill forecasts is equal to the climatological probability of the largest category.

As the suffix *d* indicates, it is recommended that the hit score be calculated not only for the categories with the highest probabilities, but also for those with the second and third highest probabilities. The ranks of the categories based on their respective probabilities are shown in columns 3–5. Full calculations are provided in Table B.6. The skill score defining the difference between the hit scores for the categories with highest and lowest probabilities ($\overline{y}_1$ and $\overline{y}_3$ respectively) is $42\% - 4\% = 38\%$, which indicates positive skill, even though the category with the second highest probability verifies most frequently. (Note that this skill score is only meaningful if the categories are climatologically equiprobable.) The very low hit score for the category with the lowest probability indicates that the forecasts have been successful at indicating what is most likely not to happen, but have been less successful at indicating the most likely outcome. This aspect of skill may be particularly valuable for some users, and would not be recognized if only the hit score for the highest category were calculated. The calculation of the hit scores for the second and highest probabilities is particularly useful in areas where forecasters

Table B.6. Example calculation of the hit scores using Eq. (8c).

i	Observation	Forecast category rank			Hits		
		B	N	A	$y_{1,i}$	$y_{2,i}$	$y_{3,i}$
1	B	1	2	3	1	0	0
2	B	1	2	3	1	0	0
3	B	2	1	3	0	1	0
4	B	=1	=1	=1	0.33	0.33	0.33
5	N	3	2	1	0	1	0
6	N	3	2	1	0	1	0
7	A	3	2	1	1	0	0
8	A	3	1	2	0	1	0
$\frac{\text{number of hits}}{\text{number of forecasts}}$					$\frac{3\frac{1}{3}}{8} \approx 42\%$	$\frac{4\frac{1}{3}}{8} \approx 54\%$	$\frac{\frac{1}{3}}{8} \approx 4\%$

have hedged (as has been evident at many of the RCOFs, for example). Where there is a tendency to assign the highest probability to the normal category, the hit score then indicates only how frequently normal occurs, and provides little indication of whether the shift in the probability distribution towards above- or below-normal was informative. In the event that the shift was predicted skilfully, the hit score for the second highest category will be high.

iv. Brier score

As with the ROC, the Brier score has to be calculated for each category individually. It is calculated using

$$BS_j = \frac{1}{n}\sum_{i=1}^{n}\left(y_{j,i} - p_{j,i}\right)^2 . \qquad \text{(B.4)}$$

where n is the number of number of forecasts, where $y_{j,i}$ is 0.0 if category j did not occur, and 1.0 if it did, and $p_{j,i}$ is the forecast probability for category j. The score is the average of the squared

differences between the index indicating whether or not the category occurred and the forecast probability for that category. Table B.7 provides an example for the above-normal category.

Table B.7. Example calculation of the Brier score for the above-normal category using Eq. (B.4).

i	$y_{3,i}$	$p_{3,i}$	$\left(y_{3,i} - p_{3,i}\right)^2$
1	0.0	0.20	0.0400
2	0.0	0.20	0.0400
3	0.0	0.25	0.0625
4	0.0	0.33	0.1111
5	0.0	0.40	0.1600
6	0.0	0.45	0.2025
7	1.0	0.45	0.3025
8	1.0	0.35	0.4225
		$\frac{1}{n}\sum_{i=1}^{n}\left(y_{3,i} - p_{3,i}\right)^2$	0.1676

The score has a range from 0.0, in the case that all the forecasts correctly indicated with 100% probability the occurrences or non-occurrences of the category in question (an average probability error of zero), to an average probability error of 1.0, given perfectly bad forecasts (those which always indicated with 100% probability the incorrect category). Thus low scores are better than high scores. The score is meant as a summary verification measure, but because it measures reliability and resolution together (as well as uncertainty), it can be a little difficult to interpret. For example, there is no guarantee that one set of forecasts, A, with a lower Brier score is more useful than a second set of forecasts, B, with a higher Brier score, even if we can be absolutely certain that the difference in the scores is not because of sampling uncertainty.

The Brier score can be used to map the quality of the forecasts (i.e. it can be calculated for each location), and a single score can be calculated using all (or subsets) of the locations. However, comparisons for scores at different locations may be complicated if there are differences in uncertainty (i.e. if the prior probabilities are not the same everywhere). Even if uncertainty is constant over the map domain, differences in the score across the map may be misleading because of the combined measurement of resolution and reliability. Discussion about the calculation of these components of the Brier score is reserved until section *viii*, since these scores are not suitable for use with small sample sizes.

v. Ranked probability score

The ranked probability score is calculated over all categories, although the cumulative values of both the observations and the forecasts is 100% for the last category, and so this one can be ignored. The score is defined as

$$RPS = \frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{k=1}^{m-1}\left(\sum_{j=1}^{k}\left(y_{j,i} - p_{j,i}\right)\right)^2, \qquad \text{(B.5)}$$

where n is the number of forecasts, m is the number of categories, $y_{j,i}$ is 1 if the i^{th} observation was in category j, and is 0 otherwise, and $p_{j,i}$ is the i^{th} forecast probability for category j (Epstein 1969; Murphy 1969, 1970, 1971). For example, $y_{1,i}$ is set to 1.0 if category 1 occurred, and to 0.0 if it did not (column 2, Table B.8), and then compared with the forecast probability for category 1 (column 3). Then, for the second category, the cumulative observation is set to 1.0 if category 1 or 2 occurred, and to 0.0 if it did not (column 5), while the probabilities for the first two categories are added together (column 6). Full calculations are provided in Table B.8.

Table B.8. Example calculation of the ranked probability score using Eq. (B.5).

i	$y_{1,i}$	$p_{1,i}$	$\left(y_{1,i}-p_{1,i}\right)^2$	$\sum_{j=1}^{2} y_{j,i}$	$\sum_{j=1}^{2} p_{j,i}$	$\left(\sum_{j=1}^{2}\left(y_{j,i}-p_{j,i}\right)\right)^2$
1	1.0	0.45	0.3025	1.0	0.80	0.0400
2	1.0	0.50	0.2500	1.0	0.80	0.0400
3	1.0	0.35	0.4225	1.0	0.75	0.0625
4	1.0	0.33	0.4444	1.0	0.67	0.1111
5	0.0	0.25	0.0625	1.0	0.60	0.1600
6	0.0	0.20	0.0400	1.0	0.55	0.2025
7	0.0	0.20	0.0400	0.0	0.55	0.2025
8	0.0	0.25	0.0625	0.0	0.65	0.4225
$\frac{1}{n}\sum_{i=1}^{n}\left(y_{1,i}-p_{1,i}\right)^2$			0.2031	$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{k}\left(y_{j,i}-p_{j,i}\right)\right)^2$		0.1551
$\frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{k=1}^{m-1}\left(\sum_{j=1}^{k}\left(y_{j,i}-p_{j,i}\right)\right)^2$						0.1791

The score has a range from 0.0 in the case that all the forecasts correctly indicated with 100% probability the verifying categories, to 1.0 given perfectly bad forecasts (those which always indicated with 0% probability the verifying category). Thus, as with the brier score, low scores are better than high scores. The score is meant as a summary verification measure, but because it measures reliability and resolution together (as well as uncertainty), and of multiple categories, it can be a little difficult to interpret. For example, there is no guarantee that one set of forecasts, A, with a lower ranked probability score is more useful than a second set of forecasts, B, with a higher ranked probability score, even if we can be absolutely certain that the difference in the scores is not because of sampling uncertainty.

The ranked probability score can be used to map the quality of the forecasts (i.e. it can be calculated for each location), and a single score can be calculated using all (or subsets) of the locations. However, comparisons for scores at different locations may be complicated if there are differences in uncertainty (i.e. if the prior probabilities are not the same everywhere). Even if uncertainty is constant over the map domain, differences in the score across the map may be misleading because of the combined measurement of resolution and reliability.

vi. Effective interest rate

The effective interest is calculated from the ignorance score, which, in turn, is calculated by taking the logarithm (to base 2) of the probability on the category that verifies:

$$\text{ignorance score} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{j,i}\log_2\left[p_{j,i}\right], \tag{B.6}$$

where n is the total number of forecasts, m is the number of categories, $y_{j,i}$ is 1 if the i^{th} observation is in category j, and is 0 otherwise, and $p_{j,i}$ is the corresponding forecast probability. The summation over j in Eq. (14) simply searches for the verifying category; if the current category is not the correct one $y_{j,i}$ is zero and so the logarithm of the probability is irrelevant. An example is shown in Table B.9, where the second column indicates the verifying category, and the third the probability for that category. The final column indicates the ignorance score for the climatological forecasts.

The effective interest rate can be obtained from the ignorance score and the score for the climatological reference forecasts, *Ign*(*ref*):

$$\text{effective interest rate} = \left(2^{Ign(ref)-Ign} - 1\right), \tag{B.7}$$

Like the Brier and ranked probability scores the effective interest rate measures a number of attributes, but has a simpler interpretation because of its relationship to investment strategies. Specifically if a user were to bet on the forecasts and received fair odds (calculated using the climatological probabilities), and to carry losses and gains forwards each time, the effective interest rate indicates the profit or loss that would be made. For example, using the data in Table B.9, and starting with \$100, the user bets \$45 on below-normal, \$35 on normal, and \$20 on above-normal, and below-normal rainfall occurs. Given fair odds, the user gets three times what was bet on the verifying category, i.e. \$135 (3 × \$45), and has made \$35 profit (or 35%). The user then bets \$67,50 (50% of \$135) on below-normal, \$40,50 (30% of \$135) on normal, and \$27 on above-normal (20% of \$135), and wins \$202,50 (3 × \$67,50), which is a 50% profit. Over the two years the user has made \$102,50 profit, which is equivalent to. These calculations are not explicitly made in Table B.9 (but can be followed in the fifth column of Table B.10).

Table B.9. Example calculation of the effective interest rate using Eqs (B.6) and (B.7).

i	j	$p_{j,i}$	$\sum_{j=1}^{m} y_{j,i} \log_2\left[p_{j,i}\right]$	$\sum_{j=1}^{m} y_{j,i} \log_2\left[0.33\right]$
1	1	0.45	-1.152	-1.585
2	1	0.50	-1.000	-1.585
3	1	0.35	-1.515	-1.585
4	1	0.33	-1.585	-1.585
5	2	0.35	-1.515	-1.585
6	2	0.35	-1.515	-1.585
7	3	0.45	-1.152	-1.585
8	3	0.35	-1.515	-1.585
		$-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{j,i} \log_2\left[p_{j,i}\right]$	1.368	1.585
			$\left(2^{Ign(ref)-Ign} - 1\right)$	$\left(2^{1.585-1.368} - 1\right) \approx 16\%$

The effective interest rate is positive for "good" forecasts, but its upper bound depends upon the prior probabilities. For a three-category forecast system with equiprobable categories, the upper bound is 200%. In this instance, fair odds pays out three times the amount bet on the verifying category, and, for perfectly good forecasts, 100% of the bet would have been placed on this category. Thus, a bet of \$100, would be paid out \$300, making a profit of \$200, which is 200% of the original bet. For a perfectly bad set of forecasts, the ignorance score becomes infinity, and so, from Eq. (B.7) the effective interest rate approaches -100%. It should be noted that the effective interest rate approaches -100% if *any* of the individual forecasts is perfectly bad (a probability of 0% is assigned to the verifying category) because the ignorance score for that one case will be infinity. The rate will be -100% even if all the other forecasts are perfectly good.

vii. Accumulated profits graph

The accumulated profits are calculated by accumulating the returns on an initial investment of, for example, \$100 or €100 (the actual currency is immaterial). The profits can be calculated using

$$b\left[\left(\prod_i \frac{p_i}{c_i}\right) - 1\right] \tag{B.8}$$

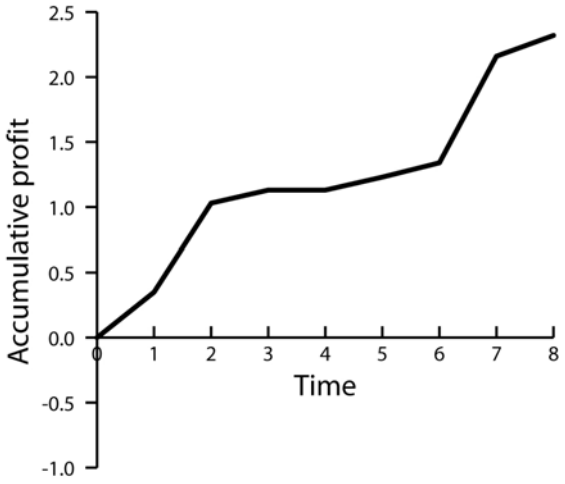

Figure B.2. Example accumulated profits graph using the data in Table B.10.

where b is the initial bet, or investment, p_i is the forecast probability for the verifying category, and c_i is its climatological probability. Eq. (B.8) successively adds the interest earned at each round to the initial investment. So for example, a 35% profit is made in round one in Table B.10 (fourth column), and so 35% is added to the initial investment by multiplying it by 1.35. Given an initial investment of $100, a profit of $35 has been made, and there is now $135 for investment in the second round (fifth column first row). In the second round, a 50% profit is made (fourth column) the $135 is multiplied by 1.50 to give $203 (fifth column), and the profit is now $103 (last column). The resulting graph for the example shown in Table B.10 is shown in Figure B.2.

Table B.10. Example calculation of the accumulated profits using Eq (B.8).

i	j	$p_{j,i}$	$\frac{p_{j,i}}{0.33}$	$\prod_i \frac{p_{j,i}}{0.33}$	$\prod_i \frac{p_{j,i}}{0.33} - 1$
1	1	0.45	1.35	1.35	0.35
2	1	0.50	1.50	2.03	1.03
3	1	0.35	1.05	2.13	1.13
4	1	0.33	1.00	2.13	1.13
5	2	0.35	1.05	2.23	1.23
6	2	0.35	1.05	2.34	1.34
7	3	0.45	1.35	3.16	2.16
8	3	0.35	1.05	3.32	2.32

As might be expected, the cumulative profits diagram is “good” if it is upward sloping, showing accumulating profits. In most cases the graph is unlikely to be monotonically increasing, and may even dip into negative territory occasionally. The fluctuations in the graph are particularly useful for emphasizing the fact that the forecasts may contribute to loss-making decisions in some years, but that in the long run the forecasts are (assuming an upwardly sloping curve) useful. It should be evident from the equation heading column 4 of Table B.10 that a profit will be made in any given year only if the forecast probability on the verifying category exceeds the climatological probability.

viii. Reliability diagram

It is not viable to construct a reliability diagram meaningfully from the data in Table B.1 because the sample size is far too small. Instead, an example is shown based on a verification of the first 10 years of PRESAO forecasts (Chidzambwa and Mason 2008). The data are shown in Table B.11, which lists the full range of possible forecast probabilities (excluding the climatological probability), and the number of forecasts for each probability (n_k) on the above-normal category, together with the number of times above-normal rainfall was observed for each of the forecast values. This third column is calculated as $\sum_{i=1}^{n_k} y_{k,i}$. The forecast relative frequency is the number of forecasts per forecast probability divided by the total number of forecasts, and so the first value is $97 \div 698 \approx 0.14$. These values are plotted as a histogram on the reliability diagram to show the sharpness of the forecasts. The observed relative frequency is calculated as the number of events divided by the number of forecasts, so the first value is $15 \div 97 \approx 0.15$. These values are plotted as the reliability curve against the forecasts in column 1 as the *x*-axis (Figure 4, top left).

To fit the regression line to the reliability curve, the values of $\overline{y}_k$ in the last column of Table B.11a are regressed against the forecast probabilities in the first column, but weighted by the forecast frequencies (column 2; using the values in column 4 will give the same result). The step-by-step calculations are shown in Table B.11b, and only the rows with non-zero frequencies are included. The regression parameters are calculated using

$$\beta_1 = \frac{\sum_{k=1}^{d} n_k \left(p_k - \overline{p} \right)\left(\overline{y}_k - \overline{y} \right)}{\sum_{k=1}^{d} n_k \left(p_k - \overline{p} \right)^2}. \qquad \text{(B.9a)}$$

and

$$\beta_0 = \overline{y} - \beta_1 \overline{x}\,. \qquad \text{(B.9b)}$$

where β_1 is the slope and β_0 the intercept of the fitted regression line, d is the number of discrete probability values (the number of rows in Table B.11b excluding the calculations), n_k is the number of forecasts for the k^{th} probability value (second column), $\overline{p}_k$ is the k^{th} probability value (first column), $\overline{p}$ is the average probability for all forecasts, $\overline{y}_k$ is the observed relative frequency for the k^{th} probability value (third column), and $\overline{y}$ is the observed relative frequency over the verification period.

The slope, β_1, can range between positive and negative infinity, but has an ideal value of 1.0, and in most cases is likely to be between 0.0 and 1.0. If the forecasts have good resolution and perfect reliability (after adjusting for over- or under-forecasting) an event should increase in frequency by 10% as the forecast probability is incremented by each 10% (e.g., from 30% to 40%, or from 40% to 50%), and the slope will be 1.0. , If the forecasts have no resolution the slope will be zero. In the example in Table B.11b, the slope indicates that observed relative frequency of above-normal rainfall increases by more than 7% when the forecast probability increases by 10%. The forecasts are therefore slightly over-confident. In general, over-confidence is indicated by a slope of between 0.0 and 1.0 (the increase in frequency will be between 0% and 10%), while under-confidence will be indicated by slopes of greater than 1.0 (increases in frequency of more than 10%). Occasionally, negative slopes may be encountered, which indicates the forecasts are "bad" in the sense that if the forecasts imply an increase in the chances of a category occurring, that category actually becomes less frequent.

Table B.11a. Example construction of a reliability diagram.

Forecast	Number of forecasts (n_k)	Number of events	Forecast relative frequency	Observed relative frequency ($\bar{y}_k$)
0.00	0	0		
0.05	0	0		
0.10	0	0		
0.15	0	0		
0.20	97	15	0.14	0.15
0.25	67	10	0.10	0.15
0.30	211	62	0.30	0.29
0.35	95	23	0.14	0.24
0.40	153	62	0.22	0.40
0.45	52	15	0.07	0.29
0.50	23	5	0.03	0.22
0.55	0	0		
0.60	0	0		
0.65	0	0		
0.70	0	0		
0.75	0	0		
0.80	0	0		
0.85	0	0		
0.90	0	0		
0.95	0	0		
1.00	0	0		
	$\sum_{k=1}^{d} n_k = 698$	$\sum_{i=1}^{n} y_i = 192$	$\bar{p} \approx 0.33$	$\bar{y} \approx 0.28$

Table B.11b. Example calculation of a weighted regression fit to the reliability curve using the data shown in Table B.11a. Note that the values in the last column were obtained by using the exact values of $\bar{p}$ and $\bar{y}$, not the approximate values indicated at the foot of the Table; the rounding errors can be quite large.

p_k	n_k	$\bar{y}_k$	$n_k (p_k - \bar{p})(\bar{y}_k - \bar{y})$	$n_k (p_k - \bar{p})^2$
0.20	97	0.15	1.49	1.58
0.25	67	0.15	0.66	0.41
0.30	211	0.29	-0.11	0.16
0.35	95	0.24	-0.07	0.05
0.40	153	0.40	1.43	0.80
0.45	52	0.29	0.09	0.78
0.50	23	0.22	-0.23	0.68
$\bar{p} \approx 0.33$		$\bar{y} \approx 0.28$	$\sum_{k=1}^{d} n_k (p_k - \bar{p})(\bar{y}_k - \bar{y}) \approx 3.26$	$\sum_{k=1}^{d} n_k (p_k - \bar{p})^2 \approx 4.46$
			$\beta_1 \approx \frac{3.26}{4.46} \approx 0.73$	$\beta_0 \approx 0.28 - 0.73 \times 0.33 \approx 0.03$

Interpreting the intercept, β_0, is difficult since it depends on the slope. However, it can be useful in diagnosing over- and under-forecasting. If the slope is close to 1.0, and the intercept is much above 0.0, under-forecasting is present, but there is over-forecasting if the intercept is much below zero. In most cases, however, the intercept is likely to be a reflection of the slope, and is likely to be somewhere between 0.0 and the observed relative frequency of the event over the verification period (which is not necessarily the same as the climatological probability). Assuming no notable over- or under-forecasting, the less resolution there is the closer the intercept will be to the observed relative frequency of the event over the verification period. Other diagnoses can be discerned from the schematic diagrams shown in Figure 3.

b. Individual forecast maps

i. Linear probability score

The linear probability score is the average of the forecasts on the verifying categories, and is calculated using

$$\text{linear probability score} = 100\% \times \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{j,i} p_{j,i} \,, \tag{B.10}$$

where n is the number of points on the map at which the forecasts are to be verified, m is the number of categories, $y_{i,j}$ is 1 if the observation at location i was in category j, and is 0 otherwise, and $p_{j,i}$ is the forecast probability for category j at location i. The summation over j represents a search for the verifying category. Using the data in Table B.2, these probabilities are shown in the third column of Table B.12.

The score is the average forecast probability assigned to the verifying categories, and it ranges between 100% for perfect forecasts (a probability of 100% was assigned to the observed category at each of the locations) to 0% for perfectly bad forecasts (a probability of 0% was assigned to the observed category at each of the locations). A "good" forecast will score more than a strategy of using the climatological forecasts, and will beat the expected score from guessing. In the example shown in Table B.12, the average probability is about 39%, which is about 6% more than climatology.

Table B.12. Example calculation of the linear probability score using Eq. (B.10).

i	Obs	$\sum_{j=1}^{m} y_{j,i} p_{j,i}$
I	B	0.45
II	B	0.50
III	B	0.35
IV	B	0.33
V	N	0.35
VI	N	0.35
VII	A	0.45
VIII	A	0.35
$100\% \times \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{j,i} p_{j,i}$		39%

ii. Average interest rate

The average interest rate is calculated from the ratios of the forecast probability on the category that verifies to its climatological probability:

$$\text{average interest rate} = \left(\frac{1}{n} \sum_{i=1}^{n} \frac{p_i}{c_i} \right) - 1, \tag{B.11}$$

where p_i is the forecast probability for the verifying category at the ith of n locations, and c_i is the corresponding climatological probability. As with the linear probability score, the summation over j in Eq. (B.11) simply searches for the verifying category. An example for the data in Table B.2 is shown in Table B.13, where the second column indicates the verifying category, and the third the probability for that category.

Table B.13. Example calculation of the average interest rate using Eq. (B.11).

i	j	$p_{j,i}$	$\frac{p_i}{c_i}$
I	1	0.45	1.35
II	1	0.50	1.50
III	1	0.35	1.05
IV	1	0.33	1.00
V	2	0.35	1.05
VI	2	0.35	1.05
VII	3	0.45	1.35
VIII	3	0.35	1.05
		$\left(\frac{1}{n} \sum_{i=1}^{n} \frac{p_i}{c_i} \right) - 1$	17.50%

The average interest rate has a similar interpretation to the effective interest rate (section *vi*): it is positive for "good" forecasts, but its upper bound depends upon the prior probabilities at each of the locations. For a three-category forecast system with equiprobable categories at all locations, the upper bound is 200%. In this instance, fair odds pays out three times the amounts bet on the respective verifying categories, and, for perfectly good forecasts, 100% of the bet at each location would have been placed on the verifying category. Thus, a bet of $100, would be paid out $300, making a profit of $200 at any location, which is 200% of the original bet. For a perfectly bad set of forecasts, the average interest rate approaches -100%. However, unlike the effective interest rate the average interest rate approaches -100% only if *all* of the individual forecasts are perfectly bad (a probability of 0% is assigned to the verifying category).

iii. Ignorance score

The ignorance score is calculated by taking the logarithm (to base 2) of the probability on the category that verifies.

$$\text{ignorance score} = -\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} y_{j,i} \log_2 \left[p_{j,i} \right], \tag{B.12}$$

where n is the total number of locations, m is the number of categories, $y_{j,i}$ is 1 if the i^{th} observation is in category j, and is 0 otherwise, and $p_{j,i}$ is the corresponding forecast probability. As with the linear probability score, the summation over j in Eq. (B.12) simply searches for the

verifying category. An example for the data in Table B.2 is shown in Table B.14, where the second column indicates the verifying category, and the third the probability for that category.

Table B.14. Example calculation of the ignorance score using Eq. (B.12).

i	j	$p_{j,i}$	$\sum_{j=1}^{m} y_{j,i} \log_2 \left[p_{j,i} \right]$
I	1	0.45	-1.152
II	1	0.50	-1.000
III	1	0.35	-1.515
IV	1	0.33	-1.585
V	2	0.35	-1.515
VI	2	0.35	-1.515
VII	3	0.45	-1.152
VIII	3	0.35	-1.515
		$-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} y_{j,i} \log_2 \left[p_{j,i} \right]$	1.368

The score ranges from 0.0 for a perfect set of forecasts, to infinity for a perfectly bad set of forecasts. In fact, if any of the forecasts has a probability of 0% for the verifying category the score will be infinity even if all the other forecasts have 100% on the verifying categories.

APPENDIX C:

Glossary

The definitions provided in this glossary are adapted from those provided by Jolliffe and Stephenson (2012), and Troccoli et al. (2008), to which the reader is referred for further details. These definitions apply primarily to how the terms are used within this document, and they may have more general meanings in other contexts. Unfortunately, within the verification literature many terms are used inconsistently: many terms are used with different meanings, while some definitions are given different terms. The interested reader is referred to the more comprehensive glossaries cited above for fuller details. Terms or phrases in italics have glossary entries.

Accuracy

An *attribute* of forecast *quality*; specifically, the magnitude of the error(s) in a single or a set of forecasts. An "accurate" forecast is one with a small error; it addresses the question "Was the forecast close to what happened?" Accuracy is usually taken as an attribute of *deterministic forecasts* and is measured in the units of the *predictand*, but here it is applied to *probabilistic forecasts* to refer to high probabilities on the verifying outcome, without specific regard to the *reliability* or *resolution* of the probabilities. For example, suppose that rainfall is observed to be above-normal, a forecast that indicated 60% probability of above-normal would be considered more accurate than one that indicated 40%. Accuracy is considered a desirable property of probabilistic forecasts for a specific target period (e.g., the seasonal forecast for January – March 2000).

Anomaly

The difference between an observed value of a meteorological variable (e.g., seasonally averaged temperature) for a single period [e.g., January – March (JFM) 2000] and its long-term average (e.g., JFM 1961–1990). In the case of seasonally averaged temperature, for example, a positive anomaly occurs when the temperature for the season in question is higher than average, and a negative anomaly occurs when the season is colder than average.

Attribute

A specific aspect of the *quality* of forecasts. Forecast *quality* is multi-faceted, and so forecasts can be described as good or bad in a number of different ways. The attributes of good *probabilistic forecasts* are *discrimination*, *reliability*, *resolution*, *sharpness*, and *skill*.

Base rate

The observed relative frequency of observations in a category as measured over a pre-defined period. If the sample period is the climatological period then the *climatological probability* for the category of interest is equivalent to the base rate. In this document the base rate is often, but not always, defined using the verification period.

Bias

A systematic difference between the forecasts and the outcomes. Biases can be *conditional* or *unconditional*.

Bootstrap

A means of estimating sampling errors in the value of a parameter (e.g., a verification score) by resampling with replacement from the original dataset. Bootstrapping is recommended for estimating the uncertainty in each verification score given that the sample size of seasonal forecasts is generally very small. The procedure involves re-calculating a verification score a large number of times and then examining the distribution of these values. Typically the distribution is summarized by identifying one of the lowest and one of the highest score

values (but not the absolute lowest and highest), and thus defining a range or "interval" between which the true score is thought to lie.

Climatological probabilities

The observed relative frequency of observations in each category as measured over a pre-defined historical period (typically, but not always, 30 years). For most seasonal forecasting systems three categories are used, and are defined so that there is an equal number of years in each category; the climatological probabilities are then about 33%.

Conditional bias

A systematic difference between the forecasts and the outcomes that is dependent upon the forecast. *Over-* and *under-confidence* are examples of conditional bias.

Confidence

A degree of belief placed by the forecaster in a forecast. A confident forecaster believes that there is less uncertainty in the outcome than an unconfident forecaster, and so the confident forecaster will issue a *probabilistic forecast* with greater *sharpness*. For example, in the three-category situation, a forecaster that says there is a 60% chance of above-normal rainfall is more confident than a forecaster who says there is a 50% chance of above-normal rainfall because 60% is a larger shift from the climatological value than is 50%. Consider another example: a forecaster who says there is a 10% chance of above-normal rainfall is more confident than one who says there is a 50% chance. The first forecaster is very confident that above-normal rainfall will not occur, and a probability of 10% is a bigger shift from 33% than is a probability of 50%.

Confidence interval

A range defining upper and lower limits between which the value of a parameter being estimated (e.g., a verification score) is likely to lie. The confidence level defines how likely it is that the interval contains the parameter value.

Consistency

A correspondence between a forecast and the forecasters' beliefs. If a forecast is consistent, it communicates what the forecaster thinks will happen, and will correctly indicate their level of uncertainty. A forecaster may want to issue a forecast that is inconsistent with their belief to avoid causing an over-reaction, for example, if there are strong indications of dry conditions.

Correctness

An *attribute* of the *quality* of *deterministic forecasts*; specifically, an exact correspondence between a forecast and an observation. For example, if a forecaster says that total rainfall will be below-normal, the forecast is correct if, and only if below-normal rainfall is observed. If above-normal rainfall occurs, the forecaster is not "less correct" than one who says that normal rainfall will occur; both forecasters are incorrect, but the second forecaster is more *accurate*.

Deterministic forecast

A forecast expressed as a specific value (e.g., total rainfall in mm) or a specific category (e.g., temperature in the below-normal, normal, or above-normal category) without any indication of uncertainty.

Discrimination

An *attribute* of the *quality* of *probabilistic forecasts*; specifically, the conditioning of the forecast on the outcome. Discrimination addresses the question: "Does the forecast differ given different outcomes?", but does not specifically address the question: "Is the forecast probability higher when an *event* occurs than when it does not occur?" If the forecast is the same regardless of the outcome, the forecasts cannot discriminate an *event* from a non-event.

Forecasts with no discrimination are useless since the forecast is, on average, the same regardless of what happens.

Event

An observation during the *target period* of a specific outcome of interest. The outcome is explicitly binary: either an event occurs during the *target period*, or it does not occur. For seasonal forecasts, an event is usually defined as the occurrence of the verifying observation in a specific category of interest. For example, if above-normal rainfall is defined as an event, an event occurs if rainfall is above-normal.

False alarm

A warning or forecast issued for an *event* that does not actually occur.

False-alarm rate (FAR)

A measure of the *quality* of *deterministic forecasts*; specifically, the number of *false alarms* divided by the number of non-events. The false-alarm rate measures the proportion of non-events that were incorrectly fore-warned, and should be distinguished from the *false-alarm ratio*, which measures the proportion of incorrect warnings.

Hedging

Issuing a forecast different to what the forecaster truly believes in order to optimize an expected benefit. The benefit may be in terms of a verification score that lacks *propriety* or may be to effect a specific response to the forecast that may be considered more desirable (e.g., trying to minimize the possibility that the forecast might be perceived as "wrong") or appropriate (e.g., avoiding over-reacting to a warning of dry conditions) than would have been realized without the hedging. Whatever the motivation, hedging results in forecasts that lack *consistency*.

Hit

A warning or forecast issued for an *event* that occurs.

Hit rate (HR)

A measure of the *quality* of *deterministic forecasts*; specifically, the number of *hits* divided by the number of *events*. The hit rate measures the proportion of *events* that were fore-warned, and should be distinguished from the *hit score*, which measures the proportion of correct warnings.

Hit score

A measure of the *quality* of *deterministic forecasts*; specifically, the number of *hits* divided by the number of warnings. The hit score (sometimes called the Heidke score) measures the proportion of correct warnings, and should be distinguished from the *hit rate*, which measures the proportion of *events* that were fore-warned. The *y*-values of the *reliability curve* measure the observed relative frequency, which can be interpreted as a conditional hit score.

Over-confidence

A tendency to over-estimate differences from climatology of the probability of an *event*, resulting in probabilities that are too high when the probabilities are increased above their *climatological* value, and too low when the probabilities are decreased. Over-confident forecasts have too much *sharpness*. Over-confidence is an example of *conditional bias*, and is diagnosed by a *reliability curve* that is shallower than 45°.

Over-forecasting

A tendency to over-estimate the probability of an *event* regardless of whether the probabilities suggest that the *event* is more or less likely to occur than climatologically. If an *event* is over-forecast it occurs less frequently than implied by the forecasts. Over-forecasting is an

example of *unconditional bias*, and is diagnosed by a *reliability curve* that is below the 45° line.

Percentile
Each part of a distribution that divides the data into one hundred equal parts.

Predictand
That which is being forecast. Predictands in seasonal forecasting are primarily seasonal rainfall total or average temperature.

Probabilistic forecasts
A forecast that is expressed as a probability or set of probabilities of one or more events occurring. Probabilistic forecasts explicitly indicate the level of uncertainty in the prediction, and communicate the level of *confidence* the forecaster has in the forecast. If a probabilistic forecast is *consistent*, the probability for any specific category can be interpreted as the probability that the forecaster thinks a deterministic forecast of that category will be correct.

Propriety
A property of a verification score for a probability forecast; specifically, a score is proper (exhibits propriety) if its value is optimized when the forecast is *consistent* with the forecaster's best judgment. For *strictly proper* scores the value is uniquely optimized.

Quality
A measure of the association between forecasts and the corresponding observations.

Quantile
Each part of a distribution that divides the data into a specified equal number of parts. A quantile is a generic term of which *tercile* and *percentile* are examples.

Reliability
An *attribute* of the *quality* of *probabilistic forecasts*; specifically, the correspondence between the forecast probabilities and the conditional observed relative frequencies of events. Forecasts are reliable if, for all forecast probabilities, the observed relative frequency is equal to the forecast probability (i.e., an event must occur on 40% of the occasions that the forecast probability is 40%, 50% of the occasions the probability is 50% etc.).

Reliability curve
A plot of conditional observed relative frequencies of events (on the *y*-axis) against forecast probability (on the *x*-axis).

Resolution
An *attribute* of the *quality* of *probabilistic forecasts*; specifically, the conditioning of the outcome on the forecasts. Resolution addresses the question: "Does the frequency of occurrence of an *event* differ as the forecast probability changes?", but does not specifically address the question: "Does the *event* become more (less) frequent as the probability increases (decreases)?" If the *event* occurs with the same relative frequency regardless of the forecast, the forecasts are said to have no resolution. Forecasts with no resolution are useless since the outcome is, on average, the same regardless of what is forecast.

Sharpness
An *attribute* of the *quality* of *probabilistic forecasts*; specifically, the degree to which the forecast probabilities differ from their *climatological* values. Sharp forecasts have probabilities that are rarely close to the *climatological probabilities*. Sharp forecasts are an indication of high *confidence*, but sharpness does not consider the outcomes, and so is not

concerned with whether the high *confidence* is appropriate. Thus *over-confident* forecasts have good sharpness even if they have poor *reliability*.

Skill

An *attribute* of forecast *quality*; specifically, a comparative measure of forecast *quality*, in which a set of forecasts has positive skill if it scores better on one or more forecast *attributes* than another set, known as the reference set. Forecast skill is usually measured against a naïve forecasting strategy, such as random guessing, perpetual forecasts of one category, or *climatological probabilities* of all categories, but can be calculated using any reference set.

Strictly proper

A property of a verification score for a probability forecast; specifically, a score is strictly proper if its value is uniquely optimized when the forecast is *consistent* with the forecaster's best judgment. Strictly proper scores discourage the forecaster from *hedging*, and are generally to be preferred to scores that lack *propriety*.

Target period

The period for which the forecast applies.

Tercile

One of two values that divides the distribution of data into three equal parts. The upper tercile is the higher of the two terciles, and is frequently used to define the lower limit of the above-normal category. The lower tercile is frequently used to define the upper limit of the below-normal category. The normal category is then bounded by the two terciles.

Training period

The period over which the forecast system was calibrated, and which was used to define the categories.

Unconditional bias

A systematic difference between the forecasts and the outcomes that is independent of the forecast. *Over-* and *under-forecasting* are examples of unconditional bias.

Under-confidence

A tendency to under-estimate differences from climatology of the probability of an *event*, resulting in probabilities that are too low when the probabilities are increased above their climatological values, and too high when the probabilities are decreased. Under-confident forecasts have insufficient *sharpness*. Under-confidence is an example of *conditional bias*, and is diagnosed by a *reliability curve* that is steeper than 45°.

Under-forecasting

A tendency to under-estimate the probability of an *event* regardless of whether the probabilities suggest that the *event* is more or less likely to occur than climatologically. If an *event* is under-forecast it occurs more frequently than implied by the forecasts. Under-forecasting is an example of unconditional bias, and is diagnosed by a *reliability curve* that is above the 45° line.

Value

A measure of the benefit achieved (or loss incurred) through the use of forecasts.

Verification

The measurement of the *quality* of a forecast or of a series of forecasts.

References

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3.

Bröcker, J., and L. A. Smith, 2007a: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661.

Bröcker, J., and L. A. Smith, 2007b: Scoring probabilistic forecasts: the importance of being proper. *Wea. Forecasting*, **22**, 382–388.

Chidzambwa, S., and S. J. Mason, 2008: Report of the evaluation of regional climate outlook forecasts for Africa during the period 1997 to 2007. *ACMAD Technical Report*, ACMAD, Niamey, 30 pp.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.*, **8**, 985–987.

Hagedorn, R., and L. A. Smith, 2008: Communicating the value of probabilistic forecasts with weather roulette. *Meteor. Appl.*, DOI: 10.1002/met.92.

Harvey, L. O., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120**, 863–883.

Hogan, R. J., and I. B. Mason, 2012: Deterministic forecasts of binary events. In Jolliffe, I. T., and D. B. Stephenson (Eds), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Wiley, Chichester, 31–59.

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

Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, Chichester, 292 pp.

Katz, R. W., and A. H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts*, Cambridge University Press, Cambridge, 222 pp.

Mason, S. J., 2012: Seasonal and longer-range forecasts. In Jolliffe, I. T., and D. B. Stephenson (Eds), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Wiley, Chichester, 203–220.

Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and levels (ROL) curves: statistical significance and interpretation. *Q. J. Royal Meteor. Soc.*, **128**, 2145–2166.

Mason, S. J., and A. P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Mon. Wea. Rev.*, **139**, 331–349.

Murphy, A. H., 1966: A note on the use of probabilistic predictions and the probability score in the cost-loss ratio decision situation. *J. Appl. Meteor.*, **5**, 534–537.

Murphy, A. H., 1969: On the "ranked probability score." *J. Appl. Meteor.*, **8**, 988–989.

Murphy, A. H., 1970: The ranked probability score and the probability score: a comparison. *Mon. Wea. Rev.*, **98**, 917–924.

Murphy, A. H., 1971: A note on the ranked probability score. *J. Appl. Meteor.*, **10**, 155–156.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600.

Murphy, A. H., 1991: Forecast verification: its complexity and dimensionality. *Mon. Wea. Rev.*, **119**, 1590–1601.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.*, **130**, 1653–1660.

Stephenson, D. B., C. A. S. Coelho, and I. T. Jolliffe, 2008: Two extra components in the Brier Score decomposition, *Wea. Forecasting*, **23**, 752–757.

Tall, A., S. J. Mason, M. K. van Aalst, P. Suarez, Y. Ait-Chellouche, A. A. Diallo, and L. Braman, 2012: Using seasonal climate forecasts to guide disaster management: The Red Cross experience during the 2008 West Africa floods. *International Journal of Geophysics*, Article ID 986016.

Tippett, M. K., and A. G. Barnston, 2008: Skill of multimodel ENSO probability forecasts. *Mon. Wea. Rev.*, **136**, 3933–3946.

Troccoli, A., M. S. J. Harrison, D. L. T. Anderson, and S. J. Mason, 2008: *Seasonal Climate Variability: Forecasting and Managing Risk*, Springer Academic Publishers, Dordrecht, 467 pp.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*, Academic Press, San Diego, 704 pp.

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

Wilson, L. J., W. R. Burrows, and A. Lanzinger, 1999: A strategy for verification of weather element forecasts from an ensemble prediction system. *Mon. Wea. Rev.*, **127**, 956–970.