

# Guidance on Verification of Operational Seasonal Climate Forecasts

2018 edition

WEATHER CLIMATE WATER



WORLD  
METEOROLOGICAL  
ORGANIZATION

WMO-No. 1220



# Guidance on Verification of Operational Seasonal Climate Forecasts

2018 edition



WORLD  
METEOROLOGICAL  
ORGANIZATION

WMO-No. 1220

#### EDITORIAL NOTE

Typefaces employed in this volume do not signify standard or recommended practices, and are used solely for legibility. The word shall is used to denote practices that are required for data representation to work. The word should denotes recommended practices.

METEOTERM, the WMO terminology database, may be consulted at <http://public.wmo.int/en/resources/meteoterm>.

Readers who copy hyperlinks by selecting them in the text should be aware that additional spaces may appear immediately following [http://](#), [https://](#), [ftp://](#), [mailto:](#), and after slashes (/), dashes (-), periods (.) and unbroken sequences of characters (letters and numbers). These spaces should be removed from the pasted URL. The correct URL is displayed when hovering over the link or when clicking on the link and then copying it from the browser.

WMO-No. 1220

© World Meteorological Organization, 2018

The right of publication in print, electronic and any other form and in any language is reserved by WMO. Short extracts from WMO publications may be reproduced without authorization, provided that the complete source is clearly indicated. Editorial correspondence and requests to publish, reproduce or translate this publication in part or in whole should be addressed to:

Chairperson, Publications Board  
World Meteorological Organization (WMO)  
7 bis, avenue de la Paix  
P.O. Box 2300  
CH-1211 Geneva 2, Switzerland

Tel.: +41 (0) 22 730 84 03  
Fax: +41 (0) 22 730 81 17  
Email: [publications@wmo.int](mailto:publications@wmo.int)

ISBN 978-92-63-11220-0

#### NOTE

The designations employed in WMO publications and the presentation of material in this publication do not imply the expression of any opinion whatsoever on the part of WMO concerning the legal status of any country, territory, city or area, or of its authorities, or concerning the delimitation of its frontiers or boundaries.

The mention of specific companies or products does not imply that they are endorsed or recommended by WMO in preference to others of a similar nature which are not mentioned or advertised.

## **ACKNOWLEDGEMENTS**

### **Lead author:**

Dr Simon J. Mason, Senior Research Scientist, International Research Institute for Climate and Society (IRI), United States of America

The *Guidance on Verification of Operational Seasonal Climate Forecasts* has been prepared under the auspices of the WMO Commission for Climatology.

### **Reviewers:**

Mr Jean-Pierre Céron (Météo France)

Mr Akihiko Shimpo (Japan Meteorological Agency)

Dr William Wang (Bureau of Meteorology, Australia)

Dr Normand Gagnon (Environment and Climate Change Canada)

Dr Richard Graham (Met Office, United Kingdom of Great Britain and Northern Ireland)

Dr Anthony Barnston (IRI, United States)



# CONTENTS

	<i>Page</i>
EXECUTIVE SUMMARY .....	vii
1. INTRODUCTION .....	1
2. FORECAST AND VERIFICATION DATA .....	3
2.1 Defining the target variable .....	3
2.2 Sizes of forecast regions .....	4
2.3 Gridding of forecasts .....	5
2.4 Verification using station data .....	5
2.5 Data inconsistencies .....	6
3. ATTRIBUTES OF “GOOD” FORECASTS .....	8
3.1 Types of forecast “goodness” .....	8
3.2 Probabilistic forecasts and forecast quality .....	9
3.2.1 Attributes of “good” probabilistic forecasts .....	10
3.2.1.1 Resolution .....	10
3.2.1.2 Discrimination .....	10
3.2.1.3 Reliability .....	11
3.2.1.4 Sharpness .....	12
3.2.1.5 Skill .....	12
3.2.2 Attributes of individual probabilistic forecast maps .....	13
4. MEASURING FORECAST QUALITY .....	15
4.1 The verification of climatological forecasts .....	15
4.2 Measuring the quality of series of forecasts .....	16
4.2.1 Measuring discrimination .....	16
4.2.1.1 Relative operating characteristics .....	16
4.2.1.2 Generalized discrimination .....	20
4.2.2 Measuring resolution .....	22
4.2.2.1 Resolution components of multi-attribute scores .....	22
4.2.2.2 Hit scores as measures of resolution .....	23
4.2.3 Measuring reliability .....	25
4.2.4 Measuring multiple attributes .....	27
4.2.5 Detailed diagnostics .....	29
4.3 Measuring the quality of individual forecast maps .....	33
4.3.1 Scoring of attributes .....	33
4.3.2 Model diagnostics .....	35
5. UNCERTAINTY OF RESULTS .....	36
APPENDIX A. WEIGHTED VERSIONS OF THE VERIFICATION SCORES .....	37
APPENDIX B. CALCULATION OF THE RECOMMENDED SCORES AND GRAPHS .....	42
APPENDIX C. GLOSSARY .....	59
REFERENCES .....	65





## EXECUTIVE SUMMARY

The purpose of this publication 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 and other forecasting centres. The recommendations are meant to complement the WMO Commission for Basic Systems Standardized Verification System for Long-range Forecasts (SVSLRF). SVSLRF defines standards for verifying model outputs from Global Producing Centres (GPCs), and so includes procedures for measuring the quality of ensemble prediction systems. In contrast, the procedures described in this publication are exclusively for verification of probabilistic forecasts, which may be model outputs, expert subjective assessments, or a combination of both. A second difference from SVSLRF is that procedures described in this publication are concerned not only with verification of a history of forecasts, but also of forecasts for one specific target period – for last year’s RCOF forecast, for example.

The recommended procedures range in complexity from simple measures for communicating forecast quality to non-specialists, to detailed diagnostic procedures for in-depth analyses of the various strengths and weaknesses of forecasts. Interpretive guides for each of the procedures are included to assist the user in understanding the verification results, and worked examples are included in Appendix B. A glossary of technical terms is also provided in Appendix C.

Ideally the multiple attributes of forecast quality should be measured individually, but some commonly used procedures measure more than one attribute at once. These procedures can lead to results that are difficult to interpret, and may lead to misleading conclusions. Alternative procedures that measure individual attributes are suggested in preference throughout this guidance publication because of their simpler interpretation and more informative results. Nevertheless, the occasional need for summary scores is recognized, and some suggestions are presented. While reliability (do the forecast probabilities give an accurate indication of the uncertainty in the outcome?) is unquestionably an important attribute, ultimately the most important attributes are resolution or discrimination. Resolution measures whether the outcome differs given different forecasts, while discrimination measures whether the forecasts differ given different outcomes. As long as there is some resolution or discrimination the forecasts contain potentially useful information, regardless of how poor the reliability is.

For detailed diagnostics of forecast quality, reliability diagrams are recommended; these diagrams measure reliability and resolution. SVSLRF recommends constructing the diagrams with points at 10% increments, consistent with typical numbers of ensemble members available for estimating the forecast probabilities. It is recommended in this publication that the diagrams be drawn with points at 5% increments rather than 10% because of the general practice of issuing seasonal forecasts with probabilities rounded to the nearest 5%.

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 provide simple visual indications of unconditional biases. For more technical use, the reliability and resolution components of the ignorance and Brier scores (BS) are suggested. Of these, the components of the ignorance score are perhaps to be preferred because of its asymmetric measurement of probability errors, which is appropriate for categories that do not have a climatological probability of 0.5.

Unfortunately, with only a few years of forecasts, it is not viable to construct meaningful reliability diagrams or calculate these component scores. Until a large history of forecasts becomes available, the diagrams will have to be constructed by pooling forecasts for many or all seasons and locations. The hit score, sometimes called the Heidke score, is an easy-to-understand measure that could be used when sample sizes are small. This score measures whether the forecasts assign highest, as opposed to high, probabilities to the verifying outcomes. The calculation of similar hit scores, but for the categories with second- and third-highest probabilities, is recommended in place of measuring the number of times the observed

category is one or two categories distant from that with the highest probability. Measuring this so-called “distance” encourages forecasters to hedge to the normal category, and so is undesirable. It is not clear which specific attribute the hit scores, although it is perhaps closest to measuring resolution. Other related problems associated with the score are that it ignores a lot of information in the forecasts, and also that it possibly promotes an overly simplistic interpretation of them.

Instead of trying to measure resolution, various discrimination scores are recommended. Discrimination can generally be measured more accurately with small samples than can resolution. The area beneath the relative operating characteristics (ROC) curve is a measure of discrimination that is already widely used and is part of SVSLRF. Construction of ROC curves is recommended for more detailed diagnostics of the ability of the forecasts to discriminate observations in each category. Separate ROC areas are calculated for each of the observational categories. A generalized version of the ROC area is suggested as a single score for assessing discrimination over all categories. This score measures the ability of the forecasts to discriminate between the wetter, or warmer, of two observed outcomes.

If there is interest in mapping the geographical distribution of forecast skill, it is necessary to use scores for individual locations, and therefore ones that can be calculated using a small sample of forecasts. Alternatively, if there is interest in comparing forecast skill for different periods, such as for different seasons or for different years, scores that can be calculated with a small sample of forecasts are again needed. The difficulties of measuring resolution and reliability were mentioned above, but these sample size problems can also complicate the interpretation of discrimination if there are differences in the relative frequencies of the observed categories between subsets of the forecasts. For example, discrimination scores may fail to give credit (or penalize) for successfully (unsuccessfully) predicting a shift toward a warmer or drier period. In such cases, scores that measure multiple attributes may be preferable. The ignorance score is recommended in place of the more commonly used Brier skill score and the ranked probability skill score (RPSS) because they lack propriety when used for typical RCOF forecasts, and for the reasons mentioned above. For communication purposes, the ignorance score could be converted to the effective interest rate and interpreted as a simple measure of forecast value.

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. There are not actually any particularly good options for measuring the quality of a specific map; hit scores could be used, although the best option is probably the ignorance score. The ignorance score measures how much additional information is needed to determine what were the actual categories, given the forecasts. For communication to non-specialists, the average interest rate is suggested, but since it can be hedged it should not be used for any formal monitoring purposes. Measures of discrimination, such as ROC, are not recommended (except possibly for forecasts over the entire globe) because of their 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. Therefore, there 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 they do not provide direct answers to the question of interest, namely, how good are the forecasts? 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.

## LIST OF RECOMMENDED PROCEDURES

### Series of forecasts

Table 1 lists the recommended verification scores and procedures for series of forecasts, together with the attributes (see section 4.2) 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 fourth column in Table 1 indicates whether the score has to be calculated on each forecast category (below, normal and above) separately, or can be calculated over all categories. The table also indicates whether or not it is viable to apply the verification procedure for each station, gridbox or 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 Wilks (2011) and Jolliffe and Stephenson (2012). Additional scores that are suggested but not specifically recommended are provided in section 4.2.

**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**

<i>Score or procedure</i>	<i>Attributes</i>	<i>Questions addressed</i>	<i>By category?</i>	<i>By location?</i>	<i>Part of SVSLRF?</i>	<i>Key references</i>
ROC graph*	Discrimination	Do the forecasts distinguish an event from a non-event?	Yes	Yes	Yes	Harvey et al. (1992)
ROC area*	Discrimination	Is the forecast probability higher when an event occurs compared to when it does not occur?	Yes	Yes	Yes	Hogan and Mason (2012)
Generalized discrimination	Discrimination	Do the forecasts distinguish higher categories from lower?	No	Yes	No	Mason and Weigel (2009)
Resolution score	Resolution	Does an event become more or less likely when the forecast probability changes?	Yes	No	No	Weijs et al. (2010)
Reliability score	Reliability	Does an event occur as frequently as implied by the forecasts?	Yes	No	No	Weijs et al. (2010)
Effective interest rate	Multiple	What is the rate of return if paid fair odds when investing on a forecast?	No	Yes	No	Hagedorn and Smith (2008)
Average profits graphs	Multiple	What is the average profit or loss if paid fair odds when investing on the forecasts?	No	Yes	No	Hagedorn and Smith (2008)

<i>Score or procedure</i>	<i>Attributes</i>	<i>Questions addressed</i>	<i>By category?</i>	<i>By location?</i>	<i>Part of SVSLRF?</i>	<i>Key references</i>
Accumulated profits graphs	Multiple	What is the accumulated profit or loss if paid fair odds when investing on the forecasts?	No	Yes	No	Hagedorn and Smith (2008)
Ignorance score*	Multiple	Given the forecast, how much additional information is needed to determine what the verifying categories were?	No	Yes	No	Roulston and Smith (2002)
Reliability diagrams*	Reliability, resolution, sharpness	How does the relative frequency of occurrence of an event depend on the forecast probability? How frequently are different probabilities issued?	Yes and no	No	Yes	Hsu and Murphy (1986)
Tendency diagrams	Unconditional bias	How has the verification period differed from the climatological period? Are the forecast probabilities systematically too high or low?	Yes	Yes	No	Mason (2012)
Slope of reliability curve	Resolution, conditional bias	Does an event become more (less) likely as its forecast probability increases (decreases)?	Yes and no	No	No	Wilks and Murphy (1998)

\* Indicates procedures that are considered to comprise a minimal set that all operational forecasting centres should strive to calculate.

### Individual forecast maps

Table 2 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 4.3.

**Table 2. List of recommended scores and procedures for individual forecast maps**

<i>Score</i>	<i>Questions addressed</i>	<i>Key references</i>
Verification maps as percentiles*	What was the verifying category? How extreme was the observed value?	
Model diagnostics	Various; in general: Did the model reproduce the observed atmospheric conditions responsible for the verifying climate anomalies?	
Hit scores for categories with highest probabilities*	How often did the category with the highest probability occur?	Mason (2012)

<i>Score</i>	<i>Questions addressed</i>	<i>Key references</i>
Hit scores for categories with second and third highest probabilities	How often did the category with the second highest probability occur? How often did the category with the lowest probability occur?	Mason (2012)
Average interest rate	What is the rate of return if paid fair odds when investing on the forecasts?	Hagedorn and Smith (2008)
Ignorance score*	Given the forecast, how much additional information is needed to determine what the verifying categories were?	Roulston and Smith (2002)

- \* Indicates procedures that are considered to comprise a minimal set that all operational forecasting centres should strive to calculate.



## 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 her/his forecasts are worth taking note of. However, the question: “Are these forecasts good?” does not usually have a simple “yes” or “no” answer for a number of reasons. First, forecast quality is not a simple binary quantity. It is perfectly reasonable to ask how good (or bad) the forecasts are rather than just whether they are good or bad. Second, forecast quality is not a simple univariate quantity. Forecasts can be “good” in a number of different ways, and so forecast A may be better than forecast 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 publication is to describe and recommend procedures for the verification of forecasts from the Regional Climate Outlook Forums (RCOFs), and of similar forecast products. More generally, the recommendations are relevant to the verification of forecasts presented as probabilities of two or more ordinal, mutually exhaustive categories. In this publication, these forecasts are described as “probabilistic forecasts”, and are distinguished from “deterministic forecasts”, which are forecasts of specific values with no indication of uncertainty. In most cases the forecasts to be verified are likely to be for three climatologically equiprobable categories (“below normal”, “normal”, and “above normal”), but the recommendations herein are applicable in more general contexts, except where indicated. 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 promote understanding of the verification results.

This guidance publication was prepared under the auspices of the WMO Commission for Climatology XIV Expert Team 3.2 on Climate Information and Prediction Services Operations, Verification and Application Services. The recommendations herein build upon the WMO Commission for Basic Systems Standardized Verification System for Long-range Forecasts (SVSLRF), which is specifically for verification of Global Producing Centre (GPC) products that are used as inputs to seasonal forecasting processes, including at RCOFs.

The SVSLRF defines standards for verifying model outputs from GPCs and so includes procedures for measuring the quality of ensemble prediction systems. The target audiences of this verification information are model developers and the immediate intended users of these products, namely forecasters from national and regional centres. In contrast, the procedures defined in the present publication are targeted partly at the forecasters, but also partly at users of the forecasts who may have no technical background in forecasting or forecast quality.

Another difference between the procedures presented in the present publication and SVSLRF is that here the specific products to be verified are series of forecasts released operationally, and so some of the difficult questions addressed by SVSLRF pertaining to the generation of hindcasts and issues of cross-validation are avoided. In addition, the limiting case of how to verify a forecast for a single period is addressed in the present guidance, which is not a concern in SVSLRF. Even when measuring the quality of all available forecasts, 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 SVSLRF are therefore recommended to indicate the estimated uncertainty in the results.

Before recommending specific procedures, some general issues pertaining to difficulties in verifying RCOF-type forecasts are discussed in Chapter 2. These issues include problems with the definition and calculation of the target variable when forecasts are presented as regions, and some potential problems with the observational data. Following this, the question of what constitutes a “good” forecast is considered in Chapter 3. Specifically, the important qualities (formally known as “attributes”) of good probabilistic forecasts are described, and

these descriptions form the basis for identifying which attributes it is most useful to measure to assess the quality of the forecasts. Using the principles established in Chapter 3, a set of recommended verification procedures is then defined, distinguishing methods used to measure a set of forecasts (section 4.2), and those for measuring the quality of a specific forecast (that is, the forecasts for one target period; section 4.3). Separate procedures are recommended for use by forecasters, and for communication to non-specialists. In Chapter 5 some procedures for indicating uncertainty in the estimates of forecast quality are considered. For all the procedures, consideration needs to be given for possible differences in the sizes of areas for which the forecasts apply. Weighting of the forecasts to account for these differences is discussed in Appendix A. Worked examples of the scores and procedures are included in Appendix B so that the step-by-step calculations can be followed. A glossary of technical terms is provided in Appendix C.

---



## 2. FORECAST AND VERIFICATION DATA

Before discussing some of the mathematical attributes of good forecasts (Chapter 3), it is necessary to consider an attribute that is more qualitative in nature: is it clear what the forecast means? To verify a forecast, it is essential to be unequivocal about exactly what is being forecast. Defining exactly what is the target variable – the predictand – is non-trivial, especially with regard to the way in which forecasts are often constructed in consensus-building approaches such as at RCOFs. This chapter discusses some problems associated with defining the target variable and arising from the nature of the forecasts and of the observational data.

### 2.1 DEFINING THE TARGET VARIABLE

Seasonal forecasts are typically 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 the forecasts 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 spatially averaged. For example, consider the idealized example shown in Figure 1, which represents 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, there were originally forecasts for two regions, whose areas are delimited by the dashed line, but these regions were combined because their 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 meaning of the forecast for region III is now not clear or has changed. The probability that the seasonal rainfall total averaged over region III will be above normal is not the same as the probabilities that the seasonal rainfall total averaged over subregion IIIa and that averaged over sub-region IIIb will each be above normal.

Why is this difference in interpretation important? Imagine a situation in which observed rainfall over subregion IIIa is above normal and over region IIIb it is below normal, and that the spatial average over the whole of region III is then normal. The original forecasts would score poorly: in subregion IIIa the category with the lowest probability (25%) occurred, and in subregion IIIb the category with the second highest probability (35%) occurred. But if the forecast is reinterpreted to refer to a regional average for the whole of region III, it would be scored as if the category with the highest probability (40%) occurred (normal), and so it would score reasonably well. Thus, by reinterpreting the forecast over region III to refer to the spatial average over the entire region, the forecasts verify as better than they really are in this instance. Combining these subregions changes the meaning of the forecast, which can have a major impact on the scoring. It is more

I	II
A 50%	A 20%
N 30%	N 35%
B 20%	B 45%
IIIa	IIIb
A 25%	
N 40%	
B 35%	

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

appropriate to verify the forecasts with the subregions as two separate regions even if the forecasts are the same. Unfortunately, in many cases the original subregions are not indicated on consensus maps, and so it may be impossible to identify where they are or how many there are.

This ambiguity in 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 GPCs, for example, or more recently by the South Asian Climate Outlook Forum), or indicating the subregions on the map as thin lines, or forecasting for stations rather than for regional indices. Forecasting for predefined homogeneous zones is another option. Until this problem is addressed, an interpretation has to be imposed upon the forecasts if they are to be verified. The fact that this interpretation may not be what was originally meant should highlight to the forecasters the importance of eliminating ambiguity in their products. Recommending a specific solution to eliminate the ambiguity at the forecast production stage is beyond the scope of this document.

## 2.2 SIZES OF FORECAST REGIONS

Figure 2 illustrates a simple example in which there are only two regions, one larger than the other. The reader is asked to imagine that above normal occurs over regions I and II. Forecaster A issued a 70% probability on the verifying category for the larger region, and a 10% probability for the verifying category for the smaller, whereas forecaster B issued a 70% probability on the verifying category for the smaller region and a 10% for the larger. Although both forecasters had one region with the highest-probability category occurring and one with the lowest-probability category occurring, surely forecaster A should get greater credit than forecaster B for having the category with the highest probability occurring over the largest area.

To address such problems, verification results should be weighted by area when the forecasts are for gridboxes or other areas. If the forecasts are for stations, then weighting the forecasts may not be appropriate unless some of the stations are clustered in space, in which case there may be some reasons for weighting the forecasts for these stations less than forecasts for more isolated stations. Whether weighting should be applied depends on whether one wants to measure forecast quality by the number of locations (in which case weighting should not be applied) or by the proportional area (in which case a weighting scheme would be desirable). The point is discussed further in section 2.4 and more details on weighting the results by area are provided in Appendix A.

Forecaster A

Region I	A 70%
	N 20%
	B 10%
Region II	A 10%
	N 20%
	B 70%

Forecaster B

Region I	A 10%
	N 20%
	B 70%
Region II	A 70%
	N 20%
	B 10%

Figure 2. Idealized example of two seasonal rainfall forecasts for two regions of different sizes

### 2.3 GRIDDING OF FORECASTS

One way to address the problems of subregions highlighted in section 2.1 is to grid the forecasts. Although gridding involves defining a set of subregions and assuming that the forecasts are valid at this new scale,<sup>1</sup> 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 preferably should be gridded/interpolated to the resolution/locations of the verification data rather than vice versa. Exceptions to this principle occur if the resolution of the verification data is substantially different from the spatial scale at which the forecasts were meant to be interpreted. There are two possibilities:

- If the verification data are of a much finer resolution than the forecasts, they may need to be aggregated to a more compatible resolution first;
- If the verification data are on a coarser resolution than the forecast, an attempt to interpolate the observations to the finer resolution may be desirable, although it must be recognized that any spatial detail in the forecasts cannot then adequately be verified.

In gridding the forecasts, some of the grids will inevitably span two or more forecast regions, and the question of what value to put down for the forecasts at such gridboxes must 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 climatological 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 the forecast representing the largest part of the gridbox should also 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 weighting scheme should be applied in these instances. Each gridbox should be weighted by its area relative to the other gridboxes. Further details on weighting the results by latitude are provided in Appendix A.

When interpolating the forecasts to station locations, it is recommended to use the following procedures, depending on the format of the forecasts:

- If the forecasts are gridded, use the forecast for the gridbox in which the station occurs;
- If the forecasts are for specific locations, use the forecast for the nearest location.

Note that these recommendations apply only if the forecasts are probabilities. More appropriate interpolation procedures may exist for other types of forecasts.

### 2.4 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 sparser areas. This weighting can be performed by constructing Thyssen polygons,<sup>2</sup> 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

<sup>1</sup> This assumption may well be invalid, but some interpretation has to be imposed on the forecast, and it is a reasonable compromise midway between treating the forecasts as representing area averages, and as representing individual locations with the same probabilities.

<sup>2</sup> 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 grid points are closest to each of the stations.

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 threshold distances in the verification analysis. No absolute value for the threshold can be recommended because of regional and seasonal 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.

## 2.5 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 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:

- (a) Calculate the mean and standard deviation of the data from the climatological dataset for the period of overlap;
- (b) 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 (a);
- (c) Calculate the mean and standard deviation of the data from the verification dataset for the period of overlap;
- (d) Multiply the results of step (b) by the standard deviation and add the mean calculated in step (c);
- (e) 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;
- (f) 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:

- (a) Calculate the parameters of a gamma distribution fitted to the climatological data for the period of overlap;
- (b) 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 (a);
- (c) Calculate the parameters of a gamma distribution fitted to the verification data for the period of overlap;
- (d) Transform the quantiles from step (b) to deviates using the gamma distribution parameters for the verification data obtained from step (c);

- (e) 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;
- (f) 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 (f) in the two algorithms above need be repeated. The aims of steps (a)–(e) 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

#### 3.1 TYPES OF FORECAST “GOODNESS”

Forecasts can be described as “good” in three different senses (Murphy, 1993):

- A forecast is “consistent” if it is a true indication of what the forecaster thinks is going to happen. In the context of probabilistic forecasts, a forecast is consistent if the probabilities are a true reflection of the forecaster’s uncertainty of the outcome. For example, 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. 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 given information that the forecaster knows is inaccurate. For example, if the forecaster thinks 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 underestimate 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”, as distinct from “strictly 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 forecast has “quality” if it corresponds to what happened. This aspect is typically measured by some form of mathematical relationship between the forecasts and the respective observations; forecasts that correspond well to 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, almost invariably the first question that is 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 one 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 more appropriate for describing the quality of probabilistic forecasts. These attributes are discussed in detail in this chapter, and procedures for measuring them are described in Chapter 4.
- A forecast has “value” if it can be used to help realize some benefit, whether economic, 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. One may consider, for example, a set of excellent forecasts that are released in a timely manner, and which inform profitable decision-making when the forecasts correspond well with the observed

outcomes. It is still possible for these forecasts to have no value if the costs incurred from the occasional “bad” forecast more than offset the benefits from the frequent “good” ones. Very good forecasts can therefore have no, or even negative, value. Conversely, forecasts with low quality can be of immense value if the occasional “good” forecast can be used to good effect. However, it is not possible to realize any benefit from forecasts that have no correspondence with the observations.<sup>1</sup>

In this publication, recommendations are made primarily for verification procedures for measuring forecast quality rather than value. However, 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.

### 3.2 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, she/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. This claim that probabilistic forecasts are always “correct” is only true in the sense that they indicated that the observed outcomes could have happened. The question of correctness of probabilistic forecasts is then so uninformative as to be essentially 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. 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 verifies. 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.<sup>2</sup> While it is not unreasonable to ask how often the category with the highest probability verifies, there are some interrelated and important problems related to such approaches. First, 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”). Second, if the probabilities are to be considered as at all meaningful, a high hit score may be considered unfavourable because it could indicate that the forecasts are unreliable. If the highest probability is 40%, one would want a “hit” only 40% (that is, less than half) of the time. Third, 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

<sup>1</sup> 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 a response to the climate, in which case forecasts that are otherwise useless may have some value. Other exceptions have to be made when the mere consideration of forecasts brings about improvements in what would be good practices anyway. For example, there may have been very limited contingency planning before forecasts were available.

<sup>2</sup> Knowledge of the base rate is necessary because the naïve expectation is that at least 50% of the forecasts should be correct. However, scores of less than 50% correct may be good if there are three or more categories, for example, while scores of greater than 50% may still be bad if any one category occurs most of the time anyway.



this scoring system does not explicitly penalize for two-category errors,<sup>3</sup> 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 most likely to be one contributor.

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, some probabilistic verification scores have undesirable properties. In addition, it is not always clear what specific attributes of forecast quality some other scores are measuring, so they can be difficult to interpret. Therefore, before recommending a set of scores, it is first necessary to define the most important attributes of good probabilistic forecasts.

### 3.2.1 **Attributes of “good” probabilistic forecasts**

#### 3.2.1.1 ***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 serve no purpose.<sup>4</sup> For probabilistic forecasts, above-normal rainfall, for example, should occur more frequently when its 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, she/he is communicating much more confidence that above-normal rainfall will occur than when she/he says there is a 20% chance. 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 3.1).<sup>5</sup> 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 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.

#### 3.2.1.2 ***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

<sup>3</sup> At some of the RCOFs, the numbers of one- and two-category errors are reported separately, which is likely to encourage hedging.

<sup>4</sup> One can imagine a hypothetical 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 attribute.

<sup>5</sup> Not to be confused with spatial or temporal resolution.



different outcomes.<sup>6</sup> Whereas resolution is concerned with whether the expected outcome differs as the forecast changes, “discrimination” is concerned with whether the forecast differs given different outcomes. 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,<sup>7</sup> 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.

### 3.2.1.3 **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. If the forecast is consistent (section 3.1), probabilistic forecasts could be interpreted as the forecaster’s expectation that a deterministic forecast of each category will be “correct”. For example, assume that a forecaster indicates probabilities of 40% for below normal, 35% for normal, and 25% for above normal. If this forecast is then interpreted as a deterministic forecast of below normal, the forecaster thinks there is a 40% chance that the deterministic forecast will be correct. Similarly, if someone were to take the same forecast as an indication of above normal, the forecaster believes there is a 25% probability she/he will be correct. Forecasts are reliable, or well calibrated, if that interpretation is correct, or equivalently if that category occurs, as frequently as the forecast implies (Murphy, 1993).

More often than not, seasonal forecasts are unreliable. The commonest situation is that the forecasts are overconfident – increases and decreases in probability are too large. Overconfidence occurs when the forecaster thinks that the probability of a specific category is increased (or decreased), but overestimates 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, she/he has correctly indicated increases and decreases in the probability, but overstated these changes, and thus was overconfident. In relatively rare cases, the forecasts may be under-confident, in which cases the increases and decreases in the probabilities are too small. Overconfidence and under-confidence are examples of conditional biases – the errors in the forecasts depend on whether the forecasts indicate increased or decreased probabilities. For overconfident 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

<sup>6</sup> As with resolution, one can imagine an abnormal 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.

<sup>7</sup> 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. An example may be 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 degrees centigrade (or on some other continuous scale), the bias could be removed, and the verification results recalculated. Some of the forecasts that previously exceeded the threshold will no longer do so, and the resolution and 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.

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

#### 3.2.1.4 **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 overconfident. In many RCOF forecasts, for example, the highest probability on any of the categories is often 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. For example, a 40% chance of above normal still implies a 60% chance of normal or below normal. RCOF forecasts consistently 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 overconfident, 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 high, there is high confidence that that category will verify; if the probability is low, there is high confidence that 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.

#### 3.2.1.5 **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 skilful compared to a second set of forecasts, but unskilful compared to a third set. 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).<sup>8</sup> Of course, it is not necessarily the case that the user knows

<sup>8</sup> Another widely used reference strategy is that of “persistence” – to assume that the most recently observed climate anomalies will continue into the target period. However, a persistence forecast cannot be expressed in probabilistic terms as simply as can climatological forecasts without invoking a relatively complex statistical model.

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. Regardless of what reference strategy is used, it is important to communicate carefully what exactly the comparison reveals about the quality of the forecasts.

### 3.2.2 **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 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 would 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 locations indicated a 10% probability of above-normal rainfall, and that above-normal rainfall occurred at 2 of the 10 (20%) locations. The temptation may be to conclude that these forecasts were overconfident 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. Instead, the forecasts were indicating only that above-normal rainfall could be expected to occur at each location on 10% of the occasions on which a 10% probability of above-normal rainfall is issued. It may well be that the one-in-ten occasion on which above-normal rainfall occurs is the same for all the locations. To try to 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 subregions have been combined (see section 2.4), 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 not think above-normal rainfall is expected over half the region, and yet is unsure of which half. Instead she/he may think that there is an equal chance that either all or none of the region will be above normal.<sup>9</sup> 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 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 quality over time, or to see how forecast quality changes with the El Niño/Southern Oscillation 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 with 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 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

<sup>9</sup> 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.

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 GPCs). 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 publication; 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 overconfidence. The multifaceted quality of a set of forecasts (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 has good reliability as well as high resolution, is well discriminated, and has high skill when measured against appropriate reference 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 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 publication.

How the various attributes of good forecasts can best be measured depends to a large extent on the sample size of the forecasts. In section 4.2, procedures are recommended for measuring these attributes given a set of forecasts for different target periods. 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 4.3, 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.

### 4.1 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 be either 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 non-climatological forecasts are not particularly useful. Therefore, 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:

- She/he does not know what is going to happen; or
- She/he thinks 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 she/he thinks that the three possible outcomes are genuinely equally likely to occur, whereas in the former case the forecaster would prefer simply to state that she/he does not have any useful information at all. The latter statement could be considered as a well-informed interpretation of all relevant information concluding that there is no basis for expecting any of the possible outcomes to be more or less likely than they are climatologically. 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 want to make a statement.

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 those procedures that explicitly require binning. These procedures include reliability diagrams (section 4.2.5) and resolution (section 4.2.3) and reliability (section 4.2.4) scores. Even in these cases it is recommended that the average of the forecast probabilities for each bin be used in place of the central probability of the bin. For all other procedures the climatological probabilities should be analysed as a separate probability value where appropriate.

## 4.2 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 whether 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 following subsections describe some 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 communication to the general public are described first, as well as 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 the forecasts can 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 overconfident 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 summarize this information for non-experts are provided.

### 4.2.1 Measuring discrimination

Resolution and discrimination are apparently very similar attributes of good probabilistic forecasts. However, 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. For discrimination we have to compare the forecasts for only three possible outcomes (each category), rather than having to compare the outcomes for each of the multitude of different forecast probabilities that may have been issued.

Given that resolution and discrimination are the most fundamental attributes of good probabilistic forecasts, but that discrimination is the easier to measure with small samples, it is this attribute that is the most logical 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) are therefore logical recommendations.

#### 4.2.1.1 *Relative operating characteristics*

The ROC can be used in forecast verification to measure the ability of the forecasts to distinguish an event from a non-event. For seasonal forecasts with three or more categories, the first problem is to define the “event”. One of the categories must be 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 and no distinction is made as to which of these two categories does occur. So, for example, if below normal is selected as the event, normal and above normal are treated equally as non-events. Given this requirement for a binary definition of outcomes, separate ROC graphs can be completed for each category. A measure



of discrimination can then be defined to indicate how well the forecasts can distinguish the selected category from the other two categories. Setting  $p_{1,j}$  as the forecast probability for the  $j$ th observed event, and  $p_{0,i}$  as the forecast probability of an event for the  $i$ th non-event, the ROC score,  $A$ , can be calculated using:

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

where  $n_0$  is the number of non-events, and  $n_1$  the number of events, and the scoring rule  $I(p_{0,i}, p_{1,j})$  is defined as:

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

If the category of interest is below normal, the score indicates the probability of successfully discriminating below-normal observations from normal and above-normal observations. It indicates how often the forecast probability for below normal is higher when below normal actually does occur compared to when either normal or above normal occurs. For example, a score of 0.7 indicates that 70% of forecasts have higher probabilities on below normal when below normal occurs compared to when below normal does not occur. The score has an intuitive scaling that is appealing to many non-specialists who might expect at least 50% of the forecasts to be “correct” – it has an expected value of 0.5 for useless forecast strategies (guessing, or always forecasting the same probabilities), and good forecasts will have a score greater than 0.5, reaching 1.0 given perfect discrimination. Scores of less than 0.5 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 0.0 given perfectly bad discrimination. Given this simple interpretation, ROC scores may be suitable for communication to non-specialists, although the name of the score is perhaps unnecessarily intimidating.

The ROC score is equivalent to the Mann-Whitney U-statistic following some simple rescaling, and this equivalency may help to understand 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 most of the forecast probabilities were higher when an event occurred compared to when not.

The score is only one component of this verification procedure. The ROC graph is a useful diagnostic that provides more detailed information than can a single score, as well as a possibly more convenient way to calculate the score. The ROC graph is constructed by calculating the ability of the forecasts to successfully identify the events. The graph can be constructed by starting with the forecasts with highest probabilities. These forecasts should point to the observations that we are most confident are events. If the selected observations are events they are called “hits”. The proportion of all events thus selected is calculated, and is known as the hit rate (HR), or probability of detection:

$$HR = \frac{\text{Number of hits}}{\text{Number of events}} \quad (2)$$

It is possible that some non-events have been selected incorrectly; these are known as “false alarms”. The proportion of non-events incorrectly selected (the false-alarm rate (FAR)) is:

$$FAR = \frac{\text{Number of false - alarms}}{\text{Number of non - events}} \quad (3)$$

The HR and FAR are commonly tabled (see example in Appendix B section B.1.1), 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 HR and FAR will be identical, but if the forecasts can discriminate the events, the HR will be larger than the FAR. Since it is unlikely that all the events were correctly selected using only the forecasts with highest probabilities, additional selections can be made using the next highest probability and the HR and FAR updated accordingly. The difference in the increments of the HR and FAR is expected to be a little less than at the first step, since there is less confidence about having correctly selected the events. These steps are continued until all the events have been selected. The HR are then plotted against the FAR. See the example in Appendix B section B.1.1 for more details.

The area,  $A$ , beneath the curve can be calculated using the trapezoidal rule:

$$A = 0.5 \times \left[ 1 + \sum_{k=0}^d (y_k x_{k+1} - y_{k+1} x_k) \right] \quad (4)$$

where  $d$  is the number of discrete probability values, and  $y_1$  and  $x_1$  are the HR and FAR, respectively, for the highest probability value only,  $y_2$  and  $x_2$  are the rates for the highest and second highest probabilities, and so on. For  $i = 0$  the HR and FAR 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. There are other ways to calculate the area beneath the curve, but the use of the trapezoidal rule is recommended rather than using the normal distribution assumption, which gives a smoother fit. The normal distribution assumption is recommended in SVSLRF, but the trapezoidal rule is recommended here for the following reasons:

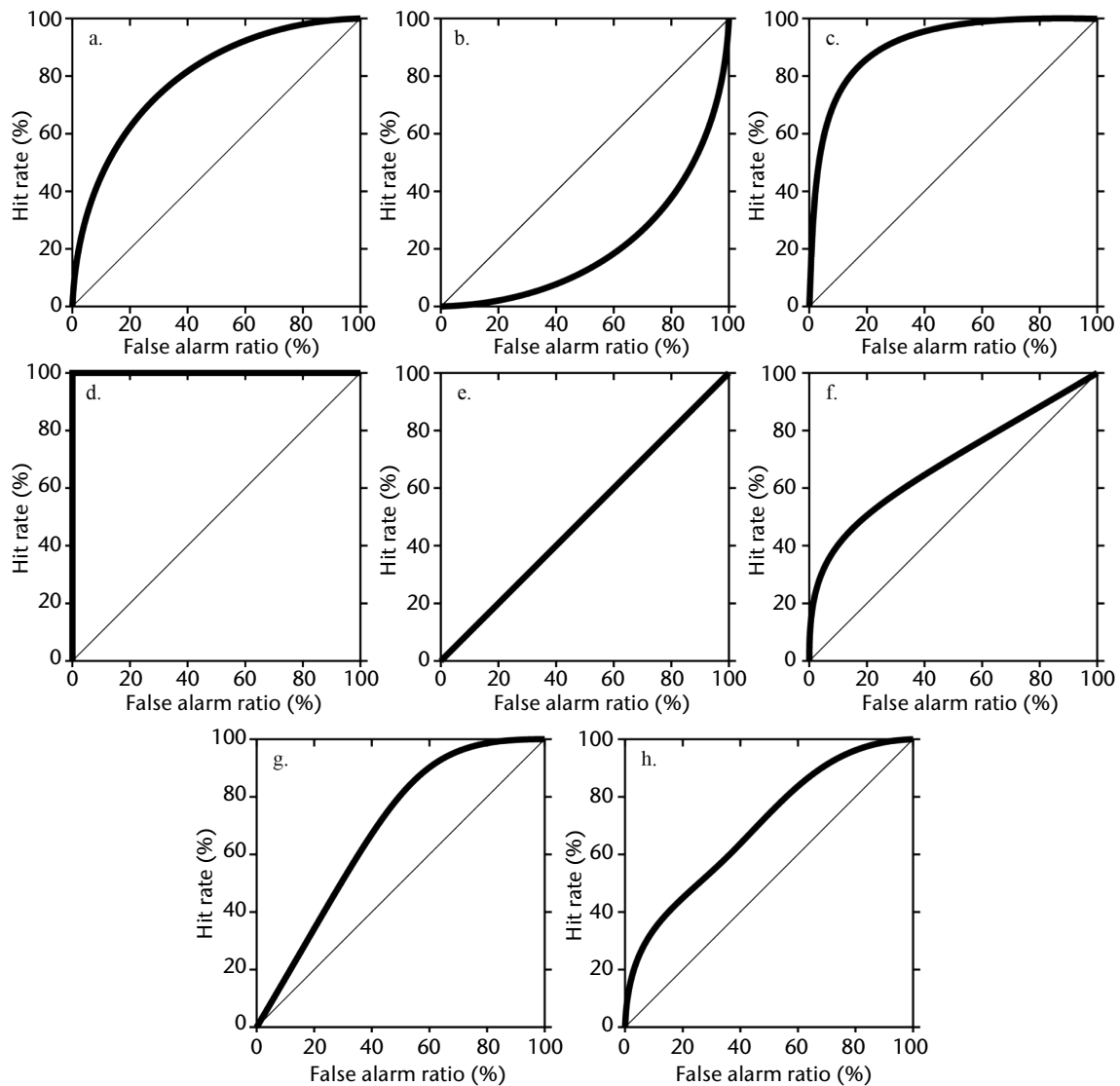
- The trapezoidal area has an intuitive interpretation – it defines the probability of successfully discriminating observations in the respective category;
- A smoothed fit is suitable for convex or concave ROC curves, which are commonly observed with shorter-range forecasts, but curves for seasonal forecasts can take a variety of different shapes;
- A 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%.

When the area is calculated using equation (4) it is equivalent to equation (1). In fact, the ROC score is often referred to as the ROC area.

It is recommended that the graph be constructed at the resolution of the forecasts, which would typically be 5%, plus an additional point for climatological probabilities. If the forecast probabilities are available at higher precision than 5% then they should be verified at this precision unless the scores are to be compared with those from forecast systems with lower precision.

The ROC curves can provide some useful diagnostics about the quality of the forecasts; some guidelines for interpreting the curves are provided by considering the idealized curves indicated in Figure 3. For good forecasts, it has just been argued that the HR will be initially much larger than the FAR, 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 3a). Conversely, if the forecasts are bad only relatively few events will be selected initially, and the curve will be concave (Figure 3b). 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 3c). If all the events are selected before any of the non-events, then the HR will reach 1.0 while the FAR is still zero, and so the area under the curve will be 1.0 (Figure 3d). Forecasts that used a guessing strategy, or one that is just as naïve as a guessing strategy, will score similar HRs and FARs, and so the curve will follow the 45° diagonal and enclose an area of 0.5 (Figure 3e). 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





**Figure 3. 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**

same time since there is no basis for selecting some over others. The HRs and FARs will then both be 1.0 if all the observations are selected, and both be 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.

More complex shapes are frequently observed for seasonal forecasts. For example, 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 3f). 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 3g). 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 3h).

The ROC areas and graphs can be calculated using forecasts pooled from all, or a regional selection of, locations. They can be calculated for individual locations and the areas mapped, although the areas are undefined if there are no occurrences of the event or of the non-event. If the verification period is very short, it is quite possible for at least one of the categories to be unobserved.

From section 3.2.2 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 primarily a sampling issue, and so when forecasts are pooled spatially and temporally the interpretation problems are diminished. For this reason, spatial pooling of forecasts is widely practised and is recommended for some of the verification procedures in CBS SVSLRF.

The graph and area are insensitive to monotonic transformations of the probabilities. There are some important implications of this insensitivity:

- The ROC will generally ignore problems such as unconditional biases – it is measuring discrimination, and discrimination alone;
- However, if there are different unconditional biases at different locations, these differences will not be ignored if the forecasts are pooled, and the score for the pooled forecasts may be quite different from a simple average of the scores for the individual locations;
- Successful forecasts of differences in climate between the verification and the climatological periods may be ignored;
- The procedure is inappropriate for measuring the quality of forecasts for a single period (see section 4.3).

In theory, the definition of the event and non-event for the ROC could be generalized so that a single curve and area could be calculated. However, 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 assessed by performing ROC on each category separately.

#### 4.2.1.2 **Generalized discrimination**

If there is need for a single measure of discrimination, the generalized discrimination score could be used (Mason and Weigel, 2009). This score provides an indication of the ability of the forecasts to discriminate wetter (or warmer) observations from drier (or cooler) ones, and can be used when there are two or more categories. In the special case of two categories, the score simplifies to the ROC score defined in equation (1), when there are three or more categories, the score can be viewed as a multi-category version of the area beneath the ROC graph. The generalized discrimination score,  $D$ , is defined as:

$$D = \frac{\sum_{k=1}^{m-1} \sum_{l=k+1}^m \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} I(\mathbf{p}_{k,i}, \mathbf{p}_{l,j})}{\sum_{k=1}^{m-1} \sum_{l=k+1}^m n_k n_l} \quad (5a)$$

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:

$$I(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) = \begin{cases} 0.0 & \text{if } F(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) < 0.5 \\ 0.5 & \text{if } F(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) = 0.5 \\ 1.0 & \text{if } F(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) > 0.5 \end{cases} \quad (5b)$$

and where:

$$F(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) = \frac{\sum_{r=1}^{m-1} \sum_{s=r+1}^m p_{k,i}(r) p_{l,j}(s)}{1 - \sum_{r=1}^m p_{k,i}(r) p_{l,j}(r)} \quad (5c)$$

where  $p_{k,i}(r)$  is the forecast probability for the  $r$ th category, and for the  $i$ th observation in category  $k$ . (If the forecast probabilities are derived directly from an ensemble counting procedure, it may be desirable to calculate the score using the method described by Weigel and Mason (2011).)

Although equation (5) may seem complicated, its interpretation is fairly simple – what is the probability of successfully discriminating the wetter (or warmer) of two observations? Equation (5a) 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 equation (5c), 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 (equation (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  $\mathbf{p}_{k,i}$ ). Its scaling is very similar to that for the ROC area – a score of 0.5 represents no skill, 1.0 indicates perfect discrimination, and 0.0 indicates perfectly bad discrimination.

As with the ROC areas, 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. However, the same potential pitfalls in interpreting and comparing ROC areas when sample sizes are small and/or forecasts are pooled also apply to the generalized discrimination score. It is possible to pool forecasts for all locations and/or seasons to increase the sample size. Note that this pooling does not normally result in a simple average of the score for the individual locations or seasons; calibration problems would be ignored when the score is calculated at individual locations or seasons, but such problems may be at least partially 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 scores for the non-pooled forecasts.

An example of the calculation of the generalized discrimination score is provided in Appendix B section B.1.2. Practitioners interested in a single summary score may wish to consider the generalized discrimination score rather than the more commonly used ranked probability skill score (RPSS). 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. 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 so overconfident forecasts frequently score badly on the RPSS.<sup>1</sup> By trying

<sup>1</sup> Some attempts have been made to correct these so-called biases in the RPSS 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 for probabilistic forecasts derived by other means. These de-biased scores are considered useful in the context of the CBS SVSLRF, which targets GPC products, but cannot be applied in the current context. In addition, the criticism remains that these scores are still abstract numbers measuring multiple attributes, and so are difficult to understand by all but specialists in forecast verification.

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.

#### 4.2.2 Measuring resolution

As defined in section 3.2.1.1, resolution measures differences in the outcomes given different forecasts.

##### 4.2.2.1 Resolution components of multi-attribute scores

A commonly used measure of resolution is the corresponding component of the Murphy (1973) decomposition of the Brier score (BS):

$$Resolution = \frac{1}{n} \sum_{k=1}^d n_k (\bar{y}_k - \bar{y})^2 \quad (6)$$

where  $n$  is the total number of forecasts,  $d$  is the number of discrete probability values (or number of probability bins),<sup>2</sup>  $n_k$  is the number of forecasts for the  $k$ th probability value,  $\bar{y}_k$  is the observed relative frequency for that value, and  $\bar{y}$  is the observed relative frequency for all forecasts. The observed relative frequency for the  $k$ th forecast probability,  $\bar{y}_k$ , is the number of times an event occurred divided by the number of times the respective probability value was forecast:

$$\bar{y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_{k,i} \quad (7)$$

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 equation (7) are often plotted on a reliability diagram (see section 4.2.5), and since it is based on the BS score separate scores can be obtained for each of the categories.

Equation (6) defines the variance of the observed relative frequencies of events 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.

An alternative measure of resolution is the Weijs et al. (2010) decomposition of the ignorance score. This score is defined as:

$$Resolution = \frac{1}{n} \sum_{k=1}^d n_k \left( \bar{y}_k \log_2 \left[ \frac{\bar{y}_k}{\bar{y}} \right] + (1 - \bar{y}_k) \log_2 \left[ \frac{(1 - \bar{y}_k)}{(1 - \bar{y})} \right] \right) \quad (8)$$

This version of the resolution score has a minimum value of 0.0 when there is no resolution and the forecasts are useless, and a maximum value of  $-\log_2 [\bar{y}(1 - \bar{y})]$ . In general, equation (8) is recommended over equation (6) because the way the ignorance score measures probability errors is preferable to that of the BS score, at least when climatological probabilities differ from 0.5, which is typically the case in seasonal forecasting. See section 4.2.4 for further discussion on this preference.

<sup>2</sup> 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 BS (Stephenson et al. 2008).

As equations (6) and (8) indicate, the forecast probabilities are not explicitly considered, and so the resolution scores do not specifically measure whether the event occurs more frequently as the forecast probability increases. Thus, although the resolution scores distinguish useless from potentially useful forecasts, they are unable to distinguish between forecasts that are useful because they are “good” (as defined in section 3.2.1.1) from those that are useful because they are so bad that one can be confident they are giving completely the wrong message. Nor can the score distinguish “good” forecasts from those for which the observed relative frequency increases and decreases randomly with changing forecast values!<sup>3</sup> For these reasons it is recommended that a skill score version of the resolution scores be interpreted with caution – positive skill could indicate either that the forecasts are better than the reference, but could equally indicate that the forecasts are worse than the reference forecasts are good.

One other problem with equations (6) and (8) is that they have 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. 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.

#### 4.2.2.2 *Hit scores as measures of resolution*

A variation on the idea of binning the forecasts and then calculating a resolution score is simply to calculate equation (7) for the bin with the highest probabilities, where the largest probability bin,  $d$ , is defined flexibly to include the highest probability for each forecast (hence there are  $n$  forecasts in this bin). This procedure amounts to calculating the hit (Heidke) score for the highest probability forecasts.<sup>4</sup> The hit score is derived from equation (7):

$$\bar{y}_d = \frac{1}{n} \sum_{i=1}^n y_{d,i} \quad (9a)$$

The hit score is usually expressed as a percentage, and can be more easily interpreted as:

$$\text{Hit score} = \frac{\text{Number of hits}}{\text{Number of forecasts}} \times 100\% \quad (9b)$$

where a “hit” is the occurrence of the category with the highest probability, (that is,  $y_{d,i} = 1$ ). 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, a half hit should be scored, 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). 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

<sup>3</sup> In the limiting case of  $d = n$  (the number of probability bins is equal to the number of forecasts), the resolution term takes on its maximum value of  $\bar{y}(1 - \bar{y})$  or  $-\log_2 [\bar{y}(1 - \bar{y})]$ , but communicates no useful information about the quality of the forecasts. This maximum value occurs regardless of the quality of the forecasts, as long as each forecast has a unique probability (which is a requirement for  $d$  to equal  $n$ ).

<sup>4</sup> In practice this score is likely to be very close to the value of  $\bar{y}_k$  for the forecasts with probabilities >35%, as suggested in the final paragraph of section 4.2.2.1. Differences would only occur in the event of highest forecast probabilities for two categories tying at 40% (or in the unlikely cases of ties at 35% or 45%).

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.

Following the arguments in section 3.2, 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 the hit score gives an extremely limited perspective on the resolution of the forecasts (as is evident from a comparison of equation (9a) with equations (6)–(8)). 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:

$$\bar{y}_j = \frac{1}{n} \sum_{i=1}^n y_{j,i} \quad (10)$$

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 equation (7): in equation (10)  $\bar{y}_j$  for  $j = 1$  is the hit score for the category with the highest probability, whereas in equation (7)  $\bar{y}_j$  for  $j = 1$  is the hit score for the lowest probability bin. An example is provided in Appendix B section B.1.3.

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. However, forecasts such as 30–40–30 in which the two outer categories have equal lowest probability are rarely given, and it is possible that the forecasts successfully point to the category least likely to verify even with 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 4.2.5.

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 equation (9) is therefore frequently calculated with the form:

$$\text{Hit skill score} = \frac{\text{Number of hits} - \text{expected number of hits}}{\text{Number of forecasts} - \text{expected number of hits}} \times 100\% \quad (11)$$

where the expected number of hits is calculated for the reference (no-resolution) forecasts.

This skill score is suitable in technical contexts, but equation (9) generally has a simpler meaning than equation (11), and so the former may be preferable for communication to non-specialists. Whenever equation (9) is used the expected number of hits should be indicated. The expected number of hits is usually calculated from a strategy with no resolution (for example, 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 equation (11), 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 3.2.1.1, 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 discussed 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.

#### 4.2.3 Measuring reliability

The ROC areas, generalized discrimination score and resolution score are all recommended because they measure the attributes of discrimination or resolution, which are considered essential measures of whether the forecasts may be potentially useful. However, in focusing only on these attributes the reliability of the forecasts is ignored. Reliability can be measured using the reliability components of the BS (Murphy, 1973) or of the ignorance score (Weijs et al., 2010). Given  $d$  discrete probability values:

$$Reliability = \frac{1}{n} \sum_{k=1}^d n_k (\bar{p}_k - \bar{y}_k)^2 \quad (12)$$

where  $n_k$  is the number of forecasts for the  $k$ th probability value ( $\bar{p}_k$ ), and  $\bar{y}_k$  is the observed relative frequency for that value (Murphy, 1973). The corresponding score from the decomposition of the ignorance score is:

$$Reliability = \frac{1}{n} \sum_{k=1}^d n_k \left( \bar{y}_k \log \left[ \frac{\bar{y}_k}{\bar{p}_k} \right] + (1 - \bar{y}_k) \log \left[ \frac{(1 - \bar{y}_k)}{(1 - \bar{p}_k)} \right] \right) \quad (13)$$

(Weijs et al., 2010).

As with resolution scores, reliability scores are a function of  $\bar{y}_k$ . As a result, reliability scores are also 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. These scores are therefore very noisy when calculated at individual locations. However, the scores can be useful for calculating the reliability of forecasts pooled from many of locations.

Reliability scores measure “errors” in the reliability of the forecasts, and so equations (12) and (13) have a minimum value of 0.0 for perfectly reliable forecasts. Equation (12) has 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). Equation (13) has no upper limit; an infinite penalty is assigned if a probability of 100% is assigned to an event that does not occur or, by implication, a probability of 0% to an event that does occur. Some forecasters consider this penalty unduly harsh, but probabilities of 0% and 100% indicate absolute certainty and so the forecasts can be defined as incorrect rather than just poorly calibrated. Notwithstanding, there are some situations in which this interpretation of certainty can seem unwarranted. In these cases, the penalty of the ignorance score will still seem unduly harsh, but a case for the infinite penalty can still be made. These situations occur when:

- Probabilities have been rounded up to 100% or down to 0%, and are not meant to indicate absolute certainty. For example, if probabilities are rounded to the nearest 5% then a forecast of 100% may actually only mean  $>97.5\%$ , while a forecast of 0% may only mean  $<2.5\%$ . In these cases a more accurate estimate of the actual probability should be used (perhaps 98.75% and 1.25% to represent the mid-points of the outer probability bins).<sup>5</sup>
- Probabilities are derived from an ensemble counting or similar procedure, and probabilities of 0% and 100% are simply a result of finite ensemble size. This argument is essentially the same as the previous one in that it implies a rounding of probabilities to the nearest  $m^{-1} \times 100\%$ , where  $m$  is the number of ensemble members. As a result a similar solution could be applied, although it may be preferable to consider a more satisfactory procedure for converting an ensemble into a probabilistic forecast.<sup>6</sup>

The measurement of reliability “errors” by the ignorance score is more appropriate than that of the BS, and so equation (13) is recommended over equation (12). The BS measures probability errors in a symmetric way: for example, if  $\bar{y}_k$  is 20%, then if  $\bar{p}_k$  is 10% the error will be the same as if it were 30%. This symmetry is reasonable if  $\bar{y}_k$  is 50%, but it is not clear why symmetry would make sense if  $\bar{y}_k$  were to take any other value; consider, for example, the errors re-expressed in terms of odds. Given that the climatological probabilities of events typically are not 50% in seasonal forecasting, the asymmetric measurement of probability errors by the ignorance score is an advantage.

A skill score version of either reliability score could be calculated with the form:

$$\text{Reliability skill score} = \left( 1 - \frac{\text{Reliability of forecasts}}{\text{Reliability of reference forecasts}} \right) \times 100\% \quad (14)$$

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. 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 3.2.1.3, 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 described in section 4.2.5.

<sup>5</sup> It is worth questioning the wisdom of rounding probabilities to the nearest 5% when the probabilities are close to 0% or 100%. Always rounding to the nearest 5% regardless of the probability may seem to be a deceptively uniform procedure, but is otherwise so. For example, rounding a probability of 2.5% up to 5% doubles the probability, while rounding from 47.5% is a much less drastic modification. It is perhaps helpful to think in terms of odds (the probability of the event happening divided by the probability of it not happening) instead of probabilities: rounding 2.5% to 5% changes the odds from 39 to 1 against to 19 to 1 against.

<sup>6</sup> There are numerous options, but if a simple procedure is wanted, then an additional ensemble member can be imagined that can be divided into each of the categories based on their climatological probabilities. So, for example, if the standard tercile-based categories are used, then each category is always assigned a starting count of one third of an ensemble member. If there are nine ensemble members and all nine are in the same category, instead of assigning that category a probability of 100% it would be assigned a probability of 93.3%.



#### 4.2.4 Measuring multiple 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 or most of the important attributes of good probabilistic forecasts. The Brier and ranked probability scores are widely used options, but for the reasons mentioned in section 4.1.1 their interpretation needs to be considered carefully. Specifically, their skill score versions may need to be avoided because they lack propriety when the climatological probability is not 0.5, but the scores themselves can be usefully mapped. However, the ignorance score is recommended in preference for the reasons detailed in section 4.2.3.

The (half-) BS<sup>7</sup> is calculated on each category,  $j$ , separately, and is defined as:

$$\text{Brier score}_j = \frac{1}{n} \sum_{i=1}^n (y_{j,i} - p_{j,i})^2 \quad (15)$$

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 B.1.4.

The ranked probability score (RPS) is defined as:

$$\text{RPS} = \frac{1}{n(m-1)} \sum_{i=1}^n \sum_{k=1}^{m-1} \left( \sum_{j=1}^k (y_{j,i} - p_{j,i}) \right)^2 \quad (16)$$

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 (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 B.1.5.

The use of the ignorance score (Roulston and Smith, 2002) is recommended over the BS and RPS. The ignorance score is defined as:

$$\text{Ignorance score} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m y_{j,i} \log_2 [p_{j,i}] \quad (17)$$

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.<sup>8</sup> 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 suggested 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):

<sup>7</sup> Strictly speaking, the BS 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, 2011).

<sup>8</sup> 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.

$$\text{Effective interest rate} = \left( 2^{Ign^{(ref)} - Ign} - 1 \right) \quad (18)$$

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 an investor would make if she/he invested on the forecasts and was paid out against odds based on the climatological probabilities. For example, given three equiprobable categories, an investor would be paid out three times the amount invested on the verifying category. Thus, consider an investor who has US\$ 100 to invest on the three categories, and who chooses to divide the investment between the three categories based on the forecast probabilities.<sup>9</sup> If the forecast indicates a 50% chance of category one, the investor invests US\$ 50 on this category, and the rest on the other categories, and then if category one occurs the investor would be paid US\$ 150, making a profit of US\$ 50 (or 50%). If the investor had invested an equal amount on each category (US\$ 33.33), then no matter which category verified she/he would be paid out US\$ 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 equation (18) yields an effective interest rate of 50%.

Given a series of forecasts, the investor may choose to invest all revenue from the previous round, dividing it according to the forecast probabilities, as before. For example, after the first round, having made a US\$ 50 profit, the investor now has US\$ 150 to invest in the second round. Assuming that the probabilities for the three possible outcomes are now 40%, 35%, and 25%, the investor would then invest US\$ 60 (40% of US\$ 150) on the first category, US\$ 52.50 on the second, and US\$ 37.50 on the third. If the first category verifies again, the investor would now have US\$ 180. The initial US\$ 100 has grown by 80%, but over two periods: 50% was made after the first round (from US\$ 100 to US\$ 150), and 20% in the second (from US\$ 150 to US\$ 180). The gain over the two rounds represents an average of about 34% per round.<sup>10</sup> The ignorance score for the forecasts after two rounds is about 1.16, and for the climatological forecasts is about 1.58. Applying equation (18) yields the average rate of return of about 34%. A more detailed example is provided in Appendix B section B.1.6.

If at any round a category verifies that was assigned a zero probability, the investor would lose all the money invested, and would 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 investor 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 equiprobable categories is 200%. An effective interest rate of zero indicates that in the long term the investor will neither profit nor lose on the forecasts, and the forecasts thus contain no useful information. Although the effective interest rate does not have an upper bound of 100%, it can be loosely 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 equation (18) 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 profits from the previous year's forecast can be invested in the subsequent year. If some of the forecasts are for different locations, then the initial investment has to be divided between each of the locations ( $s$  in equation (19)), and the effective interest rate has to be averaged using the ignorance score for each location:

$$\text{Average effective interest rate} = \frac{1}{s} \sum_{k=1}^s \left( 2^{Ign^{(ref)} - Ign_k} - 1 \right) \quad (19)$$

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 investment has to be divided

<sup>9</sup> Betting proportionally on the forecast probabilities is a strategy that maximizes the growth in the initial US\$ 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 chance of winning a very large amount of money over the long run, but at the very high risk of going bankrupt.

<sup>10</sup> The geometric mean of 50% and 20% is about 34%.

between the different periods since the returns on at least some of the initial investments are not yet available for reinvestment. The effective interest rate again has to be calculated using equation (19).

From equations (17)–(19) it can be shown that when the forecast probability on the verifying outcome exceeds the climatological probability, then  $Ign < Ign(ref)$  and so the effective interest rate for that step is positive (the investor makes a profit). However, when the forecast probability on the verifying outcome is less than the climatological probability,  $Ign > Ign(ref)$  and the investor makes a loss. Even with very good forecasts, a loss should be expected sometimes (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 profits graphs or accumulated profits graphs be drawn (Hagedorn and Smith, 2008). The accumulated graphs show a plot of:

$$\left( \prod_i \frac{p_i}{c_i} \right) - 1 \quad (20a)$$

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 B.1.7. If equation (19) is used to calculate the effective interest rate because forecasts for different locations and/or overlapping forecasts were pooled in the calculation, equation (20a) should be modified to:

$$\left( \prod_i \left( \frac{1}{s} \sum_{k=1}^s \frac{p_{k,i}}{c_{k,i}} \right) \right) - 1 \quad (20b)$$

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  $c_{k,i}$  is the corresponding climatological probability.

#### 4.2.5 Detailed diagnostics

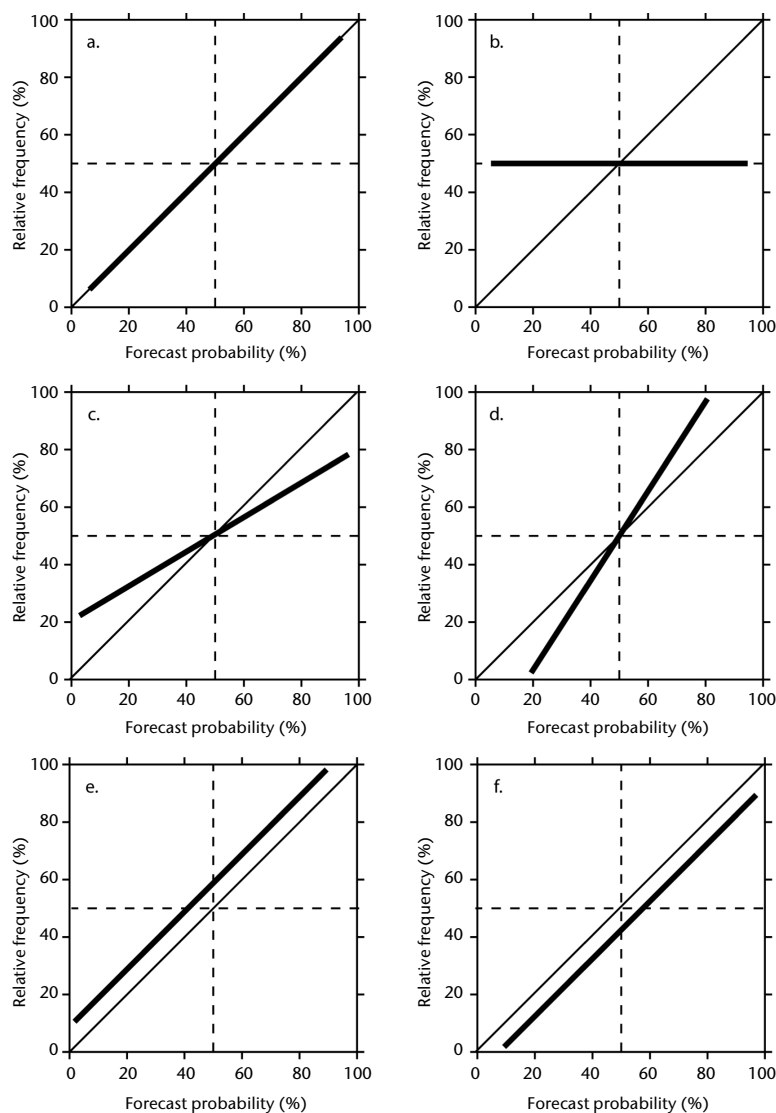
The procedures recommended thus far have provided only minimal diagnostic information. For more detailed information on forecast quality, reliability diagrams are recommended. These diagrams provide useful indications of most of the important attributes of forecast quality that were described in section 3.2.1. 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,  $\bar{y}_k$ , is the number of times an event occurred divided by the number of times the respective probability value was forecast (equation (7)). An example is provided in Appendix B section B.1.8.

The interpretation of reliability diagrams may be facilitated by considering some idealized examples as shown in Figure 4. If the forecasts are perfectly reliable then the observed relative frequency will equal the forecast probability for all values of that probability, and so the reliability curve will lie along the 45° diagonal (Figure 4a). Figure 4b illustrates the case when the event occurs with the same relative frequency regardless of the forecasts, and so the forecasts have no resolution and are useless. More typically, the forecasts have some resolution but are not perfectly reliable and will frequently show overconfidence (Figure 4c) – 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 overconfidence, the shallower is the slope of the curve. If the forecasts are under-confident, the curve is steeper than the diagonal (Figure 4d). Figure 4e illustrates the case for forecast probabilities that are consistently lower than the observed relative frequencies, indicating that the event always occurs more frequently than anticipated and so is under-forecast. Figure 4f illustrates the case when 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 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 overconfidence are represented by the two triangles in which the



**Figure 4.** Idealized examples of reliability curves showing forecast systems with (a) perfect reliability, (b) no resolution, (c) overconfidence, (d) under confidence, (e) under-forecasting, (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.

reliability curve lies in Figure 4c, whereas the areas of under confidence are represented by the two triangles in which the reliability curve lies in Figure 4d. 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 – see Figures 4e and 4f), 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 5, which are for the first 10 years of the *Prévision Saisonnière en Afrique de l'Ouest* (PRESAO) 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 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:

$$b_k = \frac{1}{n} \sum_{i=1}^n (p_{k,i} - y_{k,i}) = \bar{p}_k - \bar{y}_k \quad (21)$$

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 5, 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 5. 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:

$$\beta_1 = \frac{\sum_{k=1}^d n_k (p_k - \bar{p})(\bar{y}_k - \bar{y})}{\sum_{k=1}^d n_k (p_k - \bar{p})^2} \quad (22a)$$

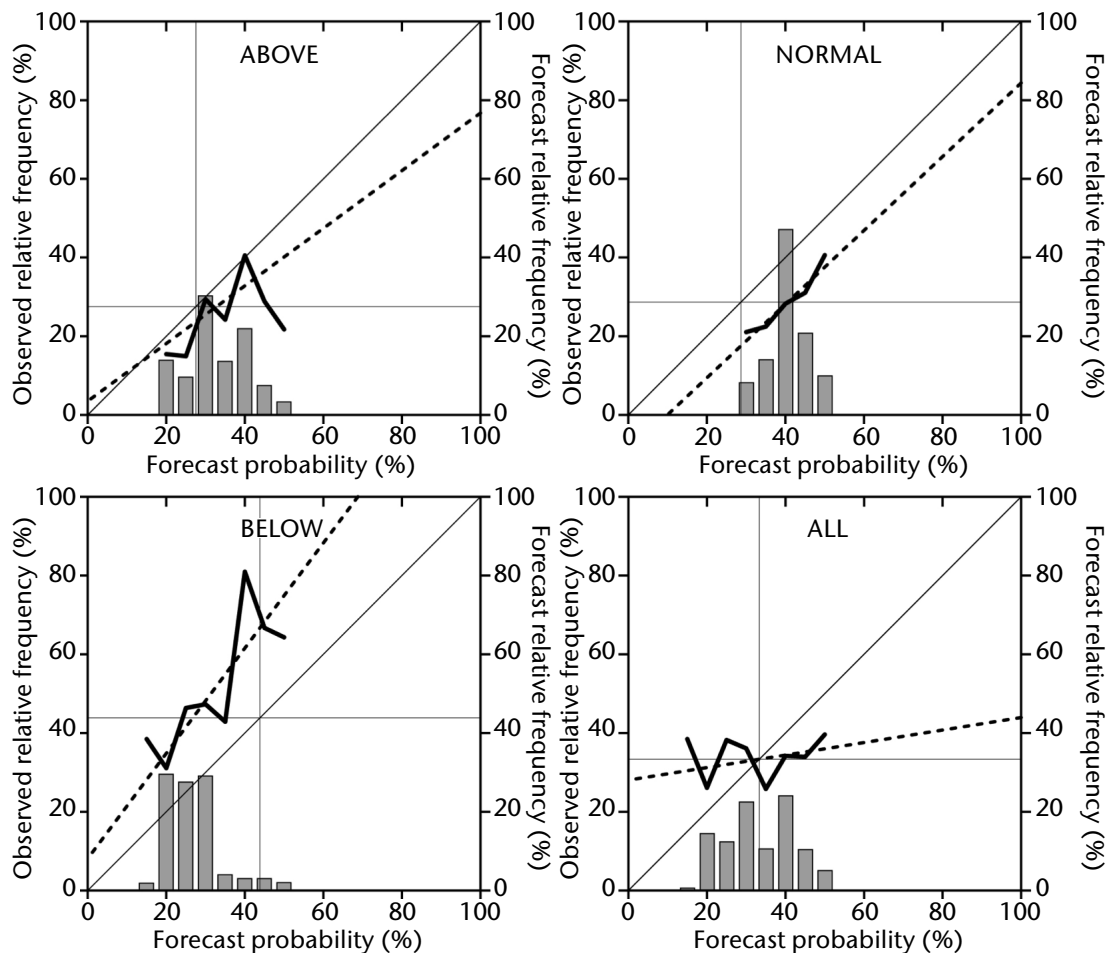
and:

$$\beta_0 = \bar{y} - \beta_1 \bar{p} \quad (22b)$$

where  $\beta_1$  is the slope and  $\beta_0$  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,  $\bar{p}_k$  is the  $k$ th probability value,  $\bar{p}$  is the average probability for all forecasts,  $\bar{y}_k$  is the observed relative frequency for the  $k$ th probability value (equation (7)), and  $\bar{y}$  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

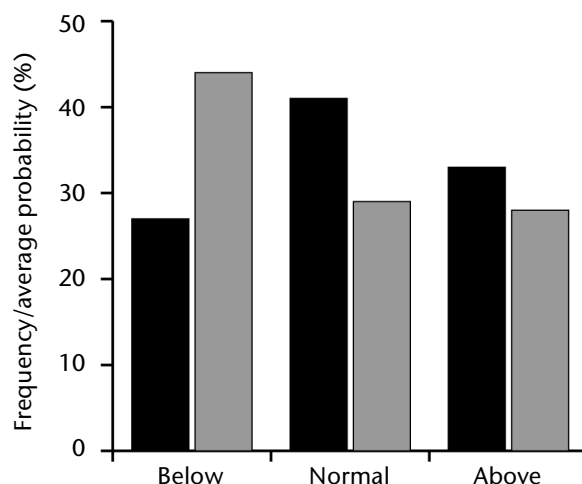
resolution an event should increase in frequency by 10% as the forecast probability is incremented by each 10% (for example, from 30% to 40%, or from 40% to 50%) and the slope will be 1.0, but if they have no resolution the slope will be zero. Overconfidence will be indicated by a slope of between 0.0 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 B.1.8.

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 (equations (22a and b)) have been suggested as two possibilities. The reliability curve itself could be communicated in tabular form rather than as a graph. The biases can be 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 diagram is shown in Figure 6, which indicates a shift towards drier conditions compared to the training period, which was not forecast.



**Figure 5. Example reliability diagrams for the first 10 years of the PRESAO seasonal rainfall forecasts for the July to September season. The bold black line shows the reliability curve, and the bold 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.**





**Figure 6. Example tendency diagram for the first 10 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.**

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 reliability diagrams will be drawn for all the forecasts available.

### 4.3 MEASURING THE QUALITY OF INDIVIDUAL FORECAST MAPS

#### 4.3.1 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 3.2.2).

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 and Hydrological Services.<sup>11</sup> Although this practice seems to have intuitive appeal, it can be problematic, as discussed in section 3.2.2. However, as discussed in section 4.2.2.2, 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

<sup>11</sup> 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” – that is, to simplify the forecast to a two-category system. Although procedures can be proposed to verify the forecasts from such a perspective, it is preferable to evaluate the information provided by the forecast as it stands, rather than to try and reinterpret the forecast and then verify what is inferred. If a two-category forecast is desired then make one!

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 the reader is referred to section 4.2.2.2 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 average interest rate is suggested as an alternative score to the hit 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 BS or RPS 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 has been described in section 4.2.4, and an example is provided in Appendix B section B.2.1. Because the forecasts are for different locations, equations (17)–(19) cannot be used. Instead, the average interest rate can be calculated using:

$$\text{Average interest rate} = \left( \frac{1}{n} \sum_{i=1}^n \frac{p_i}{c_i} \right) - 1 \quad (23)$$

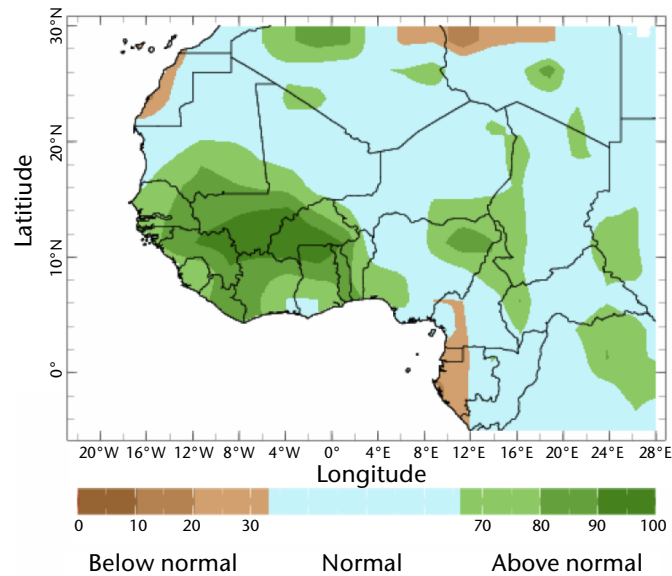
where  $p_i$  is the forecast probability for the verifying category at the  $i$ th of  $n$  locations, and  $c_i$  is the corresponding climatological probability.

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 equation (17) and is described in section 4.2.4. An example is provided in Appendix B section B.2.2.

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 maps 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 7 (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.





**Figure 7. Example rainfall anomaly map for July–August 2008 with units in percentiles using a 1971–2000 climatology**

#### 4.3.2 **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 of interest “right” for the right reasons, or if not why not. For example, in considering the quality of a specific general circulation model’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 RESULTS

Given the small sample sizes typical of seasonal forecasts, many of the procedures recommended are likely to have large sampling errors. Some uncertainty remains, therefore, 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 of 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 one can be very confident about the forecasts being good. However, calculating statistical significance is not recommended for a number of reasons. First, 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 second, 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. Third, 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 prespecified 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 (that is, 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 1 000 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 5th and 95th percentiles identified. (The 5th percentile will be the  $(n_b \times 5 \div 100)$ th ranked value, and the 95th 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, a score might be communicated as follows: “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 (that is, gridbox station, or region) is assigned a weight,  $w_i$ , based on its representative area, as discussed in Chapter 2. In each case, the verification scores are adjusted to represent the appropriate area weighting of each location. In the case of gridded forecasts, each grid cell could be weighted by its area relative to that of the other grids. Assuming a regular grid, a simple approximation to area-based weighting could be obtained by taking the cosine of the central latitude of the grid. For a more accurate calculation, the area of a gridbox is:

$$Area = 2R^2 \cos(y) \sin(0.5dy) dx, \quad (A.1)$$

where  $R$  is the radius of the Earth (6 371 km),  $y$  is the latitude at the centre of the gridbox,  $dy$  is the latitudinal resolution, and  $dx$  is the longitudinal resolution.

In most of the equations that follow, 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 multiplied by the number of time steps. For example, if there are 10 forecasts at each of 5 locations, there will be a total of 50 forecasts ( $10 \times 5$ ), 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, and so on.

### A.1 SERIES OF FORECASTS

#### A.1.1 Relative operating characteristics

If the ROC graph is to be constructed, the HRs and FARs (equations (2) and (3)) 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 equation (4) without having to worry about weights, since the weights are already built into the HRs and FARs.

Alternatively, if the ROC area is to be calculated without constructing the graph, equation (1a) must be modified to:

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

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 (1b) is unaffected.

#### A.1.2 Generalized discrimination score

The weights are applied only to equation (5a), which becomes:

$$D = \frac{\sum_{k=1}^{m-1} \sum_{l=k+1}^m \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} w_i w_j I(p_{k,i}, p_{l,j})}{\sum_{k=1}^{m-1} \sum_{l=k+1}^m \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} w_i w_j} \quad (A.3)$$

where  $m$  is the number of categories,  $n_k$  is the number of times the observation was in category  $k$ , and  $p_{k,i}$  is the vector of forecast probabilities for the  $i$ th observation in category  $k$ .

### A.1.3 Resolution score

The weighted version of the resolution scores (equations (6) and (8)) are:

$$Resolution = \frac{\sum_{k=1}^d \left( \sum_{i=1}^{n_k} w_{k,i} \right) (\bar{y}_k - \bar{y})^2}{\sum_{j=1}^n w_j} \quad (A.4)$$

and:

$$Resolution = \frac{\sum_{k=1}^d \left( \sum_{i=1}^{n_k} w_{k,i} \right) \left( \bar{y}_k \log_2 \left[ \frac{\bar{y}_k}{\bar{y}} \right] + (1 - \bar{y}_k) \log_2 \left[ \frac{(1 - \bar{y}_k)}{(1 - \bar{y})} \right] \right)}{\sum_{j=1}^n w_j} \quad (A.5)$$

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,  $\bar{y}_k$  is the observed relative frequency, and  $\bar{y}$  is the observed relative frequency for all forecasts. The denominator defines the sum of the weights for all the forecasts.

### A.1.4 Hit (Heidke) score

The hit score (equation (10)) 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:

$$\bar{y}_j = \frac{\sum_{i=1}^n w_i y_{j,i}}{\sum_{i=1}^n w_i} \quad (A.5)$$

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.

### A.1.5 Reliability score

The weighted versions of the reliability scores (equations (12) and (13)) are:

$$Reliability = \frac{\sum_{k=1}^d \sum_{i=1}^{n_k} w_{k,i} (p_{k,i} - \bar{p}_k)^2}{\sum_{j=1}^n w_j} \quad (A.6)$$

and:

$$Reliability = \frac{\sum_{k=1}^d \left( \sum_{i=1}^{n_k} w_{k,i} \right) \left( \bar{p}_k \log \left[ \frac{\bar{p}_k}{p_k} \right] + (1 - \bar{p}_k) \log \left[ \frac{(1 - \bar{p}_k)}{(1 - p_k)} \right] \right)}{\sum_{j=1}^n w_j} \quad (A.7)$$

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,  $p_{k,i}$  is the corresponding forecast probability, and  $\bar{y}_k$  is the observed relative frequency. The denominator defines the sum of the weights for all the forecasts (locations and time steps).

#### A.1.6 Brier score

The weighted version of the BS (equation (15)) is:

$$BS_j = \frac{\sum_{i=1}^n w_i (y_{j,i} - p_{j,i})^2}{\sum_{i=1}^n w_i}, \quad (A.8)$$

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  $j$ .

#### A.1.7 Ranked probability score

The weighted RPS (equation (16)) is defined as:

$$RPS = \frac{\sum_{i=1}^n w_i \sum_{k=1}^{m-1} \left( \sum_{j=1}^k (y_{j,i} - p_{j,i}) \right)^2}{(m-1) \sum_{i=1}^n w_i} \quad (A.9)$$

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  $j$ .

#### A.1.8 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 equation (19) rather than to equation (18). To incorporate the weights, equation (19) is modified to:

$$\text{Average effective interest rate} = \frac{\sum_{k=1}^s w_i \left( 2^{Ign(ref) - Ign_k} - 1 \right)}{\sum_{k=1}^s w_i} \quad (A.10)$$

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

#### A.1.9 Accumulated profits

The values on the  $y$ -axis for the accumulated profits graph (equation (20b)) can be obtained using:

$$\left( \prod_i \left( \frac{\sum_{k=1}^s \frac{w_k p_{k,i}}{c_{k,i}}}{\sum_{k=1}^s w_k} \right) \right) - 1 \quad (A.11)$$

where  $s$  is the number of stations/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.

### A.1.10 Reliability diagram

The weights must be applied when calculating the observed relative frequencies (equation (7)):

$$\bar{y}_k = \frac{\sum_{i=1}^{n_k} w_i y_{k,i}}{\sum_{i=1}^{n_k} w_i} \quad (\text{A.12})$$

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 equation (A.12), and replacing  $y_{k,i}$  with  $p_{k,i}$ , where  $p_{k,i}$  is the  $i$ th forecast probability in bin  $k$ . Similarly, the bin frequencies for the frequency should be replaced by the denominator in equation (A.12).

The weighted unconditional biases can be calculated using:

$$b_k = \frac{\sum_{i=1}^n w_i (p_{k,i} - y_{k,i})}{\sum_{i=1}^n w_i} \quad (\text{A.13})$$

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 equation (22) since the weights are already considered in the construction of the regression curve. The only modification required is to replace  $n_k$  in equation (22a) with the sum of the weights for all the forecasts in the current bin (that is, with the denominator in equation (A.12)).

## A.2 INDIVIDUAL FORECAST MAPS

### A.2.1 Hit score

The weights are included in the same way as in equation (A.5).

### A.2.2 Average interest rate

The average interest rate (equation (23)) is modified in a similar way to equation (A.10):

$$\text{Average interest rate} = \left( \frac{\sum_{i=1}^n w_i p_i}{\sum_{i=1}^n w_i} \right) - 1 \quad (\text{A.14})$$

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

### A.2.3 Ignorance score

The weighted version of the ignorance score can be calculated using:

$$Ignorance\ score = - \frac{\sum_{i=1}^n w_i \sum_{j=1}^m y_{j,i} \log_2 [p_{j,i}]}{\sum_{i=1}^n w_i} \quad (A.15)$$

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 equiprobable categories (below normal (B), normal (N) and above normal (A))**

<i>Year</i>	<i>Observation</i>	<i>Below</i>	<i>Normal</i>	<i>Above</i>
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 equiprobable categories (below normal (B), normal (N), and above normal (A))**

<i>Location</i>	<i>Observation</i>	<i>Below</i>	<i>Normal</i>	<i>Above</i>
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



## B.1 SERIES OF FORECASTS

### B.1.1 Relative operating characteristics

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

$$A = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} I(p_{0,i}, p_{1,j}) \quad (\text{B.1a})$$

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  $I(p_{0,i}, p_{1,j})$  is defined as:

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

Using the data in Table B.1 and setting the above-normal category as an event, from equation (B.1a)  $n_1 = 2$ , 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 (equation (B.1b)). A total of  $n_0 \times n_1 = 12$  comparisons are made.

These comparisons are shown in Table B.3. There are 9 out of 12 correct selections, and one tie, yielding a score of about 79%.

**Table B.3. Example calculation of the ROC area for the above-normal category using equation (2)**

$i$	$j$	$Year_{0,i}$	$Year_{1,j}$	$p_{0,i}$	$p_{1,j}$	$I(p_{0,i}, p_{1,j})$
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
$\sum I(p_{0,i}, p_{1,j})$						9.5

The ROC curve is constructed by calculating HRs and FARs for decreasing probability thresholds. The HRs and FARs are calculated using (respectively):

$$y_k = \frac{1}{n_1} \sum_{i=1}^{n_1} y_{k,i} \quad (\text{B.2})$$

and:

$$x_k = \frac{1}{n_0} \sum_{i=1}^{n_0} x_{k,i} \quad (\text{B.3})$$

where  $n_1$  is the number of events,  $n_0$  is the 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.4a. 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 HRs on the  $x$ -axis against the FARs 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.4b, and is in agreement with the area obtained from Table B.4a. The trapezoidal rule is given by:

$$A = 0.5 \times \left[ 1 + \sum_{k=0}^d (y_k x_{k+1} - y_{k+1} x_k) \right] \quad (\text{B.4})$$

where  $d$  is the number of discrete probability values, and  $y_1$  and  $x_1$  are the HRs and FARs for the highest probability value only,  $y_2$  and  $x_2$  are the rates for the highest and second highest probabilities, and so on. (Note the slight change of notation from equation (4)). For  $i = 0$  the HR and FARs 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.

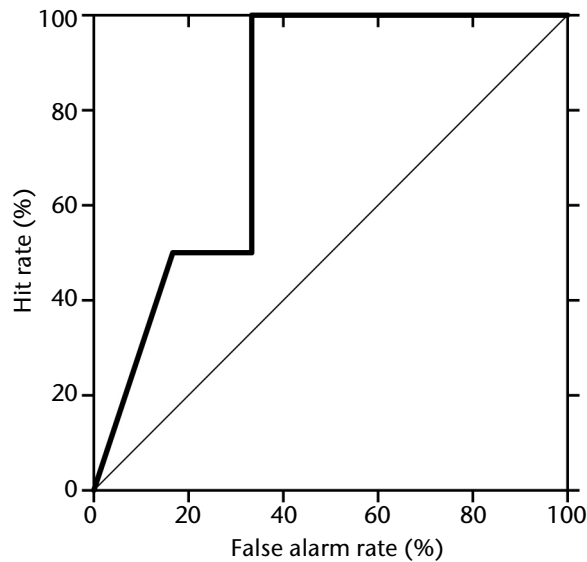


Figure B.1. Example ROC graph for the above-normal category using the HRs and FARs from Table B.4

**Table B.4a. Example calculation of the HRs and FARs 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
	FAR		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
	HR		0.50	0.50	1.00	1.00	1.00	1.00	1.00

**Table B.4b. Example calculation of the ROC area for the above-normal category using equation (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			
$\sum_{k=0}^d (y_k x_{k+1} - y_{k+1} x_k)$						0.58
$0.5 \times \left[ 1 + \sum_{k=0}^d (y_k x_{k+1} - y_{k+1} x_k) \right]$						0.79

**B.1.2 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:

$$D = \frac{\sum_{k=1}^{m-1} \sum_{l=k+1}^m \sum_{i=1}^{n_k} \sum_{j=1}^{n_l} I(\mathbf{p}_{k,i}, \mathbf{p}_{l,j})}{\sum_{k=1}^{m-1} \sum_{l=k+1}^m n_k n_l} \quad (\text{B.5a})$$

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:

$$I(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) = \begin{cases} 0.0 & \text{if } F(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) < 0.5 \\ 0.5 & \text{if } F(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) = 0.5 \\ 1.0 & \text{if } F(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) > 0.5 \end{cases} \quad (\text{B.5b})$$

and where:

$$F(\mathbf{p}_{k,i}, \mathbf{p}_{l,j}) = \frac{\sum_{r=1}^{m-1} \sum_{s=r+1}^m p_{k,i}(r) p_{l,j}(s)}{1 - \sum_{r=1}^m p_{k,i}(r) p_{l,j}(r)} \quad (\text{B.5c})$$

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 equation (B.5a)  $m = 3$ ,  $n_1 = 4$ ,  $n_2 = 2$  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 first, 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 equation (B.5c) 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 ( $r = 2$ ) 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 equation (B.5c) is shown in Table B.5, and the result of equation (B.5b) 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  $17.5/20 = 87.5\%$ .

Equation (5) 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.5. Example calculation of the generalized discrimination score**

$K$	$L$	$i$	$j$	$Year_{k,i}$	$Year_{l,j}$	Equation (B.5c)	$I(p_{k,i}, p_{l,j})$
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

$K$	$L$	$i$	$J$	$Year_{k,i}$	$Year_{l,j}$	Equation (B.5c)	$I(\mathbf{p}_{k,i}, \mathbf{p}_{l,j})$
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
$\sum I(\mathbf{p}_{k,i}, \mathbf{p}_{l,j})$							17.5

### B.1.3 Hit (Heidke) scores

The hit score is calculated using:

$$\bar{y}_k = 100\% \times \frac{1}{n} \sum_{i=1}^n y_{k,i} \quad (\text{B.6})$$

where  $n$  is the 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 1 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 (that is, no skill) the score depends on the number of categories and the climatological probabilities of those categories, and so can be difficult to understand by non-specialists. The interpretation is simplified considerably if the categories are climatologically equiprobable, in which case the score of no-skill forecasts will be 1 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. Full calculations are provided in Table B.6, with the ranks of the categories based on their respective probabilities shown in columns 3–5. The skill score defining the difference between the hit scores for the categories with highest and lowest probabilities ( $y_1$  and  $y_3$ , 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 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.

**Table B.6. Example calculation of hit scores using equation (9b)**

<i>i</i>	<i>Observation</i>	<i>Forecast category rank</i>			<i>Hits</i>		
		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
<i>Number of hits</i> <i>Number of forecasts</i>					$\frac{3\frac{1}{3}}{8} \approx 42\%$	$\frac{4\frac{1}{3}}{8} \approx 54\%$	$\frac{\frac{1}{3}}{8} \approx 4\%$

#### B.1.4 Brier score

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

$$BS_j = \frac{1}{n} \sum_{i=1}^n (y_{j,i} - p_{j,i})^2 \quad (\text{B.7})$$

where  $n$  is the number of forecasts,  $y_{j,i}$  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 BS for the above-normal category using equation (B.7)**

<i>I</i>	$y_{3,i}$	$p_{3,i}$	$(y_{3,i} - p_{3,i})^2$
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
$\frac{1}{n} \sum_{i=1}^n (y_{3,i} - p_{3,i})^2$			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 BS is more useful than a second set of forecasts, B, with a higher BS, even if we can be absolutely certain that the difference in the scores is not because of sampling uncertainty.

The BS can be used to map the quality of the forecasts (that is, 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 (that is, 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. The components of the BS are not suitable for use with small sample sizes.

### B.1.5 Ranked probability score

The RPS 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:

$$RPS = \frac{1}{n(m-1)} \sum_{i=1}^n \sum_{k=1}^{m-1} \left( \sum_{j=1}^k (y_{j,i} - p_{j,i}) \right)^2 \quad (\text{B.8})$$

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.

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 that always indicated with 0% probability the verifying category). Thus, as with the BS, 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 multiple categories, it can be a little difficult to interpret. For example, as observed for the BS in the previous section, there is no guarantee that one set of forecasts, A, with a lower RPS is more useful than a second set of forecasts, B, with a higher RPS, even if we can be absolutely certain that the difference in the scores is not because of sampling uncertainty.

**Table B.8. Example calculation of the RPS using equation (B.8)**

$i$	$y_{1,i}$	$p_{1,i}$	$(y_{1,i} - p_{1,i})^2$	$\sum_{j=1}^2 y_{j,i}$	$\sum_{j=1}^2 p_{j,i}$	$\left( \sum_{j=1}^2 (y_{j,i} - p_{j,i}) \right)^2$
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



$i$	$y_{1,i}$	$p_{1,i}$	$\left(y_{1,i} - p_{1,i}\right)^2$	$\sum_{j=1}^2 y_{j,i}$	$\sum_{j=1}^2 p_{j,i}$	$\left(\sum_{j=1}^2 \left(y_{j,i} - p_{j,i}\right)\right)^2$
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 RPS can be used to map the quality of the forecasts (that is, 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 (that is, 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.

#### B.1.6 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:

$$\text{Ignorance score} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m y_{j,i} \log_2 [p_{j,i}] \quad (\text{B.9})$$

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 equation (B.9) 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,  $Ign(ref)$ :

$$\text{Effective interest rate} = \left(2^{Ign(ref) - Ign} - 1\right) \quad (\text{B.10})$$

Like the BS and RPS, the effective interest rate measures a number of attributes, but it has a simpler interpretation because of its relationship to investment strategies. Specifically, if a user were to invest on the forecasts and received fair odds (calculated using the climatological probabilities), and to carry losses and profits forwards each time, the effective interest rate indicates the profit or loss that would be made. Using the data in Table B.9 an example can be given in which a user, starting with US\$ 100, invests US\$ 45 on below normal, US\$ 35 on normal, and US\$ 20 on above normal, and below normal occurs. Given fair odds, the user earns twice what she/he invested on the verifying category because the odds are two to one against, retains the original investment on that category, but loses the investments on the other categories. The net effect is that the user ends with three times what was invested on the verifying category, that is US\$ 135 ( $3 \times \text{US\$ } 45$ ) and has made US\$ 35 profit (or 35%). The user then invests US\$ 67.50 (50% of US\$ 135) on below normal, US\$ 40.50 (30% of US\$ 135) on normal, and US\$ 27 on above normal (20% of US\$ 135), and ends with US\$ 202.50 ( $3 \times \text{US\$ } 67.50$ ), which is a 50% profit. Over

the two years the user has made US\$ 102.50 profit, which is equivalent to 42.3% profit per year. 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 equations (B.9) and (B.10).**

$I$	$j$	$p_{j,i}$	$\sum_{j=1}^m y_{j,i} \log_2 [p_{j,i}]$	$\sum_{j=1}^m y_{j,i} \log_2 [0.33]$
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
$-\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m y_{j,i} \log_2 [p_{j,i}]$			1.368	1.585
			$\left(2^{I_{gn}^{(ref)} - I_{gn}} - 1\right) \left(2^{1.585 - 1.368} - 1\right) \approx 16\%$	

The effective interest rate is positive for “good” forecasts, but its upper bound depends on the prior probabilities. For a three-category forecast system with equiprobable categories, the upper bound is 200%. In this instance, fair odds pays out two times the amount invested on the verifying category plus the initial investment, and, for perfectly good forecasts, 100% of the bet would have been placed on this category. Thus, an investment of US\$ 100 would make a profit of US\$ 200, which is 200% of the original bet. For a perfectly bad set of forecasts, the ignorance score becomes infinity, and so, from equation (B.10) the effective interest rate approaches -100%. The effective interest rate approaches -100% if any of the individual forecasts are 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.

### B.1.7 Accumulated profits graph

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

$$b \left[ \left( \prod_i \frac{p_i}{c_i} \right) - 1 \right] \quad (\text{B.11})$$

where  $b$  is the initial investment,  $p_i$  is the forecast probability for the verifying category, and  $c_i$  is its climatological probability. Equation (B.11) 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 US\$ 100, a profit of US\$ 35 has been made, and there is now US\$ 135 for investment in the second round (fifth column, first row). In the second round, a 50% profit is

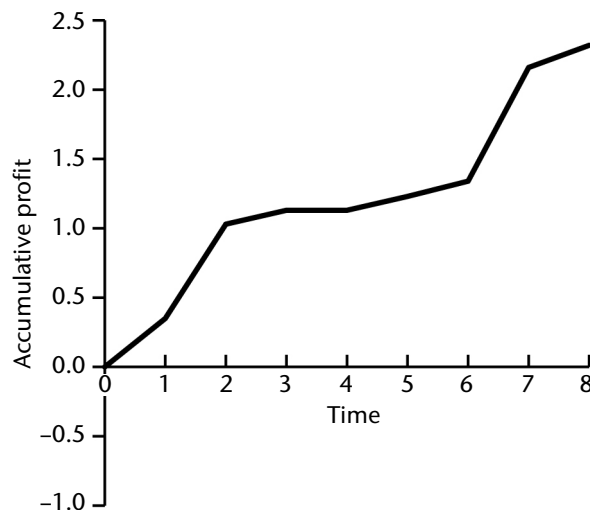
made (fourth column) and the US\$ 135 is multiplied by 1.50 to give US\$ 203 (fifth column), and the profit is now US\$ 103 (last column). The resulting graph for the example shown in Table B.10 is shown in Figure B.2.

**Table B.10. Example calculation of the accumulated profits using equation (B.11)**

$i$	$j$	$p_{ji}$	$\frac{p_{ji}}{0.33}$	$\prod_i \frac{p_{ji}}{0.33}$	$\prod_i \frac{p_{ji}}{0.33} - 1$
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.

The accumulated profits graph typically will increase or decrease exponentially, and so it can be difficult to compare the proportional gains or losses across years. As an alternative, the profits or losses made from a fixed investment each year can be plotted (Figure B.3). The calculation is shown in the last column of Table B.10. The results can be averaged across multiple locations. It is not necessarily the case that an increase or decrease in the averaged profits graph will be



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

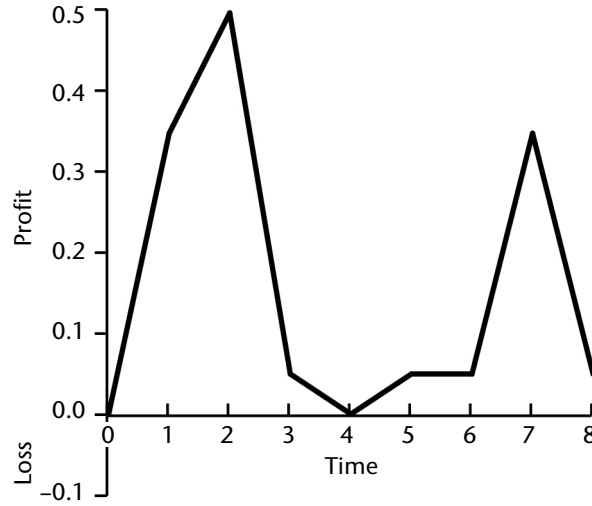


Figure B.3. Example profits graph using the data in Table B.10

matched by an increase or decrease in the accumulated profits graph, because in the latter case the results are weighted by the previous profits at each station, whereas the stations are given equal weight in the averaged graph.

#### B.1.8 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 Tables B.11a and b, which list 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  $\bar{y}_k$  in the last column of Table B.11a are regressed against the forecast probabilities in the first column, but weighted by the forecast frequencies (second column; using the values in the fourth column 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:

$$\beta_1 = \frac{\sum_{k=1}^d n_k (p_k - \bar{p})(\bar{y}_k - \bar{y})}{\sum_{k=1}^d n_k (p_k - \bar{p})^2} \quad (\text{B.12a})$$

and:

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \quad (\text{B.12b})$$

where  $\beta_1$  is the slope and  $\beta_0$  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),  $\bar{p}_k$  is the  $k$ th probability value (first column),  $\bar{p}$  is the average probability for all forecasts,  $\bar{y}_k$  is the observed relative frequency for the  $k$ th probability value (third column), and  $\bar{y}$  is the observed relative frequency over the verification period.

The slope,  $\beta_1$ , 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% (for example, 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 overconfident. In general, overconfidence 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.

Interpreting the intercept,  $\beta_0$ , 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.

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

<i>Forecast</i>	<i>Number of forecasts (<math>n_k</math>)</i>	<i>Number of events</i>	<i>Forecast relative frequency</i>	<i>Observed relative frequency (<math>\bar{y}_k</math>)</i>
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		

Forecast	Number of forecasts ( $n_k$ )	Number of events	Forecast relative frequency	Observed relative frequency ( $\bar{y}_k$ )
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 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
$\bar{p} \approx 0.33$		$\bar{y} \approx 0.28$	$\sum_{k=1}^d n_k (p_k - \bar{p})(\bar{y}_k - \bar{y}) \approx 3.26$	$\sum_{k=1}^d n_k (p_k - \bar{p})^2 \approx 4.46$
			$\beta_1 \approx \frac{3.26}{4.46} \approx 0.73$	$\beta_0 \approx 0.28 - 0.73 \times 0.33 \approx 0.03$

## B.2 INDIVIDUAL FORECAST MAPS

### B.2.1 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:

$$\text{Average interest rate} = \left( \frac{1}{n} \sum_{i=1}^n \frac{p_i}{c_i} \right) - 1 \quad (\text{B.13})$$

where  $p_i$  is the forecast probability for the verifying category at the  $i$ th of  $n$  locations, and  $c_i$  is the corresponding climatological probability. The summation over  $j$  in equation (B.13) simply

searches for the verifying category. An example for the data in Table B.2 is shown in Table B.12, where the second column indicates the verifying category, and the third the probability for that category.

**Table B.12. Example calculation of the average interest rate using equation (B.13)**

$I$	$J$	$p_{j,i}$	$p_i/c_i$
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 (Appendix B section B.1.6) – 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 two times the amounts invested on the respective verifying categories plus the original investment, and, for perfectly good forecasts, 100% of the investment at each location would have been placed on the verifying category. Thus, an investment of US\$ 100 would make a profit of US\$ 200 at any location, which is 200% of the original investment. 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).

### B.2.2 Ignorance score

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

$$\text{Ignorance score} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m y_{j,i} \log_2 [p_{j,i}] \quad (\text{B.14})$$

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. As with the linear probability score, the summation over  $j$  in equation (B.14) 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 ignorance score using equation (B.14)**

$i$	$j$	$p_{j,i}$	$\sum_{j=1}^m y_{j,i} \log_2 [p_{j,i}]$
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
$-\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^m y_{j,i} \log_2 [p_{j,i}]$			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 have 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, there is inconsistency within the verification literature – 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 (for example, the seasonal forecast for January–March 2000).

### Anomaly

The difference between an observed value of a meteorological variable (for example, seasonally averaged temperature) for a single period (for example, January–March (JFM) 2000) and its long-term average (for example, 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 predefined 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 (for example, 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 recalculating 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 predefined 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 (for example, 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 forecaster’s 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 overreaction, 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 (for example, total rainfall in mm) or a specific category (for example, 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 FAR measures the proportion of non-events that were incorrectly forewarned, 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 (for example, trying to minimize the possibility that the forecast might be perceived as “wrong”) or appropriate (for example, avoiding overreacting 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 HR measures the proportion of events that were forewarned, 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 HR, which measures the proportion of events that were forewarned. The  $\gamma$ -values of the reliability curve measure the observed relative frequency, which can be interpreted as a conditional hit score.

### **Overconfidence**

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

**Over-forecasting**

A tendency to overestimate 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 (that is, an event must occur on 40% of the occasions that the forecast probability is 40%, 50% of the occasions the probability is 50%, and so on).

**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.

### **Scoring rule**

A rule for scoring a single forecast–observation pair.

### **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, overconfident 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 to 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 underestimate 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 underestimate 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. *Monthly Weather Review*, 78:1–3.
- Bröcker, J. and L.A. Smith, 2007a: Increasing the reliability of reliability diagrams. *Weather Forecasting*, 22:651–661.
- Bröcker, J. and L.A. Smith, 2007b: Scoring probabilistic forecasts: The importance of being proper. *Weather 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.
- Epstein, E.S., 1969: A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8:985–987.
- Hagedorn, R. and L.A. Smith, 2008: Communicating the value of probabilistic forecasts with weather roulette. *Meteorological Applications*, 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. *Monthly Weather Review*, 120:863–883.
- Hogan, R.J. and I.B. Mason, 2012: Deterministic forecasts of binary events. In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (I.T. Jolliffe and D.B. Stephenson, eds.). Chichester, Wiley:31–59. Doi:10.1002/9781119960003.ch3.
- Hsu, W.-R. and A.H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, 2:285–293.
- Jolliffe, I.T., 2007: Uncertainty and inference for verification measures. *Weather and Forecasting*, 22:637–650.
- Jolliffe, I.T. and D.B. Stephenson (eds.), 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Chichester, Wiley, Chichester.
- Katz, R.W. and A.H. Murphy (eds.), 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge, Cambridge University Press.
- Mason, S.J., 2012: Seasonal and longer-range forecasts. In: *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (I.T. Jolliffe and D.B. Stephenson, eds.). Chichester, Wiley:31–59. Doi:10.1002/9781119960003.ch3.
- Mason, S.J. and N.E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quarterly Journal of the Royal Meteorological Society*, 128:2145–2166.
- Mason, S.J. and A.P. Weigel, 2009: A generic forecast verification framework for administrative purposes. *Monthly Weather Review*, 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. *Journal of Applied Meteorology*, 5:534–537.
- Murphy, A.H., 1969: On the “ranked probability score”. *Journal of Applied Meteorology*, 8:988–989.
- Murphy, A.H., 1970: The ranked probability score and the probability score: A comparison. *Monthly Weather Review*, 98:917–924.
- Murphy, A.H., 1971: A note on the ranked probability score. *Journal of Applied Meteorology*, 10:155–156.
- Murphy, A. H., 1973: A new vector partition of the probability score. *Journal of Applied Meteorology*, 12:595–600.
- Murphy, A.H., 1991: Forecast verification: Its complexity and dimensionality. *Monthly Weather Review*, 119:1590–1601.
- Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather Forecasting*, 8:281–293.
- Roulston, M.S. and L.A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130:1653–1660.
- Stephenson, D.B., C.A.D.S. Coelho and I.T. Jolliffe, 2008: Two extra components in the Brier Score decomposition. *Weather 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. *Monthly Weather Review*, 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*. NATO Science Series, Dordrecht, Springer Academic Publishers.
- Weigel, A.P. and S.J. Mason, 2011: The generalized discrimination score for ensemble forecasts. *Monthly Weather Review*, 139:3069–3074.

- Weijs, S.V., R. Van Nooijen and N. Van de Giesen, 2010: Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Monthly Weather Review*, 138:3387–3399.
- Wilks, D.S., 2011: *Statistical Methods in the Atmospheric Sciences*. Volume 100. San Diego, Academic Press.
- Wilks, D.S. and A.H. Murphy, 1998: A case study of the use of statistical models in forecast verification: Precipitation probability forecasts. *Weather Forecasting*, 13:795–810.
-



For more information, please contact:

**World Meteorological Organization**

7 bis, avenue de la Paix – P.O. Box 2300 – CH 1211 Geneva 2 – Switzerland

**Communication and Public Affairs Office**

Tel.: +41 (0) 22 730 83 14/15 – Fax: +41 (0) 22 730 80 27

Email: [cpa@wmo.int](mailto:cpa@wmo.int)

**[public.wmo.int](http://public.wmo.int)**