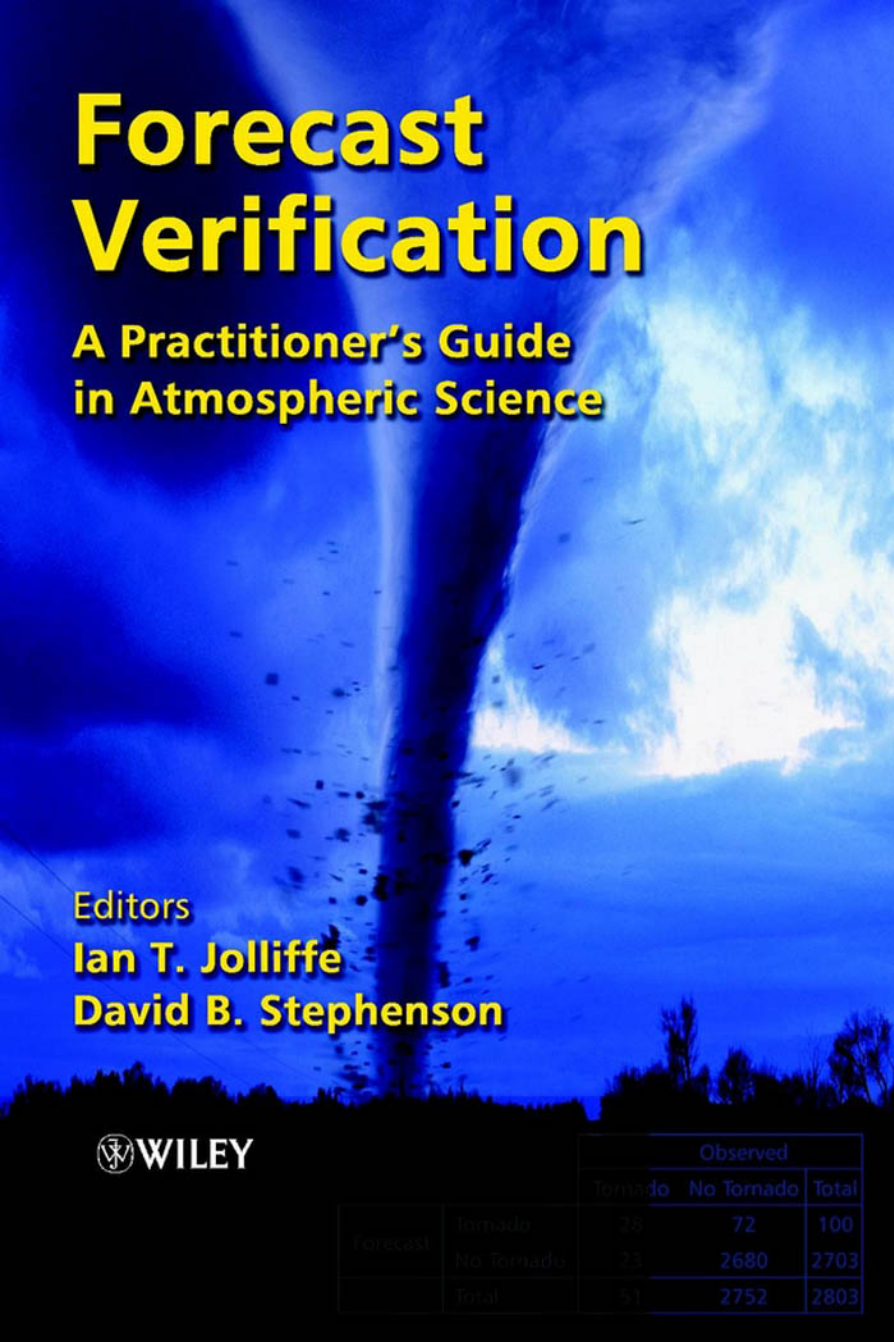
Forecast Verification

A Practitioner's Guide in Atmospheric Science

Editors
Ian T. Jolliffe
David B. Stephenson

WILEY



Forecast Verification

Forecast Verification
A Practitioner's Guide in Atmospheric Science

Edited by

IAN T. JOLLIFFE
University of Aberdeen

and

DAVID B. STEPHENSON
University of Reading

Telephone (+44) 1243 779777
e-mail (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wileyeurope.com or www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Forecast verification: a practitioner's guide in atmospheric science / edited by Ian T. Jolliffe and David B. Stephenson.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-49759-2 (alk. paper)
1. Weather forecasting–Statistical methods–Evaluation. I. Jolliffe, I. T. II. Stephenson, David B.

QC996.5.F677 2003
551.63–dc21 2002192424

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0-471-49759-2

Typeset in 10.5/13pt Times New Roman by Kolam Information Services Pvt. Ltd, Pondicherry, India
Printed and bound in Great Britain by Antony Rowe Ltd, Chippenham, Wiltshire
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.

Contents

List of Contributors

G. Candille	Laboratoire de Météorologie Dynamique, Ecole Normale Supérieure, 24 Rue Lhomond, F 75231 Paris cedex 05, France. gcandi@lmd.ens.fr
M. Déqué	Meteo-France CNRM/GMGEC/EAC, 42 Avenue Coriolis, 31057 Toulouse cedex 01, France. deque@meteo.fr
W. Drosdowsky	Bureau of Meteorology Research Centre, BMRC, PO Box 1289K, Melbourne 3001, Australia. w.drosdowsky@bom.gov.au
I.T. Jolliffe	Department of Mathematical Sciences, University of Aberdeen, King's College, Aberdeen AB24 3UE, UK. itj@maths.abdn.ac.uk
R.E. Livezey	W/OS4, Climate Services Division, Room 13228, SSMC2, 1325 East West Highway, Silver Spring, MD 20910–3283, USA. robert.e.livezey@noaa.gov
I.B. Mason	Canberra Meteorological Office, PO Box 797, Canberra, ACT 2601, Australia. ibmason@bigpond.com
J.M. Potts	Biomathematics and Statistics Scotland, The Macaulay Institute, Craigiebuckler, Aberdeen AB15 8QH, UK. j.potts@bioss.ac.uk
D. Richardson	Meteorological Office, London Road, Bracknell, Reading, RG12 2SZ, UK. david.s.richardson@metoffice.com
D.B. Stephenson	Department of Meteorology, University of Reading, Earley Gate PO Box 243, Reading RG6 6BB, UK. d.b.stephenson@reading.ac.uk
O. Talagrand	Laboratoire de Météorologie Dynamique, Ecole Normale Supérieure, 24 Rue Lhomond, F 75231 Paris cedex 05, France. talagran@lmd.ens.fr
Z. Toth	NOAA at National Centers for Environmental Prediction, 5200 Auth Rd., Room 207, Camp Springs, MD 20746, USA. zoltan.toth@noaa.gov
H. Zhang	Bureau of Meteorology Research Centre, BMRC, PO Box 1289K, Melbourne 3001, Australia. h.zhang@bom.gov.au

Y. Zhu NOAA at National Centers for Environmental Prediction, 5200 Auth Rd., Room 207, Camp Springs, MD 20746, USA. yuejian.zhu@noaa.gov

Preface

Forecasts are made in many disciplines, the best known of which are economic forecasts and weather forecasts. Other situations include medical diagnostic tests, prediction of the size of an oil field, and any sporting occasion where bets are placed on the outcome. It is very often useful to have some measure of the skill or value of a forecast or forecasting procedure. Definitions of 'skill' and 'value' will be deferred until later in the book, but in some circumstances financial considerations are important (economic forecasting, betting, oil field size), whilst in others a correct or incorrect forecast (medical diagnosis, extreme weather events) can mean the difference between life and death.

Often the 'skill' or 'value' of a forecast is judged in relative terms. Is forecast provider A doing better than B? Is a newly developed forecasting procedure an improvement on current practice? Sometimes, however, there is a desire to measure absolute, rather than relative, skill. Forecast verification, the subject of this book, is concerned with judging how good is a forecasting system or single forecast.

Although the phrase 'forecast *verification*' is generally used in atmospheric science, and hence adopted here, it is rarely used outside the discipline. For example, a survey of keywords from articles in the *International Journal of Forecasting* between 1996 and 2002 has no instances of 'verification'. This journal attracts authors from a variety of disciplines, though economic forecasting is prominent. The most frequent alternative terminology in the journal's keywords is 'forecast *evaluation*', although *validation* and *accuracy* also occur. Evaluation and validation also occur in other subject areas, but the latter is often used to denote a wider range of activities than simply judging skill or value – see, for example, Altman and Royston (2000).

Many disciplines make use of forecast verification, but it is probably fair to say that a large proportion of the ideas and methodology have been developed in the context of weather and climate forecasting, and this book is firmly rooted in that area. It will therefore be of greatest interest to forecasters, researchers and students in atmospheric science. It is written at a level that is accessible to students and to operational forecasters, but it also contains coverage of recent developments in the area. The authors of each chapter are experts in their fields and are well aware of the needs and constraints of operational forecasting, as well as being involved in research into new and improved methods of verification. The audience for the book is not restricted to atmospheric scientists – there is discussion in several chapters of similar ideas in other disciplines. For example, ROC curves (Chapter 3) are widely used in medical applications, and the ideas of Chapter 8 are particularly relevant to finance and economics.

To our knowledge there is currently no other book that gives a comprehensive and up-to-date coverage of forecast verification. For many years, the WMO publication by Stanski *et al.* (1989), and its earlier versions, was the standard reference for atmospheric scientists, though largely unknown in other disciplines.

Its drawback is that it is somewhat limited in scope and is now rather out-of-date. Wilks (1995, Chapter 7) and von Storch and Zweirs (1999, Chapter 18) are more recent but, inevitably as each comprises only one chapter in a book, are far from comprehensive. Katz and Murphy (1997a) discuss forecast verification in some detail, but mainly from the limited perspective of economic value. The current book provides a broad coverage, although it does not attempt to be encyclopaedic, leaving the reader to look in the references for more technical material.

Chapters 1 and 2 of the book are both introductory. Chapter 1 gives a brief review of the history and current practice in forecast verification, gives some definitions of basic concepts such as skill and value, and discusses the benefits and practical considerations associated with forecast verification. Chapter 2 describes a number of informal descriptive ways, both graphical and numerical, of comparing forecasts and corresponding observed data. It then establishes some theoretical groundwork that is used in later chapters, by defining and discussing the joint probability distribution of the forecasts and observed data. Consideration of this joint distribution and its decomposition into conditional and marginal distributions leads to a number of fundamental properties of forecasts. These are defined, as are the ideas of accuracy, association and skill.

Both Chapters 1 and 2 discuss the different types of data that may be forecast, and each of the next five chapters then concentrates on just one type. The subject of Chapter 3 is binary data in which the variable to be forecast has only two values, for example, {Rain, No Rain}, {Frost, No Frost}. Although this is apparently the simplest type of forecast, there have been many suggestions of how to assess them, in particular many different verification measures have been proposed. These are fully discussed, along with their properties. One particularly promising approach is based on signal detection theory and the ROC curve.

For binary data one of two categories is forecast. Chapter 4 deals with the case in which the data are again categorical, but where there are more than two categories. A number of skill scores for such data are described, their properties are discussed, and recommendations are made.

Chapter 5 is concerned with forecasts of continuous variables such as temperature. Mean square error and correlation are the best-known verification measures for such variables, but other measures are also discussed including some based on comparing probability distributions.

Atmospheric data often consist of spatial fields of some meteorological variable observed across some geographical region. Chapter 6 deals with verification for such spatial data. Many of the verification measures described in Chapter 5 are also used in the spatial context, but the correlation due to spatial proximity causes complications. Some of these complications, together with verification measures that have been developed with spatial correlation in mind, are discussed in Chapter 6.

Probability plays a key role in Chapter 7, which covers two topics. The first is forecasts that are actually probabilities. For example, instead of a deterministic forecast of ‘Rain’ or ‘No Rain’, the event ‘Rain’ may be forecast to occur with probability 0.2. One way in which such probabilities can be produced is to generate an ensemble of forecasts, rather than a single forecast. The continuing increase of computing power has made larger ensembles of forecasts feasible, and ensembles of weather and climate forecasts are now routinely produced. Both ensemble and

probability forecasts have their own peculiarities that necessitate different, but linked, approaches to verification. Chapter 7 describes these approaches.

The discussion of verification for different types of data in Chapters 3–7 is largely in terms of mathematical and statistical properties, albeit properties that are defined with important practical considerations in mind. There is little mention of cost or value – this is the topic of Chapter 8. Much of the chapter is concerned with the simple cost-loss model, which is relevant for binary forecasts. These forecasts may be either deterministic as in Chapter 3, or probabilistic as in Chapter 7. Chapter 8 explains some of the interesting relationships between economic value and skill scores.

The final chapter (9) reviews some of the key concepts that arise elsewhere in the book. It also summarizes those aspects of forecast verification that have received most attention in other disciplines, including Statistics, Finance and Economics, Medicine, and areas of Environmental and Earth Science other than Meteorology and Climatology. Finally, the chapter discusses some of the most important topics in the field that are the subject of current research or that would benefit from future research.

This book has benefited from discussions and help from many people. In particular, as well as our authors, we would like to thank the following colleagues for their particularly helpful comments and contributions: Harold Brook, Barbara Casati, Martin Goeber, Mike Harrison, Rick Katz, Simon Mason, Buruhani Nyenzi and Dan Wilks. Some of the earlier work on this book was carried while one of us (I.T. Jolliffe) was on research leave at the Bureau of Meteorology Research Centre (BMRC) in Melbourne. He is grateful to BMRC and its staff, especially Neville Nicholls, for the supportive environment and useful discussions; to the Leverhulme Trust for funding the visit under a Study Abroad Fellowship; and to the University of Aberdeen for granting the leave.

Looking to the future, we would be delighted to receive any feedback comments from you, the reader, concerning material in this book, in order that improvements can be made in future editions (see www.met.rdg.ac.uk/cag/forecasting).

1 Introduction

IAN T. JOLLIFFE[1] AND DAVID B. STEPHENSON[2]
[1]Department of Mathematical Sciences, University of Aberdeen, Aberdeen, UK
[2]Department of Meteorology, University of Reading, Reading, UK

Forecasts are almost always made and used in the belief that having a forecast available is preferable to remaining in complete ignorance about the future event of interest. It is important to test this belief *a posteriori* by assessing how skilful or valuable was the forecast. This is the topic of *forecast verification* covered in this book, although, as will be seen, words such as 'skill' and 'value' have fairly precise meanings and should not be used interchangeably. This introductory chapter begins, in Section 1.1, with a brief history of forecast verification, followed by an indication of current practice. It then discusses the reasons for, and benefits of, verification (Section 1.2). Section 1.3 provides a brief review of types of forecasts, and the related question of the target audience for a verification procedure. This leads on to the question of skill or value (Section 1.4), and the chapter concludes, in Section 1.5, with some discussion of practical issues such as data quality.

1.1 A BRIEF HISTORY AND CURRENT PRACTICE

Forecasts are made in a wide range of diverse disciplines. Weather and climate forecasting, economic and financial forecasting, sporting events and medical epidemics are some of the most obvious examples. Although much of the book is relevant across disciplines, many of the techniques for verification have been developed in the context of weather, and latterly climate, forecasting. For this reason the present section is restricted to those areas.

1.1.1 History

The paper that is most commonly cited as the starting point for weather forecast verification is Finley (1884). Murphy (1996) notes that although

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

operational weather forecasting started in the USA and Western Europe in the 1850s, and that questions were soon asked about the quality of the forecasts, no formal attempts at verification seem to have been made before the 1880s. He also notes that a paper by Köppen (1884), in the same year as Finley's paper, addresses the same binary forecast set-up as Finley (see Table 1.1), though in a different context.

Finley's paper deals with a fairly simple example, but it nevertheless has a number of subtleties and will be used in this and later chapters to illustrate a number of facets of forecast verification. The data set consists of forecasts of whether or not a tornado will occur. The forecasts were made from 10th March until the end of May 1884, twice daily, for 18 districts of the USA east of the Rockies. Table 1.1 summarizes the results in a table, known as a (2×2) contingency table (see Chapter 3). Table 1.1 shows that a total of 2803 forecasts were made, of which 100 forecast 'Tornado'. On 51 occasions tornados were observed, and on 28 of these 'Tornado' was also forecast. Finley's paper initiated a flurry of interest in verification, especially for binary (0–1) forecasts, and resulted in a number of published papers during the following 10 years. This work is reviewed by Murphy (1996).

Forecast verification was not a very active branch of research in the first half of the 20th century. A 3-part review of verification for short-range weather forecasts by Muller (1944) identified only 55 articles 'of sufficient importance to warrant summarization', and only 66 were found in total. Twenty-seven of the 55 appeared before 1913. Due to the advent of numerical weather forecasting, a large expansion of weather forecast products occurred from the 1950s onwards, and this was accompanied by a corresponding research effort into how to evaluate the wider range of forecasts being made.

For the (2×2) table of Finley's results, there is a surprisingly large number of ways in which the numbers in the four cells of the table can be combined to give measures of the quality of the forecasts. What they all have in common is that they use the joint probability distribution of the forecast event and observed event. In a landmark paper, Murphy and Winkler (1987) established a general framework for forecast verification based on such joint distributions. Their framework goes well beyond the

Table 1.1 Finley's Tornado forecasts

Forecast	Observed		
	Tornado	No Tornado	Total
Tornado	28	72	100
No Tornado	23	2680	2703
Total	51	2752	2803

(2×2) table, and encompasses data with more than two categories, discrete and continuous data and multivariate data. The forecasts can take any of these forms, but can also be in the form of probabilities.

The late Allan Murphy had a major impact on the theory and practice of forecast verification. As well as Murphy and Winkler (1987) and numerous technical contributions, two further general papers of his are worthy of mention here. Murphy (1991) discusses the complexity and dimensionality of forecast verification and Murphy (1993) is an essay on what constitutes a 'good' forecast.

Weather and climate forecasting is necessarily an international activity. The World Meteorological Organization (WMO) published a 114-page technical report (Stanski *et al.* 1989) which gave a comprehensive survey of forecast verification methods in use in the late 1980s.

1.1.2 Current Practice

Today the WMO provides a Standard Verification System for Long-Range Forecasts. This was published in February 2000 by the Commission for Basic Systems of the WMO, and at the time of writing is available at http://www.wmo.ch/web/www/DPS/SVS-for-LRF.html. The document is very thorough and careful in its definitions of long-range forecasts, verification areas (geographical) and verification data sets. It describes recommended verification strategies and verification scores, and is intended to facilitate the exchange of comparable verification scores between different centres – for related material see also http://www.wmo.ch and find Forecast Verification Systems under Search by Alphabetical Topics.

At a national level, a WMO global survey in 1997 (see WMO's general guidance regarding verification cited at the end of this section) found that 57% of National Meteorological Services had formal verification programmes. This, of course, raises the question of why the other 43% did not. Practices vary between different national services, and most use a range of different verification strategies for different purposes. For example, verification scores used by the Bureau of Meteorology in Australia range from LEPS scores (see Chapter 4) for climate forecasts, to mean square errors and S1 skill scores (Chapter 6) for short-term forecasts of spatial fields. Numbers of forecasts with absolute error less than a threshold, and even some subjective verification techniques, are also used.

There is a constant need to adapt practices, as forecasts, data and users all change. An increasing number of variables can be, and are, forecast, and the nature of forecasts is also changing. At one end of the range there is increasing complexity. Ensembles of forecasts, which were largely infeasible 20 years ago, are now commonplace. At the other extreme, a wider range of users requires targeted, but often simple (at least to express), forecasts. The nature of the data available with which to verify the forecasts is also

evolving with increasingly sophisticated remote sensing by satellite and radar, for example.

As well as its Standard Verification Systems, the WMO also provides, at the time of writing, general guidance regarding verification on its website (go to http://www.wmo.ch and find Forecast Verification under Search by Alphabetical Topics). The remainder of this chapter draws on that source.

1.2 REASONS FOR FORECAST VERIFICATION AND ITS BENEFITS

There is a fairly widely used three-way classification of the reasons for verification, which dates back to Brier and Allen (1951), and which can be described by the headings *administrative*, *scientific* and *economic*. Naturally, no classification is perfect and there is overlap between the three categories. A common important theme for all three is that any verification scheme should be *informative*. It should be chosen to answer the questions of interest and not simply for reasons of convenience.

From an administrative point of view, there is a need to have some numerical measure of how well forecasts are performing. Otherwise, there is no objective way to judge how changes in training, equipment or forecasting models, for example, affect the quality of forecasts. For this purpose, a small number of overall measures of forecast performance is usually desired. As well as measuring improvements over time of the forecasts, the scores produced by the verification system can be used to justify funding for improved training and equipment and for research into better forecasting models. More generally they can guide strategy for future investment of resources in forecasting.

Measures of forecast quality may even be used by administrators to reward forecasters financially. For example, the UK Meteorological Office currently operates a corporate bonus scheme, several elements of which are based on the quality of forecasts. The formula for calculating the bonus payable is complex, and involves meeting or exceeding targets for a wide variety of meteorological variables around the UK and globally. Variables contributing to the scheme range from mean sea level pressure, through precipitation, temperature and several others, to gale warnings.

The scientific viewpoint is concerned more with *understanding*, and hence improving the forecast system. A detailed assessment of the strengths and weaknesses of a set of forecasts usually requires more than one or two summary scores. A larger investment in more complex verification schemes will be rewarded with a greater appreciation of exactly where the deficiencies in the forecast lie, and with it the possibility of improved understanding of the physical processes which are being forecast. Sometimes there are unsuspected biases in either the forecasting models, or in the forecasters'

interpretations, or both, which only become apparent when more sophisticated verification schemes are used. Identification of such biases can lead to research being targeted to improve knowledge of why they occur. This, in turn, can lead to improved scientific understanding of the underlying processes, to improved models, and eventually to improved forecasts.

The administrative use of forecast verification certainly involves financial considerations, but the third, 'economic', use is usually taken to mean something closer to the users of the forecasts. Whilst verification schemes in this case should be kept as simple as possible in terms of communicating their results to users, complexity arises because different users have different interests. Hence, there is the need for different verification schemes tailored to each user. For example, seasonal forecasts of summer rainfall may be of interest to both a farmer, and to an insurance company covering risks of event cancellations due to wet weather. However, different aspects of the forecast are relevant to each. The farmer will be interested in total rainfall, and its distribution across the season, whereas the insurance company's concern is mainly restricted to information on the likely number of wet weekends.

As another example, consider a daily forecast of temperature in winter. The actual temperature is relevant to an electricity company, as demand for electricity varies with temperature in a fairly smooth manner. On the other hand, a local roads authority is concerned with the value of the temperature relative to some *threshold*, below which it should treat the roads to prevent ice formation. In both examples, a forecast that is seen as reasonably good by one user may be deemed 'poor' by the other. The economic view of forecast verification needs to take into account the economic factors underlying the users' needs for forecasts when devising a verification scheme. This is sometimes known as 'customer-based' verification, as it provides information in terms more likely to be understood by the 'customer' than a purely 'scientific' approach. Forecast verification using economic value is discussed in detail in Chapter 8. Another aspect of forecasting for specific users is the extent to which users prefer a simple, less informative, forecast to one which is more informative (for instance, a probability forecast) but less easy to interpret. Some users may be uncomfortable with probability forecasts, but there is evidence (H. Brooks, personal communication) that *probabilities* of severe weather events such as hail or tornados are preferred to crude *categorizations* such as {Low Risk, Medium Risk, High Risk}. Customer-based verification should attempt to ascertain such preferences for the 'customer' at hand.

At the time of writing, the WMO web page noted in Section 1.1 lists nine 'benefits' of forecast verification. Most of these amplify points made above in discussing the reasons for verification. One benefit common to all three classes of verification, if it is informative, is that it gives the administrator, scientist or user concrete information on the quality of forecasts that can be used to make rational decisions. The WMO list of benefits, and indeed this

section as a whole, is based on experience gained of verification in the context of forecasts issued by National Meteorological Services. However, virtually all the points made are highly relevant for forecasts issued by private companies, and in other subject domains.

1.3 TYPES OF FORECASTS AND VERIFICATION DATA

The wide range of forecasts has already been noted in the Preface when introducing the individual chapters. At one extreme, forecasts may be binary (0–1), as in Finley's tornado forecasts; at the other extreme, ensembles of forecasts will include predictions of several different weather variables at different times, different spatial locations, different vertical levels of the atmosphere, and not just one forecast but a whole ensemble. Such forecasts are extremely difficult to verify in a comprehensive manner but, as will be seen in Chapter 3, even the verification of binary forecasts can be a far from trivial problem.

Some other types of forecast are difficult to verify, not because of their sophistication, but because of their vagueness. Wordy or descriptive forecasts are of this type. Verification of forecasts such as 'turning milder later' or 'sunny with scattered showers in the south at first' is bound to be subjective (see Jolliffe and Jolliffe, 1997), whereas in most circumstances it is highly desirable for a verification scheme to be objective. In order for this to happen it must be clear what is being forecast, and the verification process should ideally reflect the forecast precisely. As a simple example, consider Finley's tornado forecasts. The forecasts are said to be of occurrence or non-occurrence of tornados in 18 districts, or sub-divisions of these districts, of the USA. However, the verification is done on the basis of whether a funnel cloud is seen at a reporting station within the district (or sub-division) of interest. There were 800 observing stations, but given the vast size of the 18 districts, this is a fairly sparse network. It is quite possible for a tornado to appear in a district sufficiently distant from the reporting stations for it to be missed. To match up forecast and verification, it is necessary to interpret the forecast not as 'a tornado will occur in a given district', but as 'a funnel cloud will occur within sight of an reporting station in the district'.

As well as an increase in the types of forecasts available, there have also been changes in the amount and nature of data available for verifying forecasts. The changes in data include changes of observing stations, changes of location and type of recording instruments at a station, and an increasing range of remotely sensed data from satellites, radar or automatic recording devices. It is tempting, and often sensible, to use the most up-to-date types of data available for verification, but in a sequence of similar forecasts it is important to be certain that any apparent changes in forecast quality are not simply due to changes in the nature of the data used for

verification. For example, suppose that a forecast of rainfall for a region is to be verified, and that there is an unavoidable change in the set of stations used for verification. If the mean or variability of rainfall is different for the new set of stations, compared to the old, such differences can affect many of the scores used for verification.

Another example occurs in the seasonal forecasting of numbers of tropical cyclones. There is evidence that access to a wider range of satellite imagery has led to re-definitions of cyclones over the years (Nicholls 1992). Hence, apparent trends in cyclone frequency may be due to changes of definition, rather than to genuine climatic trends. This, in turn, makes it difficult to know whether changes in forecasting methods have resulted in improvements to the quality of forecasts. Apparent gains can be confounded by the fact that the 'target' which is being forecast has moved; changes in definition alone may lead to changed verification scores.

As noted in the previous section, the idea of matching verification data to forecasts is relevant when considering the needs of a particular user. A user who is interested only in the position of a continuous variable relative to a threshold requires verification data and procedures geared to binary data (above/below threshold), rather than verification of the actual forecast value of the variable.

1.4 SCORES, SKILL AND VALUE

For a given type of data, it is easy enough to construct a numerical score that measures the relative quality of different forecasts. Indeed, there is usually a whole range of possible scores. Any set of forecasts can then be ranked as best, second best, . . . , worst, according to a chosen score, though the ranking need not be the same for different choices of score. Two questions then arise:

- How to choose which scores to use?
- How to assess the absolute, rather than relative, quality of a forecast?

In addressing the first of these questions, attempts have been made to define desirable properties of potential scores. Many of these will be discussed in Chapters 2 and 3. The general framework of Murphy and Winkler (1987) allows different 'attributes' of forecasts, such as *reliability*, *resolution*, *discrimination* and *sharpness* to be examined. Which of these attributes is most important to the scientist, administrator or end-user will determine which scores are preferred. Most scores have some strengths, but all have weaknesses, and in most circumstances more than one score is needed to obtain an informed picture of the relative merits of the forecasts.

'Goodness', like beauty, can be in the eye of the beholder, and has many facets. Murphy (1993) identifies three types of goodness:

- consistency,
- quality (also known as accuracy or skill) and
- value (utility).

Value is concerned with economic worth to the user, whereas quality is the correspondence between forecast and observations. The emphasis in this book is on quality, although Chapter 8 discusses value and its relationship to quality. Some of the 'attributes' mentioned in the last paragraph can be used to measure quality as well as to choose between scores.

Consistency is achieved when the forecaster's best judgment and the forecast actually issued coincide. The choice of verification scheme can influence whether or not this happens. Some schemes have scores for which a forecaster knows that he or she will score better on average if the forecast made differs (perhaps is closer to the long-term average or climatology of the quantity being forecast) than his or her best judgment of what will occur. Such scoring systems are called *improper* and should be avoided. In particular, administrators should avoid measuring or rewarding forecasters' performance on the basis of improper scoring schemes, as this is likely to lead to biases in the forecasts.

1.4.1 Skill Scores

Turning to the matter of how to quantify the quality of a forecast, it is usually necessary to define a baseline against which a forecast can be judged. Much of the published discussion following Finley's (1884) paper was driven by the fact that although the forecasts were correct on $2708/2803 = 96.6\%$ of occasions, it is possible to do even better by always forecasting 'No Tornado', if forecast performance is measured by the percentage of correct forecasts. This alternative unskilful forecast has a success rate of $2752/2803 = 98.2\%$. It is therefore usual to measure the performance of forecasts relative to some 'unskilful' or reference forecast. Such relative measures are known as *skill scores*, and are discussed further in several of the later chapters – see, in particular, Sections 2.7, 3.2 and 4.3.

There are several baseline or reference forecasts that can be chosen. One is the average, or expected, score obtained by issuing forecasts according to a random mechanism. What this means is that a probability distribution is assigned to the possible values of the variable(s) to be forecast, and a sequence of forecasts is produced by taking a sequence of independent values from that distribution. A limiting case of this, when all but one of the probabilities is zero, is the (deterministic) choice of the same forecast on every occasion, as when 'No Tornado' is forecast all the time.

Climatology is a second common baseline. This refers to always forecasting the 'average' of the quantity of interest. 'Average' in this context usually

refers to the mean value over some recent reference period, typically of 30 years length.

A third baseline that may be appropriate is 'persistence'. This is a forecast in which whatever is observed at the present time is forecast to persist into the forecast period. For short-range forecasts this strategy is often successful, and to demonstrate real forecasting skill, a less naïve forecasting system must do better.

1.4.2 Artificial Skill

Often when a particular data set is used in developing a forecasting system, the quality of the system is then assessed on the same data set. This will invariably lead to an optimistic bias in skill scores. This inflation of skill is sometimes known as 'artificial skill', and is a particular problem if the score itself has been used directly or indirectly in calibrating the forecasting system. To avoid such biases, an ideal solution is to assess the system using only forecasts of events that have not yet occurred. This may be feasible for short-range forecasts, where data accumulate rapidly, but for long-range forecasts it may be a long time before there are sufficient data for reliable verification. In the meantime, while data are accumulating, any potential improvements to the forecasting procedure should ideally be implemented in parallel to, and not as a replacement for, the old procedure.

The next best solution for reducing artificial skill is to divide the data into two non-overlapping, exhaustive subsets, the *training set* and the *test set*. The training set is used to formulate the forecasting procedure, while the procedure is verified on the test set. Some would argue that, even though the training and test sets are non-overlapping, and the observed data in the test set are not used directly in formulating the forecasting rules, the fact that the observed data for both sets already exist when the rules are formulated has the potential to bias any verification results. A more practical disadvantage of the test/training set approach is that only part of the data set is used to construct the forecasting system. The remainder is, in a sense, wasted because, in general, increasing the amount of data or information used to construct a forecast will provide a better forecast. To partially overcome this problem, the idea of *cross-validation* can be used.

Cross-validation has a number of variations on the same basic theme. It has been in use for many years (see, for example, Stone 1974) but has become practicable for larger problems as computer power has increased. Suppose that the complete data set consists of n forecasts, and corresponding observations. In cross-validation the data are divided into m subsets, and for each subset a forecasting rule is constructed based on data from the other $(m - 1)$ subsets. The rule is then verified on the subset omitted from the construction procedure, and this is repeated for each of the m subsets in turn. The verification scores for each subset are then combined to give an

overall measure of quality. The case $m = 2$ corresponds to repeating the test/training set approach with the roles of test and training sets reversed, and then combining the results from the two analyses. At the opposite extreme, a commonly used special case is where $m = n$, so that each individual forecast is based on a rule constructed from all the other $(n - 1)$ observations.

The word 'hindcast' (sometimes 'backcast') is in fairly common use. Unfortunately, it has different meanings to different authors and none of the standard meteorological encyclopaedias or glossaries gives a definition. The cross-validation scheme just mentioned bases its 'forecasts' on $(n - 1)$ observations, some of which are 'in the future' relative to the observation being predicted. Sometimes the word 'hindcast' is restricted to mean predictions like this in which 'future', as well as past, observations are used to construct forecasting procedures. However, more commonly the term includes any prediction made which is not a genuine forecast of a *future* event. With this usage, a prediction for the year 2000 must be a hindcast, even if it is only based on data up to 1999, because year 2000 is now over. There seems to be increasing usage of the term *retroactive forecasting* (see, for example, Mason and Mimmack 2002) to denote the form of hindcasting in which forecasts are made for past years (for example, 2000–2001) using data prior to those years (perhaps 1970–1999).

The terminology *ex ante* and *ex post* is used in economic forecasting. Ex ante means a prediction into the future before the events occur (a genuine *fore*cast), whereas ex post means predictions for historical periods for which verification data are already available at the time of forecast. The latter is therefore a form of hindcasting.

1.4.3 Statistical Significance

There is one further aspect of measuring the absolute quality of a forecast. Having decided on a suitable baseline from which to measure skill, checked that the skill score chosen has no blatantly undesirable properties, and removed the likelihood of artificial skill, is it possible to judge whether an observed improvement over the baseline is statistically significant? Could the improvement have arisen by chance? Ideas from statistical inference, namely, hypothesis testing and confidence intervals, are needed to address this question. Confidence intervals based on a number of measures or scores that reduce to proportions are described in Chapter 3, and Section 4.4, Chapter 5 and Section 6.2 all discuss tests of hypotheses in various contexts. A difficulty that arises is that many standard procedures for confidence intervals and tests of hypothesis assume independence of observations. The temporal and spatial correlation that is often present in environmental data means that adaptations to the usual procedures are necessary – see Sections 4.4 and 6.2.

1.4.4 Value Added

For the user, a measure of value is often more important than a measure of skill. Again, the value should be measured relative to a baseline. It is the *value added*, compared to an unskilful forecast, which is of real interest. The definition of 'unskilful' can refer to one of the reference or baseline forecasts described earlier for scores. Alternatively, for a situation with a finite number of choices for a decision (for example, protect or do not protect a crop from frost), the baseline can be the best from the list of decision choices ignoring any forecast (for example, always protect or never protect regardless of the forecast). The avoidance of artificially inflated value, and assessing whether the 'value added' is statistically significant are relevant to value, as much as to skill. Although individual users should be interested in value added, in some cases they are more comfortable with very simple scores such as 'percentage correct', regardless of how genuinely informative such naïve measures are.

1.5 DATA QUALITY AND OTHER PRACTICAL CONSIDERATIONS

Changes in the data available for verification have already been mentioned in Section 1.3, but it was implicitly assumed there that the data are of high quality. This is not always the case. National Meteorological Services will, in general, have quality control procedures in place that detect many errors, but larger volumes of data make it more likely that some erroneous data will slip through the net. A greater reliance on data that are indirectly derived via some calibration step, for example, rainfall intensities deduced from radar data, also increases the scope for biases in the inferred data.

When verification data are incorrect, the forecast is verified against something other than the truth, with unpredictable consequences for the verification scores. Work on discriminant analysis in the presence of misclassification (see McLachlan 1992, Section 2.5; Huberty 1994, Section XX-4) is relevant in the case of binary forecasts.

In large data sets, missing data have always been commonplace, for a variety of reasons. Even Finley (1884) suffered from this, stating that '... from many localities [no reports] will be received except, perhaps, at a very late day'. Missing data can be dealt with either by ignoring them, and not attempting to verify the corresponding forecast, or by estimating them from related data and then verifying using the estimated data. The latter is preferable if good estimates are available, because it avoids throwing away information, but if the estimates are poor, the resulting verification scores can be misleading.

Data may be missing at random, or in some non-random manner, in which particular values of the variable(s) being forecast are more prone to

be absent than others. For randomly missing data the mean verification score is likely to be relatively unaffected by the existence of the missing data, though the variability of the score will usually increase. For data that are missing in a more systematic way, the verification scores can be biased, as well as again having increased variability.

One special, but common, type of missing data occurs when measurements of the variables of interest have not been collected for long enough to establish a reliable climatology for them. This is a particular problem when extremes are forecast. By their very nature, extremes occur rarely and long data records are needed to deduce their nature and frequency. Forecasts of extremes are of increasing interest, partly because of the disproportionate financial and social impacts caused by extreme weather, but also in connection with the large amount of research effort devoted to climate change.

It is desirable for a data set to include some extreme values so that full coverage of the range of possible observations is achieved. On the other hand, a small number of extreme values can have undue influence on the values of some types of skill measure, and mask the quality of forecasts for non-extreme values. To avoid this, measures need to be robust or resistant to the presence of extreme observations or forecasts.

The WMO web page noted in Section 1.1 gives useful practical information on verification, including sections on characteristics of verification schemes, 'guiding principles', selection of forecasts for verification, data collection and quality control, scoring systems and the use of verification results. Many of the points made there have been touched on in this chapter, but to conclude the chapter two more are noted:

- Forecasts that span periods of time and/or geographical regions in a continuous manner are more difficult to verify than forecasts at discrete time/space combinations, because observations are usually in the latter form.
- Subjective verification should be avoided if at all possible, but if the data are sparse, there may only be a choice between subjective verification or none at all. In this case it can be the lesser of two evils.

2 Basic Concepts

JACQUELINE M. POTTS
Biomathematics and Statistics Scotland, The Macaulay Institute, Aberdeen, UK

2.1 INTRODUCTION

Forecast verification involves exploring and summarising the relationship between sets of forecast and observed data and making comparisons between the performance of forecasting systems and that of reference forecasts. Verification is therefore a statistical problem. This chapter introduces some of the basic statistical concepts and definitions that will be used in later chapters. Further details about the use of statistical methods in the atmospheric sciences can be found in Wilks (1995) and von Storch and Zwiers (1999).

2.2 TYPES OF PREDICTAND

The variable for which the forecasts are formulated is known as the *predictand*. A *continuous* predictand is one for which, within the limits over which the variable ranges, any value is possible. This means that between any two different values, there are an infinite number of possible values. For discrete variables, on the other hand, we can list all possible values. Variables such as pressure, temperature or rainfall are theoretically continuous. In reality, however, such variables are actually discrete because measuring devices have limited reading accuracy and variables are usually recorded to a fixed number of decimal places. *Categorical* predictands are discrete variables that can only take one of a finite set of predefined values. If the categories provide a ranking of the data, the variable is *ordinal*; for example, cloud cover is often measured in oktas. On the other hand, cloud type is a *nominal* variable since there is no natural ordering of the categories. The simplest kind of categorical variable is a *binary* variable, which has only two possible values, indicating, for example, the presence or absence of some condition such as rain, fog or thunder.

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

Forecasts of categorical predictands may be *deterministic* (e.g. *rain tomorrow*) or *probabilistic* (e.g. *70% chance of rain tomorrow*). A deterministic forecast is really just a special case of a probabilistic forecast in which a probability of unity is assigned to one of the categories and zero to the others.

Forecasts are made at different temporal and spatial scales. A very short-range forecast may cover the next 12 h, whereas long-range forecasts are issued from 30 days to 2 years ahead and may be forecasts of the mean value of a variable over a month or an entire season. Prediction models often produce forecasts of spatial fields, usually defined by values of a variable at many points on a regular grid. These vary both in their geographical extent and in the distance between grid points within that area. Meteorological data are autocorrelated in both space and time. At a given location, the correlation between observations a day apart will usually be greater than that between observations separated by longer time intervals. Similarly, at a given time, the correlation between observations at grid points that are close together will generally be greater than between those that are further apart, although teleconnection patterns such as the North Atlantic Oscillation can lead to correlation between weather patterns in areas that are separated by vast distances.

Both temporal and spatial autocorrelation have implications for forecast verification. Temporal autocorrelation means that for some types of short-range forecast, persistence often performs quite well when compared to a forecast of the climatological average. A specific user may be interested only in the quality of forecasts at a particular site, but meteorologists are often interested in evaluating the forecasting system in terms of its ability to predict the whole spatial field. The degree of spatial autocorrelation will affect the statistical distribution of the performance measures used. When spatial autocorrelation is present in both the observed and forecast fields it is likely that, if a forecast is fairly accurate at one grid point, it will also be fairly accurate at neighbouring grid points. Similarly, it is likely that if the forecast is not very accurate at one grid point, it will also not be very accurate at neighbouring grid points. Consequently, the significance of a particular value of a performance measure calculated over a spatial field will be quite different from its significance if it was calculated over the same number of independent forecasts.

2.3 EXPLORATORY METHODS

Exploratory methods should be used to examine the forecast and observed data graphically; further information about these techniques can be found in Tukey (1977); see also Wilks (1995, Chapter 3). For continuous variables boxplots provide a means of examining the location, spread and skewness of the forecasts and the observations. The box covers the *interquartile range*

(IQR) (the central 50% of the data), and the line across the centre of the box marks the *median* (the central observation). The *whiskers* attached to the box show the range of the data, from minimum to maximum. Boxplots are especially useful when several of them are placed side by side for comparison. Figure 2.1 shows boxplots of high-temperature forecasts for Oklahoma City made by the National Weather Service Forecast Office at Norman, Oklahoma. Outputs from three different forecasting systems are shown, together with the corresponding observations. These data were used in Brooks and Doswell (1996) and a full description of the forecasting systems can be found in that paper. In Fig. 2.1, the median of the observed data is 24 °C; 50% of the values lie between 14 and 31 °C; the minimum value is −8 °C and the maximum value is 39 °C. Sometimes a schematic boxplot is drawn, in which the whiskers extend only as far as the most extreme points inside the *fences*; outliers beyond this are drawn individually. The fences are at a distance of 1.5 times the IQR from the quartiles. Figure 2.2 shows boxplots of this type for forecasts and observations of winter temperature at 850 hPa over France from 1979/1980 to 1993/1994; these are the data used in the example given in Chapter 5 and are fully described there. These boxplots show that in this example the spread of the forecasts is considerably less than the spread of the observations. Notches may be drawn in each box to show approximate confidence intervals around the (sample) medians. If the notched intervals for two groups of data do not overlap, this suggests that the corresponding population medians are different. Figure 2.3 shows notched boxplots for the observed data used in Fig. 2.1 together with some artificial forecasts that were generated by adding a constant value to the actual forecasts. The notched intervals do not overlap, indicating a significant difference in the medians.

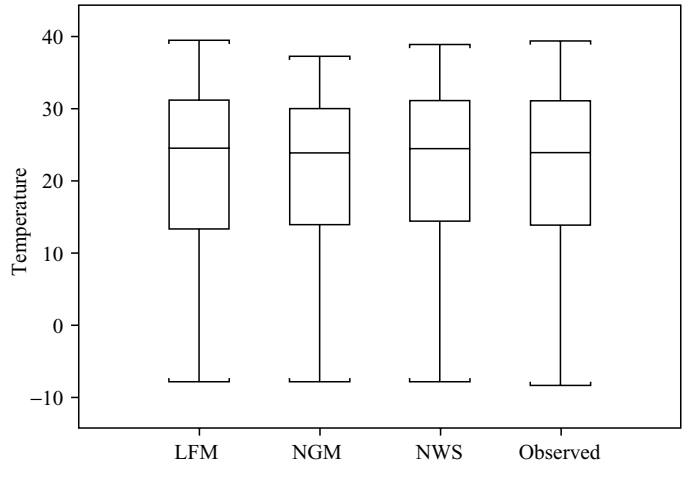


Figure 2.1 Boxplots of 12–24-h forecasts of high temperature (°C) for Oklahoma City from three forecasting systems and the corresponding observations

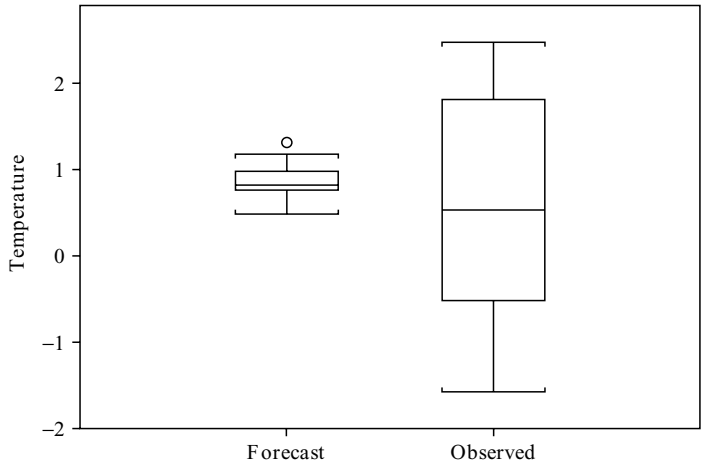


Figure 2.2 Boxplots of winter temperature (°C) forecasts at 850 hPa over France from 1979/1980 to 1993/1994 and the corresponding observations

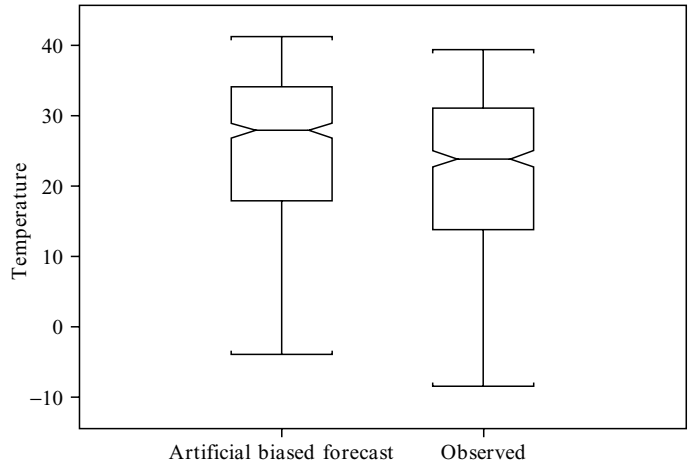


Figure 2.3 Notched boxplot of artificial biased forecasts of high temperature (°C) for Oklahoma City and the corresponding observations

Histograms and bar charts provide another useful way of comparing the distributions of the observations and forecasts. A bar chart indicating the frequency of occurrence of each category can be used to compare the distribution of forecasts and observations of categorical variables. Bar charts for Finley's tornado data, which were presented in Chapter 1, are shown in Fig. 2.4. In the case of continuous variables the values must be grouped into successive class intervals (bins) in order to produce a histogram. Figure. 2.5 shows histograms for the observations and one of the sets of forecasts used in Fig. 2.1. The appearance of the histogram may be

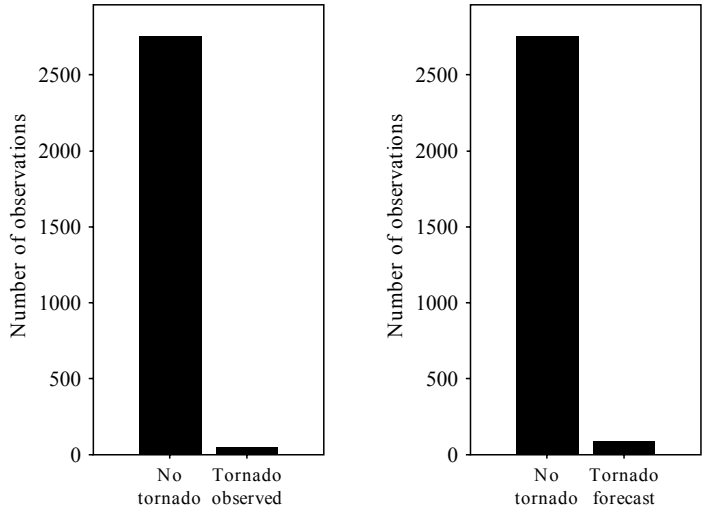


Figure 2.4 Bar charts of Finley's tornado data

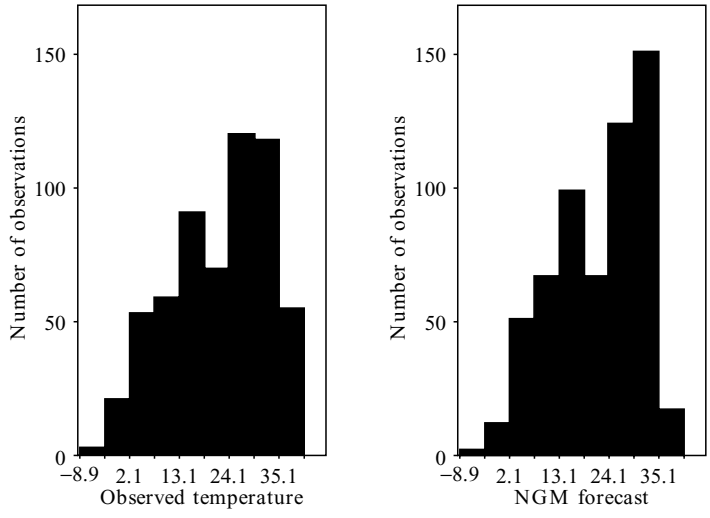


Figure 2.5 Histograms of observed high temperatures (°C) and 12–24-h forecasts for Oklahoma City

quite sensitive to the choice of bin width and anchor position. If the bin width is too small, the histogram reduces to a spike at each data point, but if it is too large, important features of the data may be hidden. Various rules for selecting the number of classes have been proposed, for example, by Sturges (1926) and Scott (1979).

Boxplots and histograms can indicate systematic problems with the forecast system. For example, the forecasts may tend to be close to the

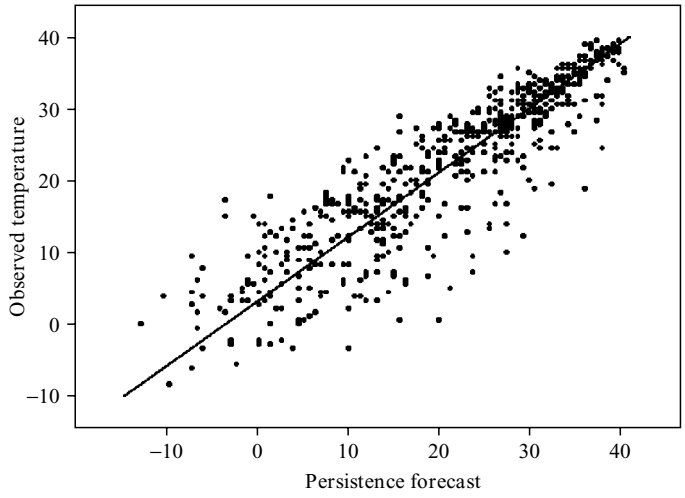


Figure 2.6 Scatterplot of observed high temperatures (°C) against persistence forecasts for Oklahoma City

climatological average with the consequence that the spread of the observations is much greater than the spread of the forecasts. Alternatively, the forecasts may be consistently too large or too small. However, the main concern of forecast verification is to examine the relationship between the forecasts and the observations. For continuous variables this can be done graphically by drawing a scatterplot. Figure 2.6 shows a scatterplot for persistence forecasts of the Oklahoma City high-temperature observations. If the forecasting system were perfect, all the points would lie on a straight line that starts at the origin and has a slope of unity. In Fig. 2.6, there is a fair amount of scatter about this line. Figure 2.7, which is the scatterplot for one of the actual sets of forecasts, shows a stronger linear relationship. Figure 2.8 shows the scatterplot for the artificial set of forecasts used in Fig. 2.3. There is still a linear relationship but the points do not lie on the line through the origin. Figure 2.9 shows the scatterplot for another set of forecasts that have been generated artificially, in this case by reducing the spread of the forecasts. The points again lie close to a straight line but the line does not have a slope of unity. In the case of categorical variables, a contingency table can be drawn up showing the frequency of occurrence of each combination of forecast and observed category. Table 1.1 showing Finley's tornado forecasts is an example of such a table. If the forecasting system were perfect all the entries apart from those on the diagonal of the table would be zero. The relationship between forecasts and observations of continuous or categorical variables may be examined by means of a bivariate histogram or bar chart. Figure 2.10 shows a bivariate histogram for the data used in Fig. 2.5.

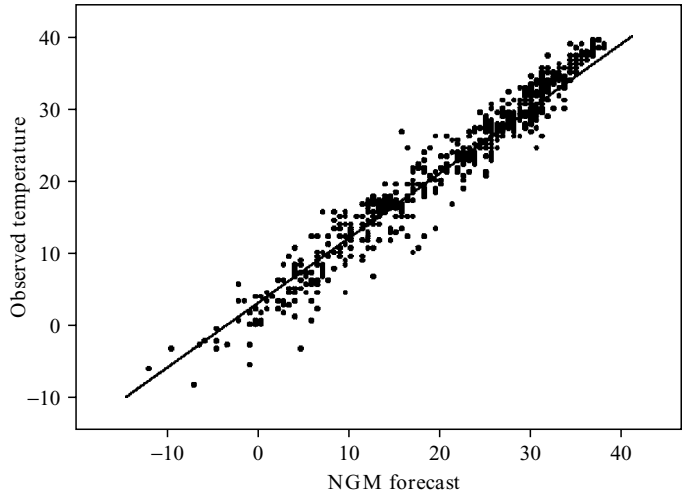


Figure 2.7 Scatterplot of observed high temperatures (°C) against 12–24-h forecasts for Oklahoma City

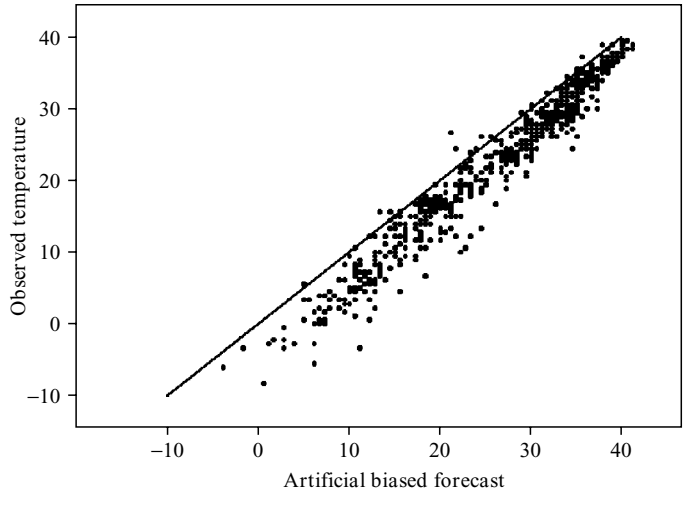


Figure 2.8 Scatterplot of observed high temperatures (°C) for Oklahoma City against artificial biased forecasts

2.4 NUMERICAL DESCRIPTIVE MEASURES

Boxplots and histograms provide a good visual means of examining the distribution of forecasts and observations. However, it is also useful to look at numerical summary statistics. Let $\hat{x}_1 \cdots \hat{x}_n$ denote the set of forecasts and $x_1 \cdots x_n$ denote the corresponding observations. The *sample mean* of the

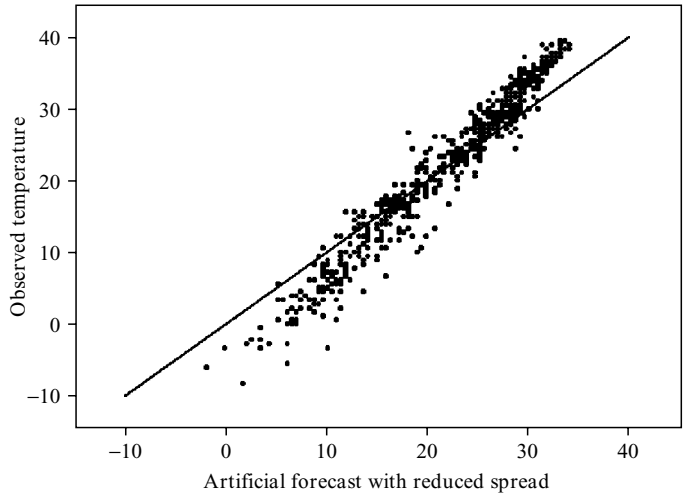


Figure 2.9 Scatterplot of observed high temperatures (°C) for Oklahoma City against artificial forecasts that have less spread than the observations

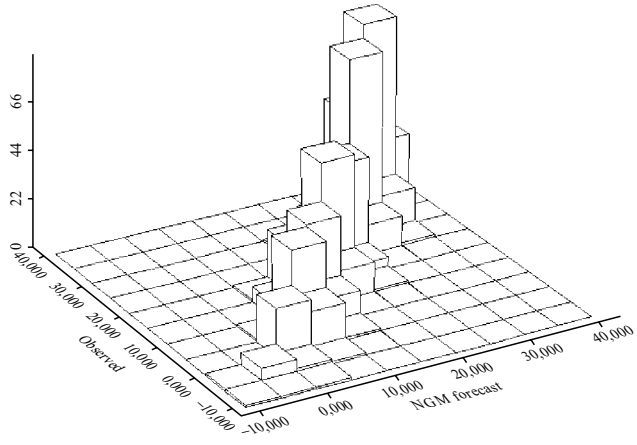


Figure 2.10 Bivariate histogram of observed high temperatures (°C) and 12–24-h forecasts for Oklahoma City

observations is simply the average of all the observed values. It is calculated from the formula

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2.1}$$

One aspect of forecast quality is the (unconditional) *bias*, which is the difference between the mean forecast $\bar{\hat{x}}$ and the mean observation $\bar{x}$. It is

desirable that the bias should be small. The forecasts in Fig. 2.3 have a bias of 4 °C.

The median is the central value; half of the observations are less than the median and half are greater. For a variable which has a reasonably symmetric distribution, the mean and the median will usually be fairly similar. In the case of the winter 850 hPa temperature observations in Fig. 2.2, the mean is 0.63 °C and the median is 0.64 °C. Rainfall, on the other hand, has a distribution that is positively *skewed*, which means that the distribution has a long right-hand tail. Daily rainfall has a particularly highly skewed distribution but even monthly averages display some skewness. For example, Fig. 2.11 is a histogram showing the distribution of monthly precipitation at Greenwich, UK, over the period 1841–1960. Positively skewed variables have a mean that is higher than the median. In the case of the data in Fig. 2.11 the mean is 51 mm but the median is only 46 mm. Other variables, such as atmospheric pressure, may be negatively skewed, which means that the distribution has a long left-hand tail. The difference between the mean and the median divided by the standard deviation (defined below) provides one measure of the skewness of the distribution. Another measure of the skewness of the observations, which is described in more detail in Chapter 5, is

$$\text{skewness} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^{\frac{3}{2}}} \tag{2.2}$$

If the data come from a normal (Gaussian) distribution (Wilks 1995, Section 4.4.2), then, provided the sample size is sufficiently large, the histogram should have approximately a symmetric bell-shaped form. For

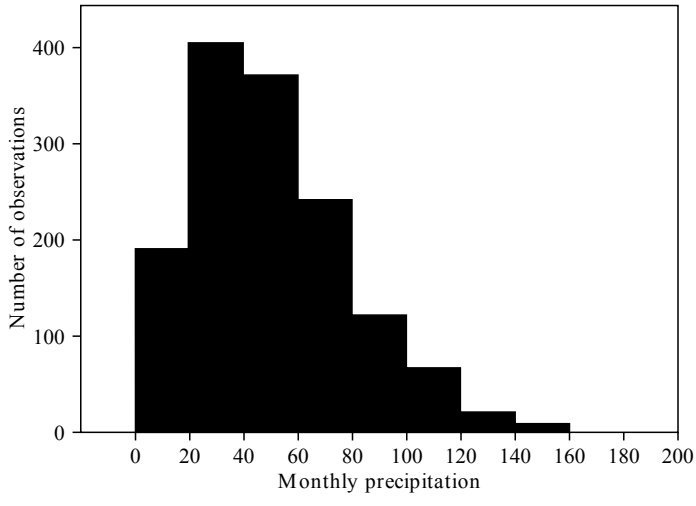


Figure 2.11 Histogram of monthly precipitation (mm) at Greenwich, UK, 1841–1960

normally distributed data, the sample mean has a number of optimal properties, but in situations where the distribution is asymmetric or otherwise non-normal, other measures such as the median may be more appropriate (Garthwaite *et al.* 2002, p. 15; DeGroot 1986, pp. 567–569). Measures that are not sensitive to particular assumptions about the distribution of the data are known as *robust* measures. The mean can be heavily influenced by any extreme values; so use of the median is also preferable if there are outliers. Measures that are not unduly influenced by a few outlying values are known as *resistant* measures. The median is more robust and resistant than the mean, but even it can sometimes display surprising sensitivity to small changes in the data (Jolliffe 1999).

The mean and median are not the only measures of the location of a data set (Wilks 1995, Section 3.2) but we now move on to consider the spread of the values. The *sample variance* of the observations is defined as

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{2.3}$$

The most commonly used measure of spread is the *standard deviation*, which is the square root of this quantity. The standard deviations for the 850 h Pa winter temperature data in Fig. 2.2 are 0.2 for the forecasts and 1.3 for the observations. A more robust measure of spread is the IQR, which is the difference between the upper and lower *quartiles*. If the data are sorted into ascending order, the lower quartile and upper quartiles are one quarter and three quarters of the way through the data, respectively. Like the median, the IQR is a measure that is resistant to the influence of extreme values and it may be a more appropriate measure than the standard deviation when the distribution is asymmetric. The Yule–Kendall index, which is the difference between the upper quartile minus the median and the median minus the lower quartile, divided by the IQR, provides a robust and resistant measure of skewness.

The median is the *quantile* for the proportion 0.5 and the lower and upper quartiles are the quantiles for the proportions 0.25 and 0.75. In general, the quantile for the proportion p, also known as the 100pth *percentile*, is the value that is 100p% of the way through the data when they are arranged in ascending order. Other quantiles in addition to the median and the quartiles may also be useful in assessing the statistical characteristics of the distributions of forecasts and observations. For example, Murphy *et al.* (1989) use the difference between the 90th percentile minus the median and the median minus the 10th percentile as a measure of the asymmetry of the distribution.

There are also summary statistics that can be used to describe the relationship between the forecasts and the observations. The *sample covariance* between the forecasts and observations is defined as

$$s_{\hat{x}x} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(\hat{x}_i - \bar{\hat{x}}) \tag{2.4}$$

The *sample correlation coefficient* can be obtained from the sample covariance and the sample variances using the definition

$$r_{\hat{x}x} = \frac{s_{\hat{x}x}}{\sqrt{s_{\hat{x}}^2 s_x^2}} \tag{2.5}$$

Further discussion of various forms of the correlation coefficient is given in Chapters 5 and 6.

2.5 PROBABILITY, RANDOM VARIABLES AND EXPECTATIONS

If observations of a categorical variable are made over a sufficiently long period of time, then the relative frequency of each event will tend to some limiting value, which is the *probability* of that event. For example, in Table 1.1 the relative frequency of the event 'tornado' is $51/2803 = 0.018$. The probability of a tornado occurring on any given day is therefore estimated to be 0.018. A *random variable*, denoted by X, associates a unique numerical value with each mutually exclusive event. For example, $X = 1$ if a tornado occurs and $X = 0$ if there is no tornado. A particular value of the random variable X is denoted by x. The *probability function* $p(x)$ of a discrete variable associates a probability with each of the possible values that can be taken by X. For example, in the case of the tornado data, the estimated probability function is $p(0) = 0.982$ and $p(1) = 0.018$. The sum of $p(x)$ over all possible values of x must by definition be unity.

In the case of continuous random variables the probability associated with any particular exact value is zero and positive probabilities can only be assigned to a range of values of X. The *probability density function* $f(x)$ for a continuous variable has the following properties:

$$f(x) \geq 0 \tag{2.6}$$

$$\int_a^b f(x)\,\mathrm{d}x = P(a \leq X \leq b) \tag{2.7}$$

where $P(a \leq X \leq b)$ denotes the probability that X lies in the interval from a to b; and

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1 \tag{2.8}$$

The *expectation* of a random variable X is given by

$$E[X] = \sum_x x\,p(x) \tag{2.9}$$

for discrete variables and by

$$E[X] = \int_{-\infty}^{\infty} x\,f(x)\,\mathrm{d}x \tag{2.10}$$

for continuous variables. In both cases $E[X]$ can be viewed as the 'long-run average' value of X, so the sample mean provides a natural estimate of $E[X]$.

The *variance* of X can be found from

$$\mathrm{var}(X) = E[(X - E[X])^2] \tag{2.11}$$

The sample variance, $s_x{}^2$, provides an unbiased estimate of var(X).

2.6 JOINT, MARGINAL AND CONDITIONAL DISTRIBUTIONS

In the case of discrete variables, the probability function for the *joint distribution* of the forecasts and observations $p(\hat{x}, x)$ gives the probability that the forecast $\hat{x}$ has a particular value and at the same time the observation x has a particular value. So in the case of the tornado forecasts: $p(1,1) = 0.010$, $p(1,0) = 0.026$, $p(0,1) = 0.008$ and $p(0,0) = 0.956$. The sum of $p(\hat{x}, x)$ over all possible values of $\hat{x}$ and x is by definition unity. In the case of continuous variables, the joint density function $f(\hat{x}, x)$ is a function with the following properties:

$$f(\hat{x},\ x) \geq 0 \tag{2.12}$$

$$\int_a^b \int_c^d f(\hat{x}, x)\ \mathrm{d}\hat{x}\ \mathrm{d}x = P(a \leq X \leq b \text{ and } c \leq \hat{X} \leq d) \tag{2.13}$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\hat{x}, x)\ \mathrm{d}\hat{x}\ \mathrm{d}x = 1 \tag{2.14}$$

The distributions with probability density functions $f(\hat{x})$ and $f(x)$, or probability functions $p(\hat{x})$ and $p(x)$ in the case of discrete random variables, are known as the *marginal distributions* of $\hat{X}$ and X, respectively. The marginal probability function $p(\hat{x})$ may be obtained by forming the sum of $p(\hat{x}, x)$ over all possible values of x. For example, in the case of the tornado forecasts

$$p(\hat{x}) = \begin{cases} p(1,1) + p(1,0) = 0.010 + 0.026 = 0.036 & \text{for } \hat{x} = 1 \\ p(0,1) + p(0,0) = 0.008 + 0.956 = 0.964 & \text{for } \hat{x} = 0 \end{cases} \tag{2.15}$$

Similarly,

$$p(x) = \sum_{\hat{x}} p(\hat{x}, x) \tag{2.16}$$

In the case of continuous variables

$$f(\hat{x}) = \int_{x} f(\hat{x}, x)\ \mathrm{d}x \tag{2.17}$$

and

$$f(x) = \int_{\hat{x}} f(\hat{x}, x)\ \mathrm{d}\hat{x} \tag{2.18}$$

The *conditional distribution*, which has probability function $p(x|\hat{x})$, gives the probability that the observation will assume a particular value x when a fixed value $\hat{x}$ has been forecast. The conditional probability function is given by the formula

$$p(x|\hat{x}) = \frac{p(\hat{x}, x)}{p(\hat{x})}. \tag{2.19}$$

A corresponding formula applies to continuous variables, with probability density functions replacing probability functions. In the case of the tornado data

$$p\,(x|\hat{X} = 1) = \begin{cases} 0.28 & \text{for } x = 1 \\ 0.72 & \text{for } x = 0 \end{cases} \tag{2.20}$$

and

$$p\,(x|\hat{X} = 0) = \begin{cases} 0.01 & \text{for } x = 1 \\ 0.99 & \text{for } x = 0 \end{cases} \tag{2.21}$$

Similarly, the probability function for the various forecast values, given that a fixed value has been observed, is given by the conditional probability function $p(\hat{x}|x)$. This function satisfies the equation

$$p\,(\hat{x}|x) = \frac{p(\hat{x}, x)}{p(x)} \tag{2.22}$$

So for the tornado data

$$p(\hat{x}|X=1) = \begin{cases} 0.55 & \text{for } \hat{x}=1 \\ 0.45 & \text{for } \hat{x}=0 \end{cases} \tag{2.23}$$

and

$$p(\hat{x}|X=0) = \begin{cases} 0.03 & \text{for } \hat{x}=1 \\ 0.97 & \text{for } \hat{x}=0 \end{cases} \tag{2.24}$$

The *conditional expectation* is the mean of the conditional distribution. In the case of discrete variables it is defined by

$$E[X|\hat{x}] = \sum_x xp(x|\hat{x}) \tag{2.25}$$

which is a function of $\hat{x}$ alone, and in the case of continuous variables by

$$E[X|\hat{x}] = \int_x xf(x|\hat{x})\ \mathrm{d}x \tag{2.26}$$

If $p(x|\hat{x}) = p(x)$ or $f(x|\hat{x}) = f(x)$, then the forecasts and observations are statistically *independent*. This would occur, for example, if forecasts were made at random, or if the climatological average value was always forecast. Forecasts that are statistically independent of the observations may perhaps be regarded as the least useful kind of forecasts. If the forecasts are taken at face value, then it is clearly possible to have worse forecasts. In the extreme case, it would be possible to have a situation in which rain was always observed when the forecast was no rain and it was always dry when rain was forecast. However, such forecasts would actually be very useful if a user was aware of this and inverted (recalibrated) the forecasts accordingly.

2.7 ACCURACY, ASSOCIATION AND SKILL

A scoring rule is a function of the forecast and observed values that is used to assess the quality of the forecasts. Such verification measures often assess the accuracy or association of the forecasts and observations. Accuracy is a measure of the correspondence between individual pairs of forecasts and observations, while association is the overall strength of the relationship between individual pairs of forecasts and observations. The correlation coefficient is thus a measure of linear association, whereas mean absolute error and mean squared error, which will be discussed in Chapter 5, are measures of accuracy.

As discussed in Chapter 1, skill scores are used to compare the performance of the forecasts with that of a reference forecast such as climatology or persistence. Skill scores are often in the form of an index that takes the value 1 for a perfect forecast and 0 for the reference forecast. Such an index can be constructed in the following way:

$$\text{skill score} = \frac{\text{score} - \text{score for reference forecast}}{\text{score for perfect forecast} - \text{score for reference forecast}} \tag{2.27}$$

The choice of reference forecast will depend on the temporal scale. As already noted, persistence may be an appropriate choice for short-range forecasts, whereas climatology may be more appropriate for longer-range forecasts.

2.8 PROPERTIES OF SCORING RULES

Murphy (1993) identified *consistency* as being one of the characteristics of a good forecast. A forecast is consistent if it corresponds with the forecaster's judgement. Some scoring rules encourage forecasters to be inconsistent (Murphy and Epstein 1967b). For example, with some scoring rules a better score is obtained on average by issuing a forecast that is closer to the climatological average than the forecaster's best judgement. A *proper* scoring rule is one that is defined in such a way that forecasters are rewarded with the best expected scores if their forecasts correspond with their judgements (both expressed in terms of probabilities). Since forecasters' judgements necessarily contain an element of uncertainty, this concept is applicable only to probabilistic forecasts. A scoring rule is *strictly proper* when the best scores are obtained if and only if the forecasts correspond with the forecaster's judgement. An example of a strictly proper scoring rule is the Brier score, described in Chapter 7.

One desirable property that applies to categorical forecasts is that scoring rules should be *equitable* (Gandin and Murphy 1992). This means that all constant forecasts of the same category and random forecasts receive the same expected score.

2.9 VERIFICATION AS A REGRESSION PROBLEM

It is possible to interpret verification in terms of simple linear *regression* models in which the forecasts are regressed on the observations and vice versa (Murphy *et al.* 1989). The book by Draper and Smith (1998) provides a comprehensive review of regression models. In the case in which the observations are regressed on the forecasts, the linear regression model is

$$x_i = \alpha + \beta\hat{x}_i + \varepsilon_i \tag{2.28}$$

where $\varepsilon_i, i = 1, \ldots, n$ are error terms. It is assumed that $\mathrm{E}[\varepsilon_i] = 0$ for all $i = 1, \ldots, n$ and that the errors are uncorrelated. The regression equation can be rewritten as

$$E[X|\hat{x}] = \alpha + \beta\hat{x} \tag{2.29}$$

Estimates $\hat{\alpha}$ and $\hat{\beta}$ of the parameters α and β can be obtained by the method of least squares. These estimates are

$$\hat{\alpha} = \bar{x} - \hat{\beta}\bar{\hat{x}} \tag{2.30}$$

and

$$\hat{\beta} = \frac{s_x}{s_{\hat{x}}} r_{\hat{x}x} \tag{2.31}$$

where s_x and $s_{\hat{x}}$ are the sample standard deviations of the observations and forecasts, respectively (see Eq. (2.3)), and $r_{\hat{x}x}$ is the sample correlation coefficient between the forecasts and the observations (see Eq. (2.5)). It is desirable that the conditional bias of the observations given the forecasts ($E[X|\hat{x}] - \hat{x}$) should be zero (unbiased). This will only be satisfied if $\alpha = 0$ and $\beta = 1$, which means that the regression line has an intercept of zero and a slope of unity. So ideally the regression line will coincide with the 45° line. Figure 2.12 shows an artificial set of Oklahoma City high-temperature forecasts for which this is the case. Note that in the literature on forecast verification, the phrase *conditional bias of the forecasts* is used to refer to both the conditional bias of the observations given the forecasts, $E[X|\hat{x}] - \hat{x}$ (type 1 conditional bias), and to $E[\hat{X}|x] - x$ (type 2 conditional bias). Since the correlation coefficient between the forecasts and the observations is always less than or equal to unity, it follows from Eq. (2.31) that the condition $\hat{\beta} = 1$ can only be satisfied if the standard deviation of the forecasts is less than or equal to the standard deviation of the observations. The points will all lie exactly on the fitted straight line only if $r_{\hat{x}x}^2 = 1$.

In the case in which the forecasts are regressed on the observations, the linear regression model can be written as

$$E[\hat{X}|x] = \gamma + \delta x \tag{2.32}$$

and estimates $\hat{\gamma}$ and $\hat{\delta}$ of the parameters γ and δ are given by

$$\hat{\gamma} = \bar{\hat{x}} - \hat{\delta}\bar{x} \tag{2.33}$$

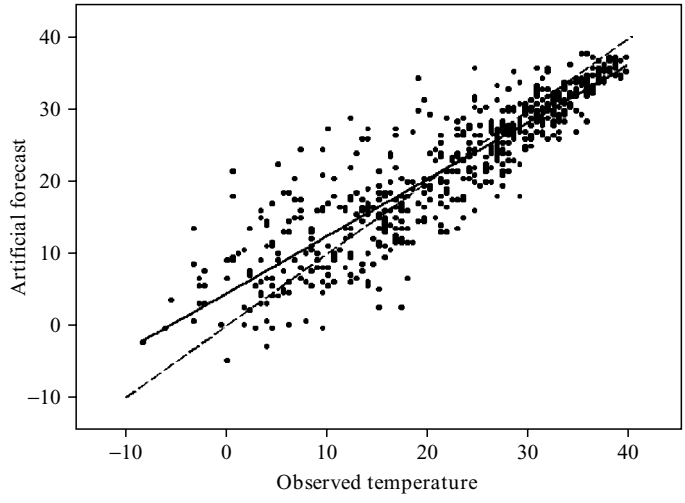


Figure 2.12 Scatterplot of artificial forecasts against observed high temperature (°C) at Oklahoma City. The solid line shows the regression line obtained by regressing the forecasts on the observations. The dashed line is the 45 ° line and the forecasts have been generated in such a way that this is the regression line obtained by regressing the observations on the forecasts

and

$$\hat{\delta} = \frac{s_{\hat{x}}}{s_x} r_{\hat{x}x} \tag{2.34}$$

In situations where $\hat{\alpha} = 0$ and $\hat{\beta} = 1$ (from which we would infer that the forecasts are conditionally unbiased), $s_{\hat{x}} < s_x$ unless the forecasts and observations are perfectly correlated (i.e. $r_{\hat{x}x} = 1$), and it follows that $\hat{\delta} < 1$ unless the forecasts are perfect. In Fig. 2.12, the regression line obtained by regressing the forecasts on the observations does not have a slope of unity even though the regression line obtained by regressing the observations on the forecasts does. Thus, forecasts that are conditionally unbiased from the perspective of regressing the observations on the forecasts will, unless they are perfect, be conditionally biased from the perspective of regressing the forecasts on the observations.

2.10 THE MURPHY–WINKLER FRAMEWORK

Murphy and Winkler (1987) outlined a general framework for forecast verification based on the joint distribution of the forecasts and observations. The traditional *measures-oriented* approach to forecast verification involves the reduction of the information from a set of forecasts and

observations into a single verification measure or perhaps a small number of performance measures. These measures are usually concerned with some overall aspect of forecast quality such as accuracy, association or skill. The alternative approach involving the use of the joint distribution of forecasts and observations is known as *distributions-oriented* verification or *diagnostic* verification. Forecast verification usually involves posterior evaluation of a sample of past forecasts and observations. In this context, the joint distribution $p(\hat{x}, x)$ is usually interpreted as being a discrete empirical relative frequency distribution. In practice even a variable that can theoretically be measured on a continuous scale is only recorded with a certain precision. For example, temperature is usually recorded to the nearest 10th of a degree Celsius. This means that only a certain number of distinct values of the forecasts and observations will be found in the verification data set. So provided that a sufficiently large data set is available, it is possible to examine the empirical relative frequency distribution even in the case of variables such as temperature that are theoretically continuous.

With a sufficiently large data set, graphical techniques such as boxplots and histograms and simple summary statistics can be applied to the conditional as well as the marginal distributions. Summary statistics such as the median and the IQR of the conditional distributions can help to identify particular values of the forecasts and observations for which the forecasting system performs especially well or especially badly. This may give insights into ways in which forecasts could be improved that would not be available through traditional overall measures of forecast performance.

It follows from Eqs. (2.19) and (2.22) that the joint distribution of the forecasts and the observations can be factored into a conditional and a marginal distribution in two different ways. The factorisation

$$p(\hat{x},\ x) = p(x|\hat{x})p(\hat{x}) \tag{2.35}$$

is known as the *calibration–refinement* factorisation. A set of deterministic forecasts is said to be perfectly *calibrated* or *reliable* if $E[X|\hat{X} = \hat{x}] = \hat{x}$ for all $\hat{x}$. The concept of a set of forecasts being completely reliable is therefore equivalent to the observations given the forecasts being conditionally unbiased. If a set of forecasts is conditionally unbiased for all forecast values, it must also be unconditionally unbiased (i.e. $E[\hat{X}] = E[X]$). Probabilistic forecasts of a binary variable are perfectly calibrated if $E[X \mid \hat{p}(1)] = \hat{p}(1)$ for all $\hat{p}(1)$, where $\hat{p}(1)$ is the forecast probability that $X = 1$. An overall measure of reliability is

$$\mathrm{REL} = E_{\hat{X}}[(\hat{X} - \mathrm{E}[X|\hat{X}])^2] \tag{2.36}$$

where the notation $E_{\hat{X}}[\cdot]$ means that the expectation is calculated with respect to the marginal distribution of the forecasts. A similar definition

applies to probabilistic forecasts of a binary variable with $\hat{X}$ replaced by the random variable $\hat{P}(1)$.

The marginal distribution $p(\hat{x})$ indicates how often different forecast values occur. If the same forecast is always issued, forecasts are said not to be *sharp*. Thus, a forecaster who always forecasts the climatological average is not sharp. Sharpness is difficult to define in the case of deterministic forecasts (Murphy *et al.* 1989) but for perfect forecasts it must be the case that $p(\hat{x})$ is equal to the marginal distribution of the observations $p(x)$. Murphy and Epstein (1967a), citing work by Bross (1953, pp. 48–52), define probabilistic forecasts as being sharp if the predicted probabilities are all either zero or unity. They suggest using the Shannon–Weaver information quantity

$$I = -\frac{1}{n}\sum_{i=1}^{n}\sum_{x=0}^{K-1}\hat{p}_i(x)\ \ln(\hat{p}_i(x)) \tag{2.37}$$

as a measure of the sharpness of probabilistic forecasts of a categorical variable, where K is the number of categories and n is the number of forecasts. This index has a minimum value of zero (maximum sharpness) when all the values of $\hat{p}(x)$ are zero or unity and a maximum value of $\ln(K)$ when all the values of $\hat{p}(x)$ are equal to $\frac{1}{K}$. The use of 'information' in forecast verification dates back to Holloway and Woodbury (1955). In the case of a binary variable, an alternative measure (Daan 1984) is

$$S = \frac{1}{n}\sum_{i=1}^{n}\hat{p}_i(1)(1-\hat{p}_i(1)) \tag{2.38}$$

This function has a minimum value of zero (maximum sharpness) when all the values of $\hat{p}(1)$ are zero or unity and a maximum value of 0.25 when all the values of $\hat{p}(1)$ are 0.5. Murphy and Winkler (1992) suggest using the variance of the forecasts $\text{var}(\hat{P}(1))$ as a measure of the sharpness of probabilistic forecasts of a binary variable, with larger values of the variance indicating greater sharpness. However, this leads to a contradiction. If the variance of the forecasts is used, a forecaster who always forecasts 'No Rain' is not at all sharp because the forecast probabilities have a variance of zero, whereas according to the definition of Murphy and Epstein (1967a) he or she is perfectly sharp, since only forecasts of zero or unity are used. On the other hand, for perfectly calibrated forecasts, the measures of sharpness are consistent with one another, as in this case

$$\begin{aligned} E[\hat{P}(1)(1-\hat{P}(1))] &= -E[(\hat{P}(1))^2] + E[\hat{P}(1)] \\ &= -E[(\hat{P}(1))^2] + E[X] \end{aligned} \tag{2.39}$$

and

$$\begin{aligned} \text{var}(\hat{P}(1)) &= E[(\hat{P}(1))^2] - (E[\hat{P}(1)])^2 \\ &= E[(\hat{P}(1))^2] - (E[X])^2 \end{aligned} \tag{2.40}$$

The terms involving $E[X]$ in (2.39) and (2.40) do not depend on the forecast and so can be ignored.

DeGroot and Fienberg (1982, 1983) introduced the concept of *refinement* as a means of comparing forecasters who are perfectly calibrated. Denoting two perfectly calibrated forecasters by A and B, A is at least as refined as B if, knowing A's predictions, it is possible to simulate B's predictions by means of an auxiliary randomisation. The mathematics of this definition is beyond the scope of this book. However, the least refined of all perfectly calibrated forecasts are those in which the forecast probabilities are simply the climatological probabilities and the most refined are those in which the forecast probabilities are always zero or unity. If the forecasts are perfectly calibrated and the forecast probabilities are all zero or unity, the forecasts must be perfect. The extension of this definition of refinement to forecasts that are not perfectly calibrated leads to the concept of *sufficiency*, discussed in Chapter 3. According to the definition given by DeGroot and Fienberg, it is not always possible to compare forecasters in terms of refinement; sometimes it is the case that neither is forecaster A at least as refined as forecaster B, nor is forecaster B at least as refined as forecaster A. So strictly speaking the concept of refinement is not the same as that of sharpness, although many authors have treated the terms as synonymous.

Sanders (1963) used the term 'sharpness' for the concept termed '*resolution*' in an earlier paper (Sanders 1958). His definition of the resolution of probabilistic forecasts of a binary variable is

$$\text{RES}_{\text{Sanders}} = E_{\hat{P}(1)}[E[X|\hat{P}(1)](1 - E[X|\hat{P}(1)])] \tag{2.41}$$

Small values of this quantity are preferable to large ones. Another measure of resolution (Murphy and Winkler 1987) is

$$\begin{aligned} \text{RES}_{\text{Murphy}} &= E_{\hat{X}}[(E[X|\hat{X}] - E[X])^2] \\ &= \text{var}_{\hat{X}}(E[X|\hat{X}]) \end{aligned} \tag{2.42}$$

For probabilistic forecasts of a binary variable, $\hat{X}$ should be replaced by $\hat{P}(1)$ in this definition. Large values of this quantity are preferable to small ones because they indicate that on average different forecasts are followed by different observations. Resolution is not equivalent to the definitions of sharpness given by Murphy and Epstein (1967a) or Murphy and Winkler (1992) since it involves the distribution of the observations as well as that of

the forecasts. However, it should be noted that for perfectly calibrated forecasts of a binary variable they become identical concepts. For perfectly calibrated forecasts $E[X\,|\,\hat{p}(1)] = \hat{p}(1)$ and $E[\hat{P}(1)] = E[X]$; so (2.41) becomes

$$\mathrm{RES}_{\mathrm{Sanders}} = E[\hat{P}(1)(1 - \hat{P}(1))] \tag{2.43}$$

and (2.42) becomes

$$\mathrm{RES}_{\mathrm{Murphy}} = E[(\hat{P}(1) - E[\hat{P}(1)])^2] = \mathrm{var}(\hat{P}(1)) \tag{2.44}$$

Both calibration and sharpness are important. Suppose the climatological average probability of rain at a particular place is 0.3. A forecaster who always forecast a 30% chance of rain would be perfectly calibrated but not sharp. A forecaster who forecast 100% chance of rain on 30% of days chosen at random and 0% chance on the other 70% of days so that $E[X\,|\hat{P}(1) = 1] = 0.3$ and $E[X\,|\hat{P}(1) = 0] = 0.3$ would be sharp but not at all calibrated. Both examples show no resolution. In the first case, the expected value of X following a forecast of *30% chance of rain* is 0.3, which is identical to the unconditional expectation of X. In the second case the expected values of X following forecasts of either *rain* or *no rain* are both equal to the unconditional expectation of 0.3.

The second factorisation

$$p(\hat{x},\ x) = p(\hat{x}|x)p(x) \tag{2.45}$$

is known as the *likelihood–base rate factorisation*. For a given forecast $\hat{x}$, the conditional probabilities $p(\hat{x}|x)$ are known as the *likelihoods* associated with the forecast. If $p(\hat{x}|x)$ is zero for all values of x except one, the forecast is perfectly *discriminatory*. When $p(\hat{x}|x)$ is the same for all values of x, the forecast is not at all discriminatory. Two measures of discrimination are (Murphy 1993)

$$\mathrm{DIS1} = E_X[(E[\hat{X}|X] - X)^2] \tag{2.46}$$

and

$$\mathrm{DIS2} = E_X[(E[\hat{X}|X] - E[\hat{X}])^2] \tag{2.47}$$

For probabilistic forecasts of a binary variable, $\hat{X}$ should again be replaced by $\hat{P}(1)$ in these definitions. Good discrimination implies that DIS1 is small and DIS2 is large. Murphy *et al.* (1989) used a measure of discrimination based on the likelihood ratio

$$\mathrm{LR}(\hat{x};\ x_i,\ x_j) = \frac{p(\hat{x}|x_i)}{p(\hat{x}|x_j)} \tag{2.48}$$

The discrimination between the observations x_i and x_j provided by the forecast $\hat{x}$ is the maximum of LR $(\hat{x};\ x_i,\ x_j)$ and 1/LR $(\hat{x};\ x_i,\ x_j)$.

The marginal distribution $p(x)$ is sometimes known as the *uncertainty* or *base rate*. It is a characteristic of the forecasting situation rather than of the forecast system. Forecasting situations involving a fairly peaked distribution are less difficult situations in which to forecast than ones in which the distribution of possible values is fairly uniform. A measure of whether a continuous distribution is peaked or not is the *kurtosis*, which is defined as

$$\text{kurtosis} = \frac{E[(X - E[X])^4]}{(E[(X - E[X])^2])^2} - 3 \tag{2.49}$$

For a normal distribution the kurtosis is zero, whereas for a uniform distribution it is -1.2. A binary variable is easier to forecast if the climatological probability is close to zero or unity than if it is close to 0.5.

By equating the two factorisations we find that

$$p(x|\hat{x})p(\hat{x}) = p(\hat{x}|x)p(x) \tag{2.50}$$

This equation can be rewritten in the form

$$p(x|\hat{x}) = \frac{p(x)p(\hat{x}|x)}{p(\hat{x})} \tag{2.51}$$

which is known as *Bayes Theorem*. An introduction to Bayesian statistical methods, which are based on this theorem, can be found in Lee (1997); see also Garthwaite *et al.* (2002, Chapters 6 and 7). If forecasts of a binary variable are perfectly calibrated and completely sharp, then the forecasts must be perfectly discriminatory. If the forecasts are completely sharp, $\hat{p}(x)$ takes only the values zero and unity. In this situation

$$\begin{aligned}\text{DIS1} = {} & P(X = 1)(P(\hat{P}(1) = 1|X = 1) - 1)^2 \\ & + P(X = 0)(P(\hat{P}(1) = 1|X = 0) - 0)^2\end{aligned} \tag{2.52}$$

It follows from (2.51) that

$$P(\hat{P}(1) = 1|X = 1) = \frac{P(X = 1|\hat{P}(1) = 1)P(\hat{P}(1) = 1)}{P(X = 1)} \tag{2.53}$$

If the forecasts are perfectly calibrated

$$P(X = 1|\hat{P}(1) = 1) = 1 \tag{2.54}$$

and

$$P(\hat{P}(1) = 1) = P(X = 1) \tag{2.55}$$

so

$$P(\hat{P}(1) = 1|X = 1) = 1. \tag{2.56}$$

Similarly,

$$\begin{aligned} P(\hat{P}(1) = 1|X = 0) &= \frac{P(X = 0|\hat{P}(1) = 1)P(\hat{P}(1) = 1)}{P(X = 0)} \\ &= \frac{(1 - P(X = 1|\hat{P}(1) = 1))P(\hat{P}(1) = 1)}{P(X = 0)} \\ &= \frac{(1 - 1)P(\hat{P}(1) = 1)}{P(X = 0)} \\ &= 0 \end{aligned} \tag{2.57}$$

Thus

$$\text{DIS1} = P(X = 1)(1 - 1)^2 + P(X = 0)(0 - 0)^2 = 0 \tag{2.58}$$

However, the converse is not true. It is possible for forecasts to be perfectly discriminatory even if they are not calibrated. For example, if forecasts of rain are always followed by no rain and forecasts of no rain are always followed by rain, the forecasts are perfectly discriminatory but not calibrated. If users are aware of the base rates and the likelihoods and use the forecasts appropriately, the fact that the forecasts are not well calibrated is irrelevant as long as they are perfectly discriminatory. However, calibration is relevant if the forecasts are taken at face value. The likelihood–base rate factorisation therefore gives information about the potential skill of the forecasts, whereas the calibration–refinement factorisation gives information about the actual skill. The base rate $p(x)$ represents the probability of an event occurring before the forecast is issued and Bayes Theorem shows how this probability should be updated when a particular forecast is issued. For example, in the case of Finley's tornado data, a forecast of *tornado* could be interpreted as *28% chance of a tornado* and a forecast

of *no tornado* could be interpreted as *1% chance of a tornado* (see Eqs. (2.20) and (2.21)).

2.11 DIMENSIONALITY OF THE VERIFICATION PROBLEM

One difficulty that arises from the use of the distributions-oriented approach is the high *dimensionality* of many typical forecast verification problems (Murphy 1991). Dimensionality is defined as the number of probabilities that must be specified to reconstruct the basic distribution of forecasts and observations. It is therefore one fewer than the total number of distinct combinations of forecasts and observations. Problems involving probabilistic forecasts or non-probabilistic forecasts of predictands that can take a large number of possible values are of particularly high dimensionality. Of even higher dimensionality are *comparative verification* problems. Whereas *absolute verification* is concerned with the performance of an individual forecasting system, comparative verification is concerned with comparing two or more forecasting systems. The situation in which these produce forecasts under identical conditions is known as *matched comparative verification*. *Unmatched comparative verification* refers to the situation in which they produce forecasts under different conditions. Whereas absolute verification is based on the joint distribution of two variables, which can be factored into conditional and marginal distributions in two different ways, matched comparative verification is based on a three-variable distribution, which has six distinct factorisations. Unmatched comparative verification is based on a four-variable distribution, which has 24 different factorisations. The number of distinct factorisations represents the *complexity* of the verification problem (Murphy 1991).

In many cases the dimensionality of the verification problem will be too great for the size of the dataset available. One approach to reducing dimensionality, which was used by Brooks and Doswell (1996), is to create a categorical variable with a smaller number of categories than the number of distinct values recorded, by dividing the values into bins. An alternative approach to reducing dimensionality, which is discussed by Murphy (1991), is to fit parametric statistical models to the conditional or unconditional distributions. The evaluation of forecast quality is then based on the parameters of these distributions. For example, fitting a bivariate normal distribution to forecasts and observations of a continuous variable would lead to five parameters: the means and variances of the forecasts and observations, respectively, and the correlation between them (Katz *et al.* 1982).

3 Binary Events

IAN B. MASON
Canberra Meteorological Office, Canberra, Australia

3.1 INTRODUCTION

Many meteorological phenomena can be regarded as simple binary (dichotomous) events, and forecasts or warnings for these events are often issued as unqualified statements that they will or will not take place. Rain, floods, severe storms, frosts, fogs, etc., either do or do not occur, and appropriate forecasts or warnings either were or were not issued. These kinds of predictions are sometimes referred to as yes/no forecasts, and represent the simplest type of forecasting and decision-making situation. The (2×2) possible outcomes (contingencies) for an event are shown in Table 3.1. There are two ways for the forecast to be correct (either a *hit* or a *correct rejection*) and two ways for the forecast to be incorrect (either a *false alarm* or a *miss*).

The search for reliable measures of the quality of deterministic binary forecasts has a long history, dating at least to 1884 when Sergeant Finley of the US Army Signal Corps published the results of some experimental tornado forecasts (Finley 1884). Murphy (1996) has given a detailed history of the so-called 'Finley affair', briefly outlined in Chapter 1. Many of the basic issues in verification were first raised in this episode, and Finley's forecasts are often used as an example of deterministic binary forecasts. The (2×2) contingency table of forecasts and observations for the whole period of Finley's experimental programme is shown as Table 1.1 of this book.

Table 3.1 The four possible outcomes for categorical forecasts of a binary event

Event forecast	Event observed	
	Yes	No
Yes	Hit	False alarm
No	Miss	Correct rejection

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

Finley measured the performance of his forecasts using *percent correct*, the proportion of correct forecasts of either kind expressed as a percentage. His forecasts attained 96.6% correct. Gilbert (1884) promptly pointed out that there would have been more correct forecasts if Finley had simply forecast no tornado every time, giving 98.2% correct. This was probably the first of many comments in the meteorological literature and elsewhere on the inadequacies of *percent correct* as a measure of forecasting performance. Gilbert (1884) proposed two new measures, one of which is known now as the *threat score* (Palmer and Allen 1949) or the *critical success index*, CSI (Donaldson *et al.*, 1975). Gilbert's second measure was rediscovered by Schaefer (1990), who recognised the historical precedence by naming it the *Gilbert skill score* (GSS). The well-known philosopher C.S. Peirce (1884) also took an interest in Finley's forecasts and proposed another measure of forecasting skill, equivalent to that most commonly known as *Hanssen and Kuipers'* (1965) score or *Kuipers' performance index* (Murphy and Daan 1985) and independently rediscovered by Flueck (1987) who named it the *true skill score*. A third paper stimulated by Finley's forecasts was published by Doolittle (1885). He proposed an association measure similar to the correlation coefficient for the (2×2) case, equal to $\sqrt{X^2/n}$, which has been used as an accuracy index for weather forecasts (Pickup 1982). Doolittle (1888) later proposed a (2×2) version of the score now known as the Heidke skill score, HSS (Heidke 1926). A number of other papers followed during the 1880s and 1890s, reviewed in Murphy (1996) and referred to as the 'aftermath' of the three methodological papers mentioned above.

It is probably fair to say that there was very little change in verification practice for deterministic binary forecasts until the 1980s, with the introduction of methods from *signal detection theory* (SDT) by Mason (1980, 1982a,b, 1989), and the development of a general framework for forecast verification by Murphy and Winkler (1987). In practice, multiple sets of binary forecasts are often produced by varying a decision threshold over a range of control values, and these sets of forecasts need to be evaluated together. In particular, it is often revealing to evaluate deterministic forecasts of a real continuous variable by considering sets of binary forecasts of exceedance events. For example, binary events defined by rainfall amounts exceeding a range of chosen thresholds, in which larger thresholds lead to fewer hits but also fewer false alarms. Alternatively, the evaluation of probabilistic forecasts of any event can be more easily evaluated by considering the set of deterministic binary forecasts obtained by choosing a range of probability decision thresholds (Mason 1979). For these kinds of multiple binary forecast problems, methods from SDT offer two broad advantages. Firstly, they provide a means of assessing the performance of a forecasting system that distinguishes between the intrinsic discrimination capacity and the decision threshold of the system. The main analysis tool that accomplishes this is the *relative* (or *receiver*) *operating characteristic*

(ROC). Secondly, SDT provides a framework within which other methods of assessing binary forecasting performance can be analysed and evaluated.

The major ideas covered by this chapter are:

- The apparent simplicity of binary event forecasts hides surprising amounts of complexity, but good methods are available for dealing with this complexity.
- A single set of binary forecasts does not provide a satisfactory basis for assessment of the quality of a forecasting system, because it shows the performance of the system at only a single decision threshold. A complete description of forecasting skill requires verification over the full range of possible thresholds.
- Murphy and Winkler's general framework for forecast verification together with SDT provide the theoretical basis for understanding the skill of binary forecasts.

The remainder of this chapter is in three parts. Section 3.2 introduces the most widely used verification measures for deterministic binary forecasts. Section 3.3 presents some basic underpinning theory mostly based on Murphy and Winkler's (1987) general joint probability framework. It includes some comments on confidence intervals (CIs) and on derivation of *optimal threshold probabilities*. Finally, Section 3.4 describes SDT and the use of SDT methods such as ROC for the verification of binary forecasts. Probabilistic forecasts of binary events are covered in more detail in Chapters 7 and 8.

3.2 VERIFICATION MEASURES

Murphy (1997) distinguishes between *verification measures*, *performance measures* and *scoring rules*. A verification measure is any function of the forecasts, the observations, or their relationship and includes, for example, the probability of the observed event (the *base rate*), even though this is not concerned with the correspondence between forecasts and observations. Performance measures constitute a subset of verification measures that focus on the correspondence between forecasts and observations, either on an individual or collective basis, for example, conditional probabilities such as the *hit rate* and the *false alarm rate*. A scoring rule is a performance measure that is not only a sample statistic but is defined for *each* individual pair of forecasts and observations and so may be used for real time case-by-case feedback to forecasters. Squared error is a scoring rule, but hit rate is not, because a number of forecasts are needed to estimate the hit rate. Most of the common verification measures for binary forecasts are discussed in this section together with some comments on the degree to which each measure satisfies screening criteria for performance measures for

binary forecasts (described in Section 3.3). Briefly anticipating Section 3.3, the main screening criteria directly relevant to performance measures for binary forecasts are *consistency*, *equitability* and *regularity*.

Table 3.2 gives the cell counts for each of the four possible combinations of forecast and observed event represented by *a*, *b*, *c* and *d*. The textbooks by Everitt (1994) and Agresti (1996) review the statistical methods that can be used to analyse categorical data presented in contingency tables. The sum of all the previous events, $n = a + b + c + d$, is known as the *sample size* and is crucial for determining the amount of sampling uncertainty in the verification statistics. The counts in Table 3.2 may be converted to relative frequencies by dividing by *n*, and then interpreted as sample estimates of joint probabilities. Note, however, that it is a good practice to always quote cell counts rather than derived relative frequencies since counts provide useful information about the number of events and are less prone to misinterpretation (Hoffrage *et al*. 2000). Table 3.3 shows the joint probabilities estimated from the counts in Table 3.2, for example, $\hat{p}(\hat{x} = 0,\ x = 0) = d/n$ is the estimated probability of joint occurrence of a forecast non-event and an observed non-event.

A summary of verification measures for deterministic binary forecasts is presented in Table 3.4 (see pp. 42–43). The table provides definitions in terms of raw counts where appropriate and in terms of Murphy and Winkler's (1987) likelihood–base rate factorisation of the joint distribution (see Chapter 2, Section 2.10).

Table 3.2 Schematic contingency table for categorical forecasts of a binary event. The numbers of observations in each category are represented by *a*, *b*, *c* and *d*, and *n* is the total

Forecast	Observed		
	Yes	No	Total
Yes	a	b	$a + b$
No	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = n$

Table 3.3 Schematic contingency table for categorical forecasts of a binary event. Joint and marginal probabilities

Forecast	Observed		
	Yes	No	
Yes	$p(\hat{x} = 1,\ x = 1)$	$p(\hat{x} = 1,\ x = 0)$	$p(\hat{x} = 1)$
No	$p(\hat{x} = 0,\ x = 1)$	$p(\hat{x} = 0,\ x = 0)$	$p(\hat{x} = 0)$
	$p(x = 1)$	$p(x = 0)$	1.0

In the following subsections, the descriptions of the various measures include some technical details that refer to the theoretical properties discussed later in Section 3.3 and to the ROC curves discussed in Section 3.4. Some readers may prefer to skip these details at first reading, and concentrate on the basic definitions of the measures.

3.2.1 Some Basic Descriptive Statistics

These are not measures of forecasting skill, but are interesting descriptive statistics. They are verification measures in Murphy's (1997) sense, being functions of the forecasts and observations, but are not performance measures since they are not concerned directly with the correspondence between forecasts and observations.

(a) *Event Probability (Base Rate)*

$$s = \frac{a+c}{n} = \hat{p}(x=1) \tag{3.1}$$

The base rate (sample climate), s, is a sample estimate of the unconditional marginal probability of occurrence of the observed event. It is purely a characteristic of the observations rather than of the forecasting system. It strictly should have no direct relevance to assessment of forecasting skill, because the forecasting system has no control over the rate of occurrence of the observed events. However, many performance measures do depend on s, and are therefore (often unduly) sensitive to variations in skill due to the natural variations in observed weather and climate. The sampling distribution of s is that of a simple binary proportion with confidence interval (CI) given by Eq. (3.73).

(b) *Probability of a Forecast of Occurrence*

$$r = \frac{a+b}{n} = \hat{p}(\hat{x}=1) \tag{3.2}$$

The marginal frequency, r, is a sample estimate of the marginal probability of a forecast of occurrence. It is the refinement component of the calibration–refinement factorisation (see Chapter 2, Section 2.10). In terms of the likelihood–base rate factors H, F and s (Section 3.3), r can be rewritten as

$$r = (1-s)F + sH \tag{3.3}$$

where H and F are the hit rate and false alarm rate, respectively (see below). Therefore,

Table 3.4 Summary of verification measures discussed in this chapter

Name of measure	Definition	Definition in terms of H, F and s (LBR factors, Section 3.3.2)	Range of values	Reference
Basic descriptive measures				
Base rate, s	$s = (a + c)/n$	s	[0,1]	
Probability of a forecast of occurrence, r	$r = (a + b)/n$	$r = (1 - s)F + sH$	[0,1]	
Bias	$\text{bias} = \frac{a+b}{a+c}$	$\text{bias} = \frac{1-s}{s}F + H$	[0,1]	Donaldson *et al.* (1975)
β	$\beta = \frac{f_1(x)}{f_0(x)} =$ slope of ROC; see text	See text	$(-\infty, +\infty)$	Green and Swets (1966)
Performance measures				
Hit rate, H	$H = a/(a + c)$	H	[0,1]	Swets (1986a)
False alarm rate, F	$F = b/(b + d)$	F	[0,1]	Swets (1986a)
False alarm ratio, FAR	$\text{FAR} = b/(a + b)$	$\text{FAR} = \left(1 + \left(\frac{s}{1-s}\right)\frac{H}{F}\right)^{-1}$	[0,1]	Donaldson *et al.* (1975)
Proportion correct, PC	$\text{PC} = (a + d)/n$	$\text{PC} = (1 - s)(1 - F) + sH$	[0,1]	Finley (1884)
Heidke skill score, HSS	$\text{HSS} = \frac{\text{PC} - E}{1 - E}$ (E is PC for random forecasts)	$\text{HSS} = \frac{2s(1-s)(H-F)}{s + s(1-2s)H + (1-s)(1-2s)F}$	[−1,1]	Heidke (1926), Murphy and Daan (1985)

Peirce's skill score, PSS	$\mathrm{PSS} = \frac{ad-bc}{(b+d)(a+c)}$	$\mathrm{PSS} = H - F$	[−1,1]	Peirce (1884), Hanssen and Kuipers (1965), Murphy and Daan (1965)
Critical success index, CSI	$\mathrm{CSI} = \frac{a}{a+b+c}$	$\mathrm{CSI} = \frac{H}{1+\frac{F(1-s)}{s}}$	[0,1]	Gilbert (1884), Donaldson *et al.* (1975)
Gilbert skill score, GSS	$\mathrm{GSS} = \frac{a-a_r}{a+b+c-a_r}$ (a_r is the number of hits for random forecasts)	$\mathrm{GSS} = \frac{H-F}{\frac{1-sH}{1-s}+\frac{F(1-s)}{s}}$	[−1/3,1]	Gilbert (1884), Schaefer (1990)
Yule's Q	$Q = \frac{ad-bc}{ad+bc}$	$Q = \frac{H-F}{H(1-F)+F(1-H)}$	[−1,1]	Yule (1900), Stephenson (2000)
Dissemination distance, d'	$\Phi^{-1}(H) - \Phi^{-1}(F)$	See Fig. 3.6	$[0,+\infty)$	Tanner and Birdsall (1958), Swets (1986a)
A_z	Area under *fitted* bi-normal ROC. See text. In the (2 × 2) case only, $A_z = \Phi\left(d'/\sqrt{2}\right)$		[0.5,1]	Swets (1986a)

$$H = -\left(\frac{1-s}{s}\right)F + \frac{r}{s} \tag{3.4}$$

and so when plotted on (F, H) axes (the ROC diagram), the isopleth of r is a straight line with slope $-(1-s)/s$, always less than or equal to 0 since s must lie between 0 and 1, and a H-axis intercept equal to r/s. When skill is perfect, $H = 1$ and $F = 0$, so $r = s$, and the probability of the forecast event equals the base rate. Forecasts with no skill have $H = F$, and therefore by Eq. (3.3) the no-skill value of r is equal to the common value of H and F, which can lie anywhere between 0 and 1. The 95% CI for the population probability corresponding to r is given by Eq. (3.73).

(c) *Frequency Bias*

Frequency bias, B, is the ratio of the number of forecasts of occurrence to the number of actual occurrences. It is referred to simply as bias when there is no risk of confusion with other meanings of the term (see glossary).

$$B = \frac{a+b}{a+c} \tag{3.5}$$

In terms of the likelihood–base rate factors,

$$B = \left(\frac{1-s}{s}\right)F + H \tag{3.6}$$

Isopleths of bias on ROC axes are given by

$$H = -\left(\frac{1-s}{s}\right)F + B \tag{3.7}$$

which is a family of straight lines with slope $-(1-s)/s$, always negative, and an intercept on the H-axis equal to bias itself. A bias of 1 is represented by a line through the upper left corner (0,1) crossing the positive diagonal of the ROC at $H = F = s$. Perfect forecasts with $H = 1$ and $F = 0$ are always *unbiased* ($B = 1$). Forecasts with no skill have $H = F$ so that bias is equal to H/s (or F/s). Bias alone conveys no information about skill, because any value can be attained with no skill, or at any intermediate level of skill, by changing the decision threshold. It is often stated that a bias of 1 is desirable, in other words the event of interest should be forecast at the same rate as it is observed to occur. Stephenson (2000) explains how biased forecasting systems can be made unbiased by randomly reassigning forecasts from the more frequently forecast category to the less frequent until a bias of 1 is achieved. Introduction of a random element in this way is likely to reduce the apparent skill of the forecasts, suggesting that bias correction

should be undertaken with caution, if at all. Furthermore, a bias of 1 is only desirable for users of the forecasts whose economically optimal threshold probability corresponds to a bias of 1 in that particular forecasting system and climate, but not necessarily for all possible users.

(d) *ROC Slope* $\beta = \mathrm{d}H/\mathrm{d}F$

The gradient β is an index of decision threshold commonly used in SDT. It is the gradient of the H versus F curve (the ROC) at the chosen operating point, and was shown by Green and Swets (1966) to be equal to the likelihood ratio $f_1(z)/f_0(z)$ at the decision threshold z that generates that point on the ROC, where $f_1(z)$ is the probability density of z before occurrences of the predictand and $f_0(z)$ before non-occurrences. β is discussed in more detail in Section 3.4.

3.2.2 Performance Measures

Performance measures are verification measures that focus on the correspondence between forecasts and observations, either on an individual or collective basis (Murphy 1997).

(a) *Hit rate, H*

This quantity is defined by

$$H = \frac{a}{a+c} = \hat{p}(\hat{x} = 1 | x = 1) \tag{3.8}$$

The hit rate, H, is the proportion of occurrences that were correctly forecast. It is a sample estimate of the conditional probability of the event being forecast *given* that the event was observed (i.e. the frequency of the event being forecast in cases when the event was observed to have occurred). It is also known as *probability of detection*, POD (Donaldson *et al.* 1975). The term hit rate is sometimes (rarely and confusingly) used to mean proportion correct (PC) (e.g. Wilks 1995, pp. 240–241). In medical statistics, H is referred to as *sensitivity*. On the ROC diagram of H versus F, constant values (isopleths) of H are horizontal straight lines. In terms of hypothesis testing, the signal detection model H is analogous to the power of a test, 1 minus the probability of a Type 2 error – see Section 3.4. The 95 % CI for its corresponding population value is given by Eq. (3.73). A threshold probability of 0 (see Section 3.3.4), meaning that occurrence is always forecast, gives $H = 1$, and a threshold probability of 1, meaning that the event is never forecast, gives $H = 0$. H is not *equitable* (see Section 3.3) because constant forecasts of either event or non-event give values for H of either 0 or 1, whereas random forecasts can give any value for H. H is not *regular* (see Section 3.3) because its isopleths on ROC axes cross the axes away from (0,0)

and (1,1). Since forecast skill depends on maximising the number of hits while minimising the number of false alarms, the hit rate alone is insufficient for measuring the skill of a forecasting system.

(b) *False Alarm Rate, F*

False alarm rate is defined by

$$F = \frac{b}{b+d} = \hat{p}\,(\hat{x} = 1 | x = 0) \tag{3.9}$$

The false alarm rate is the proportion of non-occurrences that were incorrectly forecast. It is sometimes also called *probability of false detection*, POFD. The false alarm rate must be distinguished from the *false alarm ratio*, FAR, which is the proportion of forecasts of occurrence that were not followed by an actual occurrence. The false alarm rate is the conditional probability of a false alarm conditioned (stratified) on the *event not being observed*, whereas the FAR is the conditional probability of a false alarm conditioned on the *event being forecast*. In medical statistics, the *correct rejection rate*, $1 - F$, is referred to as *specificity* and is an estimate of the conditional probability of correct rejections given that the event did not occur. Isopleths of F on the (H, F) ROC diagram are vertical straight lines. In terms of the hypothesis testing analogue to the signal detection model, F is analogous to the probability of a Type 1 error – see Section 3.4.1. The 95 % CI for its corresponding population value is given by Eq. (3.73). F is not equitable because constant forecasts of either alternative give values of either 0 or 1, whereas random forecasts can give any value. F is not regular because isopleths on ROC axes cross the axes away from (0,0) and (1,1). As for hit rate, the false alarm rate alone does not provide a measure of forecast skill.

(c) *Proportion Correct*

PC is defined in terms of cell counts by

$$\mathrm{PC} = \frac{a+d}{n} = \hat{p}\,[(\hat{x} = 1, x = 1) \text{ or } (\hat{x} = 0, x = 0)] \tag{3.10}$$

or in terms of likelihood–base rate factors

$$\mathrm{PC} = (1 - s)(1 - F) + sH \tag{3.11}$$

Isopleths of PC on ROC axes are thus given by

$$H = \frac{1-s}{s}F + \left(\frac{\mathrm{PC}}{s} - \frac{1-s}{s}\right) \tag{3.12}$$

which represents a family of straight lines with slope $(1-s)/s$ and H-axis intercept $(\mathrm{PC}/s)-(1-s)/s$. PC is often expressed in terms of percent, PC × 100 %, and is then referred to as *percent correct*. The *optimal threshold probability* that maximises PC is 0.5 (see Section 3.3.4), and hence the PC score can always be maximised by forecasting occurrence of the event whenever the observed probability of the event exceeds 0.5. For binary forecasts, 1 – PC is equal to the mean square error $(\hat{x}-x)^2$ of the binary forecast and observed variables. PC is not equitable, since forecasts without skill ($H=F$) can score anywhere between s and $1-s$ depending on the ratio of forecasts of occurrence to forecasts of non-occurrence in the sample. Constant forecasts of non-occurrence score $1-s$, and constant forecasts of occurrence score s. PC is not regular, because its isopleths on ROC axes are straight lines with slope $(1-s)/s$. This means that is possible for the same value of PC to be attained by forecasts with high levels of skill, where the isopleth of PC crosses the boundaries of the ROC unit square, implying $H>0$ for $F=0$ or $H=1$ for $F<1$, or with no skill, where the isopleth crosses the $H=F$ diagonal. PC is intuitively appealing as a measure of forecasting performance, being simply the proportion in the whole sample of correct forecasts of either kind, and was probably the first such measure used (Finley 1884). It has often been used since (e.g. Brier and Allen 1951; Ramage 1982). In medical statistics, it is sometimes taken as synonymous with accuracy (Metz 1978). Nevertheless, the dependence of PC on the base rate and the threshold probability, and its non-equitability and non-regularity make it unreliable as a performance measure. Problems were identified very soon after its first appearance, initially by Gilbert (1884). Murphy (1996) describes efforts to develop more satisfactory single number measures of forecasting performance.

(d) *False Alarm Ratio*

FAR is defined by

$$\mathrm{FAR}=\frac{b}{a+b}=\hat{p}\,(x=0|\hat{x}=1) \tag{3.13}$$

or in terms of the likelihood–base rate factors

$$\mathrm{FAR}=\left(1+\left(\frac{s}{1-s}\right)\frac{H}{F}\right)^{-1} \tag{3.14}$$

Therefore, the equation of isopleths of FAR on ROC axes is given by

$$H=F\,\frac{(1-s)}{s}\left(\frac{1-\mathrm{FAR}}{\mathrm{FAR}}\right) \tag{3.15}$$

which represents a family of straight lines through the origin with slope equal to $(1/s - 1)(1/\mathrm{FAR} - 1)$. FAR is a sample estimate of the conditional probability of a false alarm given that occurrence was forecast. It is the proportion of forecasts of occurrence that were followed by non-occurrence, and must be distinguished from F (see previous section), the proportion of non-occurrences that were forecast as occurrences. The 95 % CI for its corresponding population value is given by Eq. (3.73). FAR depends on both the base rate and the optimal threshold probability. When skill is perfect ($H = 1$ and $F = 0$), $\mathrm{FAR} = 0$, and for zero skill ($H = F$) $\mathrm{FAR} = 1 - s$. It is possible for values of FAR greater than $1 - s$ to occur in pathological forecast sets with 'negative skill' in which $F > H$, and FAR can equal 1 when $H = 0$ and $F = 1$. The negative skill situation $F > H$ actually indicates skill which can be used if the forecasts are 'recalibrated' by reversing the labels, so that forecasts of occurrence are taken as non-occurrence and vice versa. FAR by itself does not necessarily carry any information about skill because it can vary from 0 to $1 - s$ due to the variation in threshold probability alone, regardless of skill. If occurrence is never forecast, both the numerator and the denominator of Eq. (3.13) are 0, but the number of false alarms, b, is 0 and so it is reasonable in this case to define $\mathrm{FAR} = 0$. If occurrence is always forecast the number of false positives is just the sample frequency of non-occurrences, so $\mathrm{FAR} = 1 - s$. When H and F are fixed, FAR decreases with increasing base rate s – in other words, the FAR can be reduced by changing the decision threshold to have more events occurring. FAR is not a reliable performance measure unless its dependence on sample climate (base rate) and threshold probability is taken into account.

(e) *The Heidke Skill Score*

The HSS is PC adjusted to account for the proportion of forecasts that would have been correct by chance in the absence of skill. Heidke (1926) formulated this score for the general case of multi-category (more than just two binary categories) contingency tables. However, a binary forecast version of the HSS had been proposed 41 years earlier by Doolittle (1885), and so perhaps this measure should be strictly referred to as the *Doolittle skill score*. Nevertheless, to avoid the confusing proliferation of alternative names, it seems preferable to continue to refer to it as the HSS. The HSS is defined as

$$\mathrm{HSS} = \frac{\mathrm{PC} - E}{1 - E} \tag{3.16}$$

where PC is proportion correct and E is the proportion of forecasts that would have been correct if forecasts and observations were independent and assuming the same proportion of forecasts of occurrence to non-occurrence, so that for a (2×2) contingency table

$$E = \left(\frac{a+c}{n}\right)\left(\frac{a+b}{n}\right) + \left(\frac{b+d}{n}\right)\left(\frac{c+d}{n}\right) \tag{3.17}$$

which is a sample estimate of $p(x = 1)p(\hat{x} = 1) + p(x = 0)p(\hat{x} = 0)$. HSS can be rewritten in terms of likelihood–base rate factors as

$$\mathrm{HSS} = \frac{2s(1-s)(H-F)}{s + s(1-2s)H + (1-s)(1-2s)F} \tag{3.18}$$

and the equation of isopleths of HSS on ROC axes is

$$H = \left[\frac{2 + ((1-2s)/s)\mathrm{HSS}}{2 - ((1-2s)/(1-s))\mathrm{HSS}}\right]F + \left[\frac{\mathrm{HSS}/s}{(2(1-s)/s) + ((1-2s)/s)\mathrm{HSS}}\right] \tag{3.19}$$

Isopleths of HSS are therefore straight lines with slope given by the coefficient of F in Eq. (3.19) and the H-axis intercept given by the second term in square brackets. When $s = 0.5$ the slope is unity, the intercept is HSS itself, and HSS is equal to Peirce's score (see below). Perfect skill gives HSS = 1, the maximum value. The minimum value of HSS is -1, which occurs when $E = 0.5$, $H = 0$ and $F = 1$. This in fact represents perfect skill once forecasts of event and non-event are relabelled (recalibrated) as non-event and event. Any forecast set with $H < F$ and thus negative values of HSS implying 'negative' skill, can be converted to give positive skill by simply switching the labels on the forecasts. The true zero-skill value of HSS is 0, corresponding to $H = F$. HSS has a marked dependence on threshold probability, and the optimal threshold probability p^* that maximises HSS is given by (Bryan and Enger 1967)

$$p^* = s + (1-2s)\frac{\mathrm{HSS}}{2} \tag{3.20}$$

The optimal threshold probability for this score therefore depends on the value of the score and on the climatological base rate. When HSS is small, p^* is close to s, and when HSS is near its maximum value of 1, p^* is near 0.5. For very rare events ($s \to 0$), p^* tends to HSS/2. Any particular value of p^* selected on the basis of past performance will not remain optimal if there are significant changes in either base rate or skill. HSS is an equitable score since constant forecasts of occurrence or of non-occurrence score 0, as do random forecasts. However, HSS is not a regular score, because isopleths on ROC axes can cross the axes away from (0,0) and (1,1). Isopleths of HSS do not cross the no-skill diagonal, but the same value of HSS can correspond to unrealistically high skill, where the isopleth crosses the ROC axes, or an intermediate level of skill, within the ROC unit square. The dependence of HSS on the base rate and threshold probability and its non-regular ROC make it unreliable as a performance measure. Furthermore, its

sampling distribution is not known, though it could be estimated using resampling methods (Wilks 1995, Section 5.3.2).

(f) *Peirce's Skill Score*

The (2×2) form of this measure was originally proposed by the philosopher C.S. Peirce (1884). It was independently rediscovered in a multi-category form by Hansen and Kuipers (1965), and is often referred to as *Kuipers' performance index* or *Hansen and Kuipers' score*. It has also been rediscovered by Flueck (1987) who called it the *true skill statistic* (TSS). In medical statistics, an identical measure of diagnostic discrimination is known as Youden's (1950) index, and it has also been used in psychology (Woodworth 1938). In the absence of any consensus on terminology, it seems reasonable that it should be named after the earliest known discoverer. In terms of raw cell counts, PSS is defined as

$$\mathrm{PSS} = \frac{ad - bc}{(a + c)(b + d)} \tag{3.21}$$

and in terms of likelihood–base rate factors

$$\mathrm{PSS} = H - F \tag{3.22}$$

so that isopleths on ROC axes are given by

$$H = F + \mathrm{PSS} \tag{3.23}$$

which is a family of straight lines with unit slope and H-axis intercept PSS. Isopleths of this form are generated by a signal detection model in which the 'noise alone' and 'signal plus noise' probability distributions are uniform and have equal range (Swets 1986a). For rare events, the frequency of correct rejections may be orders of magnitude larger than the other cells, resulting in very low values of F and a value of PSS almost equal to H. The minimum possible value of PSS is -1, and as was the case for the Heidke score, this actually corresponds to perfect skill but completely wrong calibration. A forecast set with PSS $= -1$ can be 'recalibrated' to PSS $= +1$ by reversing the labels on the forecasts. Negative values of PSS can always be changed to positive values in this way. The true zero-skill value of PSS is 0. The optimal threshold probability for PSS is

$$p^* = \frac{a + c + 1}{n + 2} \tag{3.24}$$

which is the Bayesian sample estimate of the climatological base rate of a binary event with a uniform prior (Garthwaite *et al.* 2002, Section

6.3.3). PSS is equitable, because constant forecasts of occurrence or non-occurrence and random forecasts all score 0. It is not regular, because isopleths cross the ROC axes away from (0,0) and (1,1). Since PSS is the difference between two proportions, its sampling variance can be estimated using the normal approximation to the binomial as (Seaman *et al.* 1996; Thornes and Stephenson 2001)

$$var(\mathrm{PSS}) = \frac{H(1-H)}{a+c} + \frac{F(1-F)}{b+d} \tag{3.25}$$

In Section 8.2.1 of this book, it is shown how PSS can be related to the economic value of binary event forecasts in the cost/loss model. PSS is equitable and does not depend on sample climate. It does depend strongly on threshold probability, and is not regular. PSS is unreliable as a performance measure unless these factors are taken into account.

(g) *Critical Success Index*

This measure was first suggested by Gilbert (1884), who called it the *ratio of verification*, *v*. Schaefer (1990) noted that it has been rediscovered and renamed at least twice, by Palmer and Allen (1949), who called it the *threat score*, and by Donaldson *et al.* (1975) who called it the *critical success index*, CSI. CSI appears to be most widely used to refer to this score, and to avoid further confusion CSI is used here, although historical precedence would suggest Gilbert's *v*. CSI is defined in terms of raw cell counts by

$$\mathrm{CSI} = \frac{a}{a+b+c} \tag{3.26}$$

which in terms of the likelihood–base rate factors is

$$\mathrm{CSI} = \frac{H}{1 + \dfrac{F(1-s)}{s}} \tag{3.27}$$

Therefore, isopleths of CSI on ROC axes have the form

$$H = \frac{\mathrm{CSI}(1-s)}{s} F + \mathrm{CSI} \tag{3.28}$$

which is a family of straight lines all crossing the extension of the F-axis to the left at $F = -s/(1-s)$, with slope equal to the coefficient of F in Eq. (3.28) and H-intercept equal to CSI itself. CSI can be regarded as a sample estimate of the conditional probability of a hit given that the event of interest was either forecast, or observed, or both. CSI has been widely used as a performance measure for forecasts of rare events, because it can be

calculated without using the frequency of correct rejections. H or POD and FAR share this property, and these measures are often quoted together. Non-occurrence of rare events is usually trivially easy to forecast, and is not in general forecast explicitly. If every occasion on which a rare event is not forecast and does not occur is counted as a correct forecast of non-occurrence, the frequency of correct rejections can be orders of magnitude larger than the other elements of the contingency table, as can be seen in Finley's tornado forecasts (Table 1.1), where 95.6% of the data consists of correct rejections. Many of these forecasts would have required little or no skill, so it seems reasonable to exclude them from consideration. On the other hand, if non-occurrences are rare, correct rejections require skill. The forecasting system should then get some credit for these, but does not if CSI is used.

Perfect skill ($H = 1,\ F = 0$) gives a CSI with the maximum value of 1. The minimum value of CSI is 0, which occurs when there are no hits ($a = 0$). The value of CSI corresponding to zero skill can be anywhere between 0 ($H = F = 0$) and s ($H = F = 1$) depending on the proportion of forecasts of occurrence to non-occurrence in the sample. The optimal threshold probability for CSI is given by

$$p^* = \frac{\mathrm{CSI}}{1 + \mathrm{CSI}} \tag{3.29}$$

CSI depends strongly on threshold probability, ranging from s for $p^* = 0$ to a maximum at an intermediate value of p^* given by Eq. (3.29), then decreasing to 0 at $p^* = 1$. CSI increases with increasing base rate s, as indicated by Eq. (3.27). The variation of CSI with s and p^* was described in more detail by Mason (1989). CSI is not equitable, because constant forecasts and random forecasts may give different values of CSI, depending on the proportion of forecasts of occurrence to non-occurrence and on sample climate. CSI is non-regular, because isopleths cross the ROC axes away from (0,0) and (1,1). CSI is not a reliable performance measure unless these factors are taken into account. CSI can be regarded as a sample estimate of a probability so its sampling distribution can be modelled as binomial with sample size equal to $a + b + c$, and the methods of Section 3.3.5 apply. CSI is not equitable, its ROC isopleths are not regular, and it has a strong dependence on base rate and threshold probability. For these reasons, it is not in general a reliable performance measure.

(h) *Gilbert's Skill Score*

This score was developed by Gilbert (1884) as a modification of CSI to allow for the number of hits that would have been obtained purely by chance. It has been discussed in detail by Schaefer (1990) and Doswell *et al.* (1990) who refer to it as the *equitable threat score*. In terms of raw cell counts

$$\mathrm{GSS} = \frac{a - a_r}{a - a_r + b + c} \tag{3.30}$$

where a_r is the number of hits expected by forecasts independent of observations (pure chance) given by

$$a_r = \frac{(a+b)(a+c)}{n} \tag{3.31}$$

Note that the appearance of n in the expression for a_r means that the number of correct rejections, d, is needed to calculate GSS, whereas it was not needed for CSI. Note also that the construction of GSS from the CSI score differs from the usual construction of skill scores (see Chapter 2, Section 2.7). GSS can be written in terms of likelihood–base rate factors as

$$\mathrm{GSS} = \frac{H - F}{\frac{1 - sH}{1 - s} + \frac{F(1-s)}{s}} \tag{3.32}$$

and isopleths of GSS on ROC axes are given by

$$H = \left(\frac{1 + \mathrm{GSS}\frac{1-s}{s}}{1 + \mathrm{GSS}\frac{s}{1-s}} \right) F + \frac{\mathrm{GSS}}{1 + s(1 - \mathrm{GSS})} \tag{3.33}$$

which is a family of straight lines with slope the bracketed term and intercept the second term on the right-hand side. Therefore, GSS is not regular, because its isopleths cross the ROC axes away from (0,0) and (1,1). Perfect forecasts give $\mathrm{GSS} = 1$. Constant forecasts of either category or random forecasts give $\mathrm{GSS} = 0$, so GSS is equitable (hence the name *equitable* threat score). Since the denominator of Eq. (3.32) is always positive, forecast sets with $H < F$ have $\mathrm{GSS} < 0$. The lowest possible value of GSS, $\mathrm{GSS}_{\min}$, occurs for $H = 0$ and $F = 1$ and is given by

$$\mathrm{GSS}_{\min} = \frac{-1}{\frac{1-s}{s} + \frac{1}{1-s}} \tag{3.34}$$

$\mathrm{GSS}_{\min}$ ranges from zero when $s = 0$ and $s = 1$ to $-1/3$ when $s = 0.5$. The true zero skill for GSS is when $\mathrm{GSS} = 0$, corresponding to $H = F$. Forecast sets with $H < F$ actually have skill obtainable by reversing the labels on the forecasts. The sampling distribution of GSS is not known, but could be estimated empirically in specific verification problems using resampling methods (e.g. Wilks 1995, Section 5.3.2). The optimal threshold probability is given by

$$p^* = s\frac{1-\mathrm{GSS}}{1+\mathrm{GSS}} + \frac{\mathrm{GSS}}{1+\mathrm{GSS}} \tag{3.35}$$

The optimal threshold probability for GSS thus depends on both s and the value of the score itself. When s tends to 0, p^* tends to GSS/(1+GSS), identical to the expression for p^* for CSI. Schaefer (1990) noted that GSS is related to the HSS by the simple expression

$$\mathrm{GSS} = \frac{\mathrm{HSS}}{2-\mathrm{HSS}} \tag{3.36}$$

GSS is equitable, but its dependence on the base rate and threshold probability, and the non-regular form of its ROC, make it unreliable as a performance measure for binary forecasts.

(i) *Yule's Q (Odds Ratio Skill Score)*

This is a measure of association in (2×2) contingency tables developed by Yule (1900), named after the 19th century Belgian statistician A. Quetelet. It has been proposed as a performance measure for binary forecasts by Stephenson (2000), who refers to it as the *odds ratio skill score* (ORSS). Q has been discussed in the context of ROC analysis by Swets (1986a). In terms of cell counts

$$Q = \frac{ad-bc}{ad+bc} \tag{3.37}$$

In terms of likelihood–base rate factors, it takes the simple form independent of base rate s given by

$$Q = \frac{H-F}{H(1-F)+F(1-H)} \tag{3.38}$$

Isopleths of Q on ROC axes are given by

$$H = \left(1 + \left(\frac{1-Q}{1+Q}\right)\left(\frac{1-F}{F}\right)\right)^{-1} \tag{3.39}$$

which is a family of hyperbolas all passing through (0,0) and (1,1) (Swets 1986a) identical in shape to the ROC curves shown by Stephenson (2000) for the odds ratio. ROCs of this form are generated by a signal detection model with equal variance logistic distributions and are very similar to those produced by equal variance normal distributions (Swets 1986a). Therefore, Q's isopleths on ROC axes are regular. Q was derived by Stephenson (2000) as a transformation of the odds ratio. If the odds of a

hit are denoted by $\omega_H = H/(1-H)$ and the odds of a false alarm are denoted by $\omega_F = F/(1-F)$ then the odds ratio $\hat{\theta}$ is defined by $\hat{\theta} = \omega_H/\omega_F = ad/bc$. The odds ratio is widely used in medical statistics (Agresti 1996) and is equal to the square of the η measure developed in psychology (Luce 1963). Q is the odds ratio transformed to have range $[-1,+1]$, so

$$Q = \frac{\hat{\theta}-1}{\hat{\theta}+1} \tag{3.40}$$

The maximum value of Q, $+1$, is attained when $H = 1$ and $F = 0$, the upper left corner of the ROC unit square. The minimum value of Q is -1 and is obtained when $H = 0$ and $F = 1$, the lower right corner. This actually represents perfect discrimination and can be moved to the upper left corner by reversing the labels on the forecasts. Any point in the lower right half of the ROC can be reflected about the positive diagonal by this procedure. The zero-skill value of Q is 0, corresponding to random forecasts, which give $H = F$, except that Q is undefined for constant forecasts of a single category, which give either $H = F = 0$ or $H = F = 1$. Q is thus an equitable score only in a qualified sense that excludes the end-points of the range of H and F. It should, however, be noted that the concept of equitability was proposed for screening the restricted class of scores based on linear combinations of cell counts (Gandin and Murphy 1992). The optimal threshold probability for Q can be derived using the implied payoff matrix with the method illustrated in Section 3.3.4 but the expression is unwieldy and is left as an exercise for the reader. An approximate expression, which can be obtained by neglecting terms of order $1/n$, is

$$p^* = \frac{sH(1-H)}{sH(1-H)+(1-s)F(1-F)} \tag{3.41}$$

It is interesting to note that, while Q itself does not depend on the base rate, the optimal threshold probability that optimises Q does. A 95% CI for the corresponding population quantity based on Q can be obtained using the natural logarithm of the sample odds ratio, $\hat{\theta}$, which has an asymptotically normal distribution with standard deviation $1/\sqrt{n_h}$ where n_h is the effective number of degrees of freedom, given by $1/n_h = 1/a + 1/b + 1/c + 1/d$ (Stephenson 2000). For Finley's forecasts, log odds, $\ln\hat{\theta}$, is 3.81 and $n_h = 10.7$, so the standard deviation of log odds is 0.306 and the 95% CI is (3.20, 4.41). Converting these confidence limits for log odds to θ and then to Q using Eq. (3.40) gives the 95% CI for Q's population value as (0.922, 0.976). When either b or c is equal to 0, Q is 1, implying perfect skill. However, as Stephenson (2000) points out, when this is the case the effective degrees of freedom become 0 and the CI for Q covers all possible values in the range $[-\infty,\infty]$. Zero in any cell of

the contingency table suggests that (2×2) measures of association such as the odds ratio are no longer appropriate. Q is equitable in the qualified sense noted above, is regular and does not depend on base rate. It is therefore preferable to all of the previously discussed performance measures. However, its isopleths on ROC axes are symmetrical about the negative diagonal, so it will not adequately describe the performance of forecasting systems that produce asymmetrical ROCs.

(j) *Signal Detection Theory Performance Measures, A_z and d'*

The most widely used ROC-based measure of skill is A_z, the area under the *modelled* ROC curve $H(F)$. It is closely related to d', the separation of the means of the underlying 'noise' and 'signal plus noise' distributions, which can also be used when the two distributions have similar spread. Both these measures are discussed in detail in Section 3.4.4, and are illustrated there with a numerical example.

3.3 VERIFICATION OF BINARY FORECASTS: THEORETICAL CONSIDERATIONS

This section presents some of the basic theory behind forecast verification relevant to the verification of deterministic binary forecasts. A more general discussion is given in Sections 2.8 and 2.10 of Chapter 2.

3.3.1 A General Framework for Verification: The Distributions-oriented Approach

The conventional approach to assessment of skill in forecasting binary events prior to the mid-1980s consisted of calculating values for one or more summary measures of the correspondence between forecasts and observations, and then drawing conclusions about forecasting performance on the basis of these scores. This process is known as the 'measures-oriented' approach to verification and led to the development of the surprisingly large range of measures/scores reviewed in the previous section. While they were generally plausible, there was little in the way of an agreed background of basic theory to guide selection of appropriate measures for particular verification problems, or to systematically discuss the properties of particular measures. With the aim of remedying this situation, Murphy and Winkler (1987) proposed a general framework for verification based on the joint probability distribution of forecasts and observations (e.g. Table 3.3), and they defined forecast verification as the process of assessing the statistical characteristics of this joint distribution. This approach is known as 'distributions-oriented' or sometimes 'diagnostic'

verification. Refer to Section 2.10 of Chapter 2 for a discussion of this probabilistic framework.

The *dimensionality* of the joint probability distribution is defined by Murphy (1991) as the number of probabilities (or parameters) that must be specified to reconstruct it. In general, the dimensionality of a verification problem is equal to $MN - 1$, where M is the number of different forecast categories or values available to be used and N is the number of different observed categories or values possible. For a (2×2) contingency table, the dimensionality is three – the fourth degree of freedom being fixed by the constraint that the joint probabilities sum to unity. This implies that a full description of forecast quality in the (2×2) case requires only three parameters that contain all the information needed to reconstruct the joint distribution. Therefore, despite the many different scores for binary forecasts, there are only three independent dimensions and so the different scores are strongly interrelated with one another.

The joint distribution can be factored in two different ways into conditional and marginal probabilities that reveal different aspects of forecast quality. The calibration–refinement factorisation is given by

$$p(\hat{x},\ x) = p(x|\hat{x})p(\hat{x}) \tag{3.42}$$

For example, the probability of a hit $p(\hat{x} = 1,\ x = 1) = a/n$ is equal to $p(x = 1|\hat{x} = 1)p(\hat{x} = 1) = (a/(a + b))((a + b)/n)$. The refinement term $p(\hat{x})$ in the calibration–refinement factorisation is the marginal distribution of the forecasts, which in the case of binary forecasts, depends on the threshold for issuing a forecast of occurrence. The calibration term in the factorisation, $p(x|\hat{x})$, is the quantity usually of most interest to users of forecasts, who wish to know the probability of the weather event given that it was or was not forecast. The second way of factoring the joint distribution is known as the likelihood–base rate factorisation and is given by

$$p(\hat{x},\ x) = p(\hat{x}|x)p(x) \tag{3.43}$$

For example, the probability of a hit $p(\hat{x} = 1,\ x = 1) = a/n$ is equal to $p(\hat{x} = 1|x = 1)p(x = 1) = (a/(a + c))((a + c)/n) = a/n$. The term $p(x)$ is referred to as the *base rate* (*sample climate*) or *climatological probability* of the weather event. The likelihood term $p(\hat{x}|x)$, considered to be a function of x, is the conditional probability of the forecast $\hat{x}$ given the observation x. The calibration–refinement and likelihood–base rate factorisations are related by Bayes' theorem, obtained by equating the right-hand sides of Eqs. (3.42) and (3.43),

$$p(x|\hat{x}) = \frac{p(\hat{x}|x)p(x)}{p(\hat{x})} \tag{3.44}$$

For binary forecasts, $p(x = 1|\hat{x}) = 1 - p(x = 0|\hat{x})$ and $p(x = 1) = 1 - p(x = 0)$, and Bayes' theorem can be written concisely in terms of a relationship for obtaining the posterior *odds* from the *prior odds* using the *likelihood ratio*. Defining the *posterior odds* as

$$\omega(x|\hat{x}) = \frac{p(x|\hat{x})}{1 - p(x|\hat{x})} \tag{3.45}$$

and the *prior odds* (observed climatological odds for the event) as

$$\omega(x) = \frac{p(x)}{1 - p(x)} \tag{3.46}$$

and the *likelihood ratio* as

$$L(\hat{x}|x) = \frac{p(\hat{x}|x = 1)}{p(\hat{x}|x = 0)} \tag{3.47}$$

Bayes' theorem can be rewritten as the product

$$\omega(x|\hat{x}) = L(\hat{x}|x)\omega(x) \tag{3.48}$$

This shows that the odds on an event given the forecast is equal to the odds on the event without the forecast multiplied by the likelihood ratio, so the impact of the forecasts on the odds is summarised in the likelihood ratio. For example, if forecasts of occurrence are considered then $\hat{x} = 1$ and $L(\hat{x} = 1|x) = H/F$, so by Eq. (3.48) the odds of an occurrence given that it was forecast is equal to the climatological odds of an occurrence multiplied by H/F. It can also be seen from Eq. (3.48) that the *posterior odds ratio* $\omega(x|\hat{x} = 1)/\omega(x|\hat{x} = 0) = H/(1 - H) \times (1 - F)/F$ (see Section 3.2) is identical to the ratio of *likelihood ratios*, $L(\hat{x} = 1|x = 1)/L(\hat{x} = 0|x = 1)$, which measures the overall performance of the forecasting system when forecasting both occurrences and non-occurrences.

3.3.2 Performance Measures in Terms of Factorisations of the Joint Distribution

The hit rate, false alarm rate, and base rate are defined in terms of cell counts as:

$$H = a/(a + c) \tag{3.49}$$

$$F = b/(b + d) \tag{3.50}$$

$$s = (a + c)/n \tag{3.51}$$

These are (frequentist) sample estimates of the likelihood and base rate components of the likelihood–base rate factorisation:

$$H = \hat{p}(\hat{x} = 1|x = 1) \tag{3.52}$$

$$F = \hat{p}(\hat{x} = 1|x = 0) \tag{3.53}$$

$$s = \hat{p}(x = 1) \tag{3.54}$$

Inverting these relationships, sample relative frequencies can be expressed in terms of H, F and s as follows:

$$a/n = sH \tag{3.55}$$

$$b/n = (1 - s)F \tag{3.56}$$

$$c/n = s(1 - H) \tag{3.57}$$

$$d/n = (1 - s)(1 - F) \tag{3.58}$$

Thus, all verification measures written in terms of a, b, c and d can be expressed in terms of the three probability estimates: H, F and s.

Similarly, the calibration–refinement factors can be used to describe scores. The true positive ratio, $T = 1 - \mathrm{FAR}$, the miss ratio, M, and the probability of forecasting an occurrence, r, are defined as:

$$T = a/(a + b) \tag{3.59}$$

$$M = c/(c + d) \tag{3.60}$$

$$r = (a + b)/n \tag{3.61}$$

These are sample estimates of the calibration–refinement probabilities

$$T = \hat{p}(x = 1|\hat{x} = 1) \tag{3.62}$$

$$M = \hat{p}(x = 1|\hat{x} = 0) \tag{3.63}$$

$$r = \hat{p}(\hat{x} = 1) \tag{3.64}$$

The elements of the contingency table can be expressed in terms of these three quantities as follows:

$$a/n = rT \tag{3.65}$$

$$b/n = r(1 - T) \tag{3.66}$$

$$c/n = (1 - r)M \tag{3.67}$$

$$d/n = (1 - r)(1 - M) \tag{3.68}$$

Thus, any verification measure written in terms of a, b, c and d can be expressed in terms of the three calibration–refinement probabilities: T, M and r. It should be noted that it is not generally possible to use any three statistics in this way – for example, the familiar verification measures POD, FAR and CSI cannot be used to reconstruct the contingency table, and thus omit some information about forecast quality.

3.3.3 Metaverification: Criteria for Screening Performance Measures

Murphy (1996) coined the term *metaverification* to refer to the process of evaluating performance measures and the development of screening criteria for such measures. The three screening criteria of *equitability*, *propriety* and *consistency* have been described in Chapter 2 (Section 2.8) but will be discussed here again in relation to deterministic binary forecasts. The criteria of *sufficiency* will also be described and a fifth criterion, *regularity*, widely used in signal detection studies will also be covered.

1. *Equitability*

This criterion is based on the principle that random forecasts or constant forecasts of categories should give the same expected score (Murphy and Daan 1985; Gandin and Murphy 1992). Equitability can be assessed by using the likelihood–base rate expression for the score. Constant forecasts of occurrence give $H = F = 1$ and non-occurrence, $H = F = 0$. Random forecasts also give $H = F$, both equal to some value in [0,1] which depends on the proportion of forecasts of occurrence to non-occurrence. For example, Peirce's score is given by $\mathrm{PSS} = H - F$ and is therefore equitable, since any forecast set with $H = F$ must score 0. Proportion correct is given by $\mathrm{PC} = (1 - s)(1 - F) + sH$ and is therefore not equitable, since constant forecasts of occurrence, for which $H = F = 1$, get a value of PC equal to s, and constant forecasts of non-occurrence, $H = F = 0$, score $1 - s$. Random forecasts score somewhere between these values, so any value of PC between s and $1 - s$ is attainable without skill. If the event is rare, this covers most of the range of PC.

2. *Propriety*

This criterion applies only to scoring rules for probability forecasts, although a weaker form, *consistency* (below), applies to deterministic binary forecasts. The propriety criterion is based on a model of the forecasting process in which the forecaster has a subjective judgement of the probability of the weather event to be forecast. This personal probability may or may not be the same as the forecast actually issued. A proper score is such that its subjective expected value is maximised when the issued forecast probability is the same as the forecaster's judgement (Winkler and Murphy 1968). Proper scores thereby encourage forecasters to be 'honest' by discouraging them from *hedging* issued probabilities away from their true beliefs. A *strictly proper* score is maximised if and only if the forecast and judgement coincide (i.e. only one unique maxima – see Dawid 1986).

3. *Consistency*

Murphy and Daan (1985) identified this optimality criterion as an extension of propriety to performance measures for deterministic forecasts. It is assumed that the forecaster's judgement is represented by a probability distribution on the weather event to be forecast, but that forecasters required to issue deterministic binary forecasts follow some 'directive' (decision rule), implied or explicit, for collapsing their judgmental probability distribution onto a single category. In the case of forecasts of a quasi-continuous variable like temperature, the directive might be 'forecast the mean of the probability distribution'. The performance measure consistent with this directive is mean square error, since this is minimised by forecasting the mean. If the directive is 'forecast the median', then the consistent measure is mean absolute error (see Chapter 5 for more discussion of this). Extending Murphy and Daan's concept of consistency to binary predictands, the directive might specify a threshold probability p^* for a forecast of occurrence (Mason 1979). If the current probability for the event is less than p^*, then non-occurrence is forecast, whereas if it is greater, occurrence is forecast. It can be shown that a performance measure *consistent* with this directive is one that is optimal at this threshold probability (see Section 3.3.4). It is possible to design payoff matrices for binary forecasts that give performance measures that are consistent with any p^* (Daw and Mason 1981). Further, all binary performance measures imply a certain payoff structure and are optimised at some specific p^* (Mason 1979). When the directive is 'forecast occurrence when the forecast probability of the event exceeds the base rate', a consistent score is provided by the Peirce skill score (PSS). However, PC would not be consistent with this directive unless the base rate was exactly equal to 0.5, which can be shown to be the optimal threshold probability for PC (see Section 3.3.4). This use of the term 'consistent' as a criterion for performance measures should not be confused with a related but different meaning in reference to forecasts themselves.

Forecasts are said to be consistent if the forecasts as issued are in accord with the forecaster's personal judgement (Murphy 1993).

4. *Sufficiency*

The concept of sufficiency was introduced into forecast evaluation by DeGroot and Fienberg (1983), and developed by Ehrendorfer and Murphy (1988) and Krzysztofowicz and Long (1991a) among others. When it can be demonstrated, sufficiency provides an unequivocal ordering on the quality of forecasts. When two forecasting systems, A and B say, are being compared, A's forecasts are said to be sufficient for B's if forecasts with the same skill as B's can be obtained from A's by a stochastic transformation. Applying a stochastic transformation to A's forecasts is equivalent to randomising the forecasts, or passing them through a noisy channel (DeGroot and Fienberg 1983). A full presentation of the mathematics of the sufficiency relationship is beyond the scope of this chapter. For practical purposes if A is sufficient for B, then A's forecasts are more economically valuable than B's for all possible users (so long as the forecasts are used rationally to maximize the expected value of decisions). It has been found that it is quite unusual for one system to be sufficient for another (Murphy 1997). No performance measures for yes/no forecasts are known that are equivalent to the sufficiency relation.

5. *Regularity*

The concept of a *regular* ROC was introduced by Birdsall (1966) to mean a graph of H against F which passes through (0,0) and (1,1) is complete, i.e. for each value of F there is just one value of H, and convex; i.e. the graph will be on or above straight line segments connecting any two points on it. The term is used in a less restricted sense by Swets (1986a), and here, to mean that the graph passes through (0,0) and (1,1) and is elsewhere interior to the ROC unit square, i.e. it does not touch the axes except at these points. All empirical ROCs so far observed are regular in Swet's sense. The convexity condition of Birdsall's definition is not always satisfied, although deviations are in general slight and near the extremes of the graph. All scores for binary forecasts may be expressed in terms of hit rate, false alarm rate and sample climate, and hence have isopleths on ROC axes which imply a certain variation of hit rate with false alarm rate. ROCs generated by real forecasting systems operating at constant levels of skill have the regular form described above. Thus, it appears reasonable to demand that an acceptable performance measure should have isopleths that are regular on ROC axes. If this is not the case, then equal values of the measure do not necessarily imply equal skill and differences do not necessarily imply real differences in skill, so the measure is unreliable.

3.3.4 Optimal Threshold Probabilities

For deterministic binary forecasts produced by collapsing forecasts of probabilities $\hat{p}$ by using a decision rule such as $\{\hat{x} = 1 \text{ if } \hat{p} > p$ and $\hat{x} = 0$ otherwise$\}$, Mason (1979) showed that it is possible to find the *optimal threshold probability* $p = p^*$ that maximises any given measure's expected value. This decision-theoretic approach to score screening has strong similarities to the approaches used to maximise *forecast value* discussed in Chapter 8.

The optimal threshold probability can be derived as follows. All performance measures for binary forecasts can be regarded as average payoffs (expected loss functions) determined by some payoff matrix of the general form shown in Table 3.5.

The expected payoff (contribution to the score) for forecasting an occurrence $\hat{x} = 1$ is given by

$$E_1 = \bar{p}U_{11} + (1 - \bar{p})U_{10} \tag{3.69}$$

where $\bar{p}$ is $p(x = 1)$. The expected payoff for a forecast of non-occurrence is similarly given by

$$E_0 = \bar{p}U_{01} + (1 - \bar{p})U_{00} \tag{3.70}$$

Therefore, a higher mean score can be obtained by forecasting occurrence when $E_1 \geq E_0$, which can be seen by rearrangement of Eqs. (3.69) and (3.70) to correspond to $\bar{p}$ exceeding the optimal threshold probability p^* given by

$$\bar{p} \geq p^* = \left(1 + \frac{U_{11} - U_{01}}{U_{00} - U_{10}}\right)^{-1} \tag{3.71}$$

The forecast of $\bar{p}$ is $\hat{p}$, so $\bar{p}$ is replaced by $\hat{p}$ in Eq. (3.71).

In the case of PC, for example, the payoff matrix is given by Table 3.6, which contains the increments in PC from the nth to the $(n + 1)$th forecast for each possible contingency of forecast and observation.

Table 3.5 Schematic payoff table. U_{ij} is the reward or penalty for forecast i and observation j

Forecast	Observed	
	Yes	No
Yes	U_{11}	U_{10}
No	U_{01}	U_{00}

Table 3.6 Implied payoff table for proportion correct on the $(n + 1)$th forecast. PC is proportion correct over the previous n forecasts

Forecast	Observed	
	Yes	No
Yes	$(1 - \mathrm{PC})/(n+1)$	$-\mathrm{PC}/(n+1)$
No	$-\mathrm{PC}/(n+1)$	$(1 - \mathrm{PC})/(n+1)$

Making the appropriate substitutions in Eq. (3.71) it is straightforward to show that for PC, $E_1 \geq E_0$ when $\hat{p} \geq 0.5$, so the optimal threshold probability for PC is 0.5. Therefore, a forecaster evaluated using PC can maximise their PC score by forecasting occurrence when the probability of the event is greater than 0.5 and non-occurrence when its probability is less than 0.5. Similar considerations apply to all other performance measure for binary forecasts. The optimal threshold probability is a useful measure of consistency that can be used to screen scores to avoid hedging. For this reason, optimal threshold probabilities have been given for all the scores discussed in Section 3.2.

3.3.5 Sampling Uncertainty and Confidence Intervals for Performance Measures

A particular verification data set can be regarded as just one of many possible samples from a population with certain fixed characteristics. The various verification measures are therefore only finite sample estimates of 'true' population values, and as such are subject to sampling uncertainty. It has been unusual in weather forecast verification studies for any attempt to be made to assess this sampling uncertainty, although without some such attempt it is not possible to be sure that apparent differences in skill are real and not just due to random fluctuations. This is a deficiency in verification practice, since as Stephenson (2000) has pointed out, an estimate or measurement without some indication of precision has little meaning.

Many verification measures for binary forecasts are sample estimates of probabilities, so can reasonably be expected to have the sampling distribution of a proportion. It is possible to calculate exact confidence limits for a probability p based on a sample proportion $\hat{p}$ using the binomial distribution. The theory is discussed in most statistical textbooks and is summarised by Seaman *et al.* (1996). The CI for p obtained by inverting the binomial expression is known as the Clopper–Pearson interval. This interval, while sometimes labelled 'exact', is in fact quite conservative, containing p on more

than its nominal (e.g. 95 %) proportion of occasions, due to the discreteness of the binomial distribution. Consequently, the 'exact' interval is not regarded as optimal for statistical practice (Agresti and Coull 1998). It is common for large enough sample size n to use the normal approximation to the binomial distribution, giving the Wald CI

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}\,(1-\hat{p})/n} \tag{3.72}$$

for the true probability of success p estimated by $\hat{p}$, the fraction of 'successes' in n trials. The quantity $z_{\alpha/2}$ is the appropriate quantile of the standard normal distribution, equal to 1.96 for a 95% interval. The Wald interval tends to contain p less than 95% of the time unless the sample is quite large. As discussed by Agresti and Coull (1998), a more satisfactory CI is given by Wilson's (1927) score method,

$$\left(\hat{p} + \frac{z_{\alpha/2}^2}{2n} \pm z_{\alpha/2}\sqrt{[\hat{p}(1-\hat{p}) + z_{\alpha/2}^2/4n]/n}\right)/(1 + z_{\alpha/2}^2/n) \tag{3.73}$$

A simple modification of the Wald interval which gives results nearly as good as the score method is to replace the estimate $\hat{p}$ in Eq. (3.72) by $\tilde{p} = (y+2)/(n+4)$, i.e. add two successes and two failures, and use the Wald formula (Agresti and Coull 1998).

The validity of the above expressions relies on two assumptions, which should ideally be tested in any data set to which they are applied. These are that the time series of forecast outcomes is stationary, and that successive outcomes are independent. Stationarity means for present purposes that there is no significant change over the period of the data in the skill or decision threshold of the forecasting system or in the climatological probability of the event being forecast. Clearly, if there are trends or discontinuities in these quantities, the validity of any statistic based on the whole sample is questionable. Methods for dealing with non-stationarity are described in many textbooks (e.g. Jones 1985). The second assumption, independence, implies that in the series of forecast outcomes, the probability of an outcome at any particular step conditional on any other outcomes is equal to the unconditional probability of that outcome. Seaman *et al.* (1996) discuss this issue and comment that the independence condition seems unlikely to be completely satisfied in practice, but CIs based on the assumption are still useful, providing at least lower bounds on the uncertainty in the estimated proportions. Hamill (1999) illustrates several methods for assessing serial correlation in binary forecast verification. For situations where non-stationarity and serial correlation are present, non-parametric methods such as resampling may be more appropriate for estimating the CIs on verification scores (Wilks 1995, Section 5.3.2).

3.4 SIGNAL DETECTION THEORY AND THE ROC

There are many situations in which people or systems are required to detect some event or object, or in general to distinguish between two alternative possibilities, on the basis of information which is not enough for certainty. Weather forecasting has many examples of such situations, and others include clinical medicine, psychological testing, non-destructive testing of metals, polygraph lie detection, aptitude testing, survey research and information retrieval. A common feature of these and many other diagnostic systems is that the output may be summarised as a (2×2) array such as Table 3.1.

SDT was developed by engineers to analyse the capacity of electronic systems to detect signals in noise and was applied by psychologists as a model of human discrimination processes (Swets 1973). It has been found useful in these and other fields because it facilitates a distinction between two fundamentally different and independent processes that operate in diagnostic systems, including weather forecasting. These processes are 'discrimination' and 'decision'. The discrimination stage of a diagnostic system assesses the degree to which the current evidence favours one alternative (signal plus noise) rather than the other (noise alone). The output from this stage is a scalar quantity proportional to the current strength of the evidence. The decision process then determines whether this evidence is strong enough to assert that a signal is present in the data or if it consists of noise alone, on the basis of a threshold which partitions the 'strength-of-evidence' scale into two regions corresponding to these two alternatives. For non-probabilistic weather forecasts, the final output is an unequivocal assertion that the weather event will or will not occur.

The distinction between the discrimination and decision processes is achieved by means of an analysis tool known as the relative, or sometimes receiver, operating characteristic (ROC). The term 'relative', rather than 'receiver' is used here, following Swets (1973). The ROC is a graph of the hit rate (Y-axis) against false alarm rate (X-axis) for different decision thresholds. It should not be confused with the *operating characteristic* familiar in statistics, which is a plot of the probability of not rejecting a null hypothesis on the y-axis versus a continuous parameter on the x-axis (rather than another probability as happens in ROC plots) – see Hogg and Tanis (1997, Section 3.7). When there is only a single decision threshold available then only a single point on the ROC curve can be determined and signal detection techniques become less applicable. ROC methods provide three main benefits for the evaluation of forecasting systems:

- a pure index of accuracy, in the sense of the inherent capability of the system to discriminate one state from another;
- quantitative estimates of the probabilities of forecast outcomes (e.g. of hit and false alarm rates) for any decision threshold that the system might use

and the tradeoffs between these probabilities as the decision threshold varies;

- an index of the decision threshold which makes it possible to incorporate climatological probabilities and the values and costs of the various forecast outcomes to determine the threshold that is optimal for the forecasting system in a given situation (Swets and Pickett 1982).

In addition, SDT provides a framework for assessing the validity of performance measures. The characteristic shape of isopleths of skill on ROC axes is well established empirically, and requires two parameters for an adequate description. Measures of skill should be invariant along these isopleths, and a single-number measure cannot describe the types of variation observed (Swets 1986a).

This section gives a necessarily brief outline of SDT and ROC analysis in evaluation of binary weather forecasts. There is a very extensive literature on these methods in experimental psychology and medical diagnosis, and increasing use in other fields including meteorology. Some references in meteorology include Mason (1980, 1982a, 1989), Levi (1985), McCoy (1986), Stanski *et al.* (1989), Harvey *et al.* (1992), Buizza *et al.* (1999) (see also comments by Wilson 2000) and Mason and Graham (1999). Useful overviews and references to the literature outside meteorology can be found in Swets and Pickett (1982), Centor (1991) and Swets (1996).

3.4.1 The Signal Detection Model

In the signal detection model an occurrence of the weather event is supposed to be preceded by a 'signal' in the data, superimposed on a background of 'noise', and the data before non-occurrences is supposed to consist of 'noise alone'. On the basis of the data, the forecaster assesses the weight of evidence for the event. This weight of evidence can be represented as a point on a scalar continuum, here denoted by W. Higher values of W correspond to a higher probability of occurrence. The decision to forecast occurrence or non-occurrence of the event is then made on the basis of a predetermined threshold, denoted w, on the weight of evidence scale, so that the event is forecast if $W > w$ and not forecast otherwise. It is assumed that W has a probability density $f_0(w)$ before non-occurrences and $f_1(w)$ before occurrences. The basic model for the forecasting process is then as shown in Fig. 3.1.

The conditional probability of a hit, H, is the probability that the weight of evidence exceeds the threshold w if the event occurs, so that

$$H = \int_w^\infty f_1(w)\,\mathrm{d}w \tag{3.74}$$

Similarly, the conditional false alarm rate, F, is the probability that the weight of evidence exceeds w, when the event does not occur, so that

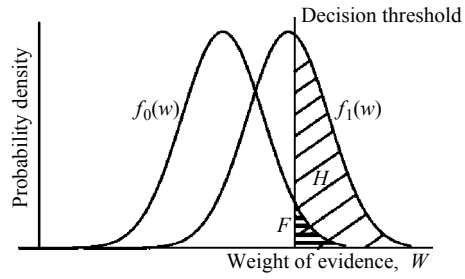


Figure 3.1 The basic signal detection model. $f_0(w)$ is the probability density of the weight of evidence variable before non-occurrences (noise alone), $f_1(w)$ the density before occurrences (signal + noise), and the areas H and F are hit and false alarm probabilities as defined in the text

$$F = \int_{w}^{\infty} f_0(w)\,\mathrm{d}w \tag{3.75}$$

Given the null hypothesis that the event will not occur, and the alternative hypothesis that it will occur, hit rate is analogous to the power of a test in statistical hypothesis testing, or 1 minus the probability of a Type 2 error (Wilks 1995, pp. 116–117). False alarm rate is analogous to the probability of a Type 1 error.

In order to proceed, it is necessary to make assumptions about the functional form of the distributions. The usual assumption, which is well supported by studies in many fields (e.g. Mason 1982a, Swets 1986b), is that the distributions are normal, or strictly speaking that they can be transformed to normality by using a possibly non-linear monotonic transformation of the W-axis. For practical purposes, normal distributions are found to give a very good fit to empirical data, although other distributions may be more appropriate in certain circumstances. Without loss of generality, it can be assumed that the W-axis is scaled so that f_0 has a mean of 0 and standard deviation of 1. The separation of the means of f_0 and f_1, conventionally denoted Δm, and the ratio of the standard deviation of f_0 to that of f_1, σ_0/σ_1, can then be estimated from verification data (Mason 1982a).

3.4.2 The Relative Operating Characteristic

The ROC is a graph of hit rate against false alarm rate as w varies, with false alarm rate plotted as the X-axis and hit rate as the Y-axis. Figure. 3.2 is a typical ROC curve generated by a model of the form of Fig. 3.1.

The location of the whole curve in the unit square is determined by the intrinsic discrimination capacity of the forecasting system, and the location of specific points on a curve is fixed by the decision threshold at which the system is operating. The points labelled low, moderate and high on the

figure represent the performance of the system at three thresholds on the weight-of-evidence axis of Fig. 3.1.

As the decision threshold w varies from low to high, in terms of Fig. 3.1 moving from left to right, H and F vary together to trace out the ROC curve for this system. Low thresholds imply that both H and F are high. In the limit when w is at $-\infty$, both H and F are 1, so the ROC point is at the upper right corner. As the threshold moves to the right in Fig. 3.1 the corresponding performance of the system is represented by ROC points moving to the left and downwards along the curve, until in the limit when w is at $+\infty$ both H and F are 0, at the lower left corner.

An empirical ROC can be plotted from forecasts issued as numerical probabilities or verbal ratings of risk by stepping a decision threshold through the forecasts, each threshold generating a (2×2) contingency table and values for H and F. The method is described in detail by Mason (1982a) and Stanski *et al.* (1989). All empirical ROCs have the basic form shown in Fig. 3.2, necessarily passing through (0,0) and (1,1), and elsewhere interior to the ROC unit square.

Perfect discrimination is represented by an ROC that rises from (0,0) along the h-axis to (0,1), then straight to (1,1). The diagonal $H = F$ represents zero skill, indicating that the forecasts are completely non-discriminatory. This can be seen from Eq. (3.48). Since $H = F$, the likelihood ratio $L(\hat{x} = 1|x)$ in Eq. (3.47) is equal to 1, so the posterior odds on an occurrence given a forecast of occurrence, $\omega(x|\hat{x} = 1)$, is equal to the prior odds on an occurrence without any forecast, $\omega(x)$. Knowing the forecast therefore makes no difference to the odds of an occurrence, and similarly for non-occurrence. Hence the forecast provides no added information.

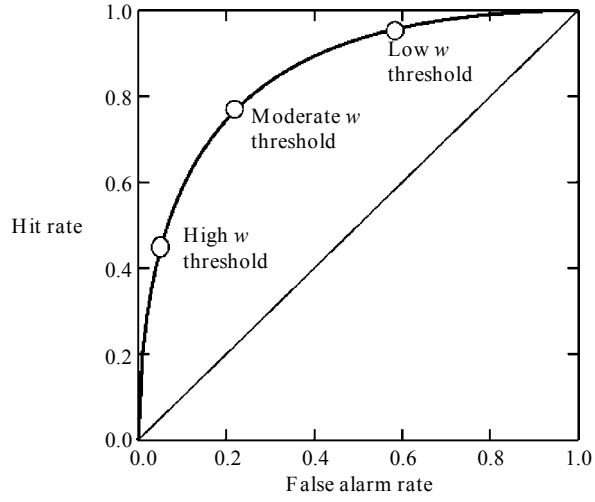


Figure 3.2 ROC generated by a model as shown in Fig. 3.1. The low threshold point is produced by a threshold well to the left in Fig. 3.1, the moderate threshold is near the crossover of f_0 and f_1, and the high threshold is well to the right

Constant forecasts of occurrence plot on ROC axes as the point (1,1), and constant forecasts of non-occurrence as the point (0,0). Forecasts could be produced by a random number generator to give any point along the $H = F$ diagonal.

ROC points below the diagonal represent the same level of skilful performance as they would if reflected about the diagonal. If a forecasting system produces ROC points in this area, the forecasts are mislabelled; forecasts of non-occurrence should be taken as occurrence, and vice versa. This can be regarded as a problem of communication with forecast users, or calibration, rather than with the intrinsic discrimination capacity of the system.

The ROC is often presented on axes transformed to standard normal deviates corresponding to H and F. More precisely, the probabilities H and F are transformed to normal variates $\Phi^{-1}(1-H)$ and $\Phi^{-1}(1-F)$, where Φ^{-1} is the inverse of the standard normal distribution function,

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp -\frac{x^2}{2}\, \mathrm{d}x \tag{3.76}$$

$\Phi^{-1}(1-H)$ gives the value exceeded by a standard normal variable with probability H, and a similar interpretation applies to $\Phi^{-1}(1-F)$.

When the underlying distributions are normal (Gaussian), or can be transformed to normality by a non-linear monotonic transformation, the ROC becomes a straight line on these axes. Mason (1982a) showed that empirical data for weather forecasts are closely linear when plotted on axes transformed in this way, supporting the use of the signal detection model with normal distributions. The separation of the means can be estimated as the intercept on the z_0 (horizontal) axis and the ratio of the standard deviations (σ_0/σ_1) as the slope of the line. Details of this procedure are beyond the scope of this chapter, but can be found in many texts, for example, Swets and Pickett (1982). When the ROC is presented in this way, it is sometimes referred to as a 'bi-normal' plot. Figure 3.3 shows the ROC in Fig. 3.2 plotted on binormal axes. The axes are labelled as 'Z-scores' corresponding to H and F:

$$Z(H) = -\Phi^{-1}(1-H) \tag{3.77}$$

and similarly for $Z(F)$.

When only a single point is available, as is the case with binary forecasts, it is not possible to determine both the slope and intercept of the ROC on binormal axes, so a slope of unity is assumed, implying that the variances of the underlying distributions are equal. On axes linear in probability the effect of non-unit slope is that the untransformed ROC is not symmetrical about the negative diagonal. Real forecasting systems often have ROCs with slopes different from one on binormal axes, ranging from about 0.7 to

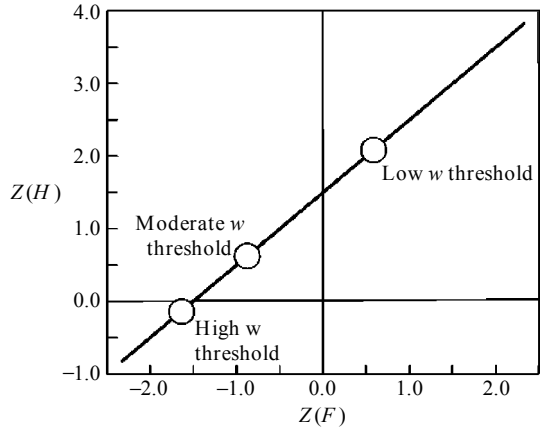


Figure 3.3 ROC on axes transformed to the standard normal deviates of H and F. The three points correspond to those representing low, moderate and high thresholds in Fig. 3.2

1.5. The separation of the means is conventionally denoted d' when equal variances are assumed.

3.4.3 Verification Measures on ROC Axes

Any verification measure defined in terms of the elements of the (2×2) contingency table can also be expressed in terms of the likelihood–base rate factors H, F and s. Since the ROC is a plot of H against F, the likelihood–base rate factorisation shows that verification measures may be plotted on ROC axes and isopleths can be compared with the form of empirical ROCs. This provides a criterion for evaluation of performance measures, since it is known that the forms of ROCs for real forecasting systems correspond closely to those generated by the signal detection model with normal probability densities, plotting as straight lines on bi-normal axes or with 'regular' shape on standard (probability) axes. If a performance measure implies an ROC with a significantly different form, it cannot be a reliable measure of skill, as constant values of the measure will not in general indicate the same level of skill, and differences may or may not indicate differences in skill. To illustrate this point, isopleths of PC on ROC axes are expressed in terms of the likelihood–base rate factors by Eq. 3.12 showing that they are straight lines with slope $(1 - s)/s$ and intercept on the H-axis equal to $\mathrm{PC}/s - (1 - s)/s$. Consider the three forecast sets in Tables 3.7–3.9 identified as A, B, and C, respectively.

B is based on a real set of rain forecasts for Canberra, Australia. A and C were constructed to have the same value of PC as B, equal to 0.80, and the same sample climate, $s = 0.18$, but different thresholds. In Fig. 3.4, all points on the straight line through A, B and C have $\mathrm{PC} = 0.80$ and

$s = 0.18$. The curved line in Fig. 3.4 is the ROC for the signal detection model with equal variance normal distributions passing through the point B.

Forecast set A shows no skill. It lies on the intersection of the PC = 0.80 line with the diagonal, where $H = F \cong 0.028$. The measure of discrimination d' is equal to 0. Forecast set B shows a moderate level of skill, with $H = 0.49$, $F = 0.13$ and $d' = 1.1$. Forecast set C shows a high level of skill, with $H = 0.95$, $F = 0.23$ and $d' = 2.4$. These are clearly very different sets of forecasts. Nevertheless, they all have the same value for PC. The best attainable value of PC for the system that produced B corresponds to the point at which the constant PC line is tangential to the ROC for the forecasting system. It can be shown that this point corresponds to a threshold probability equal to 0.5, which in this case gives PC = 0.83, corresponding to the broken line in Fig. 3.4.

A similar analysis can be made for any verification measure for binary forecasts and isopleths are discussed for each score in Section 3.2. Isopleths

Table 3.7 Forecast set *A*. See text for details

Forecast	Observed		
	Yes	No	Total
Yes	17	76	93
No	577	2617	3194
Total	594	2693	3287

Table 3.8 Forecast set *B*. See text for details

Forecast	Observed		
	Yes	No	Total
Yes	292	351	643
No	302	2342	2644
Total	594	2693	3287

Table 3.9 Forecast set *C*. See text for details

Forecast	Observed		
	Yes	No	Total
Yes	564	623	1187
No	30	2070	2100
Total	594	2693	3287

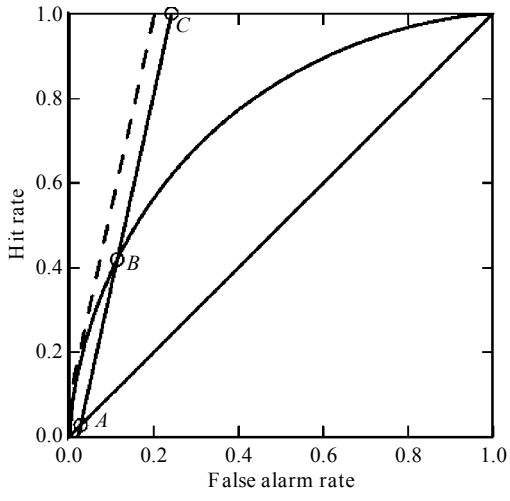


Figure 3.4 Proportion correct (PC) on ROC axes. All points on the line through *A*, *B* and *C* represent forecasts with the same PC and sample climate. The smooth curve is the theoretical equal-variance ROC through *B*. The broken line is the tangent to the ROC with the same slope as the line ABC. See text for details

of the measure can be plotted on ROC axes and compared with the appropriate regular ROC. The ROC therefore provides an effective means of evaluating verification measures.

3.4.4 Verification Measures From Signal Detection Theory

A number of measures of discrimination and decision threshold have been developed based on the ROC and SDT, discussed, for example, in Swets (1996).

(a) *The Area under the Modelled ROC Curve, A_z*

The ROC-based measure of skill most widely used is A_z, the area under the *modelled* ROC on probability axes. In other words, the straight line ROC on normal deviate axes (Fig. 3.3) is transferred to probability axes and A_z is the area under this curve (Fig. 3.5). The subscript z serves as a reminder that the measure was taken from a binormal graph (Swets 1988).

The possible range of A_z is [0,1]. Zero skill is indicated by $A_z = 0.5$, when the ROC lies along the positive diagonal, to 1.0 for perfect skill, when the ROC rises from (0,0) to (0,1) then to (1,1). A value of A_z less than 0.5 corresponds to an ROC curve below the diagonal, indicating the same level of discrimination capacity as if it were reflected about the diagonal, but wrongly calibrated.

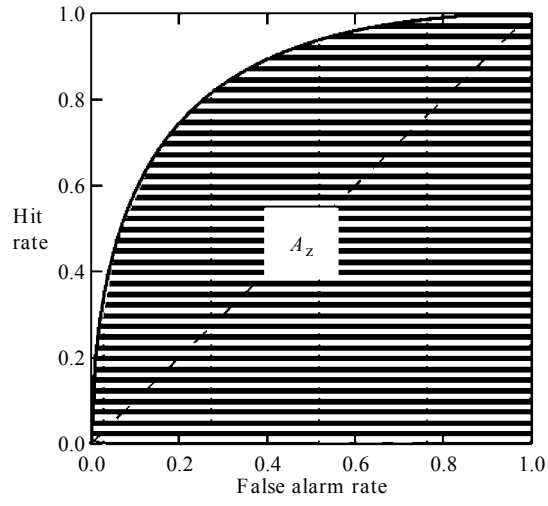


Figure 3.5 A_z, the area below the ROC based on the signal detection model with Gaussian distributions

When only a single point on the ROC is available, the variances of the underlying distributions must be assumed equal and A_z can be calculated from d' as the area under the normal distribution up to the normal deviate value equal to $d'/\sqrt{2}$ (Swets and Pickett 1982)

$$A_z = \Phi\left(\frac{d'}{\sqrt{2}}\right) \tag{3.78}$$

For example, in Finley's forecasts $d'/\sqrt{2} = 1.46$ and $\Phi(1.46) = 0.93$, which is A_z. The score A_z is equitable in the same qualified sense as Yule's Q. All points on the positive diagonal for which $H = F$, corresponding to random forecasts, have the zero-skill value $A_z = 0.5$ except that when $H = F = 0$ or $H = F = 1$, A_z is undefined. This is because all ROCs generated by a signal detection model pass through (0,0) and (1,1), so there can be no unique value of d' or A_z at these points.

It can be shown (Green and Swets 1966) that A_z is equal to the expected value of PC in an experimental design known as a two-alternative forced-choice task. In this situation, two data sets are presented to the diagnostic system in each trial. One data set contains the 'signal' while the other is 'noise alone' and the system is required to state which is which. Since A_z can be interpreted as a proportion, its sampling distribution should follow that of a proportion with n equal to the total sample size. In the case of the Finley's data the Wilson's 95 % CI (Eq. (3.73)) for the underlying true value of A_z is [0.918, 0.937] comfortably far away from the no-skill value of 0.5. This is in agreement with the odds ratio results in Section 3.2.2, and we can conclude at 95 % confidence that Finley's forecasts did have some skill!

The area under the ROC curve can also be crudely estimated by simply joining data points and the (0,0) and (1,1) corners on the ROC diagram with straight lines and then summing the area of the resulting trapezoids (Swets 1986a; Wilson 2000). In the (2×2) case with just a single point on the ROC, this estimate is equal to 0.5(1+PSS). In view of the well-established fact that empirical ROCs on probability axes are generally convex (e.g. Fig. 3.2), this approximation will underestimate the area found by tracing the full ROC. A_z therefore provides a less-biased (parametric) estimate of the performance of the forecasting system.

(*b*) *Discrimination Distance, d′*

The distance d' between the means of the 'noise' and 'signal plus noise' distributions, was first proposed as a signal detection measure by Tanner and Birdsall (1958) for systems that have equal variances. It is estimated by the difference between the standard normal deviates corresponding to $1 - F$ and $1 - H$,

$$d' = \Phi^{-1}(1 - F) - \Phi^{-1}(1 - H) \tag{3.79}$$

For example, in the Finley forecasts (Table 1.1) $H = 0.549$ and the standard normal deviate $\Phi^{-1}(1 - H) = -0.123$. This distance is shown (not to scale), as z_1 the deviation from μ_1 in Fig. 3.6. Similarly, $F = 0.0262$, and $\Phi^{-1}(1 - F) = 1.940$ is shown as z_0 in Fig. 3.6. The distance d' is therefore given by $z_0 - z_1$, equal to 2.06. The possible range of d' is $(-\infty, +\infty)$, but the range encountered in weather forecasts is generally from 0 to 4. Zero skill is indicated by $d' = 0.0$, whereas $d' > 3$ is a very high level of skill, not often encountered in operational weather forecasts. Distance d' is equitable in the same qualified sense as Yule's Q. It takes the no-skill value of 0 for all $H = F$, corresponding to random forecasts, but since Φ^{-1} is undefined for arguments of 0 or 1, d' is undefined when either H or F is 0 or 1. Isopleths of d' are shown in Fig. 3.7.

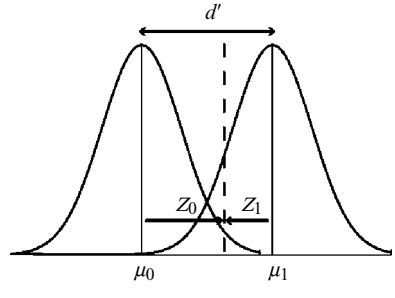


Figure 3.6 Calculation of d'. The vertical broken line is the decision threshold implied by observed hit and false alarm rates. See text for details

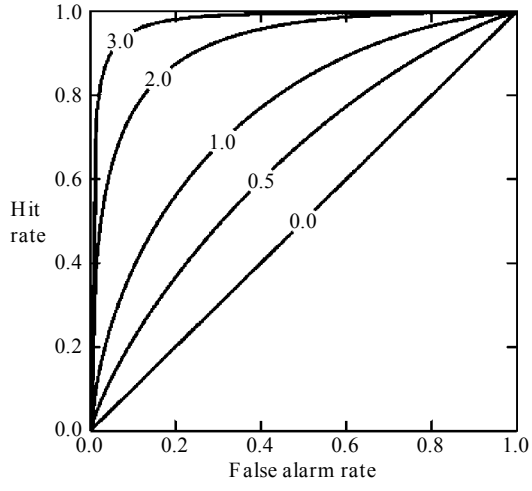


Figure 3.7 Theoretical ROCs generated by equal variance SDT models with Gaussian distributions. Labels are values of d', the separation of the means in units of the common standard deviation

(*c*) *Decision Threshold Index, ROC Slope* $\beta = \mathrm{d}H/\mathrm{d}F$

The location of the decision threshold is important in determining the performance of a forecasting system and there are indices of threshold based on the signal detection model. For a single set of binary forecasts, the slope of the ROC on probability axes at the relevant point is often used. This can be shown to be equal to the likelihood ratio $f_1(w)/f_0(w)$, the ratio of the ordinates of the density functions at the threshold w, denoted β (Green and Swets 1966). The slope β is therefore the ratio of the probability of the weight of evidence at the threshold given occurrence to the weight of evidence at the threshold given non-occurrence. For Finley's data, the implied threshold is 1.94 standard deviations to the right of the mean of $f_0(w)$, and $f_0(1.94) = 0.396$. Similarly, the threshold is 0.123 standard deviations to the left of the mean of $f_1(w)$, and $f_1(0.123) = 0.061$, and therefore $\beta = 0.396/0.061 = 6.52$. The implication is that Finley waited to issue a warning until the current weather situation was at least 6.5 times more likely given tornados than given no tornados.

β can be converted to a probability using Eq. (3.48). For example, using Finley's data again, the prior odds of a tornado are $0.018/0.982 = 0.019$, and the posterior odds are the prior odds multiplied by likelihood ratio, $0.019 \times 6.52 = 0.121$. The corresponding posterior probability, $0.121/1.121 = 0.108$, so that a warning for tornados was issued if the probability given the current weight of evidence was greater than about 11%. This may seem low but it needs to be borne in mind that the base rate of tornado occurrence in this sample was very small (0.018). Finley (unintentionally) waited until the probability of a tornado was more than 6 times greater than its climatological base rate value before issuing a warning.

4 Categorical Events

ROBERT E. LIVEZEY
Climate Services Division, Silver Spring, MD, USA

4.1 INTRODUCTION

In the previous chapter, verification of forecasts of binary events, i.e. forecasts of two categories of discrete events that are collectively exhaustive, was thoroughly examined. Such forecasts reduce to forecasts of either 'yes' or 'no' for some event. Here verification is extended to forecast problems with more than two categories or classes, again mutually exclusive and collectively exhaustive. The focus is strictly on discrete events, but these can be defined in terms of ranges of continuous variables; for example, forecasts of low, near-normal or high temperatures where the categories are defined as temperatures less than or equal to some threshold, between this threshold and a higher one, and greater than or equal to the second threshold, respectively.

One approach to verification of forecasts of three or more categories is to apply the results and procedures of signal detection theory described in Section 3.4 separately to all of the embedded two-category forecast sets. For the three-class temperature forecast example above, the embedded two-class forecast sets are for the occurrence of a low temperature or not, a near-normal temperature or not, and a high temperature or not. Dealing with multiple ROC diagrams simultaneously can be cumbersome and visualization of their joint implications difficult. Likewise the odds ratio score (Yule's Q) for binary forecasts described in Section 3.2.2 cannot be generalized simply to more than two categories. However, log-linear models whose parameters can be related to Q can be constructed for these verification situations. Discussion of these models is beyond the scope of this text, but extensive information about them can be found in Agresti (1996, Chapters 6 and 7).

The difficulties inherent in verifying multi-category forecasts have led to a plethora of proposed scores and misinformation about them in the literature. Fortunately, in the early 1990s, three related sets of equitable skill scores were proposed that embody almost all of the desirable attributes various authors have highlighted. Two of these are specific subsets of the

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

third set elegantly developed by Gandin and Murphy (1992). Their only deficiency may be that they are not regular as defined in Section 3.3.3. Nevertheless, the use of the Gandin and Murphy scores is emphasized here and other scores, like the higher-dimension extensions of the Heidke (HSS) and Peirce (PSS) skill scores, are presented only for purposes of clarification or illustration of points. All of these other scores have considerable deficiencies compared to the Gandin and Murphy scores and are not recommended.

The discussion of skill scores in Section 4.3 makes up the bulk of this chapter. It is preceded in Section 4.2 by material about contingency tables and associated measures of accuracy and followed in Section 4.4 by a discussion of the important topic of sampling variability of skill. The latter addresses the question of whether an apparently successful set of categorical forecasts represents real skill or is just the result of luck. Before proceeding to the next section, two critical issues need to be understood to approach the application of the measures and scores presented below.

The first issue was noted in Section 2.2 but needs to be reemphasized here, namely that a set of forecast event categories can either be ordinal or nominal. These terms refer to whether the order of categories does or does not matter, respectively. There are preferred attributes for scores for ordinal forecast categories in addition to those for nominal ones and the Gandin and Murphy skill scores are constrained in different ways in the two cases. These points are made where appropriate in Section 4.3.

The other important consideration is the estimation of probabilities in determining skill. An example is the application of probabilities estimated from a large sample of previous observation/forecast pairs to a verification of a new set. This requires the assumption that the new sample came from the same population as the previous sample, i.e. that the statistics of the observation/forecast pairs are stationary. In this chapter, the exclusive use of sample probabilities (observed frequencies) of categories of the forecast/observation set being verified is recommended, rather than the use of historical data. The only exception to this is for the case where statistics are stationary and very well estimated. This is rarely the case in environmental verification problems especially those that deal with forecasts on seasonal to interannual or longer time scales. Clear examples of this are presented in Section 4.2 for seasonal temperature forecasts made in the 1980s by the US National Weather Service (NWS). The use of observed current frequencies for estimating probabilities implies that forecast sample sizes are large enough for reasonably accurate estimates. For three or more class forecast problems a greater burden is placed on the sample by estimation of the contingency table than of the observed marginal frequencies. It can be argued that a sample size of the order of $10K^2$ or more 'independent' data points is required to properly estimate the K^2 possible forecast/observation outcomes in a $(K \times K)$ contingency table. Smaller sample sizes would be sufficient to estimate the marginal probabilities of the K different categories.

4.2 THE CONTINGENCY TABLE: NOTATION, DEFINITIONS AND MEASURES OF ACCURACY

4.2.1 Notation and Definitions

The basis for the discussion of verification of categorical forecasts and the complete summary of the joint distribution of forecasts and observations is the contingency table. Let $\hat{x}_i$ and x_i denote a forecast and corresponding observation, respectively, of category i ($i = 1, \ldots, K$). Then the relative sample frequency (i.e. the cell count n_{ij} divided by the total forecast/observation pair sample size n) of forecast category i and observed category j can be written as

$$\hat{p}(\hat{x}_i,\ x_j) = p_{ij}; \quad i,\ j = 1, \ldots,\ K \tag{4.1}$$

The two-dimensional array of all the n_{ij} is referred to as the contingency table. The sample probability distributions of forecasts and observations, respectively, then become

$$\begin{aligned} \hat{p}(\hat{x}_i) &= \sum_{j=1}^{K} p_{ij} = \hat{p}_i; \quad i = 1, \ldots,\ K \\ \hat{p}(x_i) &= \sum_{j=1}^{K} p_{ji} = p_i; \quad i = 1, \ldots,\ K \end{aligned} \tag{4.2}$$

The sample frequencies (4.2) are sometimes referred to as the empirical marginal distributions. Although the notation on the left-hand sides of (4.1) and (4.2) is technically desirable, indicating as it does with '^' that the quantities are *estimated* probabilities, it is somewhat cumbersome. Hence, we will use the simpler notation on the right-hand sides of the equations for the remainder of the chapter.

Examples of contingency tables showing p_{ij} in percent with accompanying marginal distributions are given for US NWS seasonal mean temperature forecasts in three categories in Tables 4.1 and 4.2. The forecasts were made at almost 100 US cities for the years 1983–1990 for February through April (FMA) and June through August (JJA), respectively, so each table is constructed from a sample of almost 800 (not entirely independent) forecasts/observations. The three classes of below-, near- and above-normal temperatures were defined in terms of class limits separating the coldest and warmest 30% of the 1951–1980 record from the middle 40% for each city and season. Thus, the three categories are not equally probable, given the observed climate. Based on the distributions of the observations (the bottom rows) in both tables it seems reasonable to conclude that the seasonal temperature climate of the United States was not stationary between 1951–1980 and 1983–1990; specifically the climate has warmed

between the two periods leading to considerable non-uniformity in the marginal distributions of the observations.

As can be seen in the right-hand columns, curiously the NWS forecasters were cognizant of this climate change for the warm season forecasts (Table 4.2) but not for the cold season forecasts (Table 4.1); the former set (JJA) is clearly skewed towards forecasts of the warm category at the expense of the cold category whereas the latter (FMA) exhibits no clear preference for either extreme category. This may be related to a tendency to issue forecasts of 'least regret' on the part of the forecasters; i.e. the forecasters may have been reluctant to predict relatively warm conditions too frequently in the winter but not in the summer. This kind of forecaster bias is based on the perception that an erroneous prediction of above-normal has more serious consequences for most users than an erroneous forecast of below-normal in the winter and vice versa in the summer.

The ability here to glean useful information from the marginal distributions of forecasts and observations is an example of the power of the distributions-oriented approach to verification first discussed in Section 2.10. Throughout this chapter the use of all of the information in the contingency table will be emphasized.

Table 4.1 Table giving relative frequencies p_{ij} in percent (total sample size $n = 788$) for US mean temperature forecasts for February through April 1983–1990

Forecast	Observed			Forecast distribution
	Below-normal	Near-normal	Above-normal	
Below-normal	7	14	14	35
Near-normal	4	9	16	29
Above-normal	4	8	24	36
Observed distribution	15	31	54	100

Table 4.2 Table giving relative frequencies p_{ij} in percent (total sample size $n = 788$) for US mean temperature forecasts for June through August 1983–1990

Forecast	Observed			Forecast distribution
	Below-normal	Near-normal	Above-normal	
Below-normal	3	8	4	15
Near-normal	8	13	18	39
Above-normal	7	14	25	46
Observed distribution	18	35	47	100

4.2.2 Measures of Accuracy

An extensive list of measures of accuracy for binary forecasts based on the entries of the contingency table and the marginal distributions was introduced in Section 3.2. Only three of these, the proportion correct (PC) (Section 3.2.2), bias (Section 3.2.1) and probability of detection (POD) or hit rate (Section 3.2.2) are both appropriate and frequently used when $K > 2$. Using the notation in Eqs. (4.1) and (4.2) they are defined for K categories, respectively:

$$\mathrm{PC} = \sum_{i=1}^{K} p_{ii} \tag{4.3}$$

$$\mathrm{bias}_i = \hat{p}_i / p_i; \quad i = 1, \ldots, K \tag{4.4}$$

$$\mathrm{POD}_i = p_{ii} / p_i; \quad i = 1, \ldots, K \tag{4.5}$$

The biases reveal whether some forecast categories are being over- or under-forecast while the probabilities of detection quantify the success rates for detecting different categorical events. These quantities along with PCs are presented in Table 4.3 for the forecast/observation sets from Tables 4.1 and 4.2.

The two seasonal mean temperature forecast/observation sets have similar proportions correct (around 40%) with little bias (value close to 1) for near-normal forecasts and only modest probabilities of detection (less than 40%). As discussed earlier, the sets are quite dissimilar for the extreme categories. The winter forecasts have a large cold bias and similar probabilities of detection for above- and below-normal categories. In contrast, for summer the below-normal class is modestly under-forecast and severely under-detected.

It should be noted that, although the next section concentrates on a small number of skill scores with desirable attributes, there are other 'measures of association' for contingency tables that could, in theory, be used as verification scores. Goodman and Kruskal (1979), which collects together four papers published by its authors between 1954 and 1972, illustrates the

Table 4.3 Measures of accuracy for US mean temperature forecasts in three categories (below-, near- and above-normal) for February through April (Table 4.1) and June through August (Table 4.2) 1983–1990

	PC	Bias			POD		
		Below	Near	Above	Below	Near	Above
FMA	0.40	2.30	0.94	0.69	0.47	0.29	0.44
JJA	0.42	0.78	1.11	0.98	0.17	0.37	0.53

range of possibilities. Many of the measures discussed by Goodman and Kruskal are for (2×2) tables (including those covered in Chapter 3) or $(2 \times K)$ tables, but some are for the $(K \times K)$ tables of this chapter.

4.3 SKILL SCORES

In this section, desirable aspects of skill scores for multi-categorical forecasts will first be discussed. Following this, the development of Gandin and Murphy's (1992) equitable scores will be outlined along with a frequently encountered special case. The convenient subset of these scores presented by Gerrity (1992) as well as the subset called LEPSCAT (Potts *et al.* 1996) will then be covered.

4.3.1 Desirable Attributes

Two of the scores introduced in Section 3.2.2, extended to forecasts for more than two categories, are useful to illustrate some of the attributes (by their presence or absence) exhibited by the Gandin and Murphy scores. They are the Heidke and Peirce skill scores given, respectively, by:

$$\text{HSS} = \left(\sum_{i=1}^{K} p_{ii} - \sum_{i=1}^{K} p_i \hat{p}_i \right) \Big/ \left(1 - \sum_{i=1}^{K} p_i \hat{p}_i \right) \tag{4.6}$$

$$\text{PSS} = \left(\sum_{i=1}^{K} p_{ii} - \sum_{i=1}^{K} p_i \hat{p}_i \right) \Big/ \left(1 - \sum_{i=1}^{K} p_i p_i \right) \tag{4.7}$$

Except for the use of different estimates for the number of correct forecasts expected by chance in their denominators, these skill scores are the same. They are both measures of the percent of possible improvement in number of correct forecasts over random forecasts.

Note first that both scores are *equitable* (Sections 2.8 and 3.3); HSS and PSS have zero expectation for random forecasts and for constant forecasts of any single category ($\hat{p}_i = 1.0$). These are highly desirable properties. The form of both of these scores explicitly depends on the forecast distribution, a less desirable property, in contrast to the other equitable scores featured in this chapter.

An example of a non-equitable score is the one formerly used by NWS to verify the forecast sets represented by Tables 4.1 and 4.2. Recall that the forecast categories were defined by class limits for three unequally distributed categories with probabilities of 0.30, 0.40 and 0.30 estimated from 1951–1980 data. The NWS replaced the estimates of expected hit frequency for random forecasts in both the numerators and denominators of Eqs. (4.6) and (4.7) with 0.34 ($= 0.30^2 + 0.40^2 + 0.30^2$), the theoretical probability of

random hits. With a stationary climate the expected PC for constant forecasts of near-normal is 0.40, leading to both positive expected numerators and hence positive expected skills in Eqs. (4.6) and (4.7). For the 1983–1990 winter forecasts and observations in Table 4.1 there is little difference between this non-equitable score and HSS and PSS. All three are between 0.09 and 0.10 in Table 4.4. However, the non-equitable score is substantially higher than the other two (0.12 versus 0.05; Table 4.4) for the 1983–1990 summer forecast/observations in Table 4.2, because the expected proportion of correct forecasts by chance is 0.38 ($= 0.18 \times 0.15 + 0.35 \times 0.39 + 0.47 \times 0.46$), compared to 0.34 for the non-equitable score. The Gerrity skill scores (GS) in Table 4.4 are defined in Section 4.3.3.

Next observe that both HSS and PSS weight all correct forecasts the same regardless of the relative sample probabilities. This encourages a forecaster to be conservative by not providing a greater reward for successful forecasts of lower probability events. By the same token a skill score should provide greater or lesser penalties for different types of incorrect forecasts if categories have different sample probabilities. For $K > 2$ neither of these scores utilize off-diagonal information, i.e. the distribution of incorrect forecasts, in the contingency table. More specifically, if the predictands are ordinal, greater discrepancy between forecast and observed classes should ideally be penalized more than a lesser difference between classes. For example, an erroneous cold forecast should be more heavily penalized if warm is observed than if near-normal is observed.

Lastly, for ordinal predictands, the magnitude of a score should be relatively insensitive to the type or number of categories when applied to forecasts made by assigning categories to objectively produced continuous forecasts. Satisfying this criterion will also ensure that scores computed from smaller dimensional categorizations embedded within the contingency table will be consistent with the score for the full table and each other. Barnston (1992) graphically illustrates how a score similar to the discontinued NWS score applied to equally probable categories does not exhibit this property (Fig. 4.1). He synthetically produced categorical forecasts for different numbers of equally probable categories from continuous forecasts with known linear association (correlation – see Chapter 2 and Section 5.4.4)

Table 4.4 Skill scores for US mean temperature forecasts in three categories for February through April (Table 4.1) and June through August (Table 4.2) 1983–1990. The scores are described in the text

	Score			
	Inequitable	HSS	PSS	GS
FMA	0.09	0.09	0.10	0.17
JJA	0.12	0.05	0.05	0.08

with the observations. Note that a forecaster can achieve a higher score with the same underlying measure of linear association simply by reducing the number of categories, or degree of resolution, of the forecasts. Further, note the (perhaps counter-intuitive) phenomenon that the correspondence between scores and underlying linear associations (correlations) becomes closer as the number of categories decreases in Barnston's experiment. Ideally, this close association should not be degraded with an increase in the number of categories.

The subsets of the complete family of Gandin and Murphy scores described in Sections 4.3.3 and 4.3.4 are constructed in ways that ensure their consistency from categorization to categorization and with underlying linear correlations. These scores likewise are equitable, do not depend on the forecast distribution, do not reward conservatism, utilize off-diagonal information in the contingency table, and penalize larger errors more when predictands are ordinal. It is the view here that the only important distinguishing factor between the two sets is the convenience of their use, with the practical advantage going to the Gerrity (1992) scores rather than LEPSCAT.

4.3.2 Gandin and Murphy Equitable Scores

In this section, the major steps and assumptions used in the axiomatic approach of Gandin and Murphy to develop their family of equitable scores

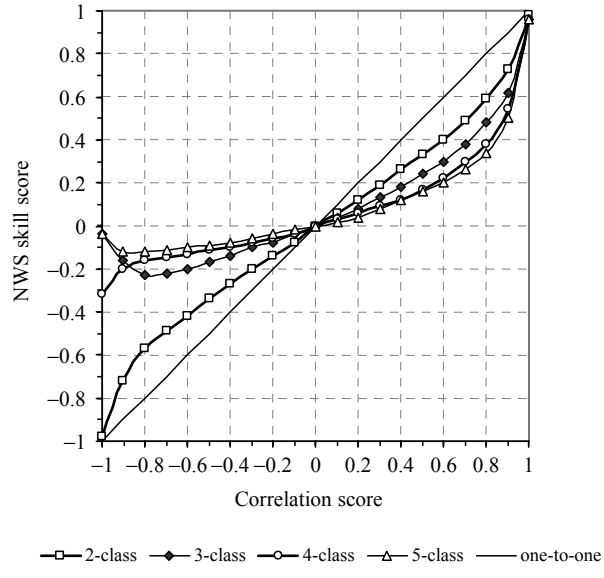


Figure 4.1 A previously used NWS score as a function of correlation score for two (shaded squares), three (open diamonds), four (shaded circles), and five (open triangles) equally likely categories (after Barnston 1992). Note that for a given underlying linear association the score can be increased by reducing the number of forecast categories

will be briefly outlined. A scoring matrix s_{ij} ($i, j = 1, \ldots, K$), is used to define a general form of a skill score using the contingency table:

$$\text{GMSS} = \sum_{i=1}^{K} \sum_{j=1}^{K} p_{ij} s_{ij} \tag{4.8}$$

The scoring matrixs is a tabulation of the reward or penalty every forecast/observation outcome represented by the contingency table is accorded. It is similar in spirit to the expense matrix of the simple cost–loss model used to assess forecast value (see Section 8.2). The remaining problem is to determine what the K^2 elements of the scoring matrix should be to ensure that GMSS is equitable. If random forecasts are arbitrarily required to have an expected score of 0 and the score for perfect forecasts is set to 1, then $K + 1$ relationships constraining the s_{ij} are:

$$\begin{aligned} \sum_{j=1}^{K} p_j s_{ij} &= 0, \quad i = 1, \ldots, K \\ \sum_{i=1}^{K} p_i s_{ii} &= 1 \end{aligned} \tag{4.9}$$

The first K relationships in Eq. (4.9) simply state that constant forecasts of any category i must have a score of 0 while the other relationship constrains a perfect score to be 1. Eqs. (4.9) are insufficient to determine the full scoring matrix: if $K = 3$, they provide only four out of nine necessary relationships and, more generally, only $K + 1$ of K^2.

The number of remaining relationships needed can be reduced by $K(K - 1)/2$ by the assumption of symmetry for S:

$$s_{ji} = s_{ij} \tag{4.10}$$

This condition states that it is no more or less serious to forecast class i and observe j than vice versa and is reasonable in the vast majority of environmental prediction problems. There may be user-related situations where the condition would not be appropriate and the development of equitable scores would proceed differently from what is presented here. For example, Jolliffe and Foord (1975) developed equitable scores for $K = 3$ in which they insist that $s_{11} = s_{22} = s_{33}$. Imposing this constraint together with $s_{13} = s_{31}$ and $s_{21} = s_{23}$ leads to $s_{12} \neq s_{21}$. Another reasonable constraint on S is to require that the reward for an incorrect forecast be less than a correct one:

$$\begin{aligned} s_{ij} &\leq s_{ii} \\ s_{ij} &\leq s_{jj} \end{aligned} \tag{4.11}$$

If the predictand categories are ordinal, a final constraint requires that the reward for an incorrect forecast be less than or equal to an incorrect forecast that misses the observed category by fewer classes, i.e. for three-category predictands a two-class error is rewarded less than or equal to a one-class error:

$$\begin{aligned} s_{i'j} &\leq s_{ij}, \quad |i' - j| > |i - j| \\ s_{ij'} &\leq s_{ij}, \quad |i - j'| > |i - j| \end{aligned} \tag{4.12}$$

Eqs. (4.9)–(4.12) completely describe the Gandin and Murphy family of equitable scoring matrices. Note that both subsets of this family presented in Sections 4.3.3 and 4.3.4, respectively, are for ordinal predictand categories.

For three-category predictands, Eq. (4.10) reduces the number of missing relationships to determine S to two. Thus, two elements of S have to be specified and Gandin and Murphy set $s_{12} = k_1$ and $s_{23} = k_2$. The elements of the scoring matrix then become:

$$\begin{aligned} s_{11} &= [p_3 + p_1(p_3 - p_2)k_1 + p_3(p_2 + p_3)k_2]/[p_1(p_1 + p_3)] \\ s_{13} &= -[1 + (p_1 + p_2)k_1 + (p_2 + p_3)k_2]/(p_1 + p_3) \\ s_{22} &= -(p_1k_1 + p_3k_2)/p_2 \\ s_{33} &= [p_1 + p_1(p_1 + p_2)k_1 + p_3(p_1 - p_2)k_2]/[p_3(p_1 + p_3)] \end{aligned} \tag{4.13}$$

To complete the construction of an equitable score for three-category predictands, choices have to be made for k_1 and k_2. These choices are limited by constraints of Eq. (4.11) and, if the predictands are ordinal, of Eq. (4.12) as well. Application of the constraints leads to the permissible values defined by the quadrilaterals in Fig. 4.2; nominal predictand categories by the larger quadrilateral and ordinal by the square.

Gandin and Murphy provide a numerical example in which $p_1 = 0.5$, $p_2 = 0.3$ and $p_3 = 0.2$, where the three categories are far from being equally probable and the categorization is highly asymmetrical. Choosing $k_1 = -0.5$ and $k_2 = -0.25$, the midpoint of the bottom of the square in Fig. 4.2 leads to the scoring matrix in the upper left section of Table 4.5. Also shown for contrast is the case for $p_1 = 0.2$, $p_2 = 0.5$ and $p_3 = 0.3$, in which the categorization is more symmetrical, and the scoring matrices in both cases for the scores developed in the next subsection (where they will be discussed). The most important thing to note about the scoring table for the asymmetrical case is that the reward for a correct forecast of event three is double that for event two and over three times that for event one. Event three is not only the least probable but also represents one of the two tails of the ordinal continuum encompassed by the three events, so the reward for correctly forecasting it should be substantially higher than forecasting event two (with only a slightly larger probability). Another interesting feature of this scoring matrix is the fact that the penalty for

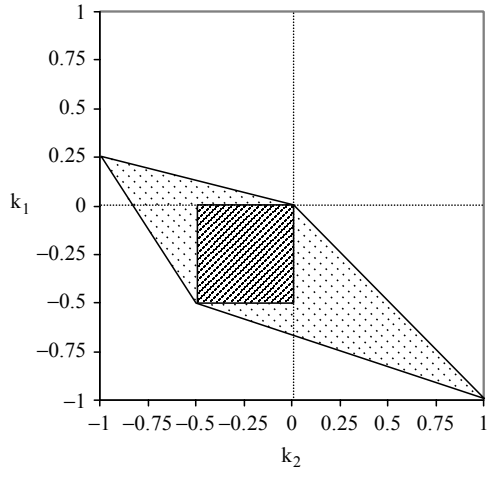


Figure 4.2 Acceptable domains for numerical values of the specified elements $k_1 (= s_{12})$ and $k_2 (= s_{23})$ (after Gandin and Murphy 1992)

Table 4.5 Equitable scoring matrices for three-category forecasts with two asymmetric sets of event probabilities

	Event probabilities (p_1, p_2, p_3)	
	(0.5,0.3,0.2)	(0.2,0.5,0.3)
Gandin and Murphy (1992)	$k_1 = -0.5,\ k_2 = -0.25$ $\frac{1}{28}\begin{bmatrix} 16 & -14 & -19 \\ -14 & 28 & -7 \\ -19 & -7 & 58 \end{bmatrix}$	$k_1 = -0.5,\ k_2 = -0.25$ $\frac{1}{60}\begin{bmatrix} 156 & -30 & -54 \\ -30 & 21 & -15 \\ -54 & -15 & 61 \end{bmatrix}$
Gerrity (1992)	$k_1 = -0.375,\ k_2 = 0.0$ $\frac{1}{8}\begin{bmatrix} 5 & -3 & -8 \\ -3 & 5 & 0 \\ -8 & 0 & 20 \end{bmatrix}$	$k_1 = -0.286,\ k_2 = -0.375$ $\frac{1}{168}\begin{bmatrix} 372 & -48 & -168 \\ -48 & 57 & -63 \\ -168 & -63 & 217 \end{bmatrix}$

incorrectly predicting event three but observing event one and vice versa is greater than the reward for correctly forecasting the highly probable event one. This is not the case for the more symmetrical categorization (upper right of Table 4.5) because events one and three have more comparable probabilities. However, the rewards for correctly forecasting these two events are quite disparate, suggesting that $k_1 = -0.5$ and $k_2 = -0.25$ are not the best choices for this verification situation. The related scores developed in the next subsection lead to a more reasonable reward/penalty matrix, so trial and error is not recommended. Before these scores are

presented the scoring matrices for a frequently encountered three-event special case are examined to conclude this section.

The special case is that for symmetric event probabilities, i.e. when $p_1 = p_3$. With this condition only one of the elements of the scoring matrix needs to be specified. Gandin and Murphy set

$$\begin{aligned} s_{12} &= s_{23} = k, \qquad \text{leading to:} \\ s_{11} &= s_{33} = (1 + 2kp_1)/2p_1 \\ s_{13} &= -[1 + 2k(1 - p_1)]/2p_1 \\ s_{22} &= -2kp_1/(1 - 2p_1) \end{aligned} \tag{4.14}$$

The range of permissible values of k is 0 to -1 inclusive, but if the events are ordinal and the categories are defined so that both events one and three can be considered equidistant from event two, then $-0.5 \leq k \leq 0$. The equidistant assumption may not always be reasonable, especially in those cases where the distribution of the underlying variable is highly skewed. An example is precipitation where the range of values for equiprobable tails will be very dissimilar, with a much narrower range for the below-normal precipitation class compared to above-normal.

When $k = -0.5$, all of the off-diagonal elements of the scoring matrix are equal to k, appropriate for nominal events where no distinction is made between incorrect forecasts. The values of all elements of S are presented in Fig. 4.3 for ranges of p_1 and k. Except for very large and very small values of p_1, S is not very sensitive to k. Scoring matrices are shown in Table 4.6 for both $p_1 = 0.33$ and $p_1 = 0.3$ with a choice of $k = -0.25$, midway in its permissible range for the ordinal case. Also included in the table for the same examples are scores developed in the next two subsections that are part of the Gandin and Murphy family of equitable scores. Note that for all S rewards for correct predictions of events one and three are considerably greater than for event two, penalties for two-class errors are considerably greater than for a miss by just one event, and that both of these large rewards and penalties increase with a decrease in p_1.

4.3.3 Gerrity Equitable Scores

Gerrity (1992) discovered the construction of a subset of the Gandin and Murphy scores for ordinal categorical event forecasts which ensures their consistency or equitability (see Sections 2.8 and 3.3.3), provides convenient and alternative means for their computation, and seems to lead to reasonable choices for various k's. These scores will be denoted by GS. The construction of their scoring matrices start with the following definition:

$$a_i = \frac{1 - \sum_{r=1}^{i} p_r}{\sum_{r=1}^{i} p_r} \tag{4.15}$$

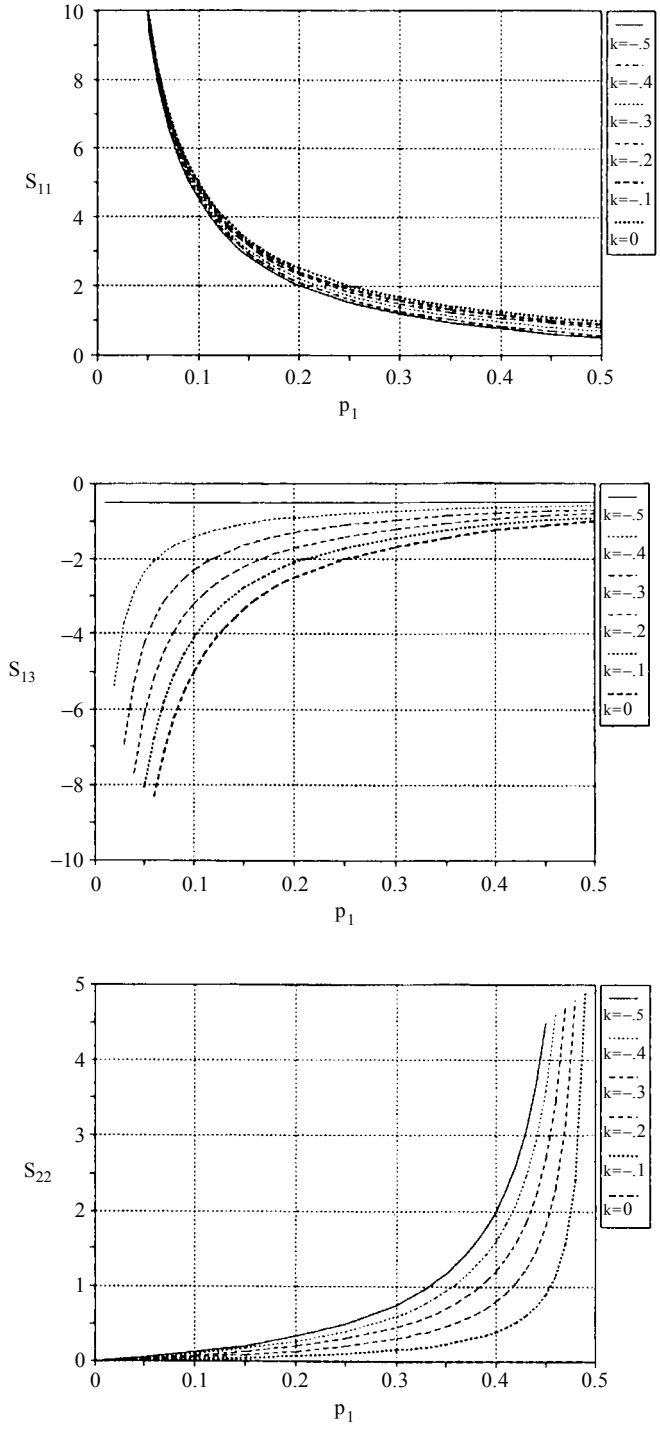


Figure 4.3 Numerical values of the elements in the scoring matrix in the special (3×3) situation as a function of the climatological probability p_1 for selected values of the specified element $k(= s_{12})$: (top) s_{11}, (middle) s_{13}, (bottom) s_{22} (from Gandin and Murphy 1992)

Table 4.6 Equitable scoring matrices for three-category forecasts with symmetric sets of event probabilities

	Event probabilities (p_1, p_2, p_3)	
	(0.33,0.33,0.33)	(0.3,0.4,0.3)
Gandin and Murphy (1992)	$k = -0.25$ $\frac{1}{24}\begin{bmatrix} 30 & -6 & -24 \\ -6 & 12 & -6 \\ -24 & -6 & 30 \end{bmatrix}$	$k = -0.25$ $\frac{1}{24}\begin{bmatrix} 34 & -6 & -26 \\ -6 & 9 & -6 \\ -26 & -6 & 34 \end{bmatrix}$
Gerrity (1992)	$k = -0.25$ $\frac{1}{24}\begin{bmatrix} 30 & -6 & -24 \\ -6 & 12 & -6 \\ -24 & -6 & 30 \end{bmatrix}$	$k = -0.286$ $\frac{1}{21}\begin{bmatrix} 29 & -6 & -21 \\ -6 & 9 & -6 \\ -21 & -6 & 29 \end{bmatrix}$
Potts *et al.* (1996)	$k = -0.167$ $\frac{1}{36}\begin{bmatrix} 48 & -6 & -42 \\ -6 & 12 & -6 \\ -42 & -6 & 48 \end{bmatrix}$	$k = -0.18$ $\frac{1}{33}\begin{bmatrix} 49 & -6 & -41 \\ -6 & 9 & -6 \\ -41 & -6 & 49 \end{bmatrix}$

The elements of S are then given by:

$$
\begin{aligned}
s_{ii} &= b\left(\sum_{r=1}^{i-1} a_r^{-1} + \sum_{r=i}^{K-1} a_r\right) \\
s_{ij} &= b\left(\sum_{r=1}^{i-1} a_r^{-1} - (j-i) + \sum_{r=j}^{K-1} a_r\right); \quad 1 \le i < j \le K \\
s_{ji} &= s_{ij} \\
b &= \frac{1}{K-1}
\end{aligned}
\tag{4.16}
$$

Recall that all Gandin and Murphy scoring matrices (including these) are symmetrical. Note also that the summations are 0 for those cases in Eq. (4.16) when the upper index is less than the lower. Finally observe in Eq. (4.16) that $s_{iK} = -1$ always.

Three-category Gerrity scoring matrices are included in Tables 4.5 and 4.6 for comparison with the Gandin and Murphy versions developed with arbitrarily selected constants. There are no substantial differences between the matrix elements for the symmetric categorizations included in Table 4.6. In fact the tables (and implied k's) are identical for the case of three equally probable categories. On the other hand large differences show up for the cases represented in Table 4.5. Both scoring matrices for the highly

asymmetrical categorization (0.5,0.3,0.2) seem reasonable, but the Gerrity matrix seems a more logical choice for the more symmetrical case with a highly probable event two (0.2,0.5,0.3). The Gerrity k's lead to less disparate rewards for correct forecasts of events one and three, with the former exceeding the latter by much less than a factor of 2 rather than by much greater.

The fact that the Gerrity scoring matrices all seem to have reasonable rewards/penalties may be related to a remarkable property of GS. The Gerrity score can be alternatively computed by the numerical average of $K - 1$ two-category scores (which are identical to PSS; when Eqs. (4.15) and (4.16) are evaluated for $K = 2$ and substituted into Eq. (4.8) we get Eq. (4.7)). These are the $K - 1$ two-category contingency tables formed by combining categories on either side of the partitions between consecutive categories. For three-category temperature forecasts two-category scores would be computed with Eq. (4.7) for $K = 2$ for (1) above-normal and a combined near- and below-normal category and (2) below-normal and a combined near- and above-normal category, and then simply averaged. This convenient property of GS also guarantees a considerable amount of consistency between the various scores computed from different partitions of the contingency table.

The Gerrity score has been computed for each of the data in Tables 4.1 and 4.2. The results are included in Table 4.4 to contrast with the non-equitable score, with HSS, and with PSS. In both cases GS is greater than both HSS and PSS, considerably so for the cold season temperature forecasts. This is mainly because the rewards ($s_{11} = 3.42$ for Table 4.1) for correct below-normal forecasts (7% of the total) by themselves outweigh the penalties for all of the two-class errors (18% of the total; well below the number expected by chance). The reward for a correct forecast of above-normal temperature ($s_{33} = 0.51$) is only about 1/7 of that for a below-normal forecast. However, there are so many of the former (exceeding the expected number) that they also contribute meaningfully to GS. The other scores clearly are deficient in rewarding the NWS forecasters for successfully predicting the less likely below-normal category and for making relatively few large forecast errors. Because of its convenience and built-in consistency, the family of GS is recommended here as equitable scores for forecasts of ordinal categorical events.

4.3.4 LEPSCAT

Another subset of the equitable Gandin and Murphy scores for ordinal predictands that are consistent by construction are those introduced by Ward and Folland (1991) and later refined by Potts *et al.* (1996) called LEPSCAT. LEPS is the acronym for 'linear error in probability space' and CAT stands for 'categorical', because LEPS has been developed for continuous predictands as well. In fact, the latter is given by

$$L = 3\left(1 - |F_{\hat{x}} - F_x| + F_{\hat{x}}^2 - F_{\hat{x}} + F_x^2 - F_x\right) - 1 \tag{4.17}$$

where the F's are the cumulative distribution functions, respectively, of the ordinal forecasts and observations. The family of LEPS scores is based on the linear distance between the forecast and the observation in their sample probability spaces, the second term in L. The other terms ensure that L is equitable and does not exhibit certain pathological behavior at its extremes. Normalization is provided by the factor 3. To develop the scoring matrix S for categorical forecasts, Eq. (4.17) is averaged over all possible values of forecasts and observations for each outcome of the contingency table.

The results for equally probable three-category events and the symmetrical categorization (0.3,0.4,0.3) are shown for comparison in Table 4.6. The LEPSCAT matrices are renormalized to facilitate comparisons with corresponding ones in Table 4.6; all but the corner entries match for matrices with the same categorization. It is clear that the contrasts between rewards/penalties for forecast/observation pairs involving event two and other pairs involving only events one and three are much greater in the LEPSCAT matrices than in either the corresponding Gandin and Murphy or Gerrity scores, perhaps too much so. In particular, LEPSCAT (1) rewards correct forecasts of event two much less and those of events one and three somewhat more, and (2) penalizes one-category misses (e.g. event one forecast/event two observed) much less and two-category misses (e.g. event one forecast/event three observed) much more.

LEPSCAT scores are part of the Gandin and Murphy family of scores because they satisfy Eqs. (4.9)–(4.12). Substitution of the LEPSCAT k's, respectively, into Eq. (4.14) results in exactly the LEPSCAT scoring matrices in Table 4.6. This set of scores provides the opportunity to illustrate the consistency of the two constructed Gandin and Murphy scores and their correspondence with a linear correlation measure of association. Potts (personal communication) performed the same experiment for LEPSCAT scores for equally probable categorical events as that by Barnston described in Section 4.3.1 for the non-equitable NWS score. The result was practically no difference between scores for two through five categories for underlying linear correlations greater than -0.6, with both sets of scores only moderately different from the correlation in this range (Fig. 4.4).

4.3.5 Summary Remarks on Scores

Based on all factors, the Gerrity scores (Section 4.3.3) are an appropriate choice for ordinal categorical event forecast verification problems. They have all the desirable properties outlined in Section 4.3.1, their scoring matrices or the scores themselves are easy to calculate, and in every instance examined here produced reasonable reward/penalty matrices. Nevertheless, LEPSCAT scores provide an excellent alternative. Other major sources for information on verification of forecasts of more than two categorical events

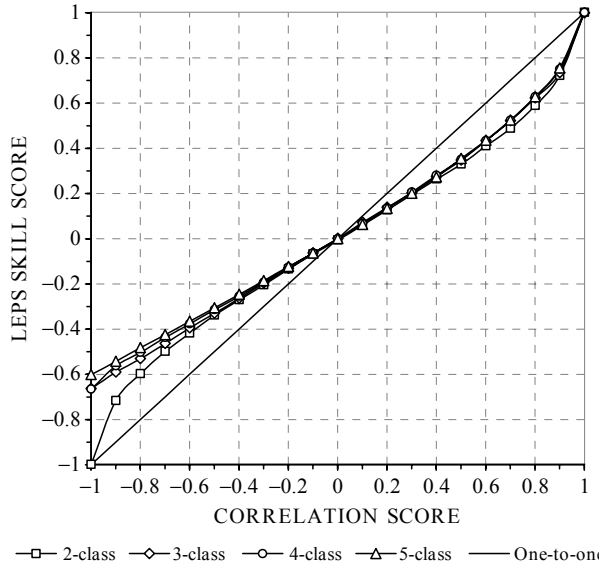


Figure 4.4 LEPSCAT as a function of correlation score for two (shaded squares), three (open diamonds), four (shaded circles) and five (open triangles) equally likely categories. In contrast to Fig. 4.1 there is little difference between the four curves and only moderate differences from the underlying linear correlations over most of their range

are Murphy and Daan (1985), Stanski *et al.* (1989) and Wilks (1995, Section 7.2). Each of these sources is rich in information, but all three are out of date. Wilks does mention LEPS and Gandin and Murphy (1992) scores but does not fully explore their relationships and implications.

4.4 SAMPLING VARIABILITY OF THE CONTINGENCY TABLE AND SKILL SCORES

Two related questions about Tables 4.1 and 4.2 and the associated measures of accuracy and skill in Tables 4.3 and 4.4, respectively, are (1) whether there is sufficient confidence that they reflect non-random forecasts and (2) given sufficient confidence, are the computed accuracy and skill measures reliable enough? With only one cautious exception, classical statistics offers little assistance in answering these questions. Nevertheless, with reasonable insight into the spatial–temporal characteristics of the forecast/observation set, various resampling strategies are available that can lead to satisfactory answers.

The standard way to test the null hypothesis that an unevenly populated (or 'extreme') contingency table was the result of independent, randomly matched forecast/observation pairs for categorical events is to perform a 'chi-squared' (χ^2) test. In the notation of Section 4.2 (with n denoting the independent sample size) the test statistic takes the form

$$X^2 = n \sum_{i,\ j=1}^{K} (p_{ij} - \hat{p}_i p_j)^2 / \hat{p}_i p_j \tag{4.18}$$

This expression, called the 'Pearson' chi-squared statistic, is equivalent to the sum over every cell of the $(K \times K)$ contingency table of the squared difference between actual occurrences and the number expected by chance for the cell divided by the number expected by chance. An alternative to Eq. (4.18) with the same asymptotic null distribution is the 'likelihood ratio' chi-squared:

$$G^2 = 2n \sum_{i,\ j=1}^{K} p_{ij} \log\left(p_{ij} / \hat{p}_i p_j\right) \tag{4.19}$$

In a sense G^2 is more fundamental than X^2 as it is a likelihood ratio statistic. However, the two are asymptotically identical and an advantage of the Pearson chi-squared (4.18) over the likelihood ratio (4.19) is that the sampling distribution of the former tends to converge more rapidly to the asymptotic distribution as n increases for fixed K (Agresti 1996, Section 2.4.7 – see also Stephenson 2000).

A crucial thing to note about Eqs. (4.18) and (4.19) is that they are directly proportional to the sample size; two tables of identical relative frequencies will have different X^2 and G^2 if they are constructed from different size samples. The asymptotic distribution of X^2 or G^2 for different degrees of freedom m is well known and is tabulated in a large number of sources. The degrees of freedom are the number of cells in the contingency table (K^2) minus the number of restrictions on the counts expected by chance in the cells. For the tables discussed in this chapter, each row and column is constrained to sum to its respective observed marginal total. But one of these $2K$ totals is dependent on the rest because it can be determined easily from them. Thus, there are $2K - 1$ restrictions and $m = (K - 1)^2$. For Tables 4.1 and 4.2 $m = 4$. For this case the values X^2 or G^2 must exceed, for their chance probability to be less than, respectively, 0.1, 0.01 and 0.001, are listed in Table 4.7 along with the computed values of X^2 and G^2 for each table for $n = 788$, the actual number of forecasts used to construct both tables, and for $n = 100$.

The results given by Tables 4.1 and 4.2 are statistically significant at least at the 0.1% level if it can be assumed that all 788 forecasts are independent of each other. In fact, for both the seasonal mean temperature observations and forecasts it is well known that this assumption is not warranted. Each seasonal forecast was made at almost 100 locations over the contiguous United States. The cross-correlations among these sites for seasonal mean temperatures are very strong, so that the 100 stations behave statistically like only about 10 independent locations. This implies that there are perhaps only the equivalent of about 80 independent data points contributing

Table 4.7 Values for X^2 and G^2 for US mean temperature forecasts in three categories for February through April (Table 4.1) and June through August (Table 4.2) 1983–1990 for two different independent sample sizes, 100 and 788 (The 10%, 1% and 0.1% critical values of the relevant χ^2 distribution are 7.78, 13.28, 18.47, respectively.)

		$n = 100$	$n = 788$
FMA	X^2	5.07	39.98
	G^2	5.12	40.41
JJA	X^2	3.92	30.92
	G^2	3.97	31.25

to Tables 4.1 and 4.2. Hence, the values corresponding to $n = 100$ in Table 4.7 indicate that the results in Tables 4.1 and 4.2 are probably not even statistically significant at the 10% level.

There is also some serial correlation between one forecast year and the next but this is weak in this example compared to the uncertainties introduced by the spatial correlations. In some applications temporal correlations in either the forecasts or observations may be strong and introduce additional ambiguities into the application of a chi-square test.

For these and other situations where spatial (cross) and/or temporal (serial) correlations cannot be ignored, strategies based on resampling and randomization techniques can often be employed to obtain satisfactory answers to the two questions posed at the beginning of this section. Livezey (1999), von Storch and Zwiers (1999, Section 6.8) and Wilks (1995, Section 5.4) all discuss these situations (with examples) and outline methods for dealing with them. So-called 'bootstrap' techniques (Efron and Tibshirani 1993) will be outlined below for the US seasonal mean temperature examples used throughout this chapter to illustrate the application of these methods, although there is no particular reason to prefer them over other resampling approaches.

Let the vectors containing mean temperature observations and forecasts, respectively, for US locations for particular February through Aprils be denoted by $X^i, \hat{X}^i$ ($i = 1983, \ldots, 1990$). The objective is to develop a null distribution (from random forecasts) for either X^2, a measure of accuracy, or a skill score to compare against an actual computed value. This null distribution must preserve the non-trivial spatial and temporal correlations in the observation and forecast samples. In this example, only the spatial cross-correlations are important so the sample is built up by randomly mismatching maps (the vectors) of the observations with maps of the forecasts. This is done by randomly selecting a forecast map from the eight

possibilities (1983–1990) to match against the 1983 observed map, replacing the selected forecast map in the pool, and repeating the process for each subsequent year's map of observed February through April mean temperatures. Symbolically, each X^i ($i = 1983, \ldots, 1990$) is paired with a randomly selected $\hat{X}^j$ ($j = 1983, \ldots, 1990$). This random selection *with replacement* is what distinguishes the bootstrap technique from other randomization techniques, like permutation methods. There are 16,777,216 different selections for the bootstrap compared with 40,320 distinct permutations. The various statistics, measures, or scores of interest are computed from this sample and saved. The whole process is then repeated enough times (1000 is common) to produce smooth probability distributions for testing actual computed values.

The test is conducted by determining the proportion of the bootstrap sample that is (in the case of X^2 or a skill score) larger than the value being tested. Based on the size of this proportion a decision can be made whether or not to reject the null hypothesis.

If a null hypothesis can be reasonably rejected for whatever statistic or measure is being tested, for example, the Gerrity score, then another bootstrap procedure can be used in this example to determine confidence limits for the corresponding population score. To do this a large number of bootstrap samples of eight correctly paired forecast and observation maps (X^i, $\hat{X}^i$) are constructed, each by eight random draws from the full pool of eight February through April maps. For each of these samples the quantity of interest is computed, resulting in a large enough sample for a smooth distribution from which confidence intervals can be determined.

If serial correlation is important, these approaches can be appropriately modified to produce applicable distributions as long as the original sample is not too severely constrained in time. Whatever the case, unless a very large number of independent forecast/observation pairs are available or computed accuracies and skills are very large the analyst should view his/her verification results critically and not make unwarranted conclusions about forecast performance.

5 Continuous Variables

MICHEL DÉQUÉ
Meteo-France CNRM/GMGEC/EAC, Toulouse, France

5.1 INTRODUCTION

This chapter will present methods for the verification of real continuous scalar quantities such as temperature, pressure, etc. In practice, not all values are possible for physical variables (e.g. negative precipitation, or temperatures below absolute zero), but it is simpler to consider that all real values in the range $[-\infty, \infty]$ can be reached with the restriction that the probability density is zero outside the physically achievable range. Continuous real variables are commonly produced (e.g. by partial differential equations), and categorical forecasts, which were considered in the previous two chapters, are often obtained by applying thresholds to continuous variables. Computers always represent real numbers at finite precision, and therefore produce, from a mathematical point of view, discrete representations of continuous real variables. However, for high enough machine precision (e.g. 32 bits), the continuous assumption is reasonable. One can also argue that user-related decision problems are often categorical, e.g. a user who wants to prevent frost damage does not care whether the predicted temperature is exactly 10 or 11 °C. This example highlights the difference between forecast verification, which does not have to impose thresholds, and is thus more general, and the assessment of forecast value, which requires a more categorical decision-based approach.

This chapter is structured as follows. Section 5.2 presents the two forecasting examples used in the rest of the chapter. Verification criteria based on first-order and second-order moments are described in Sections 5.3 and 5.4, respectively. Section 5.5 introduces scores based on the cumulative frequency distribution. Concluding remarks are given in Section 5.6.

5.2 FORECAST EXAMPLES

In this chapter, two examples are taken from the PROVOST project seasonal forecast experiments (Palmer *et al.* 2000). The PROVOST project

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

used four climate models to produce ensembles of 4-month mean forecasts for the winters 1979–1993. Each ensemble consists of nine model forecasts starting at the best estimate of observed situations, each lagged by 24 h. The use of ensemble forecasts is discussed in Chapter 7. Here, we will consider the mean of all the forecasts as a single deterministic forecast. The sea surface temperatures and initial atmospheric conditions were provided by the European Centre for Medium-range Weather Forecast (ECMWF) reanalyses (Gibson *et al.* 1997). The 15-year period 1979–1993 was the longest available at the time of the PROVOST project, having homogeneous 3-dimensional atmospheric fields over the globe. Therefore, there are ensembles of $m = 36$ multi-model forecasts for $n = 15$ winters in our examples. The aim of the PROVOST project was to test the feasibility of seasonal forecasting using efficient ocean data assimilation and simulation. The use of *a posteriori* observed instead of *ex-ante* forecast sea surface temperature implies that the predictive skill obtained in these forecasts is a potential maximum skill rather than an achievable skill. Using four different models helps take account of structural model uncertainties in the individual models. In most cases, the ensemble mean forecast performs at least as well if not better than the best of the four individual models.

The first example consists of winter mean (January–March, forecasts with lead-times 2–4 months) 850 hPa temperature averaged over France. Forecast and observed values are shown in Fig. 5.1. ECMWF reanalyses are used as best estimates of the observed truth. The reason for ignoring the first month of the forecast (December in this case) comes from the need to evaluate the potential of the general circulation model (GCM) to respond to the sea surface temperature forcing rather than to initial conditions.

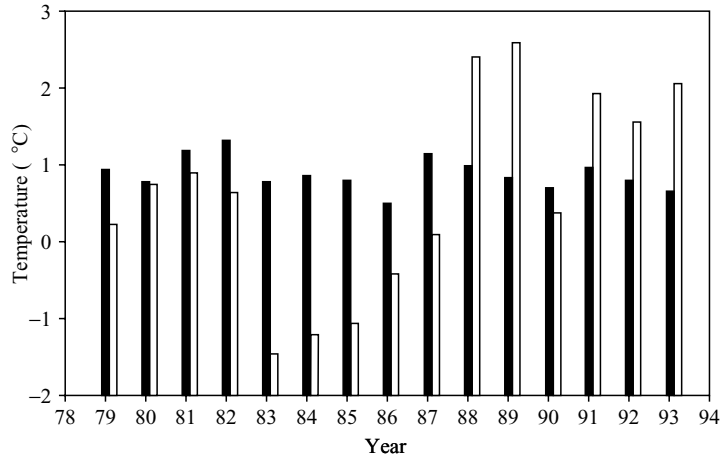


Figure 5.1 Mean winter temperature forecasts (solid bar) and mean observations (empty bar) over France at 850 hPa (close to 1500 m altitude) from 1979/1980 to 1993/1994

Including the first month artificially inflates the scores, simply because the first 10 days can be predicted with good skill (persistence).

The second example concerns forecasts of summer (June–September mean) precipitation over the tropical Atlantic region (French West Indies). Forecast and observed values are shown in Fig. 5.2. Observed values were based on merged gauge and satellite precipitation estimates (Xie and Arkin 1996). In PROVOST, only three models each with ensembles of nine forecasts ($m = 27$) were used for these forecasts.

As can be seen from the figures, the forecast values do not match closely the observed values. This is quite normal, given the long lead-time of these seasonal forecasts. The fact that the skill is low makes the use of verification criteria very important. Such criteria help answer questions about whether the scores are superior to those of trivial forecast methods, whether any differences in skill are statistically significant, and whether it might be possible to improve the forecasts by post-processing. If the forecasts matched closely the observations (e.g. as is the case for 12 h ahead short-range weather forecasts), quantitative verification criteria would be of more use to forecasters (e.g. for inter-model 'beauty contests' and evaluation of improvements), but of less use to the forecast users who have to make decisions based on the forecasts.

5.3 FIRST-ORDER MOMENTS

5.3.1 Bias

The forecasts and the observations in Figs. 5.1 and 5.2 do not have the same mean levels. In the case of climate prediction with numerical GCMs,

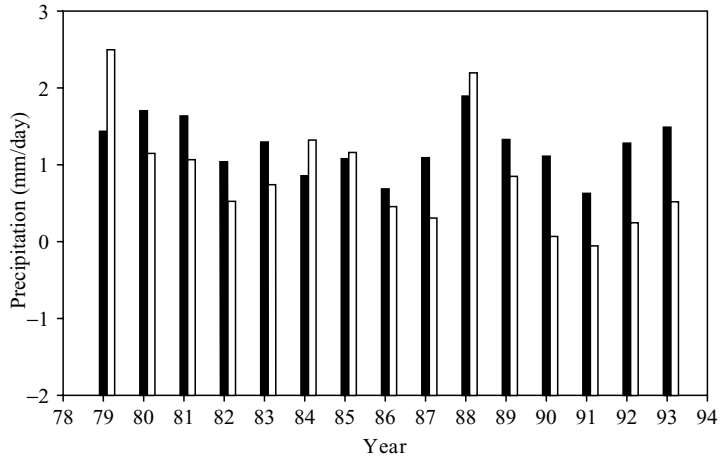


Figure 5.2 Mean summer precipitation forecasts (solid bar) and mean observations (empty bar) over the tropical Atlantic (French West Indies) from 1979 to 1993

contrary to unbiased statistical predictions, there is little chance that a model mean climatology equals exactly that of the observed climatology. When averaging several models, there is some cancellation of the mean bias. However, as shown by multi-model comparison experiments, models often contain related systematic errors and so multi-model averages can never completely remove all the bias in the mean.

To avoid causing problems for forecast users (who often take the forecasts at face value), it is important to correct the bias before delivering a forecast. This can be done simply by removing the mean bias estimate over a set of previous forecasts. Since the mean bias is only an estimate based on a finite sample of past forecasts, it contains sampling noise, which may degrade the forecast. If a model has a small bias and only a few past forecasts are available, it is better not to correct the forecast (Déqué 1991). Instead of this *a posteriori* bias correction, one can also add an incremental empirical correction inside the forecasting equations (*a priori* correction, Johansson and Saha 1989, Kaas *et al.* 1999). This can be physically more satisfactory since the predicted evolution then remains close to the evolution of the actual climate over a wide range of forecast lead-times. However, this is more difficult to set up, and statistically less efficient and transparent than post-processing the forecasts.

In the case of short-range forecasts, bias correction is rarely applied, since biases tend to be small, and bias correction would necessitate that any change in the forecasting system would require the 'reforecasting' of a set of past situations. In the case of longer-lead forecasts, bias correction is essential for correcting model drift in the forecasts.

Another way to consider the mean systematic error is to deal with *anomalies* instead of raw fields. An anomaly is a deviation from the normal value represented by the long-term climatological mean (i.e. a variation centred about the mean). The climatological mean is estimated from past data and should exclude information from the forecast that is being bias corrected. When a hindcast project like PROVOST is performed, the climatological mean is calculated from the 15 year dataset, excluding the target year so that only 14 years are averaged for each year (a cross-validation approach – see Efron and Tibshirani 1993). Forecast anomalies are calculated by subtracting the mean of the other forecast values, and the observation anomaly is calculated by subtracting the means of the observations for the same periods.

The mean systematic error or bias $E(\hat{X} - X)$ can be estimated from the sample statistic

$$b = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i) \tag{5.1}$$

where $\hat{x}_i$ is the forecast value and x_i is the observed value at time i (sometimes referred to colloquially as *the verification*). For example, the bias is

0.26 °C for our temperature forecasts – in other words, the forecasts are too warm on average. Note that calculating the bias for each year with $n-1$ years, i.e. excluding the year in question, and then averaging the n resulting biases is equivalent to calculating the bias for all n years. Our temperature bias is small compared to the interannual standard deviation in observed temperatures of 1.35 °C – the bias only represents about 20 % of the standard deviation in temperature. For the precipitation forecasts, the bias is 0.51 mm/day and so the forecasts are generally too wet on average. Compared to the mean precipitation amount of 4.03 mm/day, the bias in the forecasts of 0.51 mm/day is small relative to the mean climatology. However, the bias is equal to 50 % of the observed interannual standard deviation of 1.06 mm/day and so is a substantial bias in the forecasts, whose main objective is to explain the temporal variations not the long-term mean.

5.3.2 Mean Absolute Error

The mean systematic error is an inadequate measure of skill since negative errors can compensate positive errors. The simple bias correction of using anomalies described above cancels out all the mean error. However, the corrected forecast can still be far from perfect. A simple way to avoid the compensation of positive and negative forecast errors is to consider the *mean absolute error* (MAE) defined as the mean of the absolute values of the individual forecast errors, $E(|\hat{X} - X|)$. This can be estimated using the sample statistic

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{x}_i - x_i| \tag{5.2}$$

Although MAE is computationally less expensive to compute (no multiplication) and more resistant to outlier errors, the mean squared error (MSE, see Section 5.4.1) is more often used in practice. For our forecasting examples, the MAE is 1.09 °C for temperature and 0.87 mm/day for precipitation. This error includes systematic terms that are due to the model being too warm or too moist.

5.3.3 Bias Correction and Artificial Skill

Subtracting the estimated model bias is a simple method of forecast recalibration that can be used *a posteriori* to reduce forecast errors. When bias correcting past forecasts in this way, care should be taken to exclude each target year when estimating the mean biases; otherwise misleading overestimates of skill can easily be obtained (artificial skill). For example, when the overall mean bias of 0.26 °C bias is removed from all the temperature forecasts, a new smaller MAE of 1.04 °C is obtained, whereas when the bias is calculated and subtracted separately for each year a larger (yet more

realistic) MAE of 1.11 °C is obtained. This example also shows the detrimental effect of bias correction for small samples of forecasts – the cross-validation bias correction slightly increases the MAE compared to that of the original forecasts. For the example of precipitation forecasts, the MAE is 0.71 mm/day after cross-validation bias correction, and so the correction helps improve the mean score. Note, however, that whereas MSE is minimized by removing the *mean error* from the forecasts, the MAE is minimized by removing the *median error*, i.e. the median of the $\hat{x}_i - x_i$ errors. Therefore, when we correct the systematic error by removing the median systematic error instead of the mean systematic error, smaller values for MAE are obtained: 1.05 °C for temperature and 0.62 mm/day for precipitation. As precipitation is less well approximated by a normal (Gaussian) distribution than temperature, using the median rather than the mean for the bias correction is more effective. When the forecast events are not independent (e.g. consecutive daily values), a cross-validation estimation of bias can be obtained by excluding a sliding mean centred on the target day with a sufficient width that the values at the edges of the windows are independent of the ones at its centre (a typical width is one week in meteorological applications).

For the verification of seasonal forecasts over a long period (e.g. 50 years), the independence assumption may also be violated due to long-term trends caused by climate change (e.g. global mean temperature at the end of the 20th century is generally warmer than at the beginning). In this case, one should remove the long-term trend from both the observations and the forecasts before calculating any score. This procedure will avoid artificial skill caused by any concurrent long-term trends in the forecasts and the observations. Skill due to long-term trends is not a useful forecast skill for a user who is interested in knowing in advance year-to-year fluctuations.

5.3.4 Mean Absolute Error and Skill

An important thing to ascertain is whether bias-corrected MAE values of 1.05 °C and 0.62 mm/day actually signify any real skill. To do this, these scores must be compared with scores of an alternative low-skill forecasting method. The two most commonly used reference methods are *persistence forecasts* and *mean climatology forecasts*. Both these forecasts require no model development (but do require a database of past observations). Persistence can be biased by persistence due to the annual cycle, and so bias correction is mandatory. In our example, we will take the monthly mean for the month just before issuing the forecast, i.e. November for temperature, and May for precipitation. The MAE of the persistence forecasts is worse than the model MAE: 1.50 °C for temperature and 1.12 mm/day for precipitation. However, the comparison is not really fair, since the persistence forecast has a larger interannual variability than the multi-model ensemble mean forecast since it is based on a single monthly mean

rather than the mean of 36 months. If we apply a linear regression to take account of the extra variance in the persistence forecast, the MAE then becomes 0.98 °C for temperature and 0.85 mm/day for precipitation – hence, the multi-model ensemble mean forecast for temperature has a larger MAE than the persistence forecasts.

In long-range forecasting, climatological mean forecasts provide more skilful reference forecasts than persistence. The climatological mean forecast is obtained by constantly forecasting the same sample mean of the past observations for every event. For our examples, the climatological mean forecasts give an MAE of 1.17 °C for temperature and 0.91 mm/day for precipitation. Hence, the multi-model ensemble forecasts perform better than climatological mean forecasts for both temperature and precipitation.

5.4 SECOND AND HIGHER-ORDER MOMENTS

5.4.1 Mean Squared Error

The MSE $E[(\hat{X} - X)^2]$ is perhaps one of the most widely used forecast scores. It is estimated by the sample statistic

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^2 \tag{5.3}$$

The square root of this quantity, the RMSE, has the same units as the forecast variable. Because of the square in Eq. (5.3), the MSE is more sensitive to large forecast errors than is MAE. For example, a single error of 2 °C contributes exactly the same amount to MAE as is contributed by two errors of 1 °C, whereas for MSE the larger error would contribute twice as much to the score. From a user's point of view, it may seem natural to penalize larger errors and to be more indulgent towards small errors. However, the MSE is unduly sensitive to outlier errors in the sample (caused, e.g. by data corruption, atypical events, etc.) rather than being representative of the forecasts as a whole. MAE is more *resistant* to outliers than is MSE. The scores, MAE and MSE, are specific $p = 1$ and $p = 2$ cases of the more general L_p Minkowski norm

$$\left(\frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i)^p\right)^{1/p} \tag{5.4}$$

As p tends to infinity, this norm tends to the maximum error in the sample. The L_∞ norm provides an upper error bound for a user, but this will increase indefinitely as the sample size of past forecasts increases with time. Unfortunately, this score depends on only one value in the sample

(the one with largest error), and so for a system with (say) 1000 perfect forecasts and only one erroneous forecast, the score will be entirely determined by the error of the single erroneous forecast.

For our forecasts, the temperature RMSE is 1.27 °C and the precipitation RMSE is 0.97 mm/day. As mentioned in Section 5.4.1, the MSE is minimized when the mean error is subtracted from the forecast, or, equivalently, the forecast and observation data are centred about their respective sample means. This result is strictly only true for infinite sample sizes (i.e. the population) and is not always guaranteed to improve the MSE for small samples of forecasts due to the presence of sampling uncertainty. Indeed, with our examples, the RMSE after *a posteriori* correction is 1.33 °C for temperature and 0.88 mm/day for precipitation – a similar behaviour as was obtained with the MAE. Therefore, in this case, it is wise to bias correct precipitation, but bias correcting temperature requires more than 15 years of previous forecasts and observations to be effective. It can be shown mathematically that MAE is never greater than RMSE, and they are only equal in the unlikely situation when the forecast error is a constant for all cases. For our forecasts, it can be seen that the usual relationship holds for MAE and RMSE.

In order to assess predictability, one must compare the RMSE with that obtained using low-skill alternative forecasting methods. The persistence of the anomaly of the latest monthly mean available at the time of the forecast yields an RMSE of 1.86 °C for temperature and 1.32 mm/day for precipitation. The climatological mean forecast (the forecast that minimizes the MSE among all the forecasts which are statistically independent of the observations) gives an RMSE of 1.35 °C for temperature and 1.06 mm/day for rainfall. As was also the case for MAE, the model outperforms both the climatological and persistence forecasts.

5.4.2 MSE Skill Score

To judge the skill of an RMSE score one must compare it to the RMSE of a low-skill forecast, e.g. climatological RMSE. Murphy and Epstein (1989) proposed the MSE *skill score* defined by

$$\mathrm{MSESS} = 1 - \frac{\mathrm{MSE}}{\mathrm{MSE}_{\mathrm{clim}}} \tag{5.5}$$

where $\mathrm{MSE}_{\mathrm{clim}}$ is the MSE for climatological mean forecasts. When the mean bias of the forecasts is zero, MSESS is identical to the square of the product moment correlation coefficient between the forecasts and the observations (Section 5.4.4). The basic idea of skill scores is discussed in Section 2.7 of Chapter 2. A value of 1 is the maximum value for MSESS and indicates a perfect forecast; a value of 0 indicates a model forecast equivalent to a climatological forecast. A negative value implies that the

model is worse than climatology, in terms of MSE, although this does not necessarily imply that the model has no skill at all. When the MSESS is aggregated over time or space, it is important to accumulate separately the numerator and the denominator of Eq. (5.5) rather than average the local values of MSESS. Unlike MAE and MSE, MSESS is dimensionless, and increases with forecast skill. In our two examples, the MSESS is 0.12 for temperature and 0.16 for precipitation (before bias correction). After bias correction, the precipitation forecasts have an increased MSESS of 0.31.

Once a sample estimate of a score has been obtained, the next question is to ascertain its statistical significance. Parametric tests can be used but they are not always very appropriate because of underlying restrictive assumptions. Therefore, non-parametric tests offer more flexibility for checking whether the scores for two forecasting systems are significantly different. A simple way to do this is by Monte Carlo permutation techniques: the 15 forecasts are reassigned at random among the 15 years. For 15 years, there are about 10^{12} permutations, but 1000 random drawings are generally sufficient to estimate a 95% prediction interval for a score based on a no-skill null hypothesis. We use the term prediction interval rather than confidence interval since sample scores are random sample statistics not probability model parameters.

With the above permutation procedure, the 95% prediction interval for the RMSE of the temperature forecasts is [1.25,1.48], which yields a 95% prediction interval on the MSESS of [−0.20, 0.14]. The skill score prediction interval includes MSESS = 0.12 and so at the 5% level of significance the null hypothesis that the forecasts have no real skill cannot be rejected. For the precipitation forecasts, the skill score is significantly different from zero at the 5% level of significance, since the MSESS interval is [−0.77, 0.28] and so does not include the MSESS = 0.31 obtained for the bias-corrected precipitation forecasts. The interval for the MSESS is not symmetrical for a random forecast, especially for precipitation, but this can be explained by the fact that the MSESS can take values over the range from minus infinity to one.

5.4.3 MSE of Scaled Forecasts

Removing the mean bias is one possible way of improving the MSE. Rescaling the forecast anomaly is another simple calibration technique for improving forecasts. In general, it is possible to correct biases in the mean *and* the variance by performing a linear regression, $\hat{X}' = E(X|\hat{X}) = \beta_0 + \beta_1\hat{X}$, of the observations on the forecasts (see Section 2.9). A simple geometric interpretation of this operation will be given in Section 5.4.4. For hindcasts, the regression coefficients should be estimated using data that excludes the target year. Since this recalibration amounts to a linear transformation of the forecasts, it cannot change the product moment correlation coefficient between the forecasts and the observations. Hence, for

forecasts with no mean error, it cannot improve the MSESS defined in Eq. (5.5). For our examples, the MSESS of the forecasts recalibrated in this manner are −0.04 for temperature and 0.22 for precipitation. These MSESS values are less than those of the original uncalibrated forecasts, which indicates that in this case the sample size of 15 years is too short to obtain reliable estimates of the regression coefficients. A similar phenomenon was noted in Section 5.4.1 when correcting the mean bias in the temperature forecasts. For the persistence forecasts, the MSESS for the regression-recalibrated forecasts is −0.22 and −0.20 for temperature and precipitation, respectively. If we apply the permutation procedure to the regression-recalibrated forecasts, the 95% prediction interval for the MSESS becomes [−0.33, 0.27] for temperature and [−0.39, 0.24] for precipitation. Hence, recalibration in this particular case has unfortunately not produced forecasts with any significant skill at the 5 % level. One should therefore be very careful when attempting statistical post-processing of forecasts based on small samples of past forecasts. Post-processing is currently used with success in short- and medium-range forecasting. However, in seasonal forecasting, small sample sizes present a strong limiting factor, and even the bias correction of the mean must be applied with caution.

5.4.4 Correlation

Because of its invariance properties, the correlation coefficient is considered as the 'king of all scores' in many predictive sciences, and in particular in weather and climate forecasting. The product moment correlation coefficient (also known as Pearson's correlation) is defined as

$$\rho = \mathrm{cor}(x, \hat{x}) = \frac{\mathrm{cov}(x, \hat{x})}{\sqrt{\mathrm{var}(x)\mathrm{var}(\hat{x})}} \tag{5.6}$$

where $\mathrm{cov}(x, \hat{x})$ is the covariance between the observations and forecasts and $\mathrm{var}(x) = \mathrm{cov}(x, x)$ and $\mathrm{var}(\hat{x}) = \mathrm{cov}(\hat{x}, \hat{x})$ are the variances of the observations and forecasts, respectively. The covariance can be estimated from the sample of past forecasts and observations, e.g. $\mathrm{cov}(x, \hat{x})$ is estimated by the sample statistic

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_i)(\hat{x}_i - \bar{\hat{x}}_i) \tag{5.7}$$

Correlation is a dimensionless and positively oriented score, and its square is equal to MSESS for forecasts having no bias in the mean. Correlation is invariant to shifts in the mean and rescaling of either the forecasts or observations; hence mean bias correction and linear regression post-processing of the forecasts do not change the correlation of the forecasts

with the observations. A practical benefit is that there is no need to remove the bias or correct the amplitude of the forecast anomaly. In fact, one can calculate anomalies with respect to *any* mean climatology. Because correlation is invariant to changes in scale, one can correlate observations measured in Celsius directly with forecasts in Fahrenheit (i.e. conversion of units will not change the result). A correlation of $+1$ or -1 implies a perfect linear association between the forecast and observations, whereas a correlation of zero signifies that there is no linear association between the variables. It should be noted that two *uncorrelated* variables are not necessarily *independent* since they can be non-linearly rather than linearly related to one another.

From a geometrical point of view, the correlation coefficient can be seen as the cosine of an angle in sample space. Figure 5.3 shows three points representing the origin (O) which corresponds to a climatological mean forecast, the bias-corrected forecast (F), and the verification observation (V) in a projection of the 15-dimensional sample space. Since we assume that the bias is corrected, the mean of the forecasts equals the mean of the observations. The correlation is the cosine of the angle (FOV), the RMSE is the distance OF, and the climatological RMSE is the distance OV. The optimization of RMSE by scaling consists of replacing F by F'. Figure 5.3 shows that the squared correlation and the maximum MSESS are identical for these forecasts with no bias in the mean. For this reason, one can say that correlation measures the potential skill of unbiased forecasts rather than the actual MSE skill that includes contributions from bias in the mean (Murphy and Epstein 1989).

For our examples, the correlation is 0.16 for temperature and 0.55 for precipitation in the case of the model forecasts, and 0.16 and 0.46, respectively, for persistence forecasts. Because correlation is restricted to be less

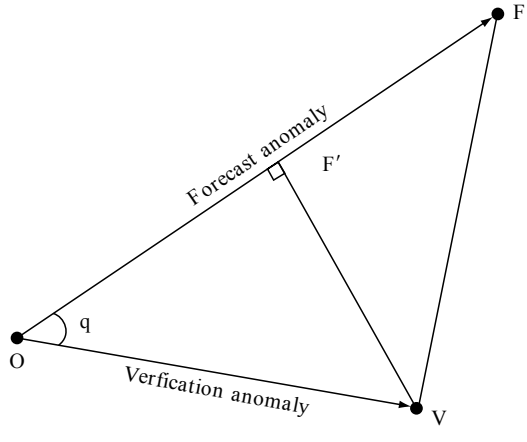


Figure 5.3 Scheme representing the relative position of origin (O), forecast anomaly (F) and verification anomaly (V). F' is the best rescaled forecast

than 1, improvements in large correlations are necessarily smaller than those in correlations closer to zero. For this reason, and another one discussed below, it is better to consider a non-linear transformation of correlation known as the Fisher z-transform

$$z = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) \tag{5.8}$$

This z-score can take any value in the range minus infinity to positive infinity and is better approximated by a normal distribution than is correlation. The statistical significance of the difference between two correlations can be most easily tested using a 2-sample Z-test with $Z = (z_1 - z_2)/s$ and the asymptotic (large sample) sampling error $s = \sqrt{2/(n-3)}$. For the model forecasts, the z-scores are 0.16 and 0.62 for temperature and precipitation, respectively, and for testing the significance of a single z-score the asymptotic standard error is $1/\sqrt{15-3} = 0.29$, hence only the precipitation forecasts have correlations (marginally) significantly different from zero at the 5% level of significance (i.e. a z-score more than 1.96 standard errors away from zero).

Table 5.1 summarizes the different scores obtained for temperature and precipitation for the main criteria. Like MSESS, but unlike MAE or RMSE, the correlation coefficient is an increasing function of the skill. The correlation coefficient of the climatological forecast is not uniquely defined, since it involves a ratio of zero by zero. However, by symmetry, one can consider it to be zero, and so a positive correlation is then interpreted as a forecast better than the climatology. The crucial question is the threshold above which a correlation can be considered as statistically significant, since a sample estimate of a correlation coefficient is never exactly zero.

The real skill of forecasts can be assessed by statistically testing whether or not the correlation is significantly different from zero. Both parametric and non-parametric approaches can be used. A simple parametric approach is to try to reject the no-skill null hypothesis that the forecasts and observations are independent and normally distributed. Under this null hypothesis, the sample statistic

Table 5.1 Summary of main scores for seasonal forecasts of temperature and precipitation. Values in parentheses correspond to bias-corrected forecasts

	Temperature forecasts	Precipitation forecasts
Bias	0.26 °C	0.51 mm/day
MAE	1.09 (1.05) °C	0.87 (0.62) mm/day
RMSE	1.27 (1.33) °C	0.97 (0.88) mm/day
MSESS	0.12 (0.03)	0.16 (0.31)
Correlation	0.16	0.55

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{5.9}$$

is distributed as a Student's t-distribution with $n-2$ degrees of freedom. Inverting this expression for $n = 15$ forecasts gives a 95% prediction interval for a no-skill correlation of [–0.57, 0.57]. Both our temperature and precipitation forecasts have correlations that lie within this interval and so at the 5% level of significance are *not inconsistent with being no-skill forecasts*. Note that we can only say that the previous data are *not inconsistent* with the null hypothesis not that the null hypothesis *is* true (i.e. *the forecasts have no-skill*) – more data in the future may allow us to reject the null hypothesis and then say that the forecasts have skill. To avoid the need for the normality assumption, we can use a non-parametric (distribution-free) permutation procedure similar to that described earlier for the MAE and MSE scores. For our forecasts, the permutation method gives a slightly narrower 95% prediction interval for correlation of [−0.51, 0.51]. Although it may appear sensible to exclude permutations in which one or more years is unchanged, such censoring is a bad idea since it biases the interval estimate towards negative values. For example, censoring of permutations gives intervals of [−0.53, 0.50] and [−0.53, 0.47] for the temperature and precipitation forecasts, respectively. Based on the permutation interval, it can be concluded that our precipitation forecasts (but not the temperature forecasts) have marginally significant skill at the 5% level of significance, which is in agreement with the results obtained for the MAE and the RMSE scores.

Another important issue for correlation is how best to aggregate several correlations obtained, e.g. by pooling over a geographical region. Because of its additive properties, it is well known that one should average the MSE score for all the cases before taking the square root, rather than averaging all the RMSE scores. Nevertheless, it is not unusual in meteorological studies to calculate time averages of spatial correlations or spatial averages of time correlations. Alternatively, some studies average the Fisher z-scores, which then give more weight to correlations further from zero. A better less-biased approach is to calculate the components of the 2×2 (co-)variance matrix between forecasts and observations by averaging over all cases, and then calculate the overall correlation coefficient using these four values (Déqué and Royer 1992; Déqué 1997). This aggregated covariance approach has several advantages. Firstly, the aggregation of the covariances over space and time can be done in either order and so it treats spatio-temporal aggregation in a symmetrical manner, which is not the case when taking the spatial average of correlations over time or time averages of correlations over space (see discussion of ACC in Section 6.3). Secondly, the resulting correlation corresponds to a weighted average of the individual correlation coefficients, with larger weights being given to forecast and

observation anomalies with larger magnitudes. This makes sense from the user's perspective since large magnitude anomalies are the ones that have the most impact. Thirdly, the geometrical interpretation remains applicable and the squared overall correlation is equal to the MSESS for unbiased forecasts.

5.4.5 An Example: Testing the 'Limit of Predictability'

To illustrate how correlation can be used to test the skill of forecasts, we will demonstrate its use in this section on daily temperature forecasts from the PROVOST experiment. Figure 5.4 shows the correlation between daily temperature forecasts and observations as a function of the lead-time – days 1–30 correspond to the days in November, the first month of forecasts. With only a small sample of 15 winters, the correlation curve does not exhibit perfectly smooth monotonic decay due to sampling fluctuations. The correlations lie within the 95% no-skill prediction interval of $[-0.57, 0.57]$ (shaded area) for lead-times longer than 6 days. The limit of predictability for daily temperature forecasts (crudely) estimated from these 15 years is therefore 6 days. Beyond this time scale, there is no significant skill for daily temperature values but there can be skill in forecasting statistical quantities such as monthly and seasonal mean temperatures.

5.4.6 Rank Correlations

The product moment correlation discussed above measures the strength of the *linear association* between forecasts and observations. Linear and non-linear *monotonic association* between forecasts and observations can be measured using statistics based on the ranks of the data (i.e. the position

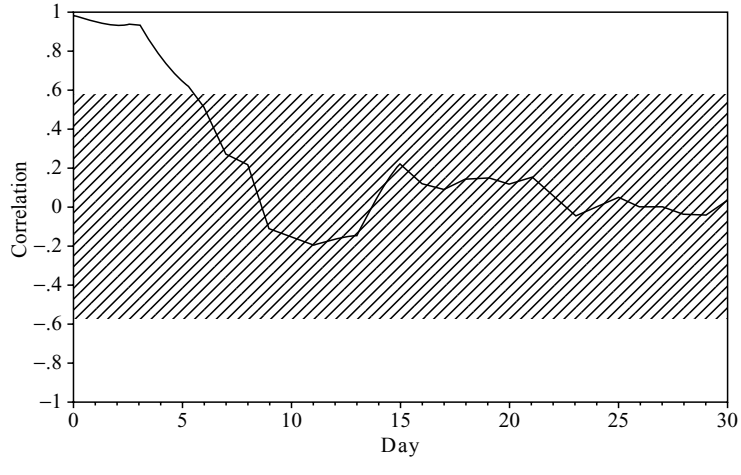


Figure 5.4 Correlation of daily temperature forecasts as a function of lead-time; the shaded area corresponds to the 95% no-skill prediction interval

of the values when arranged in increasing order). For example, the Spearman rank correlation coefficient is simply the product moment correlation coefficient of the ranks of the data (see Wilks 1995). A rank correlation of one means that the forecast values are an increasing function of the observations – i.e. there is a perfect monotonic association between the forecasts and the observations. The rank correlation coefficient is also more resistant (less sensitive) to large outlier values than is the product moment correlation coefficient. Since the mean of the ranks is $(n+1)/2$ and the variance of the ranks is $(n^2-1)/12$, the Spearman rank correlation coefficient can be expressed as

$$r_s = \frac{12}{n(n^2-1)} \sum_{i=1}^{n} \hat{R}_i R_i - \frac{3(n+1)}{n-1} \tag{5.10}$$

where $\hat{R}_i$ and R_i are the ranks of the ith forecast and observation, respectively. In the asymptotic large sample limit for normally distributed data, the product moment correlation coefficient and the rank correlation coefficient can be shown (Saporta 1990) to be related by

$$r \approx 2 \sin\left(\frac{r_s \pi}{6}\right) \tag{5.11}$$

For our examples, the rank correlation is 0.20 for temperature and 0.46 for precipitation. The fact that these values are close to the product moment correlation coefficients (0.16 and 0.55) suggests that the association between forecasts and observations has a dominant linear component. Tests based on the rank correlation are distribution-free (robust) since they make no assumption about the marginal distributions of the data. For large samples of independent forecasts and observations (the asymptotic no-skill null hypothesis), the rank correlation coefficient is normally distributed with a mean of zero and a variance of $1/(n-1)$. For a sample of forecasts this result gives a 95% prediction interval of $[-0.53, 0.53]$ for the rank correlation coefficient (the exact calculation yields $[-0.52, 0.52]$). Both temperature and precipitation forecasts have a rank correlation inside this interval and so do not have significant skill at 95% confidence.

Dependency can also be tested non-parametrically using Kendall's tau correlation statistic. For each pair of times i and j, a new sign variable s_{ij} is created that takes the value 1 when $(x_i - x_j)(\hat{x}_i - \hat{x}_j)$ is positive, i.e. when the forecast and observation evolve in the same direction, and -1 otherwise. The average of the sign variable can then be used to construct the correlation

$$\tau = 1 - \frac{4}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{i-1} s_{ij} \tag{5.12}$$

When the forecasts and observations are independent, the distribution of τ can be tabulated without any assumptions about the distribution of the forecast and observation. It can be demonstrated that τ is asymptotically normally distributed with a mean of zero and a variance of $2(2n+5)/(9n(n-1))$. One advantage compared to Spearman's rank correlation is that the convergence to the asymptotic normal distribution is faster for Kendall's tau correlation, which is approximately normally distributed for sample sizes as small as eight. For our examples, Kendall's correlation is 0.18 and 0.40 for the temperature and precipitation forecasts, respectively. This is a little less than Spearman's correlation. The 95% no-skill asymptotic prediction interval is $[-0.38, 0.38]$ which is narrower than the interval for Spearman's correlation. Therefore, the precipitation forecasts over the tropical Atlantic have significant skill at the 5% level of significance whereas the temperature forecasts over Europe are not significantly skilful.

As explained in Saporta (1990), the three correlation coefficients (Pearson, Spearman and Kendall) are in fact particular cases of the so-called Daniels' correlation:

$$r_{\mathrm{d}} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}\,\hat{d}_{ij}}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 \sum_{i=1}^{n}\sum_{j=1}^{n} \hat{d}_{ij}^2}} \tag{5.13}$$

where d_{ij} is a distance between the objects i and j defined for the three correlations as follows:

- Pearson $d_{ij} = x_i - x_j$.
- Spearman $d_{ij} = R_i - R_j$, where R_i is the rank of x_i.
- Kendall $d_{ij} = \mathrm{sgn}(x_i - x_j)$.

From Eq. (5.13), it is obvious that Daniels' correlations are invariant under any recalibrations (i.e. bias correction or rescaling) that transform $d_{ij} \rightarrow \delta + \gamma d_{ij}$. It is interesting to note that the persistence forecast for precipitation that has a rather high, yet non-significant, Pearson correlation for precipitation (0.46), has almost zero values of Spearman's (0.08) and Kendall's (0.04) correlations. This feature is explained by the scatter plot of these forecasts and observations shown in Fig. 5.5. The first year with a persistence forecast of 7.7 mm/day and an observed value of 6.3 mm/day is a clear outlier from the cloud of other points, and thereby contributes excessively to the Pearson correlation but not to the more resistant Spearman and Kendall correlations.

This example emphasizes the need to use resistant scores when judging small samples of forecasts.

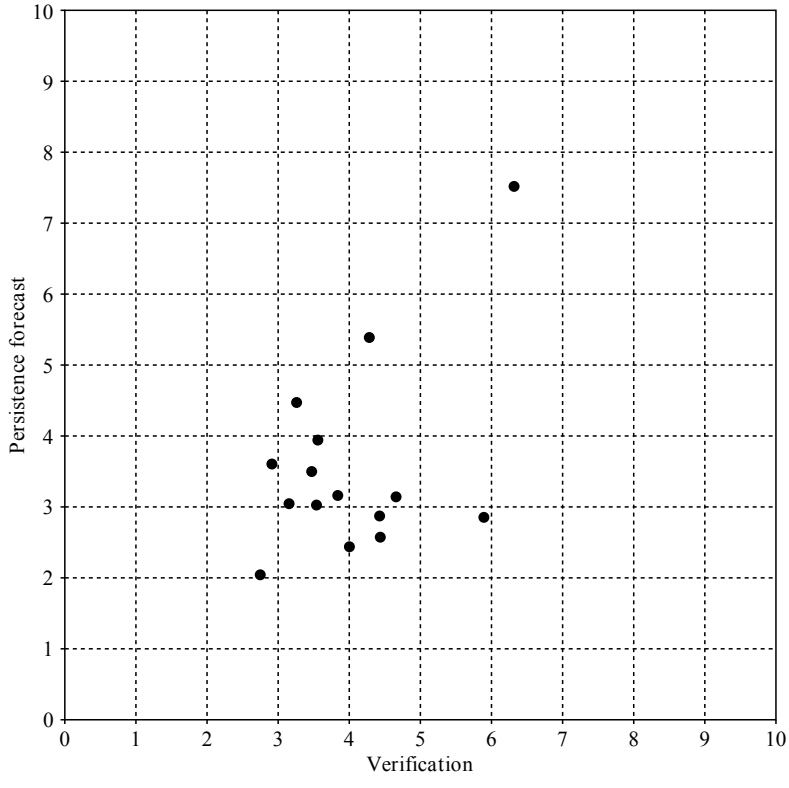


Figure 5.5 Scatter diagram of precipitation persistence forecasts versus observations (mm/day)

5.4.7 Comparison of Moments of the Marginal Distributions

In forecast evaluation, it is desirable that the overall distribution of forecasts is similar to that of the observations, irrespective of their case-to-case relationship with the observations. A high correlation can misleadingly be obtained with forecasts that have a very different statistical distribution from the observations. If forecasts are used at face value as input variables for impact models, e.g. of crop production, or are used as a basis for a search for analogues, the results can be misleading. Therefore, in addition to estimating association, one must also compare properties of the marginal distributions of the forecasts and the observations.

The first-order moments of the marginal distributions are the means of the forecasts and observations and these can be compared by taking the difference to obtain the bias discussed in Section 5.3.1. The statistical significance of the bias can be tested with a 2-sample paired t-test (Wilks 1995). The second-order moments of the marginal distributions are the variances of the forecasts and the observations. These can be tested for equality by performing an *F*-test on the ratio of variances. However, this test is not widely used since the assumption of equal variances is almost never rejected

for small samples even when it is not true (i.e. the test has low power). With sample sizes of 15, the ratio of variances must be greater than 2.5 in order to reject the null hypothesis at the 5% level of significance. The variance of the forecasts and observations are more easily interpreted when expressed in terms of standard deviations that have the same units as the predictand. For ensemble mean forecasts, the variance of the mean forecast will invariably be less than that of the observations due to the averaging that has taken place to create the mean of the ensemble of forecasts. For example, the standard deviation of our temperature forecasts is 0.22°C compared to 1.35°C for the observations. For the precipitation forecasts, the standard deviation of the forecasts is 0.53 mm/day compared to 1.06 mm/day for the observations. Before the forecasts are delivered to unsuspecting users, it is important to rescale (inflate) them. The scaling factor that yields the correct interannual variance (6.1 for temperature and 2.0 for precipitation) is invariably not the best choice for minimizing the MSE (0.2 for temperature and 0.4 for precipitation). However, rescaling does not change measures of association such as the correlation between the forecasts and observations.

Comparisons of higher-order moments of the marginal distributions such as skewness and kurtosis can also be revealing for forecasts and observations that are not normally distributed. The asymmetry of the distributions can be estimated using the third-order moment about the mean known as the moment measure of skewness:

$$b_1 = \frac{1}{ns^3}\sum_{i=1}^{n}(x_i - \bar{x})^3 \tag{5.14}$$

where $\bar{x}$ is the sample mean and s is the sample standard deviation. Skewness is a dimensionless measure of the asymmetry of the distribution. Positive skewness indicates that the right tail of the distribution is fatter than the left tail and that more values lie below the mean than above. Symmetric distributions such as the normal distribution have zero skewness, but zero skewness does not imply that the distribution has to be symmetric. Recalibration procedures such as removing the mean bias or rescaling do not change the skewness. More resistant measures of skewness also exist and are more reliable than the moment measure of skewness when dealing with small samples (see Wilks 1995). The fourth-order moment about the mean is known as kurtosis and is defined by the dimensionless non-negative quantity

$$b_2 = \frac{1}{ns^4}\sum_{i=1}^{n}(x_i - \bar{x})^4 \tag{5.15}$$

Kurtosis does not measure the spread of the variable, since it is insensitive to scaling or bias correction. It measures the frequency of rare events, with

respect to the standard deviation. If the tails of the probability density decrease rapidly, kurtosis is small, whereas if the tails decrease slowly (e.g. as a negative power law), then kurtosis is large. The normal distribution has a kurtosis of 3 and exhibits a rapid decrease in the tails, e.g. the probability 0.05 of a value being more than two standard deviations from the mean is much less than the upper probability bound of $1/2^2 = 0.25$ defined by the Chebyshev inequality (Hogg and Tanis 1997, Section 12.4). Skewness and kurtosis can be used to test whether a variable is normally distributed (normality test). However, both these higher moments of the distribution are very sensitive to large outlier values, and more resistant approaches based on the cumulative distribution are preferable.

5.5 SCORES BASED ON CUMULATIVE FREQUENCY

5.5.1 Linear Error in Probability Space

The *linear error in probability space* (LEPS) score is defined as the mean absolute difference between the cumulative frequency of the forecast and the cumulative frequency of the observation:

$$\text{LEPS}_0 = \frac{1}{n}\sum_{i=1}^{n} |F_X(\hat{x}_i) - F_X(x_i)| \tag{5.16}$$

where F_X is the (empirical) cumulative distribution function of the observations. The basic idea is that an error of 1 °C in temperature in the tail of the distribution is less important than the same error nearer the mean where it corresponds to a greater discrepancy in cumulative probability. For uniformly distributed forecasts and observations in the range [0,1], the LEPS_0 score becomes equivalent to the MAE score. LEPS_0 is the probability of obtaining a value between the forecast and the observation. LEPS_0 is a negatively oriented score with values in the range [0,1]. It is only zero when the forecasts are perfect. For no-skill forecasts in which the forecasts are completely independent of the observations, the population expectation of the score can be written as

$$E\ (\text{LEPS}_0) = \frac{1}{2} - E_{\hat{X}}[F_X(\hat{X})(1 - F_X(\hat{X}))] \tag{5.17}$$

which is minimized by repeatedly forecasting the median of the observations $\hat{x} = F^{-1}(0.5)$ as was the case for the MAE score. LEPS_0 can be used for discrete continuous as well as variables (see Section 4.3.4). Potts *et al.* (1996) introduced a positively oriented score based on LEPS_0:

$$\text{LEPS} = 2 - 3(\text{LEPS}_0 + \overline{F_X(\hat{x})(1 - F_X(\hat{x}))} + \overline{F_X(x)(1 - F_X(x))} \tag{5.18}$$

where the overbar indicates the mean over a previous sample of forecasts and matched observations.

Unlike LEPS_0, this score is equitable (Gandin and Murphy 1992) – when the forecast is independent of the verification (random or constant forecast), the score is zero. Each correct forecast contributes on average $1/n$ to the score, whereas correct forecasts of extreme events contribute $2/n$. For our example forecasts, due to the small sample size, the mean LEPS score of climatological mean forecasts is 0.04 and is 1.07 for perfect forecasts. LEPS_0 is 0.24 and 0.28 for the temperature and precipitation forecasts, respectively, while LEPS is 0.12 and 0.17. The significance of the LEPS scores can be tested by a permutation procedure and all the above values are inside the 95 % no-skill prediction interval. As was noted for the MAE score, the LEPS score is sensitive to the model bias and is generally best when the median of the forecasts is close to the median of the observations. With such a bias correction, LEPS decreases for temperature (0.01) and increases for precipitation (0.21) in agreement with similar behaviour noted for the RMSE score.

5.5.2 Quantile–Quantile Plots (q–q Plots)

The *quantile–quantile plot* (q–q plot or qqplot) is a visual way for comparing the marginal cumulative probability distribution of a sample with either that of a theoretical distribution or with that of another sample of data (see Wilks 1995). For comparison with a theoretical distribution (e.g. the normal distribution), the method consists of plotting the empirical quantiles of a sample against the corresponding quantiles of the theoretical distribution (i.e. the points $(x_{[i]},\ F^{-1}(i/(n+1))$, where $x_{[i]}$ is the ith value of the sequence of x values sorted into ascending order). If the points lie along the line $x_{[i]} = F^{-1}(i/(n+1))$, then the cumulative probability distribution of x is equal to the theoretical distribution $F(x)$. The panels in Fig. 5.6 shows plots of normal quantiles versus the quantiles for standardized temperature and precipitation forecasts (solid circles) and observations (empty circles). The temperature and precipitation forecast and observed data have been standardized to have zero mean and unit variance, so that they can be directly compared with standard normal variables. All the points lie close to the 45 ° line, which indicates that the forecasts and observations are approximately normally distributed (thereby justifying some of our earlier normality assumptions in this chapter). However, the points do deviate below the line for large anomalies in the data, which indicates some presence of positive skewness especially in precipitation.

5.5.3 Conditional Quantile Plots

The conditional quantile plot is an extension of the above method to forecast verification. The empirical conditional distribution of the observations for

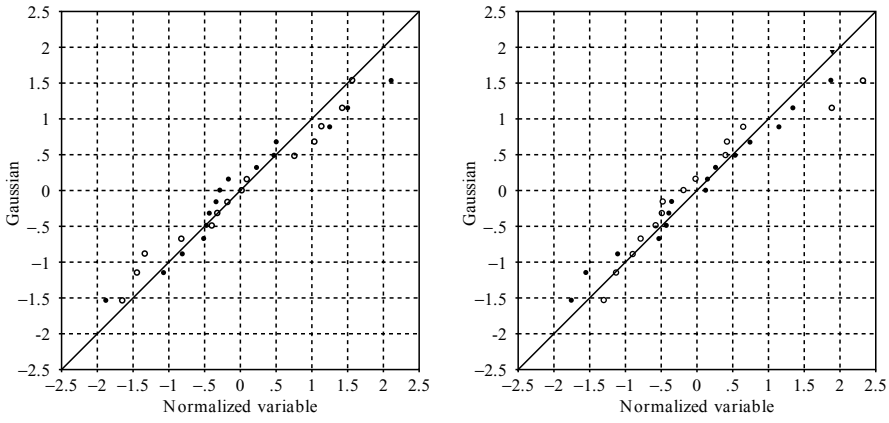


Figure 5.6 Quantile–quantile plots of standardized temperature (left) and precipitation (right) forecasts. The filled circles indicate the 15 ranked forecasts and the empty circles the 15 ranked observations

given forecast values is estimated and the 25 %, 50 % and 75 % percentiles are then plotted against the forecast values. Mathematically, the conditional quantile plot is a scatter plot of the conditional observation quantiles $F^{-1}_{X|\hat{X}=\hat{x}}(q)$ versus the conditioning forecast values for a chosen probability, e.g. $q = 0.25$. If the sample of observations is large enough, rare percentiles such as 10 % and 90 % can then be reliably estimated and plotted (Wilks 1995). A very good forecast shows the points lying close to the 45 ° diagonal. An easy to improve forecast shows the percentiles close to each other, but far from the 45 ° diagonal. In this situation, a good recalibrated forecast can be obtained by using the 50 % percentile of the observation conditioned on the forecast instead of the original forecast. The worst case is when the 25 % and 75 % percentiles are far from each other, even if the 50 % percentile is along the 45 ° line. However, in this case, there may be some forecasts for which the range is narrower, so that there is some skill in some instances.

Constructing conditional distributions is in fact the best way (in a least-squares sense) to produce a probabilistic forecast from a deterministic forecast (Déqué *et al.* 1994). However, a practical difficulty arises from the finite size of the sample of past forecasts and observations. If the forecast is only crudely binned into 10 distinct categories, one needs at least 100 pairs of previous forecasts and observations to guarantee an average of at least 10 forecast–observation pairs in each bin. With our example based on only 15 forecasts, it is not possible to consider more than a maximum of three bins (e.g. below normal, normal, above normal) in order to be able to have at least the very minimum of five observations in each bin needed to estimate quantiles. The conditional quantile plots illustrated in Fig. 5.7 were constructed by using a larger data sample obtained by regional pooling of data from all model grid points located in wider

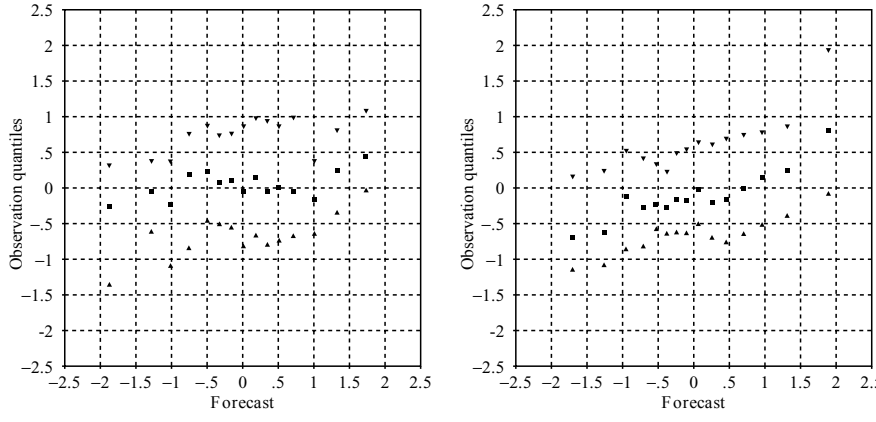


Figure 5.7 Conditional quantile plots for standardized forecasts of temperature (left) and precipitation (right). The downward pointing triangles indicate the 75% percentiles; the squares indicate the 50% percentiles, and the upward pointing triangles indicate the 25% percentiles

boxes: Europe (35N–75N, 10W–40E) for temperature, and the western tropical Atlantic (10S–30N, 100W–50W) for precipitation. Each local grid point variable is standardized to have zero mean and unit variance, in order to avoid artificial effects due to geographical contrasts (e.g. a good correspondence between forecast and observed cold areas).

Figure 5.7 shows for 15 ranked subsets of the forecasts (the number 15 is arbitrarily chosen to be equal to the number of years, but a different number could easily have been taken), the conditional median (50 % percentile) and the conditional 50 % interval about the median (the 25 % and 75 % percentiles) of the corresponding observations. The geographical domain has 352 grid points, so that 5280 (15×352) forecast values have been sorted, and the quantiles have been estimated with subsamples of 352 observations. The conditional quantiles in Fig. 5.7 do not lie on perfectly smooth curves due to the presence of sampling variations. The total variation of the 50 % quantile with forecast value is comparable to the width of the conditional interquartile range, which confirms the lack of strong dependency between the observations and these forecasts, previously noted in this chapter. For a climatological mean forecast, the observation interquartile range would be the 25 % and 75 % percentiles of the observed distribution [−0.7,0.7] no matter what value the forecast took. The fact that some slight trends can be seen in the conditional quantiles in Fig. 5.7 suggests some potential for making (low-skill but not no-skill) probability forecasts on seasonal time scales.

5.6 CONCLUDING REMARKS

We have seen in this chapter that there is no unique way of measuring the skill of a forecast of a continuous variable. Although the product moment correlation coefficient may appear to be a good choice for measuring the association between forecasts and observations, this score is sensitive to outliers, is not adapted to finding non-linear relationships between predictor and predictand, and is difficult to test analytically for variables that are not normally distributed. For these reasons, many verification measures have been developed and the most widespread have been presented in this chapter: bias, MAE, MSE, product moment correlation, rank correlation, Kendall's τ, Daniels' correlation, and LEPS.

We have considered here predictands consisting of a single continuous scalar variable. However, meteorological as well as climate forecasts often consist of fields of numbers representing spatial maps. However, from a specific user's point of view, the predictand of interest is often local, and the verification of a forecast at a specified location is of utmost importance. In fact, forecast evaluation over a region can be considered the aggregation of local scores. When the number of local forecasts is small, spatial aggregation can sometimes help to increase the sample size and thereby reduce the sampling uncertainties in the scores (provided that the area is large enough and the spatial dependency weak enough). In this situation, one can then test whether forecasts over a whole region are skilful without being able to determine whether local forecasts have any real skill. This rather paradoxical situation is often encountered in seasonal forecasts, which frequently have very small sample sizes (e.g. 15 winter forecasts).

6 Verification of Spatial Fields

WASYL DROSDOWSKY AND HUQIANG ZHANG
Bureau of Meteorology Research Centre, Melbourne, Australia

6.1 INTRODUCTION: TYPES OF FIELDS AND FORECASTS

In the preceding chapters the set of forecasts and observations have been functions of time only, that is the forecasts have been made a number of times for the same single parameter, such as precipitation or temperature, at a single geographic location. The predictands in these forecasts have either been discrete events (for two categories or binary events in Chapter 3, or for more than two categories in Chapter 4) or continuous variables (Chapter 5), defined at a single spatial point. Although it has not always been explicitly stated, the verification measures described have usually been assumed to involve averaging over time.

Forecasts of spatial fields involve the same parameter over a range of geographic locations. As with forecasts for a single point, these fields can be generated by statistical or dynamical methods, or a combination of these. The predictands can be either continuous or discrete, and the forecasts expressed as a definite (deterministic) statement or in terms of probabilities. Forecast fields generated by numerical weather prediction (NWP) or numerical climate models generally consist of deterministic forecasts of continuous variables, for example, mean sea level pressure (MSLP) or temperature over a region. These forecasts are spatially coherent, since the forecast values at different grid points are related to one another through the dynamic relationships embodied in the models.

In contrast, long-range outlooks are generally probabilistic forecasts of discrete variables, for example, the probability of rainfall or temperature in one of three equally likely categories: below normal, near normal and above normal. These forecasts are most often generated by statistical or empirical methods, although ensembles of numerical model simulations are becoming more common for both short- or medium-range weather forecasts and long-range seasonal outlooks. The verification of these ensemble forecasts is covered in Chapter 7. Statistical or empirical forecasts may have some coherent spatial structure if the predictands involve large-scale patterns such as those generated through data reduction methods such as principal

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

component analysis (empirical orthogonal function (EOF) analysis), or canonical correlation analysis (CCA). However, in many cases they are often just a collection of individual forecasts, possibly generated by different models and with different predictors, applied at individual grid points.

A possible additional problem with spatial fields is the need to match the forecast field on a regular grid with verifying observations taken at irregularly spaced locations. Operational NWP models are run through a cycle of data assimilation, data analysis and forecasts, which ensures that a verifying field is available on the same grid, with the same spatial and temporal resolution as the forecast field. The data assimilation and analysis phases of this process have been extensively documented, for example, in the textbooks of Daley (1991) and Emery and Thompson (1998). However, one important predictand field, precipitation, is not well handled by this assimilation and analysis procedure, and must usually be treated separately. The short spatial scales involved, especially in regions with significant topography, and the highly variable station location and density of most rainfall observation networks, complicate the analysis of rainfall data.

Matching irregularly spaced verifying data with statistical or empirical forecasts on a regular grid can be done in two ways. Either the original station data are used and the forecasts are interpolated from the regular grid, or the data are interpolated to the grid and compared with the forecasts. In either case, the verifying observations should be available in the same format as the data used to develop and evaluate the hindcast skill of the model. An example of the first approach is given by Barnston (1994) who developed a CCA forecast system for temperature and precipitation at 59 irregularly spaced United States stations. Interpolation to a regular grid is performed for display of the forecasts and skill estimates on a spatial map. In contrast, Drosdowsky and Chambers (2001) and Jones (1998) describe a discriminant analysis forecast scheme for Australian seasonal rainfall and temperature in which the original station data are first interpolated to a regular grid (Weymouth *et al.* 1999). McBride and Ebert (2000) verified NWP forecasts of daily rainfall over the Australian continent using a daily version of the analysed gridded rainfall data set. In this case, both the predictand and the verifying field represent grid box averaged rainfall. Additionally, area averaged estimates of accumulated rainfall from satellite and radar remote sensing systems (Xie and Arkin 1996) are becoming available. These estimates have different characteristics to, and may not be suitable for verifying, individual station or single point forecasts.

Since the forecasts and observations of spatial fields are functions of both space and time, the summation of the various scores or measures described in the previous chapters can be done in a number of ways by partitioning the full data set into various subsets. Three different partitions will be described in this chapter; the chosen method will depend on the nature of the forecast, and the purpose of the verification statistics generated.

The simplest procedure is to ignore the difference between the temporal and spatial dimensions, and simply treat the set of forecast and observed values as a combined ensemble over both space and time. A classic example of this approach, for a categorical forecast of a discrete predictand, is the set of Finley tornado forecasts. Finley's forecasts were issued over a three-month period at two times of the day for 18 districts in the continental US, east of the Rockies, and are essentially a collection of categorical forecasts of a discrete predictand (occurrence or non-occurrence of a tornado) at a number of spatial locations. Most accounts of the verification of these forecasts pool all the forecasts and observations into one data set (for example, Table 1.1, Wilks 1995, pp. 241–242, Table 7.1 and Murphy 1996, Table 3) to give one overall score or skill measure. This procedure is often necessary for forecasts of relatively rare events such as tornados or severe storms to ensure sufficient numbers of forecast and observed events in the verification data. Verification measures for this type of forecast have been discussed extensively in Chapter 3. The same procedure could also be used for continuous predictands using measures such as the mean squared error (MSE – Chapter 5), with the summation over both space and time. The major disadvantage of this methodology is of course the loss of information on the geographic and temporal variability in the forecast skill.

In some cases the single overall score, referred to as the 'grand mean' by Wigley and Santer (1990), can be a useful measure. In all forecast systems, skill will generally not be uniform, either spatially or temporally. Some regions will show high levels of skill while others will show none; similarly some periods, corresponding to particular weather or climate regimes may be more predictable than others. In many cases there will be insufficient data available to test that these differences in skill between models at individual grid points, or at particular times are significantly different. In comparing different forecast schemes or different versions of numerical models, an overall measure of the skill of the forecast system is often desired. For additive measures such as the MSE it makes no difference as to how (or in what order) the averaging over all space and time points is performed. However, if some intermediate results involve non-linear operations such as taking square roots to convert MSE into a root mean square (RMS) measure, or calculating skill scores, then the results will generally be different. For example, the monthly mean RMS obtained by calculating the RMS for each of n_t spatial maps and averaging these (Eq. (6.1)), will generally be different to that produced by calculating the RMS at each of the n_s grid points and averaging these (Eq. (6.2)), and both will be different to the monthly mean RMS obtained by averaging over all space and time points before taking the square root (Eq. (6.3)):

$$\overline{\mathrm{RMS}} = \frac{1}{n_t}\sum_{t=1}^{n_t}\sqrt{\frac{1}{n_s}\sum_{s=1}^{n_s}(\hat{x}_{st} - x_{st})^2} \tag{6.1}$$

$$\overline{\mathrm{RMS}} = \frac{1}{n_s}\sum_{s=1}^{n_s}\sqrt{\frac{1}{n_t}\sum_{t=1}^{n_t}(\hat{x}_{st} - x_{st})^2} \tag{6.2}$$

$$\overline{\mathrm{RMS}} = \sqrt{\frac{1}{n_s}\sum_{s=1}^{n_s}\frac{1}{n_t}\sum_{t=1}^{n_t}(\hat{x}_{st} - x_{st})^2} \tag{6.3}$$

Saha and van den Dool (1988) define an anomaly correlation (see Section 6.3) using a summation similar to Eq. (6.3) over both space and time. This can be further complicated if the averaging is over temporal or spatial subsets of the forecast and observed data, such as monthly averages over a hemispheric or regional domain.

More useful verification measures, giving information on the spatial or temporal variability of the forecast skill, can be obtained by partitioning the data into subsets grouping together (a) all forecasts and observations at different times at the same spatial location, to be described in Section 6.2, or (b) all forecasts and observation for the entire spatial array at a single time (Section 6.3). This partitioning is similar to that discussed by Wigley and Santer (1990). A somewhat different approach to the verification of the full spatio-temporal variability is described in Section 6.4, where use is made of data reduction techniques such as EOF analysis. These techniques are becoming more common in climate (change) model evaluation. While they may lose specific detail at individual points or times they do provide overall measures of the model's skill in reproducing the observed (space and time) variability.

Finally, Section 6.5 briefly discusses rainfall forecasts, which pose particular problems and which have been the subject of much recent research.

6.2 TEMPORAL AVERAGING

The first type of partition treats the data as subsets of forecasts over time at each individual location in the spatial domain, that is, each location (station or grid point) is treated as a different variable. This procedure, which treats each spatial location as a separate variable, can be viewed as a multivariate version of the discussion in the previous chapters. This is a common and useful procedure in operational (weather or climate) forecast verification, where a primary requirement of the verification process is an estimate of the forecast skill at each individual grid point or spatial location. This is especially true of verification carried out for economic or administrative purposes, where the user is often most interested in a single or a few key locations, even though the forecast for these locations is generated from a more comprehensive spatial forecast.

The usual scores and skill measures described in the preceding chapters depending on the type of forecast (deterministic or probabilistic) can be applied on a point-by-point basis. Although the resulting values are often plotted or displayed on the geographic grid of the spatial locations, this is essentially a set of individual scores. As such, in this verification the spatial structure of the forecast and observations, that is the relations between individual points, is not explicitly taken into account. An example is given in Fig. 6.1, which shows ROC scores, defined as the area under the ROC curve (see Section 3.4.4), for forecasts of the probability of exceeding median seasonal rainfall, evaluated on a point-by-point basis over the Australian continent.

Maps such as Fig. 6.1 show the local forecast skill. If we have a collection of a large number of individual points, some percentage of these may be 'significant' or skilful simple due to chance. This problem of assessment of 'global' or 'field significance' has been noted in diagnostic or exploratory climate studies where points with, for example, significant correlations with a potential predictor, may form distinctive teleconnection patterns which reflect the spatial relationships within the observed data, and not necessarily the relationship with the external variable (Livezey and Chen 1983). As discussed by Livezey and Chen (1983) and Wilks (1995, Section 5.4), the assessment of field significance therefore involves two steps; firstly, the

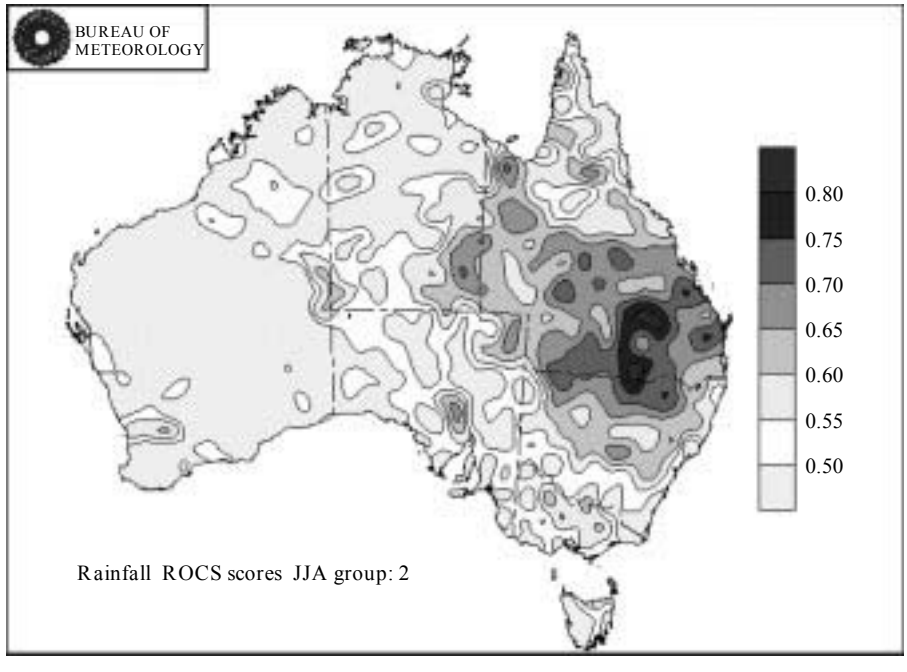


Figure 6.1 ROC scores, defined as the area under the ROC curve, for 50 cross-validated hindcasts over the period 1950–1999 of the probability of above median winter (JJA) rainfall over the Australian continent

determination of local significance at each grid point or location, and secondly, the field significance. The latter step involves two factors; the multiplicity problem due to the large number of individual local points, and the interdependence of these local points due to spatial correlation in the verifying observations.

In diagnostic studies involving correlations or differences in means and variances, the local significance can usually be easily evaluated using standard statistical techniques. While this is not usually the case with skill measures, as discussed in Section 4.4, confidence intervals and local significance levels may be determined using resampling methods such as the bootstrap (Wilks 1995, p. 145). The field significance can then be evaluated as in the case of diagnostic studies. The multiplicity problem is most often analysed using the binomial distribution, which is used to determine the probability of at least the observed number X of N grid points or stations passing the local significance test. The effect of spatial correlation is more difficult to assess and is usually done by permutation or bootstrap procedures – see Section 4.4 and Livezey and Chen (1983).

6.3 SPATIAL AVERAGING

The unique characteristics of the verification of spatial fields are revealed by the partitioning of the data into subsets of forecasts at each location for a given time. This gives a score for a particular time calculated over all spatial points. This is also a common procedure in NWP or climate model evaluation where scores are produced for each day and then often (but not always) further averaged over time to give monthly or seasonal means. Scores can also be calculated directly for seasonal forecasts. As in Section 6.2, most or all of the standard verification measures can be applied, with the summation now being over all spatial points rather than all times.

6.3.1 Measures Commonly Used in the Spatial Domain

Although these measures are defined here in a spatial context, several have already been met in Chapter 5. The only difference is that their summations and averages were with respect to time rather than spatial position.

Mean Error (ME) or Bias

As just noted, this is defined in a similar manner to the time averaged counterpart, except that the summation is over the spatial variable, which may be individual grid points or stations:

$$\mathrm{ME} = \frac{1}{n}\sum_{i=1}^{n}(\hat{x}_i - x_i) \tag{6.4}$$

In weather forecasting and climate models the observed and forecast fields often diverge over the time of the model run, and the difference between the mean observed and forecast fields is therefore sometimes referred to as the *model bias* or *climate drift*.

Mean Absolute Error (MAE)

This is again defined in a similar manner to the time averaged counterpart, but with the averaging over the spatial variable, indexed by *i*:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{x}_i - x_i| \tag{6.5}$$

Mean Squared Error

This is one of the most commonly used verification measure in NWP and/or model evaluation. Again the definition corresponds to the time averaged form:

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\hat{x}_i - x_i)^2 \tag{6.6}$$

Murphy (1988) decomposed the MSE for the time averaged case into three components; a bias or mean error term, the sample variances of the forecasts and observations, and a term involving the covariance or correlation between the forecasts and observations. Murphy and Epstein (1989) performed a similar decomposition on the space averaged MSE (Eq. (6.6)). Denoting the forecast field by $\hat{\mathbf{x}}$, the observed verifying field by $\mathbf{x}$, and a suitable climatological field by $\mathbf{x}_c$, we can define *anomaly* fields by:

$$x_i' = x_i - x_{ci} \tag{6.7}$$

$$\hat{x}_i' = \hat{x}_i - x_{ci} \tag{6.8}$$

By adding and subtracting the forecast and observed field spatial means at each grid point, Eq. (6.6) can be written as

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} [(\hat{x}_i' - \bar{\hat{x}}') - (x_i' - \bar{x}') + (\bar{\hat{x}}' - \bar{x}')]^2 \tag{6.9}$$

The MSE is unchanged by this procedure, and is therefore insensitive to the choice of climatological field. Squaring and then rearranging terms leads to

$$\mathrm{MSE} = (\bar{\hat{x}}' - \bar{x}')^2 + s_{\hat{x}'}^2 + s_{x'}^2 - 2s_{\hat{x}'x'} \tag{6.10}$$

The first term is the square of the mean error of the anomaly fields,

$$s_{\hat{x}'}^2 = \frac{1}{n}\sum (\hat{x}_i' - \bar{\hat{x}}')^2 \tag{6.11}$$

is the forecast anomaly field sample variance, and $s_{x'}^2$ is the corresponding observed anomaly field variance. The final term is twice the covariance between the observed and forecast anomaly fields. Dividing the covariance by the forecast and observed anomaly field standard deviations this term becomes the 'anomaly correlation' described in the next section. Note that there is no temporal variability, and the variance and covariance terms of the anomaly fields are relative to the spatial means of these anomaly fields. Each of the last three terms in Eq. (6.10) contains some element of the climatological field, and is therefore sensitive to the choice of this field.

Anomaly Correlation Coefficient

Although the anomaly correlation coefficient (ACC) is one of the most widely used measures in the verification of spatial fields, there are a number of different definitions available in the literature. These differences arise from two sources, namely, the choice of climatological field and whether the correlation is centred or uncentred. Following the derivation in the previous section, and in Murphy and Epstein (1989), the ACC is the centred correlation between forecast and observed anomalies. That is

$$\mathrm{ACC} = \frac{\sum_{i=1}^{n} (\hat{x}_i' - \bar{\hat{x}}')(x_i' - \bar{x}')}{n s_{\hat{x}'} s_{x'}} \tag{6.12}$$

where the various terms have the same meaning as in the previous section. This expression is similar to the Pearson product moment correlation and involves the cross-product of the deviations of the forecast and observed anomalies about the sample means, which, in the case of spatial fields, are the spatial mean of the forecast and observed anomaly fields. These anomalies are themselves departures of the actual forecast and observed values from a 'normal' or climatological value, which is different at each grid point. This form of the ACC is used at a number of operational centres, as referenced by Hollingsworth *et al.* (1980) at ECMWF, Livezey *et al.* (1995) at NCEP and Mullenmeister and Hart (1994) at BMRC, among others.

The earliest reference to 'the correlation for the anomaly' appears to be by Miyakoda *et al.* (1972). Their definition, which is also used by Miyakoda *et al.* (1986), is not the usual Pearson product moment shown in Eq. (6.12), but instead is an uncentred cross-product between the forecast and observed anomalies. Using notation consistent with Eq. (6.12), the Miyakoda *et al.* (1972) version of the ACC (their Eq. (11)) can be written as

$$\mathrm{ACC_u} = \frac{\sum_{i=1}^{n} (\hat{x}_i')(x_i')}{n s_{\hat{x}'} s_{x'}} \tag{6.13}$$

In this expression the 'standard deviation' terms in the denominator are also uncentred. Potts *et al.* (1996) perform a decomposition of the MSE similar to that of Murphy and Epstein, but based on this uncentred version of the ACC. This decomposition (Potts *et al.* 1996, Eq. (3b)) does not include the bias term in Eq. (6.10). This form of the ACC has been used in some theoretical studies that examine predictability (for example, Saha and van den Dool 1988, Anderson and van den Dool 1994), and is quoted in the two most widely referenced overviews of the field, Stanski *et al.* (1989) and Wilks (1995).

Does it matter which definition is used? Wilks (1995, p. 278) discusses the difference in the two forms and notes that they will be identical if the spatial mean of both the forecast and observed anomaly fields is zero. Wilks suggests that this may be true for global or hemispheric fields, but will not necessarily hold for relatively small areas. However, while this condition is more likely to be met for the observed field, it will not be true for forecast fields that exhibit significant departures from climatology, even those of global or hemispheric scale. Unfortunately, this is the case for many fields generated by NWP models and coupled climate models. For example, most NWP models exhibit a cold temperature bias (relative to climatology) in the upper troposphere and lower stratosphere that grows with increasing forecast lead-time. The spatial mean of the anomaly (from climatology) of this field will not be close to zero, and the two different expressions for the anomaly correlation may give significantly different results.

The situation may be worse in climate models where there is significant climate change, and both forecasts and observations differ or 'drift' substantially from a climatology based on an earlier period. Both versions of the anomaly correlation have been used in signal detection analysis of climate change model experiments (see Section 6.4). In this regard the uncentred version of the ACC can be made arbitrarily large, even approaching unity if both the forecast and observed fields are sufficiently far from the chosen climatology, so that all the anomalies are relatively large, and of the same sign. In the NWP field this sensitivity of both versions of the ACC to the climatology used makes it difficult to compare the verification of different models by different operational centres.

S1 Score

This measure was introduced by Teweles and Wobus (1954), and its use has been restricted to forecasts of pressure or height fields. The score is based on a comparison of gradients between grid points or stations, on the assumption that winds which are directly related to pressure or height gradients are

more significant or useful quantities to verify than the pressure or height fields themselves. The score was developed at a time when forecast charts were manually prepared, and usually limited to the pressure or height fields. The score appears to be maintained at a number of operational forecast centres for historical reasons, with verification on the same grid extending back many decades. For example, Fig. 6.2 shows monthly averages of S1 scores calculated over the same set of grid points over the Australian region from 1976 to 2000, a period spanning a number of model changes.

The score is defined as

$$\mathrm{S1} = 100 \frac{\sum_i |\Delta \hat{x}_i - \Delta x_i|}{\sum_i \max\{|\Delta \hat{x}_i|,\ |\Delta x_i|\}} \tag{6.14}$$

where the sums are evaluated over a predefined set of pairs of adjacent points or stations.

$\Delta \hat{x}_i$ is the forecast pressure difference between the ith pair of adjacent points, and Δx_i is the corresponding observed pressure difference. The numerator is therefore a measure of the mean absolute error in the pressure

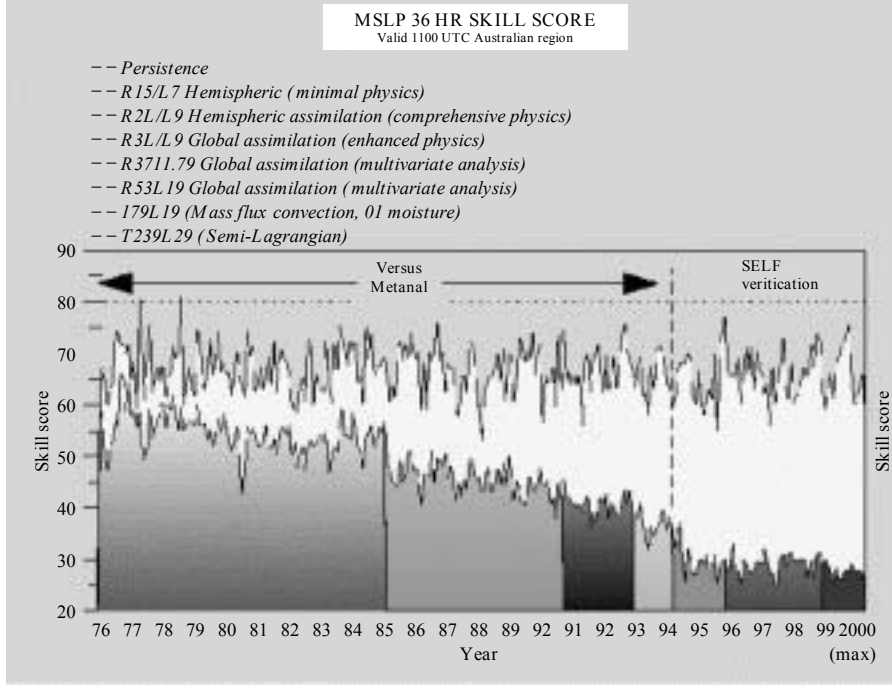


Figure 6.2 Monthly averaged S1 scores for 36 h forecasts of MSLP over the Australian region from 1976 to 2000. Note the steady downward trend compared with the unchanged scores for persistence

gradients, while the denominator, which is the maximum of the observed or forecast pressure gradient, is a normalising factor.

Small values of S1 are desirable. Figure 6.2 shows a consistent level of S1 between 60 and 70 for persistence forecasts while model-based forecasts have values that decrease from over 50 to less than 30 over the period. Some of the decrease is gradual showing improvements during the years in which a particular model was used. Other changes are more abrupt, at the vertical lines corresponding to model changes.

The S1 score has a number of disadvantages. Firstly, the score is highly dependent on the size of the forecast domain and the set of grid points used for verification. In particular, the score can fluctuate rapidly when evaluated over very small regions, relative to the size of typical weather systems, or those containing few grid point pairs.

The denominator was introduced by Teweles and Wobus (1954) to prevent human forecasters from attempting to optimise their scores by forecasting weak systems, thereby lowering the sum of absolute pressure gradient forecast errors in the numerator. However, this correction term tends to be larger in winter when weather systems are more intense, and hence introduces a seasonal trend of lower winter and higher summer values into the S1 score. Since the S1 score is a corrected sum of absolute pressure gradient errors, it should be related to, and behave similarly to, the mean absolute error of the wind field.

LEPS

The score based on linear error in probability space (LEPS), which has been discussed for categorical and continuous data in Sections 4.3.4 and 5.5.1, respectively, can equally well be applied to spatial data – see, for example, Potts (1991), Potts *et al.* (1996).

6.3.2 Map Typing and Analogue Selection

The spatial field verification discussed here is closely related to the problem of map typing and analogue selection. In measures-orientated forecast verification, we are measuring the closeness between the forecast and observed spatial fields. Map typing or analogue selection involves the matching of pairs of similar or closely related observed spatial maps. Most of the measures discussed here, as well as many others, have been used for this purpose. Toth (1991) compared nine different similarity measures including the MAE, the MSE, the (uncentred) anomaly correlation and a modified S1 score. The best measures were judged not just on their ability to select analogues between individual maps, but also to select those which remained close to the initial map over a forecast period beyond the initial selection period or day.

In Toth's (1991) study neither the uncentred correlation nor S1 performed well; nor did a measure based on vorticity. The best measures

were functions of gradients, but defined differently from S1, while MSE and a weighted version of MAE also did quite well. Relative performance of the various measures depends on the criteria by which they are compared. For example, S1 and anomaly correlation, which fared badly in Toth's comparisons, performed well in a study conducted by Potts (1991) of the power of measures to detect differences between fields.

6.3.3 Accounting for Spatial Correlation

Apart from the measures based on gradients, like S1, the measures discussed so far do not actually consider the spatial structure of the forecast and observed fields (despite the sometimes used term 'pattern correlation' for the ACC) in the same sense that the usual correlation between two time series says nothing about the temporal 'pattern' of the two time series. For example, two time series may be positively correlated if both contain in-phase relatively high-frequency variations, or both contain a similar long-term trend. These temporal patterns of time series, and their relationships can be elucidated by spectral or cross-spectral (coherence) analysis. In a similar manner, to examine the correspondence between spatial patterns we could perform a similar decomposition using a two-dimensional Fourier analysis over the spatial dimensions. Huth (1996) examined a wider range of techniques used in circulation classification studies, included cluster analysis and PCA, as well as the correlation and MSE measures. Stephenson (1997) showed that a generalised anomaly correlation based on standardised principal components (PCs) defined with the Mahalanobis metric was invariant under linear transformations, such as linear interpolation to another grid.

6.4 ASSESSMENT OF MODEL FORECASTS IN THE SPATIO-TEMPORAL DOMAIN

Weather and climate variability cover a broad range of temporal and spatial variations. Thus, the assessment of the skill of model simulations and/or forecasts requires more than the calculation of traditional measures such as the MSE aggregated over time (Section 6.2) or over space (Section 6.3). In recent years the availability of long time integration of the global atmospheric models used in seasonal forecasts and climate change simulations have led to a number of studies using different statistical methods to assess the full spatio-temporal variability of these models.

6.4.1 Principal Component Analysis (EOF Analysis)

PCA, commonly known in atmospheric science as EOF analysis, is a well-known dimension reducing technique (Jolliffe 2002) which can be used to compact the structure of a set of spatial maps into a number of leading

modes. Hence, the original spatial variables are replaced by a smaller number of PCs. An advantage of verifying PCs, as discussed by Stephenson (1997), is that the PC time series are uncorrelated so that each PC contains separate information about model skill. However, this is offset by the loss of detailed information at local spatial points.

An example of this technique is given by Renshaw *et al.* (1998), who assessed a climate model's ability to simulate the leading modes of low-frequency weather variability. The leading modes from observations and from the model simulations are obtained via rotated EOFs of daily filtered 500-h Pa heights. This analysis allows the verification of a number of aspects from the model results: (i) whether the dominant spatial patterns from model simulations and observations have similar features; (ii) whether the relationship between the leading modes and imposed tropical sea-surface-temperatures (SSTs) forcing has similar features in the model results and in the observations; (iii) by comparing the time series of the leading spatial modes from model results and observations, it illustrates the similarity of the variability of those leading modes in the model simulations and observations. Using a similar approach, Peng *et al.* (2000) studied the model simulation skill of the SST forced global climate variability.

Wang and Rui (1996) reported a method for assessing the model forecasting skill in ensemble seasonal and climate forecasts. In this approach, they apply PCA to the spatial fields, and then calculate the spread of model ensemble forecasts in the leading PCs. Thus, results from this method display the dominant spatial patterns from the ensemble forecasts and the uncertainty attached to them.

Peng *et al.* (2000), however, have noted that having a realistic representation of the leading modes of variability in the model does not necessarily translate into forecasting skill because the model must also exhibit the response at the appropriate times. Poor model forecasting and/or simulation skill may be the result of two different problems: the model modes do not resemble the observed modes (in structure and/or location) or the timing of the principal responses may be incorrect.

6.4.2 Combining Predictability with Model Forecast Verification

In recent years, there have been a number of seasonal climate forecasts experiments employing global climate models forced by either the observed or forecasted SSTs (for example, Kumar *et al.* 1996; Hunt and Hirst 2000; Frederiksen *et al.* 2001). The rationale behind these experiments is that the atmospheric seasonal mean state at some locations and seasons is dominated by the slow-varying boundary forcing. Statistical studies have demonstrated strong correlations between tropical SST variations and tropical and extra-tropic climate anomalies (for example, Ropelewski and Halpert 1986, 1987; Nicholls 1989). However, the seasonal mean state is also partially

attributed to the random weather noise, which is due to the chaotic nature of atmospheric process. Thus, in recent years, there have been a number of studies aimed at investigating the potential predictability of atmospheric variability from seasonal (for example, Rowell 1998; Zheng and Frederiksen 1999) to decadal scale (for example, Rowell and Zwiers 1999). These model forecasts can be evaluated in two different ways: one by calculating the difference between modelled and observed seasonal and climate anomalies; and the other by focusing on the assessment of the model performance on the part of variation which is potentially predictable. Verification of just the potentially predictable component leads to improved understanding of the model forecasts and the natural processes of the climate system.

There are a number of approaches proposed in the predictability studies; however, all are based on the analysis of variance (ANOVA) approach. Rowell (1998) decomposed the total atmospheric variance from the ensemble atmospheric model integrations forced with observed SST and sea-ice into two components: a component due to the slow-varying SST and sea-ice boundary conditions; and the other the random internal variability component due to the chaotic nature of the atmospheric process. Potential predictability is then defined by the ratio of SST forced variance to the total variance. Zheng and Frederiksen (1999) proposed an approach that further decomposes the total variance from an ensemble of model simulations or from the observations into three components, namely, the forced component, the internal source component, and the weather noise component. Statistical formulae for the estimated variability of model-simulated data are derived in Zheng and Frederiksen (1999) including the variances of the forced component, the weather noise, and the internal source component.

With the assumptions that the ratio of the forced component variance to the variance of the internal source component is correctly simulated by the model, the variance of the forced component in the observational data can be calculated, together with the variability of the weather noise component. Then a series of correlations can be calculated to measure the skill of the total forecasts, and each of its components. The advantage of such a verification approach is apparent. Such correlation indices clearly demonstrate the success or failure of the forecast model in simulating or predicting the seasonal climate anomalies due to the slow-varying boundary SST forcing component, and the anomalies due to the internal process; and to what extent the observed seasonal anomalies are attributed to stochastic weather noise which is unpredictable beyond several days.

6.4.3 Signal Detection Analysis

A major difficulty in climate change studies arises from trying to separate a weak climate signal that represents the human impacts, if any, on climate from the noisy background representing the natural climate variations. A number of statistical approaches, as reviewed in Santer *et al.* (1995a)

and Zwiers (1999), have been proposed to detect the climate change signal from observations based on the climate model simulations, identify the causes of the observed climate changes and verify the reliability of current climate model predictions of future climate and climate change. Such approaches, described briefly below, may have the potential to be applied to other spatial forecasting verification problems.

Santer *et al.* (1995b) applied pattern similarity statistics to assess the resemblance of the modelled and observed climate change signals. Supposing $\hat{x}_i$ is the equilibrium climate change signal from a model simulation and $\mathbf{x}_{it}$ is a series of time-evolving observed climate anomaly patterns relative to a temporally smoothed reference state, then two kinds of spatial pattern correlations are calculated to detect the climate change signals from observations:

$$R(t) = \frac{\sum_{i=1}^{n} (x_{it} - \bar{x}_t)(\hat{x}_i - \bar{\hat{x}})}{n s_x s_{\hat{x}}} \tag{6.15}$$

$$C(t) = \frac{\sum_{i=1}^{n} x_{it} \hat{x}_i}{\sum_{i=1}^{n} \hat{x}_i^2} \tag{6.16}$$

where i is the spatial index and t the temporal index, an overbar indicates a spatial average, and s_x and $s_{\hat{x}}$ are the spatial variances of the anomalies of the observed time series and the climate change signal in the model.

$R(t)$ and $C(t)$ are very similar to the centred and uncentred spatial anomaly correlations, respectively, as described in Section 6.3. The major difference is the use of a slowly evolving climatology from which the anomalies are calculated. As noted in Santer *et al.* (1995b) each of these statistics has its advantages. $C(t)$ is a measure of the strength of the model signal in the observations, and its trend over time can only be attributed to the increase or decrease of the similarity between model and observed changes. $R(t)$ focuses on the similarity of the patterns of change between the model and observation, and is useful in discriminating between different forcing mechanisms with different signature patterns. Other sophisticated signal detection methods have been developed or proposed in the climate change literature (North *et al.* 1995; Hasselmann 1997 Zwiers 1999).

6.5 VERIFICATION OF SPATIAL RAINFALL FORECASTS

One of the most challenging verification tasks is the verification of spatial precipitation forecasts. Two main problems are associated with the analysis and evaluation of either rainfall rate or rainfall amount. First, precipitation is a highly discontinuous variable in the sense that there are substantial segments both in space and time where rainfall is zero; whilst in other segments it is positive and continuous. Second, as already noted

in Section 6.1, there is often a mismatch between the form of the forecasts and the verifying data. The forecasts are presented as spatial maps, usually derived from values on a regular spatial grid. On the other hand, verifying observations are likely to be made at an irregularly spaced network of stations, or estimated indirectly, for example, from radar reflectances. In addition, rainfall amounts have a highly skewed distribution, ruling out procedures that make implicit or explicit use of Gaussian assumptions.

Many standard verifIcation procedures are used, with varying degrees of success, on spatial rainfall data. This section concentrates on techniques not discussed elsewhere in the book. Measures such as root mean square error and correlation coefficients give poor scores when a spatial rainfall forecast is correct in intensity and areal extent, but is displaced in space or time from its correct position. Part of the problem arises because of the highly skewed nature of rainfall amounts, with extreme values having undue influence on values of standard measures. To address this problem, alternative scores such as the *root mean squared factor* have been introduced (Golding 1998).

More sophisticated approaches decompose skill, or lack of skill (forecast–observation disagreement), into parts due to displacement of rainfall patterns, differences in amount of rainfall, and residual disagreement. Ebert and McBride (2000), Hoffman and Grassotti (1996) and Hoffman *et al.* (1995) introduce two feature-based methods based on such decompositions. The displacement disagreement is obtained by translating the forecast rainfall features over the observed features until a 'best fit criterion' is satisfied. This criterion consists of minimising MSE (Ebert and McBride 2000) or a spectral coefficients cost function (Hoffman and Grassotti 1996). Ebert and McBride (2000) and Ebert (2001) also introduce a novel contingency table, based on displacement and amount disagreement categories, and use it to assess Australian forecasts.

Briggs and Levine (1997) developed a multi-valued verification score to assess forecasts at different spatial scales. A wavelet decomposition of the forecast and observed fields is performed, and 'noise' is removed from both fields by applying a threshold on the wavelet coefficients. Then forecast and observation fields are reconstituted for each spatial scale and verification scores are evaluated for each scale. The percentages at which each spatial scale contributes to the overall score are also evaluated. As noted earlier, at the time of writing this is an active area of research.

7 Probability and Ensemble Forecasts

ZOLTAN TOTH[1], OLIVIER TALAGRAND[2], GUILLEM CANDILLE[2] AND YUEJIAN ZHU[3]
[1]NOAA National Centers for Environmental Prediction, Camp Springs, MD, USA
[2]Laboratoire de Météorologie Dynamique, Paris cedex, France
[3]SAIC at NOAA National Centers for Environmental Prediction, Camp Springs, MD, USA

7.1 INTRODUCTION

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

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

Probabilistic forecasts can also be based on a set of deterministic forecasts valid at the same time. Assuming the forecasts are independent realizations of the same underlying random process, an estimate of the forecast probability of an event is provided by the fraction of the forecasts predicting the event among all forecasts considered. This technique, known as *ensemble forecasting* (see Leith 1974; Ehrendorfer 1997; Stephenson and Doblas-Reyes 2000; and references therein), can produce probabilistic forecasts

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

based on a set of deterministic forecasts, without relying on past verification statistics. In certain fields of environmental science, such as meteorology and hydrology, the ensemble forecasting technique is now becoming widely used. Therefore, this chapter will also present some of the methods that have been developed to directly evaluate a set of ensemble forecasts, before they are interpreted in probabilistic terms.

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

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

7.2 MAIN ATTRIBUTES OF PROBABILISTIC FORECASTS

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

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

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

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

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

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

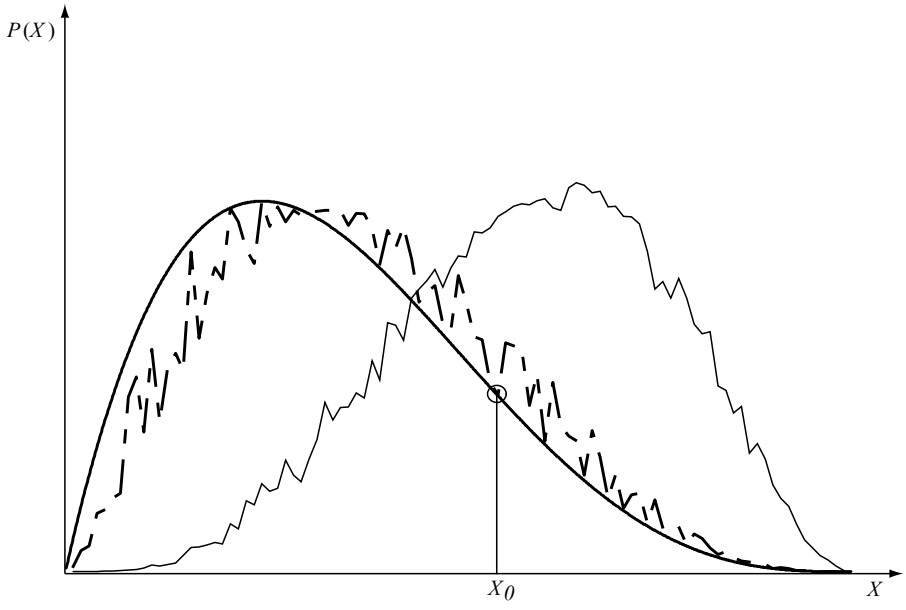


Figure 7.1 A hypothetical forecast probability density function $\hat{p}(x)$ (full curve) for a one-dimensional variable x, along with a verifying observed value x_0 for a single case. The additional two curves represent possible distributions for the verifying values observed over a large number of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

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

A probabilistic (or any other) forecast system, as pointed out above, can be validated only in a statistical sense. Therefore, similarly to the case of probabilistic forecasts for a given event discussed above (there is a 40 % probability that it will rain tomorrow), one has to wait until a particular probability distribution $\hat{p}(x)$ has been predicted a number of times. Let us denote by $p_o(x)$ the frequency distribution of observations corresponding to the cases when $\hat{p}(x)$ is forecast. If $p_o(x)$ is similar to $\hat{p}(x)$ (as is the case for distribution $p_1(x)$ shown by the dash-dotted curve in Fig. 7.1), then the prediction $\hat{p}(x)$ can be described as statistically consistent with the observations. If, however, $p_o(x)$ is distinctly different from $\hat{p}(x)$ (as is the case for the

distribution $p_2(x)$ shown by the dashed curve in Fig. 7.1), then the forecast $\hat{p}(x)$ is statistically *inconsistent* with observations.

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

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

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

The two main attributes of probabilistic forecasts, *reliability* and *resolution*, are a function of both the forecasts and the verifying observations. Resolution was defined above as the variability of the observed frequency distributions associated with different forecast scenarios around the climatological p.d.f. Another property, *sharpness*, measures the variability of the *forecast* (and not the corresponding observed) probability distributions around the climatological p.d.f. Note that in a perfectly reliable forecast

system the forecast probability values, by definition, are identical to the corresponding frequency of verifying observations. For a reliable forecast system sharpness is therefore identical to resolution.

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

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

7.3 PROBABILITY FORECASTS OF BINARY EVENTS

In Sections 7.3.1–7.3.3, we will consider verification methods for the simplest conceptual case of probability forecasts of binary events. Such events, denoted by A, can be defined in different ways. One can use an inequality of the form $\{A: X > u\}$, where X is a scalar variable for which a probabilistic forecast is made, and u is a given threshold value. Examples of this type include the occurrence or not of a particular binary event A such as 'the temperature at a given location x at forecast lead-time t will be greater than 0 °C', or 'the total amount of precipitation over a given area and a given period of time will be more than 50 mm'. Other events, like 'Tropical storm

Emily will hit land', or 'Electric power distribution will be disrupted by thunderstorms', cannot be easily expressed in terms of a meteorological parameter exceeding a certain threshold, yet are equally interesting. This section is devoted to the verification of probabilistic forecasts of binary events, regardless of how they are defined.

Since a collection of probabilistic forecasts of discrete variables or categories having multiple values can also be considered as a set of probabilistic binary events, the methods of this section have a more general importance in probabilistic forecast verification. Probabilistic forecasts of binary events are, therefore, of fundamental importance in the verification of probability forecasts.

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

7.3.1 The Reliability Curve

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

The solid line in Fig. 7.2 shows the *reliability curve* obtained by plotting values of $f(q)$ against q, for the forecast system and event described above. Although rather close to the line $f(q) = q$ (i.e., perfect reliability), the reliability curve does show some significant deviations from it. In particular, the slope of the reliability curve in Fig. 7.2 is below that of the $f(q) = q$ diagonal. Note that deviations from the diagonal are not necessarily indicative of true deviations from reliability but can also be due to sampling

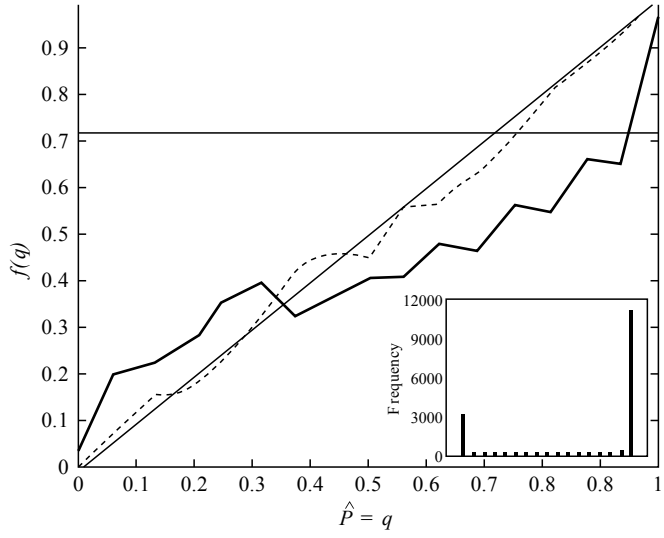


Figure 7.2 Reliability diagram for the NCEP Ensemble Forecast System (see text for the definition of the event A under consideration). Full line: reliability curve for the operational forecasts. Dashed line: reliability curve for perfect ensemble forecasts where 'observation' is defined as one of the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

variations. When statistics, as in our example, are based on a finite sample, the reliability curve for even a perfectly reliable forecast system is expected to exhibit sampling variations around the diagonal. The amount of sampling variability can be easily assessed by plotting reliability curves for the same forecast system except now using a randomly chosen member of the ensemble of forecasts in place of the verifying observations. By definition, the forecast system should be perfectly reliable in this case, and so deviations from the diagonal are due to sampling variations. When compared to the diagonal, the difference between the perfect (dashed line in Fig. 7.2) and operational ensemble curve (solid line) reflects the true lack of reliability in the forecast system, given the size of the verification sample. Bootstrap methods (see Efron and Tibshirani 1993) could easily be developed to quantify the sampling uncertainty in these estimates of reliability.

The graph in the lower right corner of Fig. 7.2 is known as a sharpness diagram which shows the frequencies for the forecast probabilities dividing by the total number of forecasts gives sample estimates of the marginal probability distribution of the forecast probabilities q. Since the probabilities of zero or one are used in 90 % of the forecast cases, the forecast system exhibits a considerable degree of sharpness, as defined above. This is due to the small spread in temperatures in these short-range 2-day lead-time forecasts, resulting in either none or all of the forecasts often falling 4 °C below the climatological mean value.

7.3.2 The Brier Score

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

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

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

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

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

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

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

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

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

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

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

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

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

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

where the expectation over X is given by

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

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

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

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

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

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

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

$$BSS_{\mathrm{c}} = 1 - B_{\mathrm{rel}} - B_{\mathrm{res}} \tag{7.10}$$

In our operational NCEP forecasting example, the Brier score for the system represented by the solid curve in Fig. 7.2 is equal to 0.066. The base rate for the event under consideration is equal to 0.714, which yields 0.677 for the Brier skill score BSS_{c}. The corresponding values of the components defined in Eq. (7.9) are $B_{\mathrm{rel}} = 0.027$ and $B_{\mathrm{res}} = 0.296$. These values are typical of the values produced by present-day operational short- and medium-range weather forecasting systems. It is often found that the reliability term is significantly (typically one order of magnitude less) smaller than the resolution term.

Figure 7.3 shows the BSS defined using the climatological forecast system BSS_{c}, full curve) and its two components B_{rel} (thin-dashed curve) and B_{res}

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

Finally, we note again that if both the forecast and observed systems are stationary in time and there is a sufficiently long record of their behavior

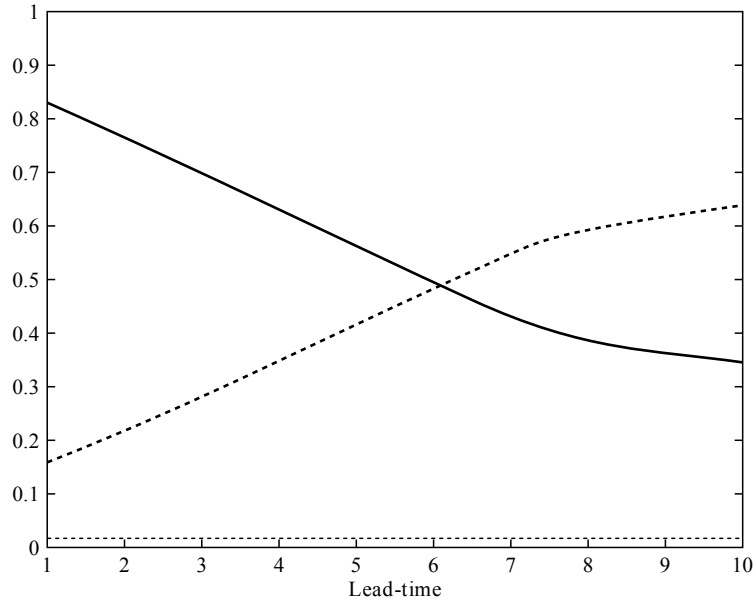


Figure 7.3 Brier skill score (BSS_c, full curve, positively oriented), and its reliability (B_{rel}, thin dash), and resolution (B_{res}, thick dash, both negatively oriented) components, as a function of forecast lead-time (days) for the ECMWF Ensemble Prediction System (see text for the definition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

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

7.3.3 Verification Based on Decision Probability Thresholds

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

As explained in detail in Section 3.4 of Chapter 3, a continuous set of deterministic binary forecasts can be verified using signal detection techniques. One of the most powerful tools is the *relative operating characteristic* obtained by plotting the hit rate versus the false alarm rate for each possible

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

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

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

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

The integral in Eq. (7.11a) is the average of $f(q)$ for those circumstances when $q > p_t$. When $f(q)$ is a strictly monotonically increasing function of p_t, the threshold inequality $q > p_t$ becomes equivalent to $f(q) > f(p_t)$, and the comparison with the threshold can be done on the *a posteriori* calibrated probabilities $q' = f(q)$ as well as on the directly predicted probabilities q. The same argument equally applies to the integral in (7.11b), which means that the ROC curve is invariant to *a posteriori* calibration $q' = f(q)$ (where the points on the reliability curve are moved horizontally to the diagonal). Thus, if the reliability curve is strictly monotonically increasing, the ROC curve, just like the resolution component of the Brier score, depends only on the *a posteriori* calibrated probabilities $q' = f(q)$ (and their probability distribution). The ROC curve, in these cases, is therefore independent of reliability, and measures the resolution of the forecasting system. The resolution component of the Brier score and the ROC curve therefore often provide very similar qualitative information.

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

7.4 PROBABILITY FORECASTS OF MORE THAN TWO CATEGORIES

7.4.1 Vector Generalization of the Brier Score

The Brier score was defined in Eq. (7.1) for the verification of probability forecasts of binary events. However, Brier (1950) gave a more general definition that considered multiple categories of events. Let us consider an

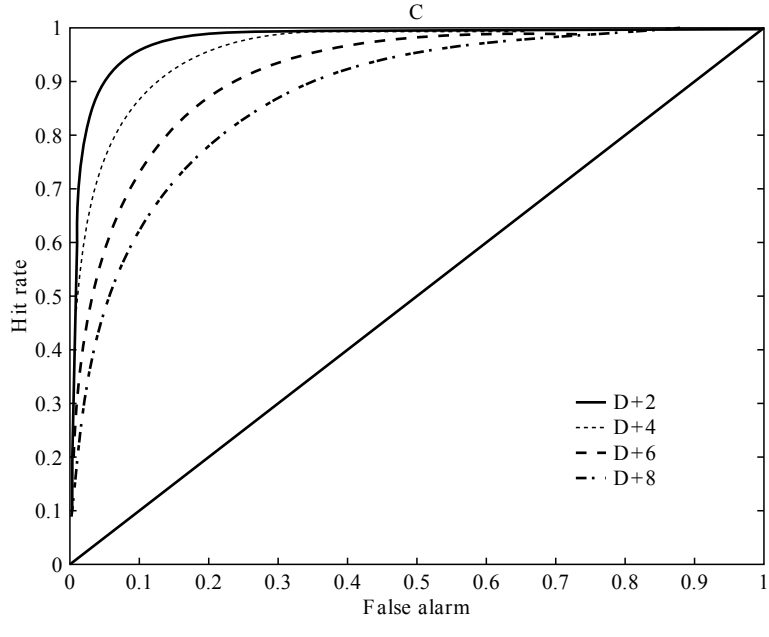


Figure 7.4 ROC curves for the same event and predictions as in Fig. 7.3, for four different forecast lead-times (2, 4, 6, 8 days)

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

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

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

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

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

where $\underline{p}_\mathrm{c}$ is the sequence (s_k). Similarly to Eq. (7.8), Eq. (7.13) provides a decomposition of B_K into reliability, resolution, and uncertainty terms.

7.4.2 Information Content as a Measure of Resolution

It has been argued that given a suitably large sample of previous forecasts and matching observations, probabilistic forecasts can be made more reliable by calibration. Unlike reliability, the resolution of a forecasting system cannot be changed by calibration and so represents the (invariant) ability of the forecasting system to resolve future events. Various measures of reso-

lution have been proposed for probability forecasts including ones based on information theory measures such as *information content* (entropy) (see Section 2.7; Toth *et al.* 1998; Stephenson and Doblas-Reyes 2000; Roulston and Smith 2002). One possible definition of the information content (I) of a forecast of the probabilities for K mutually exclusive, climatologically equiprobable, and exhaustive categories is given by

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

where $0 \leq \hat{p}_i \leq 1$ is the forecast probability for the ith category that satisfies $\sum_{i=1}^{K} \hat{p}_i = 1$. When forecasts are perfectly reliable, the mean information content over all forecasts can be considered to be another measure of resolution. Under these conditions, the information content ranges between zero for a uniform probability forecast that forecasts $\hat{p}_i = 1/K$ for all categories, and one for a deterministic forecast that forecasts $\hat{p}_i = 1$ for only one category and 0 for all others. Figure 7.5 shows the mean information content of a 10-member 0000 UTC subset of the NCEP ensemble forecasts for 500 hPa geopotential height as a function of lead-time. For not perfectly reliable forecasts, a more general and invariant measure of resolution can be obtained by considering the information content of the calibrated forecasts $I[f(\hat{p})]$. It should be noted that information content defined in this way is equal to the likelihood ratio statistic for goodness-of-fit tests (Garthwaite

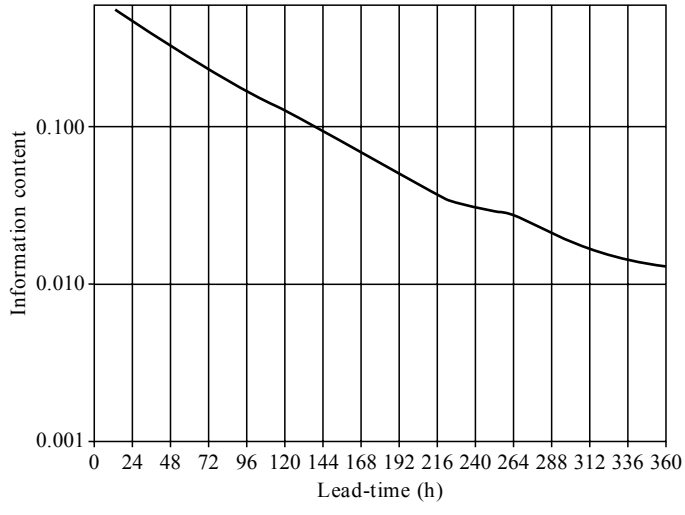


Figure 7.5 Information content as defined in text for calibrated probabilistic forecasts (with near perfect reliability) based on a 10-member subset of the NCEP ensemble. Forecasts are made for 10 climatologically equally likely intervals for 500 hPa geopotential height values over the Northern Hemisphere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

et al., 2002, Section 8.4.1) of which the G^2 measure of association for $(K \times K)$ contingency tables (Chapter 4, Section 4.4) is a special case. G^2 is asymptotically equivalent to the better known X^2 (Pearson) measure of association (Stephenson 2000), and so the latter goodness-of-fit measure may also provide a good overall measure of resolution for probability forecasts.

7.5 PROBABILITY FORECASTS OF CONTINUOUS VARIABLES

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

7.5.1 The Discrete Ranked Probability Score

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

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

where B_k is the Brier score for the event $A_k = \{X \leq x_k\}$. The RPS is similar to the multiple category Brier score in Eq. (7.12), but, as its name implies, it takes into account the ordered nature of the variable X. Here the events A_k are not mutually exclusive, and A_j implies $A_{j,j'>j}$. Consequently, if X is for instance forecast to fall in an interval $[x_j,\ x_j + 1]$ with probability one, but is observed to fall into another interval $[x_j',\ x_j' + 1]$, the RPS increases with the increasing absolute difference $|j - j'|$.

7.5.2 The Continuous Ranked Probability Score

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

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

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

7.6 SUMMARY STATISTICS FOR ENSEMBLE FORECASTS

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

Section 7.2 pointed out that the inherent value of forecast systems lies in their ability to distinguish between cases when an event has a higher or lower than climatological probability to occur in the future (resolution). As Figs. 7.3 and 7.4 demonstrate, resolution decreases rapidly with lead-time (due to the loss of information in the flow). This is because in fluid systems such as the atmosphere and oceans, naturally occurring instabilities amplify initial and model related uncertainties. Even though skill is reduced and eventually lost, forecasts can remain (or can be calibrated to remain) statistically consistent with observations (reliable). An ensemble forecast system that is statistically consistent with observations is often called a perfect ensemble in a sense of perfect reliability. An important property of a perfectly reliable ensemble is that the verifying analysis (or observations)

should be statistically indistinguishable from the forecast members. Most of the verification tools specifically developed for and applied to ensemble forecasts are designed to evaluate the statistical consistency of such forecasts. These additional measures of reliability, as we will see below, can reveal considerably more detail as to the nature and causes of statistically inconsistent behavior of ensemble-based probabilistic forecasts than the reliability diagram (Section 7.3.1) or the single measure of the reliability component of the Brier score (Section 7.3.2). By revealing the weak points of ensemble forecast systems, the ensemble-based measures provide important information for the developers of such systems that can eventually lead to improved probability forecasts.

7.6.1 Ensemble Mean Error and Spread

If the verifying analysis is statistically indistinguishable from the ensemble members, then its mean distance from the mean of the ensemble members (ensemble mean error) must equal the mean distance of the individual members from their mean (ensemble standard deviation or spread) – see Buizza (1997) and Stephenson and Doblas-Reyes (2000). Figure 7.6 compares the root mean square error of the ensemble mean forecast and the mean spread of the NCEP ensemble forecasts as a function of lead-time. Initially,

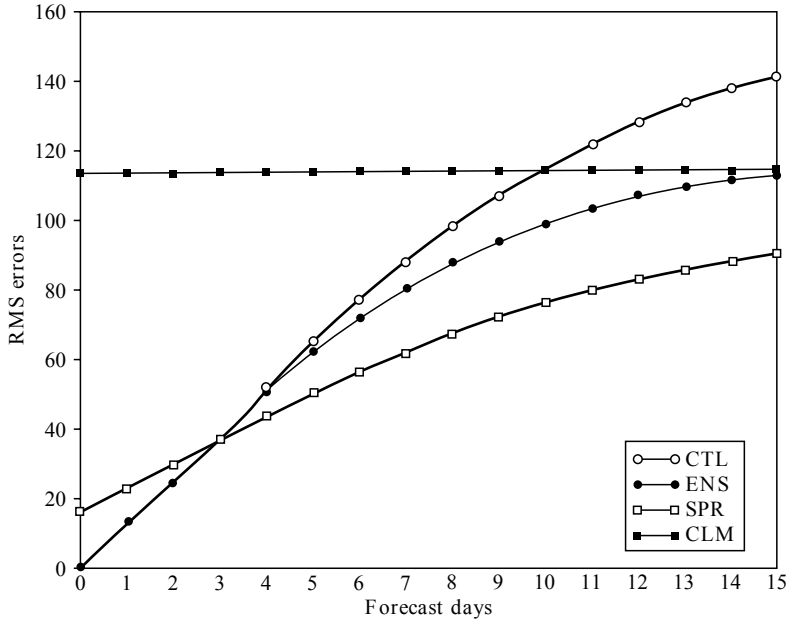


Figure 7.6 Root mean square error of 500 hPa geopotential height NCEP control (open circle), ensemble mean (full circle), and climate mean (full square) forecasts, along with ensemble spread (standard deviation of ensemble members around their mean, open square), as a function of lead-time, computed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

the ensemble spread is larger than the ensemble mean error, indicating a larger than desired initial spread. The growth of ensemble spread, however, is less than that of the error, which then leads to insufficient spread at later lead-times. This is typical behavior in current ensemble forecast systems that often tend to underestimate ensemble spread due to not accounting for all possible sources of model related uncertainty (e.g., structural uncertainty caused by the model parameterizations being incorrect).

Since for perfectly reliable forecast systems the spread of the ensemble forecasts is equal to the error in the ensemble mean, for such systems the spread can also be considered as a measure of resolution (and therefore forecast skill in general). For example, an ensemble with a lower average ensemble spread can more efficiently separate likely and unlikely events from one another (and so has more information content). It is worth mentioning that the skill of the ensemble mean forecast is often compared to that of the single *control forecast* obtained by starting with the best initial conditions (control analysis). Once non-linearity becomes pronounced, the mean of an ensemble that properly describes the case-dependent forecast uncertainty is able to provide a better estimate of the future state of the system than the control forecast (see Toth and Kalnay 1997). In a good ensemble forecasting system, the ensemble mean error should therefore be equal to or less than the error of the control forecast (see Fig. 7.6). It follows that in a reliable ensemble the spread of the ensemble members around the mean will be less than that around the control forecast.

7.6.2 Equal Likelihood Frequency Plot

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

Whether all ensemble members are equally likely or not is in itself neither a desirable nor an undesirable property of an ensemble prediction system. When ensemble forecasts are used to define forecast probabilities, however, one must know if all ensemble members can be treated in an indistinguishable fashion. This can be tested by generating a frequency plot showing the number of cases (accumulated over space and time) when each member was the forecast closest to the verifying diagnostic (see Zhu *et al.* 1996). Information from such a frequency plot can be useful as to

how the various ensemble members must be used in defining forecast probability values. Equal frequencies indicate that all ensemble members are equally likely and can be considered as independent realizations of the same random process, indicating that the simple procedure used in Section 7.3.1 for converting ensemble forecasts into probabilistic information is applicable.

Figure 7.7 compares the frequency of NCEP ensemble forecasts being closest to the verifying analysis value, averaged for 10 perturbed ensemble members, with that of an equal and a higher horizontal spatial resolution unperturbed control forecast as a function of lead-time. Note first in Fig. 7.7 that the higher resolution control forecast, due to its ability to better represent nature, has an advantage over the lower resolution members of the ensemble. This advantage, however, is rather limited. As for the low resolution control forecast, at short lead-times, when the spread of the ensemble around the control forecast is too large (see Fig. 7.6), it is somewhat more likely to be closest to the verifying analysis. At longer lead-times, when the spread of the NCEP ensemble becomes underestimated due to under-representation of model related uncertainty, the control forecast becomes less likely to verify best. When the spread is too low at longer lead-times, the ensemble members are clustered too densely and the verifying analysis often lies outside of the cloud of the ensemble. In this situation,

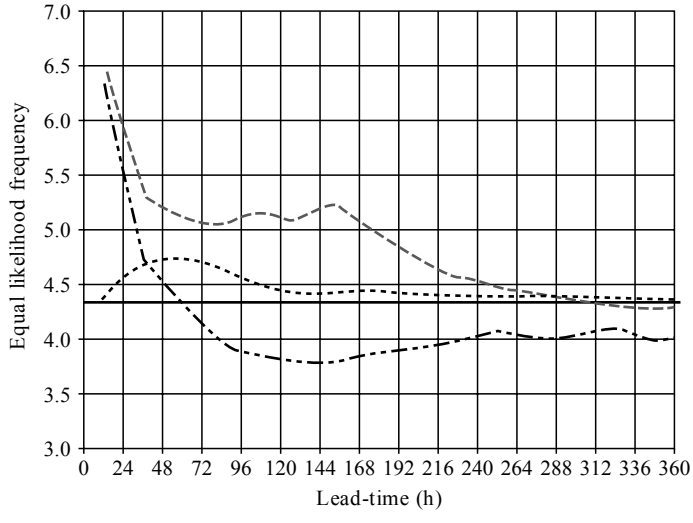


Figure 7.7 Equal likelihood diagram, showing the percentage of time when the NCEP high (dashed) and equivalent resolution control (dash-dotted), and any one of the 10-member 0000 UTC perturbed ensemble 500 hPa geopotential height forecasts (dotted) verify best out of a 23-member ensemble (of which the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

since the control forecast is more likely to be near the center of the ensemble cloud than the perturbed members, a randomly chosen perturbed forecast has a higher chance of being closest to the verifying observation than the control. The opposite is true at short lead-times characterized by too large spread. The flat equal likelihood values at intermediate lead-times (i.e., the 48-h perturbed and equal resolution control forecasts have the same likelihood in Fig. 7.7) thus are indicative of proper ensemble spread (cf. Fig. 7.6), and hence good reliability.

7.6.3 Analysis Rank Histogram

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

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

7.6.4 Multivariate Statistics

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

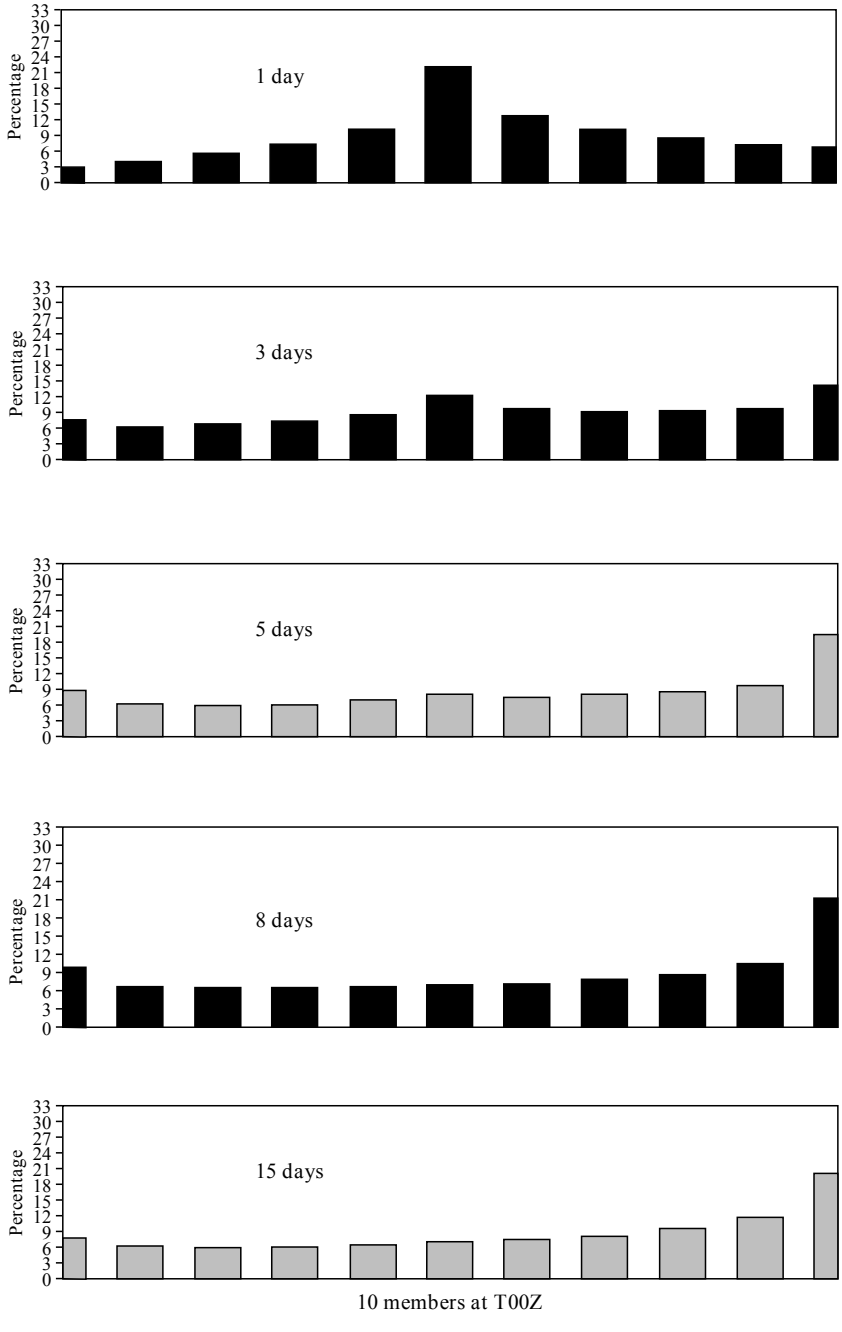


Figure 7.8 Analysis rank histogram for a 10-member 0000 UTC NCEP ensemble of 500 hPa geopotential height forecasts over the Northern Hemisphere extratropics during December 2001–February 2002

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

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

7.6.5 Time Consistency Histogram

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

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

7.7 LIMITATIONS OF PROBABILITY AND ENSEMBLE FORECAST VERIFICATION

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

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

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

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

7.8 CONCLUDING REMARKS

Reliability and resolution are the two main attributes of forecast systems in general. For probabilistic forecasts, reliability is defined as the statistical consistency between forecast probability values and the corresponding

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

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

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

The relationships among the different verification scores such as the ranked probability skill score, the relative operating characteristic, and the information content are not clearly understood. Which scores are best suited for certain applications is not clear either. It is important to mention in this respect that the value of forecasts can also be assessed in the context of their use by society. Some of the verification scores discussed above have a clear link with the economic value of forecasts. For example, the resolution component of the BSS, and the ROC-area, two measures of the resolution of forecast systems, are equivalent to the economic value of forecasts under certain assumptions (Murphy 1966; Richardson 2000). The economic value of forecasts has such significance that an entire chapter in this book (Chapter 8) is devoted to this topic.

8 Economic Value and Skill

DAVID S. RICHARDSON
UK Meteorological Office, Bracknell, UK

8.1 INTRODUCTION

Three types of forecast 'goodness' were identified by Murphy (1993): *consistency, quality* and *value* (Chapter 1, Section 1.4). Consistency and quality have been the main focus of much of this book on forecast verification. However, this chapter will now consider economic value (or utility) and its relationship to quality measures such as forecast skill. Page limitations prevent us from giving a comprehensive review of the economic value of forecast information in a single chapter. Rather, the aim is to introduce the basic concepts of the value of forecast information to users and to explore some of the fundamental implications for forecast verification.

The main aspects of deriving economic benefits from forecasts are incorporated in so-called decision-analytic models (e.g. Murphy 1977; Katz and Murphy 1997b). A decision maker (forecast user) has a number of alternative courses of action to choose from, and the choice is to some extent influenced by the forecast. Each action has an associated cost and leads to an economic benefit or loss depending on the weather that occurs. The task of the decision maker is to choose the appropriate action that will minimise the expected loss (or maximise the expected benefit). In this chapter, we focus on the simplest of these economic decision models known as the (static) cost–loss model (Ångström 1922; Thompson 1952; Murphy 1977; Liljas and Murphy 1994).

A simple decision model is introduced in Section 8.2. It is applied first to deterministic forecasts; then it is used to show how probability forecasts can be used in the decision-making process. The benefit of probability forecasts over deterministic forecasts is then assessed.

A particularly useful feature of this simple model is that it provides a link between user value on the one hand and more standard verification measures on the other. In Section 8.3, the relationship between value and the relative operating characteristic (ROC) is explored, and in Section 8.4 a measure of overall value (over all users) and the link with the Brier score are

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

considered. The contrasting effects of ensemble size on the ROC area and Brier skill scores are examined in Section 8.5, and the differences in behaviour of the two scores is interpreted in terms of user value.

The chapter will be illustrated with examples taken from the operational Ensemble Prediction System (EPS) of the European Centre for Medium-Range Weather Forecasts (ECMWF) (Palmer *et al.* 1993; Molteni *et al.* 1996; Buizza *et al.* 2000). More detailed evaluations of the economic value of these forecasts in specific situations are given by Taylor and Buizza (2001) and Hoffschildt *et al.* (1999). Further recent examples using the cost–loss model to evaluate ensemble forecasts can be found in Richardson (2000), Zhu *et al.* (2002) and Palmer *et al.* (2000).

8.2 THE COST/LOSS RATIO DECISION MODEL

Consider a decision maker who is sensitive to a specific adverse weather event A. For example, A may be 'the occurrence of ice on the road' or 'more than 20 mm of rain in 24 h'. We assume that the decision maker will incur some loss L if this bad weather event occurs and no action is taken. The forecast will be useful only if the decision maker can take some action to prevent or limit the expected loss due to bad weather. We take the simplest possible situation where the user has just two alternative courses of action – either do nothing (carry on as normal) or take some form of protective action to prevent loss. This action will cost an amount C (additional to the normal expenditure). Examples could be gritting roads to prevent the formation of ice, or arranging to move an outdoor event inside if heavy rain is expected.

There are four possible combinations of action and occurrence with the net cost depending on what happened and what action was taken. If action is taken, then the cost is C irrespective of the outcome. However, if action is not taken, the expense depends on the actual weather that occurs: if event A does not occur there is no cost, but if event A does occur, then there is a loss L. The situation is summarised in Table 8.1, sometimes known as the *expense matrix* (payoff matrix). Note that all expenses are taken relative

Table 8.1 The expense matrix: costs and losses for different outcomes in the simple cost–loss decision model

Action taken	Event occurs	
	Yes	No
Yes	C	C
No	L	0

to the 'normal event' of no action and no adverse weather, and that we have (once again) taken the simplest scenario where the protective action completely prevents the potential loss. This may appear to be an oversimplification of the general case of different expenses in each cell of the expense matrix (Section 3.3.4), but the expression for economic value developed in the following is essentially the same for both situations (Richardson 2000). Table 8.1 captures the salient features of the more general cost–loss model.

Assume the decision maker aims to minimise the average long-term loss by taking the appropriate action on each occasion. We assume for simplicity that the weather is the only factor influencing the decision. To set a baseline for economic value, we first consider the reference strategies available in the absence of forecast information. If there is no forecast information available, then there are only two possible choices: either always protect or never protect (we will ignore the third possible option of making decisions randomly). If the decision maker always protects, the cost will be C on every occasion, so the average expense is

$$E_{\text{always}} = C \tag{8.1}$$

On the other hand, if action is never taken, there will be some occasions with no expense and other occasions with loss L. Over a large number of cases, let s be the fraction of occasions when event A occurred (the climatological base rate; see Chapter 3, Section 3.2.1). The average expense is then given by

$$E_{\text{never}} = sL \tag{8.2}$$

In general, E_{always} and E_{never} are not equal, and so to minimise losses the decision maker should choose the strategy with the smallest average expense. The optimal strategy (decision rule) is to always take protective action if $E_{\text{always}} < E_{\text{never}}$ and never take protective action if $E_{\text{always}} > E_{\text{never}}$. For this optimal strategy, the mean expense can be written as

$$E_{\text{climate}} = \min(C,\ sL) \tag{8.3}$$

This will be referred to as the *climate expense* because the user needs to know the climatological base rate probability (s) of the event in order to know whether to always protect or not. For events having base rates greater than the cost/loss ratio C/L the decision maker should always take protective action (assume the event is going to happen), whereas for rarer events with base rates less than C/L, the decision maker should never take protective action (assume the event will not happen). For events with a base rate equal to the cost/loss ratio, the expense is the same for both strategies. The climate expense provides a baseline for our definition of forecast value.

It has been assumed that the base rate used to determine the choice of default action will stay the same in the future (stationarity assumption). Given perfect knowledge of the future, the decision maker would need to take action only when the event was going to occur. The mean expense would then be

$$E_{\text{perfect}} = sC \tag{8.4}$$

The fact that this is greater than zero for non-zero base rates reminds us that some expense is unavoidable. The aim of using forecast information in the decision process is to optimally reduce the climate expense from E_{climate} towards E_{perfect}. However, the mean expense can never be completely reduced to zero unless either the base rate is zero or there is no cost for taking preventative action.

The value, V, of a forecast system can be defined as the reduction in mean expense relative to the reduction that would be obtained by having access to perfect forecasts:

$$V = \frac{E_{\text{climate}} - E_{\text{forecast}}}{E_{\text{climate}} - E_{\text{perfect}}} \tag{8.5}$$

This definition is equivalent to the standard definition of a skill score with climatology as the reference (Chapter 2, Section 2.7). Similar to results for equitable skill scores, zero value will be obtained by constantly forecasting either the event if $s > C/L$ or non-event if $s < C/L$. A maximum value of one will be obtained for systems that perfectly forecast future events. When V is greater than zero, the decision maker will gain some economic benefit by using the forecast information in addition to using the base rate information.

It should be noted that the value is sometimes defined in an absolute sense as the saving $S = E_{\text{climate}} - E_{\text{forecast}}$ (e.g. Thornes and Stephenson 2001). This definition has some advantage for a specific decision maker in that it gives a direct measure of the amount of financial benefit (e.g. in units of currency). However, the definition of value used here, sometimes referred to as *relative value*, is more useful in a general context where we wish to compare the forecast value for a range of different users. The perfect saving $S_{\text{perfect}} = E_{\text{climate}} - E_{\text{perfect}}$ gives an absolute upper bound on how much can be saved.

8.2.1 Value of a Deterministic Binary Forecast System

A deterministic binary forecast system gives a simple yes/no prediction of whether the weather event A will occur or not. The decision maker takes protective action when the forecast is for A to occur and does nothing otherwise. The value of such forecasts over a set of previous events can be

estimated using the cell counts in a contingency table accumulated over previous events (Table 8.2; see also Chapter 3).

The sample mean expense using the forecasts is easily obtained by multiplying the expenses in Table 8.1 by the corresponding relative frequencies in Table 8.2:

$$E_{\text{forecast}} = \frac{a}{n}C + \frac{b}{n}C + \frac{c}{n}L \tag{8.6}$$

It is convenient to re-express this expense in terms of sample estimates of the likelihood–base rate conditional probabilities: hit rate $H = a/(a+c)$, false alarm rate $F = b/(b+d)$ and base rate $s = (a+c)/n$. This yields

$$E_{\text{forecast}} = F(1-s)C - Hs(L-C) + sL. \tag{8.7}$$

Substitution into Eq. (8.5) then gives an expression for the relative value of the forecasts:

$$V = \frac{\min(\alpha,\ s) - F(1-s)\alpha + Hs(1-\alpha) - s}{\min(\alpha,\ s) - s\alpha} \tag{8.8}$$

where $\alpha = C/L$ is the specific user's *cost/loss ratio*. Eq. (8.8) shows that value depends not only on the quality of the system (H and F), but also on the observed base rate of the event (s) and the user's cost/loss ratio (α). The introduction of the forecast user into the verification process brings an extra dimension into the problem: the value of forecast to the user is very much user-dependent. However, it is only the cost/loss ratio that is important rather than the individual values of C and L. For simplicity, we assumed that the cost C provides full protection against the potential loss L (Table 8.1). It should be emphasised that the general case, allowing different costs or losses in each cell of the expense matrix, still leads to Eq. (8.8), with 'C/L' now representing a more general cost–loss ratio (Richardson 2000). For a given event, the base rate s is fixed; and for a particular forecast system H and F are also fixed, and so V is then only a function of the user's cost/loss ratio. We can show this variation of value with user by plotting V against

Table 8.2 Contingency table showing counts for a single event

Event forecast	Event observed		
	Yes	No	Marginal totals
Yes	a	b	$a+b$
No	c	d	$c+d$
Marginal totals	$a+c=ns$	$b+d=n(1-s)$	$a+b+c+d=n$

C/L. Since there would be no point in taking action if the protection cost C is greater than the potential loss L, we need only consider the range $0 < C/L < 1$.

Figure 8.1 shows the value of ECMWF deterministic forecasts of European precipitation for three different amount thresholds over the whole range of possible cost/loss ratios. The control forecast is a single forecast made using the best estimate of initial conditions. An ensemble of other forecasts is generated by perturbing the initial conditions of the control forecast – the value of the whole ensemble of forecasts will be discussed below. It is clear that the value of the forecasts varies considerably for classes of hypothetical users with different cost/loss ratios. For example, while some users gain up to 50% value for the 1 mm event, users with cost/loss ratios greater than 0.7 or less than 0.2 will find no benefit; for these ranges of cost/loss ratio the forecasts are less useful than climatology (negative values are not shown on the figure). The shape of the curves is similar for the different thresholds, although the maximum value is shifted towards lower cost/loss ratios for the rarer higher precipitation events. In other words, users with small cost/loss ratios benefit the most from forecasts of rare events, for example, users who incur especially large losses when the rare event happens and no protection has been taken.

It is straightforward to determine the location and magnitude of the maximum value from the expression for V in Eq. (8.8). When $\alpha < s$, Eq.

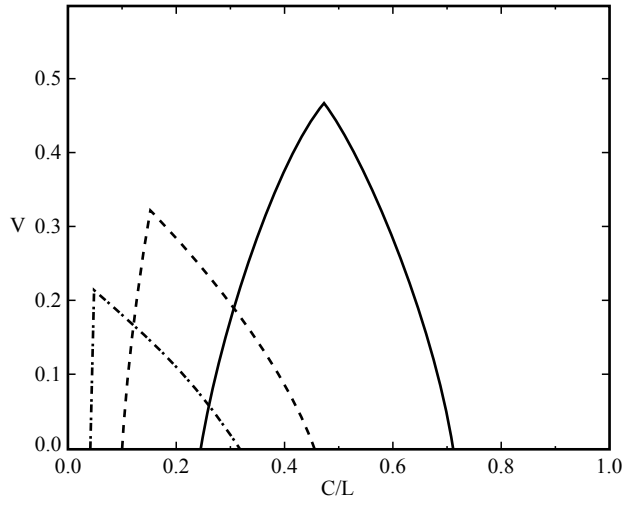


Figure 8.1 Value of ECMWF EPS control deterministic forecast of 24-h total precipitation over Europe at day 5 for winter 1999/2000. Curves show value V as a function of user cost–loss ratio C/L for three different events, defined as precipitation exceeding 1 mm (solid line), 5 mm (dashed line) and 10 mm (chain-dashed line)

(8.8) becomes

$$V = (1 - F) - \left(\frac{s}{1-s}\right)\left(\frac{1-\alpha}{\alpha}\right)(1 - H) \tag{8.9}$$

and so the value increases for increasing cost/loss ratios. When $\alpha > s$, then

$$V = H - \left(\frac{1-s}{s}\right)\left(\frac{\alpha}{1-\alpha}\right)F \tag{8.10}$$

and the value decreases for increasing cost/loss ratios. Hence, the maximum value will always occur when $\alpha = s$ (i.e. when the cost/loss ratio equals the base rate). At this point, the expense of taking either of the climatological options (always or never protect) is the same: climatology does not help the decision maker and the forecast offers the greatest benefit. As the cost/loss ratio approaches zero or one, the climatological options become harder to beat.

The maximum value obtained by substituting $\alpha = s$ in Eq. (8.8) is given by

$$V_{\max} = H - F \tag{8.11}$$

This is identical to the Peirce skill score described in Chapter 3 (Section 3.2.2). Hence, *maximum* economic value is related to forecast skill, and the Peirce skill score can be interpreted as a measure of *potential* forecast value as well as forecast quality. Two of the three aspects of goodness identified by Murphy (1993) are therefore related. However, the Peirce skill score only gives information about the maximum achievable value, and gives no information about the value for a specific user having a cost/loss ratio different to the base rate. Strictly, the maximum value should be the absolute value of the Peirce skill score since forecasts with negative Peirce skill scores can always be recalibrated (by relabelling the forecasted event as non-event, and vice versa) to have positive skill.

We can also determine the range of cost/loss ratios over which there is positive value – i.e. the class of users who can derive benefit from using the forecasts. From Eqs. (8.9) and (8.10), V is positive when

$$\frac{1-H}{1-F} < \left(\frac{\alpha}{1-\alpha}\right)\left(\frac{1-s}{s}\right) < \frac{H}{F} \tag{8.12}$$

This is more conveniently expressed as a range for α using the cell counts in Table 8.2 instead of H and F:

$$\frac{c}{c+d} < \alpha < \frac{a}{a+b}. \tag{8.13}$$

The lower and upper limits for α are sample estimates of calibration–refinement probabilities obtained by conditioning on the forecasts, namely the upper limit $a/(a+b)$ is one minus the false alarm ratio (Chapter 3, Section 3.2.2). Therefore, the range of cost/loss ratios for which the forecasts have positive value is determined by the conditional probabilities of the event occurring given the forecast. There is no value for users with cost/loss ratios sufficiently close to either zero or one unless the forecasts are almost perfect. The condition for the system to have value for at least one user (i.e. non-zero range in Eq. (8.13)) is that the event is more likely to occur following a 'yes' forecast than following a 'no' forecast.

The difference between the upper and lower limits of cost/loss ratios for which the forecasts have value is equal to $(ad - bc)/(a+b)(c+d)$, which is the Clayton skill score – a score similar to the Peirce skill score except conditioned on the forecasts rather than the observations (H. Brooks and M. Wandishin, personal communication). The ratio of the upper and lower limits in Eq. (8.13) is equal to $(\theta + a/b)/(1 + a/b)$, where $\theta = ad/bc$ is the odds ratio (Chapter 3, Section 3.2.2). Therefore, to have positive range and so value for at least one user, the odds ratio must exceed one (D.B. Stephenson, personal communication). This sets the absolute minimum standard for a forecast system to have any practical value, and provides another example of how a quality measure (the odds ratio) provides a necessary condition for extracting value from forecasts (Stephenson 2000). The Peirce skill score and the Clayton skill score quantify two of the most important features of the value curves (i.e. the maximum value and the range of cost/loss ratios that can gain value from the forecasts), and therefore are useful measures for quantifying both skill and value of binary forecasts.

8.2.2 Probability Forecasts

A potential stumbling block to the use of probability forecasts, particularly for those more familiar with deterministic forecasts, is the perception that probabilities have no place in the real world where hard yes/no decisions are required. The decision framework of the cost/loss model provides a simple illustration of the importance of the informed use of probabilities in maximising the value of forecast information.

When presented with forecast information in the form of probabilities, the question facing the decision maker is how high does the probability need to be before the threat is great enough to warrant protective action being taken. The decision maker needs to set a *threshold probability* p_t so that action is taken only when the forecast probability exceeds p_t. In this way, the probability forecast is converted to a deterministic binary forecast that can be evaluated using the standard methods presented in Chapter 3. By varying the decision threshold probability p_t over the range zero to one, a sequence of hit rates and false alarm rate pairs $(F(p_t),\ H(p_t))$ can be plotted that

traces out the ROC for the system (Chapter 3, Section 3.4.2). Different value curves such as those shown in Fig. 8.1 can be plotted for each distinct point on the ROC diagram (i.e. each value of threshold probability).

Figure 8.2 shows a selection of value curves obtained for an ensemble of 50 forecasts – see Chapter 7 for details of how probability forecasts can be obtained from an ensemble of forecasts. A different value curve can be produced for each of the 50 possible probability threshold values $p_t = \{1/51,\ 2/51, \ldots,\ 50/51\}$. For visual clarity, curves are displayed only for the subset of probability decision thresholds $p_t = \{1/51,\ 5/51,\ 10/51,\ 15/51, \ldots,\ 50/51\}$ corresponding to 1, 5, 10, 15, ..., 50 members. The envelope curve (heavy solid line) shows the optimum maximum value of the forecasting system, obtained when each user chooses the probability threshold that maximises the value for their specific cost/loss ratio. The envelope curve is never less than any of the individual curves, which shows that economic value for any particular user is maximised by choosing the optimal probability threshold for their particular cost/loss value. Therefore, no single threshold probability (i.e. a deterministic forecast) will be optimal for a range of users with different cost/loss ratios – this is a strong motivation for providing probability rather than deterministic forecasts.

The importance of choosing the correct probability threshold for each user is shown in Fig. 8.3. A user with $C/L = 0.2$ would have relatively large potential losses and would therefore benefit by taking action at a relatively low probability. But a different user, with $C/L = 0.8$ say (relatively high

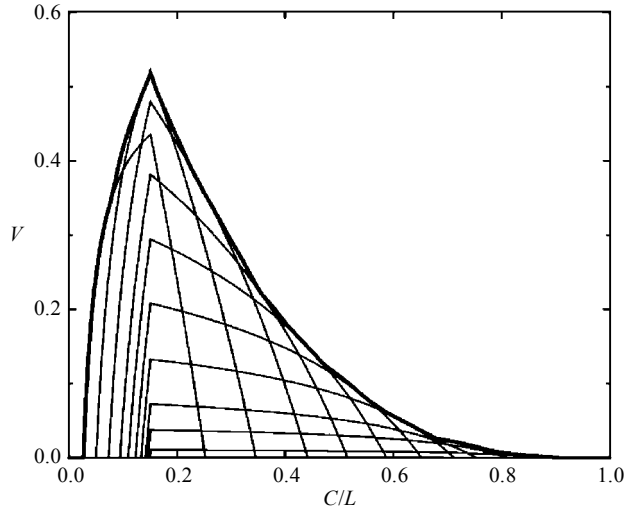


Figure 8.2 Value of ECMWF EPS probability forecasts of 24-h total precipitation exceeding 1 mm over Europe at day 5 for winter 1999/2000. Thin curves show value V as a function of cost–loss ratio C/L for different choices of probability threshold ($p_t = 0.02, 0.1, 0.2, \ldots$); heavy solid line shows the envelope curve of optimal value

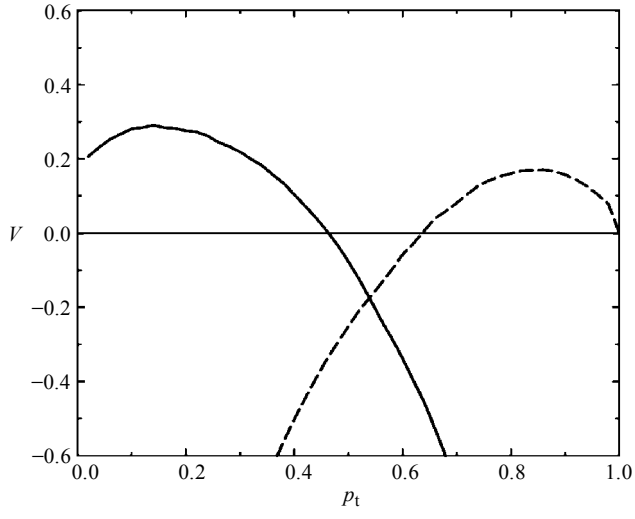


Figure 8.3 Variation of value V with probability threshold p_t for the EPS probability forecasts of Fig. 8.2 for users with $C/L = 0.2$ (solid line) and $C/L = 0.8$ (dashed line)

costs) would wait until the event was more certain before committing to expensive protective action. If either user took the decision threshold appropriate to the other user, value would be reduced and in this case even negative. The naively sensible choice of $p_t = 0.5$ would give no value to either user!

The main advantage of probability forecasts is that different probability thresholds are appropriate for different users. Deducing the appropriate probability at which to act is straightforward and follows similar reasoning to that at the beginning of the chapter for deciding whether to act or not with only climate information. Consider only those occasions where the forecast probability $\hat{p}$ is one particular value, $\hat{p} = q$, say. Should the user act or not? The average expense of taking action is of course

$$E_{yes}(q) = C \tag{8.14}$$

The mean expense of not acting, averaged over all cases with $\hat{p} = q$, will be

$$E_{no}(q) = p'(q)L \tag{8.15}$$

where $p'(q)$ is the fraction of times the event occurs when the forecast probability is q. Hence, users with $\alpha < p'(q)$ should take action, while those users with larger α should not. In general, users should act if $p'(\hat{p})$ is greater than their cost/loss ratio.

For reliable forecasts (Chapter 7), $p'(\hat{p}) = \hat{p}$, so the optimal strategy is to act if $\hat{p} > \alpha$ and not act if $\hat{p} < \alpha$. In other words, the appropriate probability threshold for a given user is $p_t = \alpha$ i.e. their own cost/loss ratio. If the

forecasts are not reliable, then the threshold should be adjusted so that $p'(p_t) = \alpha$. The calibration procedure discussed in Chapter 7 makes this adjustment to the forecast probabilities so that users of calibrated forecasts can then use the threshold probability equal to their cost/loss ratio.

The value curves shown in the rest of this chapter all plot the envelope curve of optimal value, i.e. they assume the forecasts are correctly calibrated. In this sense they are a measure of potential value rather than actual value, although the difference is small for reliable forecasting systems that are reasonably well calibrated.

8.2.3 Comparison of Deterministic and Probabilistic Binary Forecasts

Figure 8.4 shows the value obtained for ECMWF probability forecasts at four different precipitation thresholds compared to the value obtained for the control and ensemble mean deterministic forecasts. Since maximum value occurs at the climate frequency, the curves for the more extreme

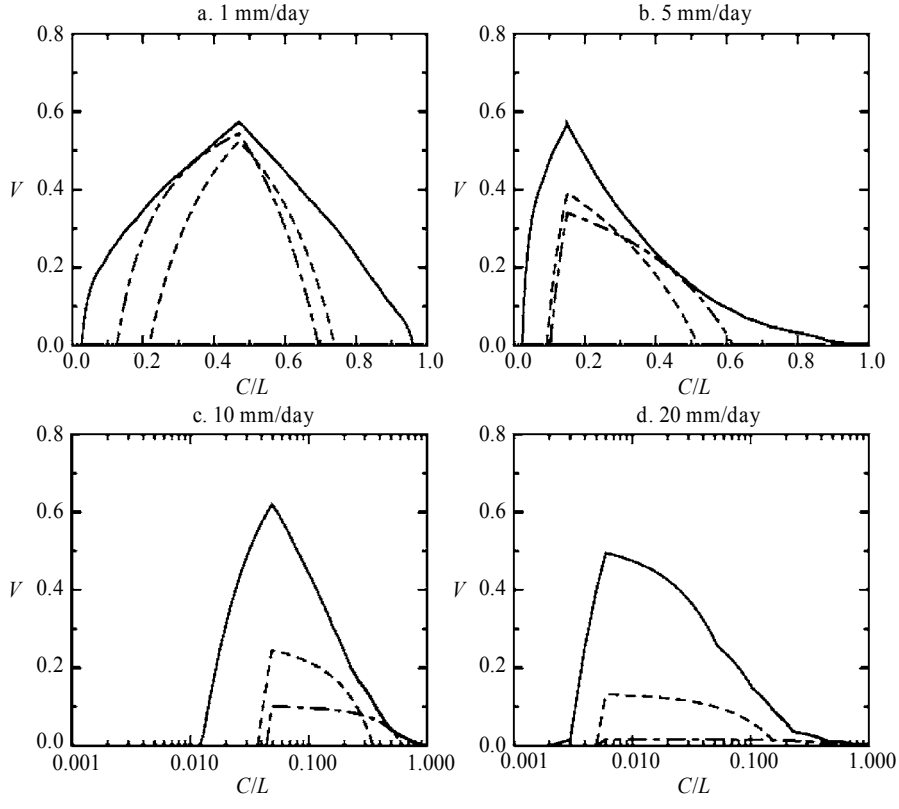


Figure 8.4 Value of ECMWF EPS forecasts of 24-h total precipitation exceeding 1, 5, 10 and 20 mm over Europe at day 5 for winter 1999/2000. Curves show value *V* as a function of cost–loss ratio *C*/*L* for EPS probability forecasts (solid line) and for the deterministic control (dashed line) and ensemble mean (chain-dashed line) forecasts

events are concentrated around lower values of cost/loss ratio; to see the differences between the curves more clearly the lower two plots are plotted with a logarithmic axis for the cost/loss ratio. The additional value (benefit) of the probability forecasts is clear for each precipitation event. For the 1 mm event (Fig. 8.4a), maximum value is similar for all three forecasts, but the probability forecasts have greater value for a wider range of users. For the higher precipitation thresholds, there is increased benefit using the probability forecasts compared to using the deterministic forecasts.

Comparing the ensemble mean and control forecasts, users with low cost/loss ratios will gain more from the control forecast while large cost/loss users will prefer the ensemble mean (although for the 1 mm threshold this is reversed). The ensemble mean is an average field and is therefore less likely to contain extreme values than the individual members. If the ensemble mean forecasts heavy precipitation, it is likely that the majority of members also forecast large amounts. In other words, the ensemble mean forecasts implies relatively high probability – which is more optimal for users with high cost/loss ratios. In contrast, the single control forecasts, taken alone, must be treated with more caution – it is a relatively lower probability indication of heavy precipitation and therefore more likely to benefit low cost/loss ratio users. The generally low value for the higher precipitation amounts for the ensemble mean is another indication of the difficulty for an averaged field to produce intense precipitation events.

The ensemble mean value curve coincides with the probabilistic forecast value curve at some point where users will find the ensemble mean has equal value to the probability forecasts. However, it is apparent that for almost all cost/loss ratios, the probability forecasts have considerably greater value than the ensemble mean deterministic forecast. The biggest advantage of probability forecasts is their ability to provide this more valuable information. The added benefit of the deterministic ensemble mean forecast is less apparent and in some situations may even be less valuable than that of the control forecast.

While for some users the ensemble mean forecast has the same value as the probability forecasts, the control value curve is always below the probability forecast value curve and therefore all users will gain greater benefit by using the probability forecasts. This is related to the control forecast (F,H) point lying below the probability forecast ROC curve on the ROC diagram (Chapter 3, Section 3.4), whereas the ensemble mean (F,H) point lies on the probability forecast ROC curve. The link between value and the ROC is considered in more detail in the following section.

8.3 THE RELATIONSHIP BETWEEN VALUE AND THE ROC

There are clearly links between the economic value analysis of the cost/loss model and the ROC analysis discussed in Section 3.4 of Chapter 3. Value is

a function of both H and F, and for a probability forecast the appropriate choice of threshold probability p_t is important. The ROC curve is a two-dimensional plot of $(F(p_t),H(p_t))$ for each probability threshold. The value curve can be calculated from the ROC data as long as the base rate s for the event is known. Note that in this case of probability forecasts, the base rate of the event is independent of the choice of threshold. One benefit of the value approach is that it shows how different aspects of the ROC relate to the economic value of the forecasts. This section explores the links between the two approaches.

Figure 8.5 shows the value curves for the 5 mm precipitation event of Fig. 8.4b plotted alongside the corresponding ROC curves. One additional curve is shown on the plots of Fig. 8.5 – the ensemble mean 3-day ahead forecast, which is an example of a forecast with higher intrinsic quality than the forecasts used for the other plots. The deterministic control and ensemble mean forecasts are each represented by single points on the ROC diagram. In Fig. 8.5a, these points are shown connected with straight lines to the corners (0,0) and (1,1); the interpretation of these straight-line segments will be discussed below. The ROC curve for the probability forecasts consists of the set of points (H_k,F_k) determined by taking action when at least k ensemble members predict the event. There are m points on this curve for an m-member ensemble. The points are joined by straight-line segments and connected to the corners (0,0) and (1,1).

As remarked in the previous section, the value of the control forecast is less than that of the probability forecasts for users with all possible cost/loss ratios. Correspondingly, the control point on the ROC diagram falls below the probability forecasts ROC curve, again indicating that the control forecast performance is worse than that of the probability forecasts. The benefit of the probability forecasts over the deterministic control forecast is unequivocal since all users receive greater value from the probability

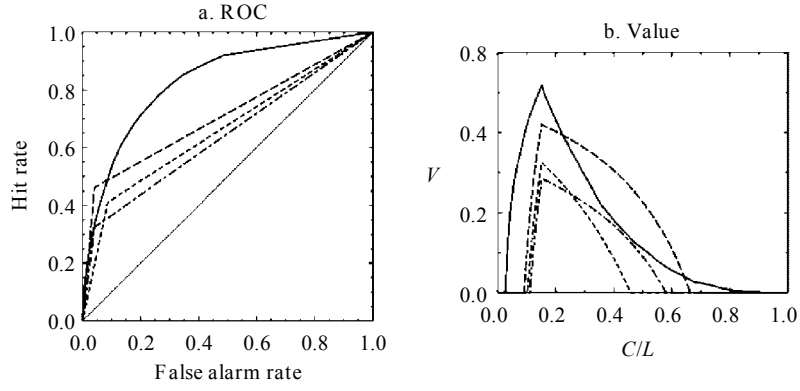


Figure 8.5 ROC and value plots for the 5 mm precipitation event of Fig. 8.4. Solid line – EPS probability forecasts; dashed line – control forecast; chain-dashed line – ensemble mean; long dashed line – ensemble mean 3-day ahead forecast

forecasts than from the control forecast: the probability forecasts are sufficient for the deterministic control forecast in the sense described in Section 3.3.3 of Chapter 3. However, it is worth noting that some users will not benefit from either system and that the differences will be more significant for some users than for others.

The ensemble mean point lies on the probability forecast ROC curve. This suggests that the basic system performance is the same. However, the benefit of the probability forecasts depends on the user: on the value curve it can be seen that the value of the two configurations is the same for some users although for the majority of cost/loss values the probability forecasts are better because of the flexibility allowed by the range of threshold probabilities.

The 3-day ahead ensemble mean forecast point lies above the 5-day ahead probability forecast ROC curve – 3-day ahead deterministic forecasts have intrinsically higher quality than 5-day ahead probability forecasts. This translates into higher value for at least some cost/loss ratios. Nevertheless, there may still be a significant proportion of users for whom the probability forecasts have greater value. This is an example where neither system can be deemed sufficient for the other; the relative benefit again is dependent on the user.

The most common summary skill measure for the ROC diagram is the area under the ROC curve (Chapter 3, Section 3.4.4). The area A is the fraction of the unit square below the ROC curve that can be most simply (under-)estimated by summing the area below the points on the curve using the trapezium rule. For a deterministic forecast, the single point can be joined directly to the corners (0,0) and (1,1) as shown in Fig. 8.5 and the area of two resulting trapeziums can then be summed. The area A will generally be smaller than the parametric A_z described in Section 3.4.4, but A and the connecting straight-line segments have useful interpretations. Since the diagonal $H = F$ represents no skill (Section 3.2.1), an area-based skill score can be defined as

$$\mathrm{ROCSS} = 2A - 1 \tag{8.16}$$

which varies between zero for points on the diagonal (no skill) to one for perfect forecasts. For a deterministic forecast, it is straightforward to show that $\mathrm{ROCSS} = H - F$, in other words the Peirce skill score that gives the maximum attainable value $V_{\max}$ (Eq. (8.11)). So the skill measure ROCSS based on the 'straight-line' area gives an estimate of the maximum value that can be obtained from the deterministic forecast.

For the probability forecasts, the maximum value over the set of probability thresholds (indexed by k) is given by

$$V_{\max} = \max_k (H_k - F_k) \tag{8.17}$$

On the ROC diagram this is the maximum distance (horizontal or vertical) between the ROC points and the diagonal line $H = F$. For the probability forecasts, ROCSS is larger than $V_{\max}$. In terms of usefulness, this reflects the greater benefit provided to a wider range of users by using different probability thresholds for different users. When more points are added to an ROC curve (by having a greater number of ensemble members/probability thresholds), the area will increase and so will the value for some users.

Neither measure, though, can be used to determine the value of a forecast system to a general user or group of users. For example, in Fig. 8.5 the control forecast has a larger ROCSS or maximum value (0.35) than the ensemble mean (0.28), but for $\alpha > 0.25$, the ensemble mean forecast is the more valuable forecast. Summary measures of forecast value are discussed further in the following two sections.

The straight-line segments joining the ROC points are also informative to the user. The slope of the line joining the deterministic ROC point (F,H) to the lower left corner (0,0) is simply H / F, while the slope of the connecting line to the upper right corner (1,1) is $(1 - H)/(1 - F)$. The ratio of the two slopes gives a simple geometric method for calculating the odds ratio. In Section 8.2, these two ratios were shown to determine the range of cost/loss ratios for which the forecasts have value (Eq. (8.12)). The steeper the slope of the line between (F,H) and (0,0), the more users with higher cost/loss ratios will benefit; and the shallower the slope of the line between (F,H) and (1,1), the more users with lower cost/loss ratios will benefit. In Fig. 8.5a, the line joining the origin to the ensemble mean ROC point is steeper than the corresponding line for the control forecast, and in Fig. 8.5b, it can be seen that more users with higher cost/loss ratios benefit from the ensemble mean forecast than from the control forecast. Conversely, the line between the control forecast ROC point and (1,1) is shallower than the equivalent ensemble mean forecast line and low cost/loss ratio users therefore benefit more from the control forecast.

For the probability forecasts, the slopes of the lines joining the first and last points to the top-right and bottom-left corners again determine the range of users for whom the system will have positive value. Similar reasoning shows that the slopes of the straight lines connecting the intermediate points indicate the range of cost/loss ratios for which each probability threshold is optimal. If $V_k(\alpha)$ is the value associated with the ROC point (F_k,H_k), then from Eq. (8.12), $V_k(\alpha) > V_{k+1}(\alpha)$ and $V_k(\alpha) > V_{k-1}(\alpha)$ when

$$\frac{H_{k-1} - H_k}{F_{k-1} - F_k} < \left(\frac{\alpha}{1-\alpha}\right)\left(\frac{1-s}{s}\right) < \frac{H_k - H_{k+1}}{F_k - F_{k+1}} \tag{8.18}$$

So the range of α is determined by the slopes of the lines joining (F_k,H_k) to the adjacent points. From the definition of the hit rate and false alarm rates,

the gradient of the line joining the two ROC points (F_{k-1},H_{k-1}) and (F_k,H_k) can be written as

$$\frac{H_k - H_{k+1}}{F_k - F_{k+1}} = \left(\frac{p'_k}{1-p'_k}\right)\left(\frac{1-s}{s}\right) \tag{8.19}$$

where p'_k is the observed frequency of the event given forecasts are in class k (k out of m members predict the event for the m-member ensemble forecasts). By comparing Eqs. (8.18) and (8.19), it can be seen that $V_k(\alpha)$ will be greater than $V_{k-1}(\alpha)$ and $V_{k+1}(\alpha)$ for values of α between p'_{k-1} and p'_k. For positively calibrated forecast systems, we expect p'_k to increase with p_k (the higher the forecast probability, the more likely it should be that the event occurs). On the ROC, this corresponds to the slope of the lines joining the ROC points increasing monotonically as k increases (moving from the upper-rightmost point towards the lower left corner); this is generally the case for a large enough sample of data, although ROCs generated from small data samples may be more variable. So long as the monotonicity holds, the limits for α given above hold not just for $V_k(\alpha)$ compared to $V_{k-1}(\alpha)$ and $V_{k+1}(\alpha)$, but extend to all points: the value V_k associated with forecast probability p_k (and the corresponding H_k and F_k) is optimal for users with $p'_k < \alpha < p'_{k-1}$.

In summary, the slopes of the first and last line segments joining the ROC points to the endpoints (0,0) and (1,1) give the range of cost/loss ratios for which value is positive. Extending the ROC by adding more points at either end will thus increase the range of users who will receive positive value, while adding additional points between the existing ones will benefit by providing finer resolution of probability threshold. Both these measures will help to increase the total area under the ROC curve, and hence overall value of the forecasting system.

The empirical ROCs presented in this section demonstrate the value that can be obtained with the available set of probability thresholds (up to 50 for the 50-member ensemble forecasts). Parameterising the ROC curves as in Section 3.4 of Chapter 3 can be used to demonstrate the potential value that would be achieved if all possible probability thresholds could be used (i.e. an infinite ensemble of forecasts). This could provide a useful method for estimating the benefit that could be obtained by using a larger ensemble of forecasts (Richardson 2000).

8.4 OVERALL VALUE AND THE BRIER SKILL SCORE

The previous section explored the relationship between value and the ROC. Maximum value is easily deduced from the ROC, as is the range of users for whom value will be positive. ROC area increases as more points are added, consistent with the increase in value to a wider range of users. However,

forecast value can vary greatly between users and there is no simple relationship between any single overall measure (such as ROCSS or $V_{\max}$) and the value to specific users.

Nevertheless, it is often desirable to have an overall summary measure of performance. The specification of such a measure is the subject of this section. We will find that the familiar Brier skill score can be interpreted as one such measure, given a particular distribution of users. The implication of this will be explored both here and in the next section.

If we knew the costs and losses for every user, we could calculate the total savings over all users and produce a measure of overall value. This would then provide a representative skill measure based on the overall benefit of the forecast system.

Since we do not know the distribution of users appropriate to a given event, and this may vary between events, we use an arbitrary distribution of users. Assume the probability distribution of cost/loss ratios for users is given by $u(\alpha)$. We can then derive a measure of overall value based on the total saving made by all users. For this general derivation we assume that a user will take the probability forecasts at face value so that each user will take action when the forecast probability $\hat{p}$ is greater than their cost/loss ratio α.

Consider the occasions when the forecast probability is equal to a given value $\hat{p} = q$. Users with cost/loss ratio $\alpha < q$ will take action and hence incur cost α (per unit loss). All other users will not act and will incur a loss (assumed without loss of generality to be $L = 1$) when the event occurs. The total mean expense for all users for this particular forecast probability $\hat{p} = q$ is then given by

$$T(q) = \int_0^q u(\alpha)\alpha \, \mathrm{d}\alpha + p'(q) \int_q^1 u(\alpha) \, \mathrm{d}\alpha = p'(q) + \int_0^q u(\alpha)(\alpha - p'(q)) \, \mathrm{d}\alpha \qquad (8.20)$$

where $p'(q)$ is the observed frequency of the observed event for cases when the forecast probability $\hat{p} = q$. The total expense is obtained by taking the expectation over all possible forecast probabilities. For an ensemble of m forecasts, this amounts to a summation over all classes of probability, weighted by the fraction of occasions when the forecast probability is in each class (g_k). With some rearrangement this can be written as

$$T_{\mathrm{F}} = T_{\mathrm{C}} + \sum_{k=0}^{m} g_k \int_{p'_k}^{p_k} u(\alpha)(\alpha - p'_k) \, \mathrm{d}\alpha - \sum_{k=0}^{m} g_k \int_{s}^{p'_k} u(\alpha)(p'_k - \alpha) \, \mathrm{d}\alpha \qquad (8.21)$$

where T_{c} is the total expense if all users act using the climatological base rate as decision threshold:

$$T_C = \int_0^s u(\alpha)\alpha \, d\alpha + \int_s^1 u(\alpha)s \, d\alpha \tag{8.22}$$

as explained by Richardson (2001).

The second and third terms on the right-hand side of Eq. (8.21) are generalised forms of the *reliability* and *resolution* components of the Brier score weighted by the probability distribution of cost/loss ratios (Section 7.3.2). The third term on the right-hand side of Eq. (8.21) is the maximum reduction in expense that would be achieved if all users were to act when the forecast probability of the event p'_k exceeds their particular cost/loss ratio α. This benefit increases as the forecast probabilities differ from the base rate (note s is a limit in the integral). The potential benefit is reduced if users act on forecast probabilities that are not completely reliable. The second term on the right-hand side of Eq. (8.21) gives the additional expense incurred. This reliability term depends on the difference between p_k and p'_k (the limits in the integral) and decreases as this difference reduces since there are then fewer occasions on which the incorrect choice of action is made.

For perfect forecasts, the expense per unit loss for a given user is $s\alpha$ (Eq. (8.4)) and so the total mean expense for all users is given by

$$T_P = \int_0^1 u(\alpha)s\alpha \, d\alpha = s\bar{\alpha} \tag{8.23}$$

i.e. the mean cost/loss ratio multiplied by the base rate. A measure of overall value can then be defined as

$$G = \frac{T_C - T_F}{T_C - T_P} \tag{8.24}$$

To see the relationship of this with the Brier score, imagine a set of users with a uniform distribution of cost/loss ratios, i.e. $u(\alpha) = 1$. Eq. (8.21) then becomes

$$\begin{aligned} T_F &= \frac{1}{2}\sum_{k=0}^{m} g_k(p_k - p'_k)^2 - \frac{1}{2}\sum_{k=0}^{m} g_k(p'_k - s)^2 + \frac{s(1-s)+s}{2} \\ &= \frac{1}{2}(B_{rel} - B_{res} + B_{unc}) + \frac{s}{2} \end{aligned} \tag{8.25}$$

where the bracketed term on the right-hand side is the Brier score expressed in the standard decomposition in terms of reliability, resolution and uncertainty (Chapter 7, Section 7.3.2). It is easily seen that for this distribution of

users, $G = \text{BSS}$ and so the Brier skill score is the overall value in the special case when the users have a uniform distribution of cost/loss ratios.

It is not easy to determine the distribution of cost/loss ratios appropriate for a given forecast system. For most real world studies of economic value, the weather sensitivity and decision-making processes are more complicated than in the simple cost/loss model. There are relatively few investigations that have attached financial costs to the simple model used in this chapter. Where figures are available, they tend towards lower cost/loss ratios: for instance, 0.02–0.05 for orchardists (Murphy 1977), 0.01–0.12 for fuel-loading of aircraft (Leigh 1995), 0.125 for winter road-gritting (Thornes and Stephenson 2001). This agrees with expectation from general economic considerations (Roebber and Bosart 1996), while for large potential losses the issue of risk-aversion (not considered here) would also mitigate towards lower cost/loss ratios.

Rather than trying to find a definitive user distribution, Murphy (1966) and Roebber and Bosart (1996) have used beta distributions to parameterise various probability distributions of cost/loss ratio in order to investigate the impact on value. The indication that small cost/loss ratios are more likely than large ones, suggests that the Brier score is unlikely to represent any real-world mean overall value. In the following section, we consider this point further and demonstrate the dependence on the distribution of users of the overall value measure G.

8.5 SKILL, VALUE AND ENSEMBLE SIZE

The previous two sections have examined the relationships between value and the ROC and Brier scores. In this section, we examine a case study where the ROC and BSS present different indications of the usefulness of the forecasting systems. The cost/loss model allows us to interpret these different conclusions from the perspective of potential users. Our case study aims to demonstrate the effect of varying ensemble size on the value provided to users. We compare the overall value of an ensemble of 50 forecasts of heavy precipitation (more than 10 mm in 12 h) against that of an ensemble having only 10 ensemble forecasts. The ensemble of 10 forecasts is a random subsample of the original ensemble of 50 forecasts.

Figure 8.6 shows the ROC area skill score (ROCSS) and the Brier skill score (BSS) for the two ensemble systems. From the BSS, we can conclude that the impact of ensemble size is generally small and that neither configuration has any useful skill at lead-times beyond 5 days. In contrast, ROCSS shows skill throughout the forecast range with substantial additional benefit from the 50-member ensemble forecasts.

Figure 8.7 shows the value curves for both configurations at day 5 and day 10. Since we are considering an uncommon event ($s = 0.014$), most value is obtained for users with low cost/loss ratios. For larger cost/loss ratios, the

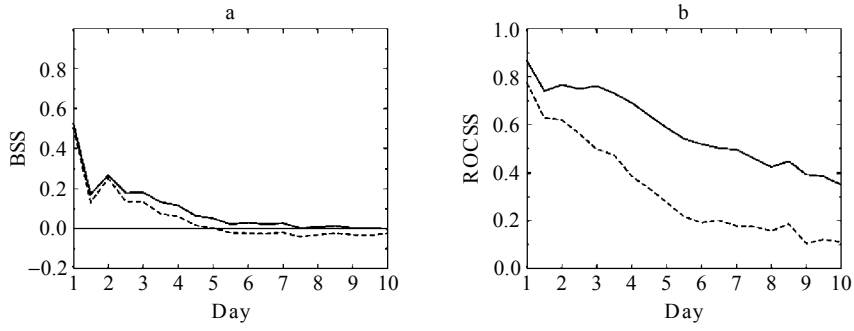


Figure 8.6 Brier (BSS) and ROC area (ROCSS) skill scores for ECMWF EPS probability forecasts of 12-h total precipitation exceeding 10 mm over Europe for winter 1996/1997. Curves show skill as a function of forecast lead-time for the operational 50-member EPS (solid line) and a 10-member EPS made from a subset of the operational members (dashed line)

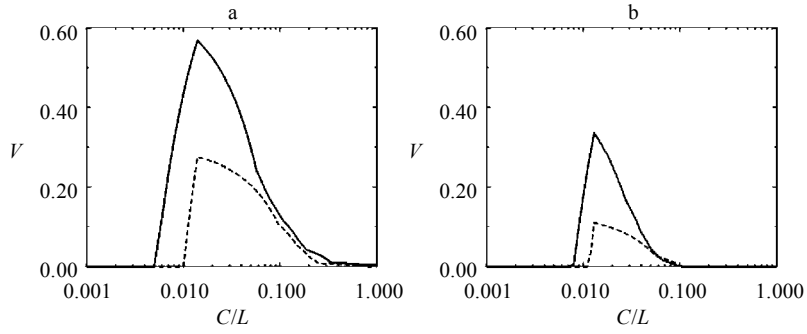


Figure 8.7 Value for day 5 (a) and day 10 (b) for the ensemble forecasts of Fig. 8.6. Solid line: 50-member ensemble; dashed line: 10-member ensemble

value is substantially smaller; by day 10, neither set of ensemble forecasts has any value for users with cost/loss ratios greater than 0.1. BSS, being a measure of the mean overall value for users with uniformly distributed cost/loss ratios gives little relative weight to the changes in value for low cost/loss ratios. In contrast, ROCSS is greater than $V_{\max}$ and so is still positive.

Differences in value between the 10- and 50-member ensemble forecasts are also greatest for low cost/loss ratios. Ideally, users should take action at a probability threshold $p_t = \alpha$, so to provide maximum benefit the ensemble must be large enough to resolve these required probability thresholds. The 50-member ensemble gives a better 1/50 resolution in probability thresholds while the small 10-member ensemble has a poorer resolution of 1/10 and so lacks the threshold discrimination needed for small cost/loss ratios. If the probabilities from the 50-member ensemble were restricted to the same 1/10 intervals as the 10-member ensemble, the large differences in value and ROCSS between the two configurations would be greatly reduced (not shown).

To summarise the dependence of overall value on the cost/loss distribution of users, Fig. 8.8 shows plots of the overall value G for various sets of users. As discussed already, BSS represents overall value for a set of users with cost/loss ratios uniformly distributed from 0 to 1. The first three panels of Fig. 8.8 illustrate the more probable situation of users with relatively low C/L. For cost/loss ratios uniformly distributed in the restricted range (0,0.2) the overall value is substantially increased compared to BSS, especially in the first 5 days, but the differences in G for the 10- and 50-member ensemble forecasts is still generally small (Fig. 8.8a). If we concentrate on a narrower band of cost/loss ratios, (0.02–0.05) for example, representing users with very large potential loss from an extreme event of heavy precipitation, the overall value again increases and the effect of ensemble size becomes more important (these users have low cost/loss ratios and need comparably low probability thresholds to achieve maximum benefit). Narrowing the cost/loss distribution towards small values both increases overall skill and emphasises the difference between the different ensemble systems. The overall mean values in Fig. 8.8c are approaching the levels of the ROCSS of Fig. 8.6b; remember that ROCSS is always greater than $V_{\max}$ and can therefore

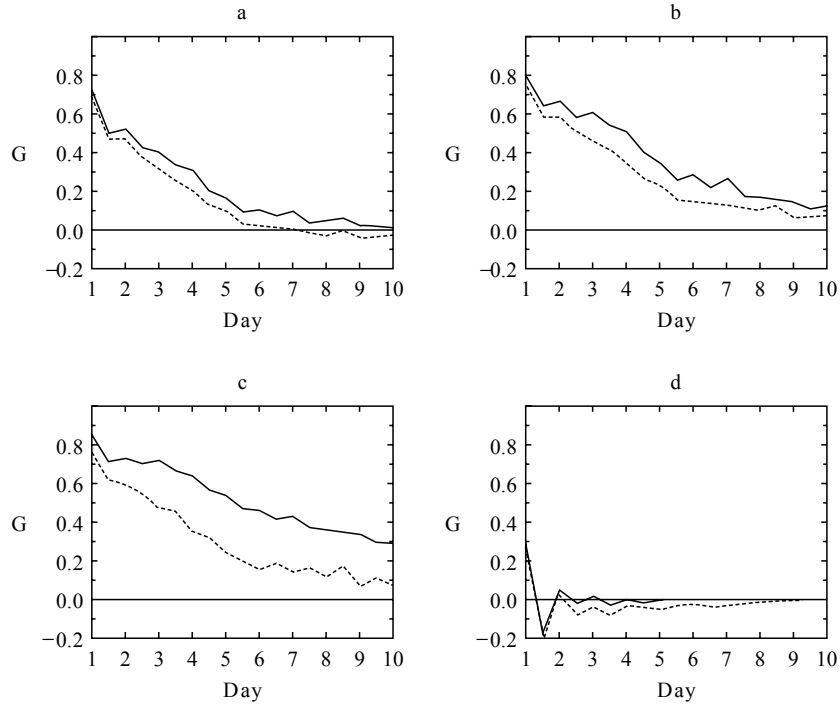


Figure 8.8 Overall value G for the ensemble forecasts of Fig. 8.6 for four example user distributions. Users are assumed to be distributed uniformly across the C/L interval of (a) 0.0–0.2, (b) 0.02–0.05, (c) 0.01–0.02 and (d) 0.5–0.8. Curves show G as a function of forecast lead-time for the operational 50-member EPS (solid line) and a 10-member EPS made from a subset of the operational members (dashed line)

never be exceeded for any distribution of users. Finally, Fig. 8.8d shows the contrasting situation for the less likely case of a set of users with relatively high cost/loss ratios (0.5–0.8); for these users there is no economic value in using either forecast system.

Comparing Figs. 8.8 and 8.6 shows that in most realistic situations, BSS and ROCSS can be considered lower and upper bounds for overall forecast value respectively. Exactly where the true value lies, between these limits, depends on the users' distribution of cost/loss ratios.

Finally, the above results depend on the event being considered. For more frequent events, greatest value will be achieved at correspondingly larger cost/loss ratios than here, and the diffeernces between BSS and ROCSS will be smaller. The benefit of larger ensembles of forecasts will, however, remain greatest for those users with the lowest cost/loss ratios, for example users who suffer very large losses when rare extreme events occur. Further discussion of the effects of ensemble size on value can be found in Richardson (2000, 2001).

8.6 SUMMARY

This chapter had two principal aims. Firstly, to introduce the concept of the economic value of forecasts in the context of a simple decision-making process; and secondly, to explore the relationship between economic value and some of the common verification measures used for forecast performance.

The simple cost/loss model gives a straightforward example of how forecast information can be used in decision-making situations. Probabilistic forecasts, when calibrated and used appropriately, are inherently more valuable than deterministic forecasts because they are adaptable to the differing needs of different users. The principal benefit of ensemble forecasting is the ability to produce reasonably reliable probability forecasts, rather than, for example, the deterministic ensemble mean. Forecast value depends not only on the quality of the forecasting system but also on the weather sensitivity of the user. Different users will gain to differing extents from the same forecasts. Indeed while one user may gain substantial benefit from a forecast system, another user may well derive no additional economic value.

This chapter has shown that several measures of forecast quality provide useful insight into the mean economic value of forecasts for a general set of users having a range of different cost/loss ratios. The Peirce skill score gives the maximum possible value that can be obtained from the forecasts, which occurs only for those users who have a cost/loss ratio exactly equal to the base rate. The Clayton skill score gives the range of cost/loss ratios over which value can be extracted by using the forecasts (H. Brooks and M. Wandishin, personal communication). The odds ratio exceeding one is a necessary and sufficient condition for at least some user to get value from

the forecasts. The Brier skill score is an estimate of the mean overall value of the forecasts for a population of users having a uniform distribution of cost/loss ratios. The area under the ROC curve provides a more optimistic estimate of mean overall value for a population of users. The two aspects of forecast 'goodness', quality and utility, are therefore intimately related. In the same way that probability may be defined subjectively as the price of a fair bet, forecast quality may be considered to be the expected utility of forecasts for a population of unknown users.

Commonly used measures of forecast performance such as the Brier skill score or ROC area skill score often give differing impressions of the value of forecasts and of the differences between competing systems. Viewing these scores from the perspective of economic value allows these differing results to be interpreted from the user perspective. The Brier skill score is equal to the overall mean value for a set of users having a uniform distribution of cost/loss ratios ranging from zero to one. If, as studies have shown, most real users generally operate with small cost/loss ratios, the Brier skill score will tend to give a pessimistic view of the overall value. It is possible for BSS to be zero or even negative and yet for the forecasts to have substantial value for a significant range of users (Palmer *et al.* 2000; Richardson 2001). Conversely, it can be argued that ROCSS gives perhaps an overly optimistic view of the value of forecasts since it is always somewhat greater than $V_{\max}$.

Given a distribution of cost/loss ratios, the overall mean value can easily be calculated. In the absence of detailed information on the distribution of cost/loss ratios, it is as well to be aware of the assumptions about users implicit in these common skill measures. As an example, while the area under the ROC curve is relatively sensitive to ensemble size, BSS is much less so. BSS would not, then, be an appropriate measure for the evaluation of extreme event forecasts for low cost/loss users; the likely benefit for these particular users from increasing ensemble size would probably not be discernible in the BSS.

Finally, no single summary measure of performance (including the overall value G) can be taken as representing the specific benefit to any individual user. Although an increase in G does, by definition, mean that value will increase over the group of users taken as a whole, the benefits to different individual users may vary significantly. It is quite possible that while overall value increases, the value to some users may actually decrease. Such skill–value reversals have been discussed by Murphy and Ehrendorfer (1987), Ehrendorfer and Murphy (1988) and Wilks and Hamill (1995). Only in the exceptional circumstance of one forecasting system being sufficient for all others, will all users be sure to prefer the same system (cf. discussion in Chapter 3, Section 3.3).

9 Forecast Verification: Past, Present and Future

DAVID B. STEPHENSON[1] AND IAN T. JOLLIFFE[2]
[1]Department of Meteorology, University of Reading, Reading, UK
[2]Department of Mathematical Sciences, University of Aberdeen, Aberdeen, UK

9.1 INTRODUCTION

The preceding chapters present a contemporary review of the wide range of forecast verification methods that have been developed and are currently being employed in weather and climate forecasting centres around the world. Since the burst of activity caused by Finley's forecasts in the 1880s, forecast verification has been undergoing exciting new developments with ever more measures/scores and techniques continually being invented (and re-invented!). The increasing amounts of weather and climate forecast products and verification data will inevitably continue to drive the demand for more verification. Forecasters will continue to ask 'How can the forecasting system be improved?'; users will continue to ask 'How useful are these forecast products?'; and administrators will surely continue to ask 'Has the forecasting performance of our institution improved?' In addition to these driving forces, the abundance of and increased reliance on forecasts in other disciplines opens up an exciting opportunity for innovative cross-disciplinary future work in forecast verification.

This final chapter will highlight some of the most important key concepts that have been introduced in previous chapters, will briefly discuss some of the different methods developed in other disciplines, and will finally outline some promising areas for future development.

9.2 REVIEW OF KEY CONCEPTS

What is forecast verification? A suitably general definition might be that *forecast verification is the exploration and assessment of the quality of a forecasting system based on a sample or samples of previous forecasts and*

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

corresponding observations. But what is meant by *quality?* Murphy (1997) explained that forecast quality has many different attributes that can be estimated using a wide variety of different sample statistics. Despite the obvious appeal, it is clear that no unique score such as mean squared error (MSE) can fully summarise the joint probability distribution between pairs of previous matched observations and the respective forecasts. Forecast verification is therefore a multi-dimensional problem with many possible scores/measures. It is this richness (or curse of dimensionality!) that makes the subject so perplexing yet so fascinating. At the risk of going into an infinite regress, meta-verification screening measures have even been invented such as *propriety*, *equitability*, *consistency*, etc., for scoring the quality of verification scores (Murphy 1997).

At first sight, it is easy to become completely bewildered by the multitude of possible verification scores and the associated philosophical issues in forecast verification. It is therefore helpful to go back to basics and reconsider the fundamental reason why we make forecasts: *to reduce our uncertainty about the* (*unknown*) *future state of a system*. Since the dawn of civilisation, humankind has attempted to cope with the uncertain knowledge about what will happen in the future by searching for clues in the past and present that help to reduce the range of possible outcomes in the future. By doing so, it is hoped that the uncertainty about the future will be reduced. This concept of reducing uncertainty about the future by *conditioning* on existing clues (forecasts) is the key to understanding the quality of forecasts. It is the dependency between observations x and forecasts $\hat{x}$ that makes forecasts useful – forecasts that are completely independent of observations are of absolutely no use in predicting the future. The dependency is best quantified by considering the *conditional probability* of the observations given the forecasts $p(x|\hat{x}) = p(x, \hat{x})/p(\hat{x})$. This conditional probability is the *calibration* factor in the *calibration–refinement factorisation* of the joint probability introduced by Murphy and Winkler (1987). The *conditional uncertainty* can be measured most easily by the variance, $\text{var}(X|\hat{X})$, of previous observations conditioned/stratified on a particular given value of forecast – the variance of the observations in the subset/class of previous cases in which the issued forecast was exactly equal to $\hat{X}$. A simple and revealing identity can be derived (see DeGroot 1986, Section 4.7) relating the mean conditional variance to the total unconditional variance of the observations:

$$E_{\hat{X}}[\text{var}_X(X|\hat{X})] = \text{var}(X) - \text{var}_{\hat{X}}[E_X(X|\hat{X})] \tag{9.1}$$

The mean of the variance of the observations given forecast information, $E_{\hat{X}}[\text{var}_X(X|\hat{X})]$, is equal to the unconditional variance $\text{var}(X)$ of the observations (total uncertainty) minus the variance of the conditional means $\text{var}_{\hat{X}}[E_X(X|\hat{X})]$ of the observations given in the forecasts. In other words, the mean uncertainty given forecast information is equal to the *uncertainty*

of the observations minus the *resolution* of the forecasting system. So for a forecasting system to reduce uncertainty it must have non-zero resolution (see Chapters 2 and 7 for more discussion about the importance of resolution). This is why resolution is one of the most important aspects in forecast verification. For the linear regression model of observations on forecasts (see Section 2.9), the resolution divided by the variance of the observations is r^2, the fraction of variance explained, and so the variance given the forecasts is reduced by a factor of $1 - r^2$ compared to the original variance of the observations. Similar conditioning arguments apply to the case of probability forecasts with the sole difference being that the conditioning variable (the forecast probability) is now a real number from 0 to 1 (or a vector of $K - 1$ probabilities for $K > 2$ categories) rather than being the same type of variable as the observation.

The identity in Eq. (9.1) also helps to explain why the MSE score (or Brier score for probability forecasts) and its partitioning are useful quantities to consider. By conditioning on the forecasts, the MSE can be decomposed into the sum of a mean variance component $E_{\hat{X}}[\mathrm{var}_X(X|\hat{X})]$ and a conditional bias (reliability) component $E_{\hat{X}}[(E_X(X|\hat{X}) - \hat{X})^2]$ as follows:

$$\begin{aligned} E[(X - \hat{X})^2] &= E_{\hat{X}}[E_X\{(X - \hat{X})^2|\hat{X}\}] \\ &= E_{\hat{X}}[\mathrm{var}(X|\hat{X})] + E_{\hat{X}}[(E_X(X|\hat{X}) - \hat{X})^2] \end{aligned} \tag{9.2}$$

Substituting for the mean variance component using the identity in Eq. (9.1) then gives the calibration–refinement decomposition of MSE:

$$E[(X - \hat{X})^2] = \mathrm{var}(X) - \mathrm{var}_{\hat{X}}[E_X(X|\hat{X})] + E_{\hat{X}}[(E_X(X|\hat{X}) - \hat{X})^2] \tag{9.3}$$

In other words, the MSE is the sum of an *uncertainty* term, a negated *resolution* term, and a *reliability* term. The first two terms are exactly the same as those that appear in identity equation (9.1) and so can be used to explain how much variance is explained by conditioning on the forecasts. The third term, reliability, is of less importance since it can be reduced to very small values by recalibrating (bias correcting) the forecasts, $\hat{X}' = E_{\hat{X}}(X|\hat{X})$, using a sufficiently large sample of previous forecasts and the assumptions of stationarity in the observations and the forecasting system. By conditioning instead on the observations rather than the forecasts, the MSE can also be expressed as a sum of a sharpness term $\mathrm{var}(\hat{X})$, a discrimination term $\mathrm{var}_X[E_{\hat{X}}(\hat{X}|X)]$, and a reliability (of the 2nd kind) term $E_X[(E_{\hat{X}}(\hat{X}|X) - X)^2]$. However, the calibration–refinement decomposition given in Eq. (9.3) is more natural for interpreting the reduction of uncertainty obtained by conditioning on forecast information. These additive properties of the quadratic MSE score make it particularly appealing for interpreting the reduction in uncertainty. However, it should be noted that this discussion has focussed on only 1st and 2nd moment quantities (means

and variances) and so does not necessarily capture all the possible features in the joint distribution of forecasts and observations (Murphy 1997). The relevance of 1st and 2nd order quantities is especially debateable for the verification of forecasts of non-normally distributed meteorological quantities such as precipitation and wind speed. Furthermore, care also needs to be exercised in practice since sample estimates of the MSE are unduly sensitive to outlier forecast errors (non-resistant statistic) as will be discussed in more detail in Section 9.4.

9.3 FORECAST EVALUATION IN OTHER DISCIPLINES

It is perhaps not surprising that meteorologists faced with the job of making regular daily forecasts over the last century have invented and applied many different verification approaches. However, meteorologists are not the only people who have to make forecasts or who have considered the forecast verification problem. This section will briefly review some developments in other fields and provide some references to the literature in these areas. It is hoped that this will stimulate a more progressive cross-disciplinary approach to forecast verification.

9.3.1 Statistics

Forecast verification is inherently a statistical problem – it involves exploring, describing, and making inferences about data sets containing matched pairs of previous forecasts and observations. Nevertheless, much of the verification work in atmospheric sciences has been done by meteorologists without the collaboration of statisticians. Notable exceptions have produced major breakthroughs in the subject – e.g. the papers mentioned throughout this book by Allan Murphy and the decision-theorist Robert Winkler.

Although statisticians have been heavily involved in time series forecasting, verification in these studies has often not gone beyond applying simple measures such as MSE (see Chatfield 2001, for a clearly written review). Chatfield (2001) classifies forecasts of continuous real valued variables into *point forecasts* (deterministic forecasts), *interval forecasts* (prediction intervals/limits in which the observation is likely to occur with a high specified probability) and *density forecasts* (the whole probability density function). Interval forecasts are to be preferred to point forecasts since they provide an estimate of the likely uncertainty in the prediction whereas point forecasts do not supply this information. Rather than consider deterministic forecasts as perfectly sharp probability forecasts in which the uncertainty is zero, it is more realistic to consider point forecasts as forecasts in which the uncertainty is unknown. In other words, deterministic point forecasts are not perfectly sharp forecasts with probabilities equal to 1 and 0, but instead they

should be assigned unknown sharpness. As reflected in the contents of the previous chapters, the intermediate case of interval forecasts has been largely overlooked by meteorologists and would benefit from some attention in the future. One exception is the article by Murphy and Winkler (1974) that discussed the performance of *credible interval* (Bayesian prediction interval) temperature forecasts. Interval forecasts for normally distributed variables can be most easily constructed from point forecasts by using the MSE of previous forecasts to estimate the interval width (Chatfield 2001). A problem with many interval and probability forecasts is that they often assume that the identified forecast model is perfect, and therefore only take into account *parametric uncertainty* (sampling errors on estimated model parameters) rather than also including estimates of *structural model uncertainty* (i.e. the model formulation may be incorrect). A clear discussion of how to take account of model uncertainty in forecasts can be found in Draper (1995) and the ensuing discussions. The main emphasis in the statistical literature when dealing with forecast uncertainty is generally not to attempt to verify the forecasts, but to provide an envelope of uncertainty around the forecast, which incorporates parameter uncertainty, random error and possibly model uncertainty.

Several interesting articles have been published in the statistical literature that address the verification/evaluation of (weather) forecasts. Gringorten (1951) presents the principles and methods behind the evaluation of skill and value in weather forecasts, and concludes by explaining his motivation for publishing in a statistical journal (universality of the verification problem). Slonim (1958) describes the trentile deviation scoring method for deterministic forecasts of continuous variables – a crude forerunner of the LEPS approach (Potts *et al.* 1996). Goodman and Kruskal (1959) review measures of association for cross-classification and provide a fascinating history and perspective on the 2×2 canonical verification problem raised by the Finley affair.

Not surprisingly, probability forecasts have also received attention from statisticians. Inspired by questions raised by the meteorologist Edward S. Epstein, Winkler (1969) wrote an article on likelihood-based scoring rules for probability assessors (subjective probability forecasts). He points out that under certain conditions the log-likelihood score becomes equivalent to Shannon's measure of entropy/information. This important fact was also mentioned in Stephenson (2000) for the case of binary forecasts. Winkler published several other single-author articles around this time on how best to evaluate subjective probability assessors. In a pivotal article published in *Applied Statistics*, Murphy and Winkler (1977) demonstrated that subjective probability forecasts of US precipitation and temperature were well calibrated. This meteorological example of the calibration of sequential probability forecasts stimulated much subsequent work on Bayesian calibration by Dawid and others. By considering sequential probability forecasts, Dawid (1982) demonstrated that recalibration of sequential probability

forecasts leads to incoherent (not the price of a fair bet) probability statements. In a profound yet difficult paper to read, DeGroot and Fienberg (1983) address this problem and discuss proper scoring rules for comparing and evaluating pairs of probability forecasters. They explain how the concepts of calibration and refinement relate to the concept of *sufficiency* that underpins much of statistical inference. In their review of probability forecasting in meteorology, Murphy and Winkler (1984) present an interesting and complete historical account of the subject. Probability forecasting and associated scoring rules were later succinctly reviewed from a statistical perspective by Dawid (1986). In a clearly written article, Schervish (1989) presents a general decision theory method for comparing two probability assessors and relates it to the calibration problem. Testing the validity of sequential probability (prequential) forecasts was revisited more recently in Seillier-Moiseiwitsch and Dawid (1993).

This short review of verification articles in the statistical literature is by no means complete, yet hopefully provides some useful starting points for further studies in the subject.

9.3.2 Finance and Economics

Financial analysts and economists are most likely the next biggest producer of forecasts issued on a regular frequent basis. Much interesting material related to financial and economics forecasting can, at the time of writing, be found on the 'Forecasting Principles' web site by Professor J. Scott Armstrong at the Wharton School, University of Pennsylvania: http://www-marketing.wharton.upenn.edu/forecast/. Point, interval and more recently density forecasts are routinely produced and their performances are evaluated using various criteria and scores.

The verification (or *ex post evaluation* as it is referred to in this discipline) of (*ex ante*) point forecasts has been recently reviewed by Wallis (1995), Diebold and Lopez (1996) and Armstrong (2001). Loss functions that are commonly used to score point forecasts include:

- mean squared (forecast) error, MS(F)E;
- mean absolute error, MAE;
- mean absolute percentage error, MAPE;
- relative absolute error, RAE.

Several studies have shown that MSE is not a reliable or resistant sample statistic for evaluating samples of *ex ante* forecasts (Armstrong and Fildes 1995; and references therein). For example, the sensitivity of MSE to outlier errors (non-resistance) makes it unsuitable for reliably ranking forecasts taking part in forecast comparisons/competitions (Armstrong and Collopy 1992). MAE is a more resistant score as discussed in Chapter 5 of this book. MAPE is obtained by taking the sample mean of $|\hat{x}_t - x_t|/|x_t|$ and helps to

account for variations in variance related to changes in mean (heteroskedasticity). For this reason, MAPE may be a useful measure to adopt in the verification of precipitation forecasts. RAE is the sample mean of the ratio of absolute errors of two different sets of forecasts and is therefore a comparative mean skill measure. Unlike skill scores based on mean scores, RAE takes the ratio of errors before averaging over the whole sample and so accounts for changes in predictability of events that take place throughout the sample period.

Various methods have been developed and employed for evaluating financial and economic interval forecasts (Christoffersen 1998; Taylor 1999). Typical methods are based on comparing the number of times the observations fall in the prediction intervals with the number of times expected given the stated probability of the interval. Formal tests of skill have been developed based on the binomial distribution. Motivated by the growth in risk management industries, there has been much interest recently in extreme one-sided intervals defined by values falling outside specified rare quantiles (e.g. below the 5% or above the 95% quantile). The forecasted rare quantile is known as *Value-At-Risk* (VaR) and gives information about possible future large losses in portfolios, etc. (see Dowd 1998; and the VaR web site http://www.GloriaMundi.org/). Various methods are used to verify VaR systems but many have low power and so fail to reject the no-skill hypothesis even for systems that do have real skill. Unlike point forecasts, these types of interval forecast need good predictions of future volatility (standard deviation) in order to be accurate and economists have developed many methods for forecasting variance that could be of potential use to the meteorological community.

The most complete description of the future value of a real variable is provided by forecasting its future probability density function. In a special edition of the *Journal of Forecasting* completely devoted to density forecasting in economics and finance, Tay and Wallis (2000) present a selective survey of density forecasting in macroeconomics and finance and discuss some of the issues concerning evaluation of such forecasts. They describe the *fan chart* method for displaying density forecasts obtained by grey shading time series plots of a set of regularly spaced quantiles. Tay and Wallis (2000) point out that no suitable loss functions have been defined for assessing the whole density and that in general the loss function would depend on the specific forecast user. Recent attempts to quantitatively evaluate density forecasts have used the *probability integral transform*

$$\hat{p}_t = \hat{F}(x_t) = \int_{-\infty}^{x_t} \hat{p}(x_t')\mathrm{d}x_t' \tag{9.4}$$

to transform the observed value x_t at time t into predicted cumulative probabilities $\hat{p}_t$ that should be uniformly distributed from 0 to 1 given

perfect density forecasts $p(\hat{x}_t)$. Either histograms or cumulative empirical distributions of $\hat{p}_t$ can be plotted and compared to that expected for uniformly distributed probabilities. Formal tests such as the Kolmogorov–Smirnov and likelihood-ratio tests are used to test uniformity. This approach is the basis for the analysis rank histogram diagram consistency method discussed in Chapter 7. Methods for multivariate density evaluation and calibration are reviewed in Diebold *et al.* (1999).

9.3.3 Environmental and Earth Sciences

In addition to meteorological events that cause some of the largest insured losses on the planet, our terrestrial environment is also full of other potential hazards such as floods, earthquakes, volcanoes, etc. Due to the impact on human lives and property, forecasting/prediction systems have been set up in many countries to give advanced warning about such events.

Hydrological services and environment agencies in many countries frequently issue regional flood warnings, often at very short notice (several hours lead) – see http://www.cig.ensmp.fr/∼iahs/nreps.htm for a list of national representatives. The warnings are often categorical forecasts such as the UK Environment Agency's four categories consisting of *severe flood warning*, *flood warning*, *flood watch* and *all clear*. Such forecasts are difficult to verify for a number of reasons. Firstly, the events at a particular location are often extremely rare, e.g. a severe flood might have occurred only once or not at all since historical records began. The number of joint events of severe flood warning and a severe flood happening are therefore likely to be extremely small and so contingency tables are full of very small numbers, which makes statistical analysis almost impossible. The problem can sometimes be alleviated by pooling (regional analysis) of data from similar yet independent catchments. Another major problem with flood verification is in defining the actual event – flooding is often extremely sensitive to small local elevation differences that are not easily measured. One possible way to avoid this problem is to focus instead on the resulting damage as measured by insurance claims but this then brings in other nuisance factors such as population density, housing prices, etc. A sensible way to evaluate flood warnings is in terms of how the information improves decision making. Krzysztofowicz (1995) reviews recent advances associated with flood forecast and warning systems and presents a Bayesian decision theory approach to evaluating flood-warning systems. Krzysztofowicz (2001) presents a compendium of reasons for moving away from deterministic to probabilistic forecasting of hydrological variates. In some sense, flood-warning categories should perhaps be interpreted as crude probability statements about a single event rather than as deterministic categorical forecasts.

Earthquake prediction is a controversial subject in seismology. The International Association for Seismology and Physics of the Earth's Inter-

ior (IASPEI) panel for the evaluation of precursors found that no unambiguous precursors have been identified but around seven have been placed on a list of 'potential precursors', awaiting further evaluation (Professor Ian Main, University of Edinburgh, personal communication). Few seismologists believe that deterministic forecasts of earthquakes before they occur have much skill or public value. However, there is some time dependence in the aftershock sequence that could be used as a conditional probability in time-dependent seismic hazard calculations for design and construction of buildings and infrastructure. The IASPEI panel found that a false impression of skill had been created in the literature by the practice of publishing only successful forecasts out of the huge number of precursors tested at locations all around the world. By selecting only the successful cases, such a reporting bias can easily lead to misleading confidence in the true skill of forecasts. The moral of this story is that it is very important to also publish results when forecasting systems do not perform well – this then gives a more balanced view of the whole subject. Seasonal climate forecasting is in danger of falling into a similar trap by isolating *posterior only specific events* in certain years that could have been forecast with skill (e.g. Dong *et al.* 2000). It is therefore important for forecasters to provide not only forecasts of future events but also provide clear information on past forecast performance and ideally also an uncensored sample of previous forecasts and observations.

9.3.4 Medical and Clinical Studies

The most common form of verification in medical studies occurs when there is a diagnostic test whose outcome is used to infer the likely presence of a disease. This corresponds to the classic (2×2) contingency table, as exemplified by Finley's tornado example, which was discussed in detail in Chapter 3. In disease diagnosis the 'forecast' is 'disease present' if the diagnostic test is positive and 'disease absent' if the test is negative. This set-up is given in Table 9.1, using the same notation a, b, c, d, for the table as elsewhere in the book.

A number of properties derived from such a table are in common use (Bland, 1995, Section 15.4):

Table 9.1 Diagnostic medical tests – possible outcomes

Forecast	Observed		
	Disease present	Disease absent	Total
Positive test – disease present	a	b	$a+b$
Negative test – disease absent	c	d	$c+d$
Total	$a+c$	$b+d$	n

- *sensitivity* $= a/(a + c)$;
- *specificity* $= d/(b + d)$;
- *positive predictive value* $= a/(a + b)$;
- *negative predictive value* $= d/(c + d)$.

It can be seen that three of these properties are related to quantities defined in Chapter 3, with different names. Thus sensitivity is simply the hit rate, specificity = 1 – false alarm rate and positive predictive value is 1 – false alarm ratio. Negative predictive value does not have a direct analogue in Chapter 3, perhaps because for most circumstances in meteorology it is more important to correctly forecast the occurrence of a meteorological event than it is to forecast its absence. In the case of medical diagnosis, however, it is crucial that negative diagnoses are correct most of the time. Otherwise, an opportunity for medical intervention at an early stage of a disease may be missed.

Altman and Royston (2000) discuss an *index of separation* (PSEP), which in the (2×2) case equals (positive predictive value + negative predictive value -1). More generally, with K diagnostic categories, but only two outcomes (disease presence or absence) PSEP is equal to the difference between the proportion of diseased individuals for the least and most favourable diagnostic categories. The idea is similar to that of *resolution* (see Section 2.10).

For many diagnostic tests, the values of sensitivity and specificity can be varied by adjusting some threshold that determines whether the test result is declared positive or negative. Plotting sensitivity against (1 – specificity) as the threshold varies gives an ROC curve (Section 3.4). A number of variations on the basic ROC curve have been developed in a medical context. For example, Venkatraman (2000) constructs a permutation test for comparing ROC curves, Rodenberg and Zhou (2000) show how to estimate ROC curves when covariates are present, and Baker (2000) extends the ROC paradigm to multiple tests.

9.4 FUTURE DIRECTIONS

This final section will attempt to highlight some (but not all) of the areas of forecast verification that we believe could benefit from more attention in future studies. The identification of areas for development is based on our insight gained in reviewing and editing all the chapters in this book and in our writing of Chapters 1, 9 and the glossary. Additional ideas have been stimulated by the interesting talks presented at the recent 'Workshop on Making Verification more Meaningful' held at the National Center for Atmospheric Research, Boulder, Colorado, August 2002 (at the time of writing many of the presentations were available at www.rap.ucar.edu/research/verification/pres.html).

Forecast verification is based on sample statistics calculated from samples of previous matched pairs of forecasts and observations. Since the sample size is always *finite*, the resulting verification statistics are always prone to sampling errors/uncertainties. Some scores such as MSE are more sensitive (less resistant) to outlier errors than are other scores and are therefore prone to larger sampling errors. For small samples such as those obtained for seasonal forecasts, it is important to be aware of these issues and consider more resistant measures (e.g. MAE). The resistance of scores to outlier errors needs to be taken into account when screening scores for suitability. Sampling uncertainty can also sometimes be reduced by taking advantage of the large number of spatial degrees of freedom to pool forecasts over suitable regions. Regional pooling could be a useful (essential) approach for increasing the effective sample size for verification in cases when there are few samples in time. Note, however, that pooling will not provide much benefit in cases where there are large spatial correlations (e.g. teleconnection patterns).

There is a need to move beyond purely descriptive sample statistics in forecast verification. Verification aims to make *inferences* about the *true* skill of the forecasting system based on the sample data available from previous sets of forecasts and observations. Therefore, sample scores should only be considered as finite sample *estimates* of the true scores of the system (i.e. the scores given an infinite number of previous forecasts). The sample estimates differ from the true values because of sampling uncertainties and so it is important not only to quote scores but also to provide estimates of sampling error. Few studies in the verification literature actually provide any form of error estimate (or confidence intervals) on their quoted scores – the scores are implicitly taken at face value as being perfectly correct which amounts to incorrectly assuming that the verification data set is infinite. Approximate sampling errors can be estimated either analytically by finding the sampling distribution of the score from the null distribution of forecasts and observations, or by computational resampling methods such as bootstrap (Efron and Tibshirani 1993). Often simple insightful analytic expressions can be found without having to resort to less transparent computational methods. For example, it is easy to derive analytic expressions for approximate sampling errors for, or confidence intervals based on, ratios of counts of independent events such as hit rates, etc., by using the binomial distribution (see Section 3.3.5 and stephenson 2000). In addition, the neglected topic of statistical power would benefit from further research, as would the verification of complete probability density functions.

Another area that could benefit from more research is in statistically modelling the joint distribution of forecasts and observations. The high $K^2 - 1$ dimensionality of the verification problem estimated by considering the counts in a $K \times K$ contingency table is more apparent than real since not all the counts in the K^2 cells are independent of one another. In fact, when the forecasts and observations are normally distributed (as is

sometimes a reasonable approximation, e.g. temperatures), then the problem has only five dimensions related to the five parameters of the joint bivariate normal distribution (e.g. two means, two variances and one correlation). So in this case the binning approach (a crude non-parametric estimate of the joint probability) would overestimate the dimensionality whenever more than two categories were used. It should also be noted that five dimensions is less than the total number of distinct aspects of quality identified by Murphy (1997) and so not all aspects of quality are independent in these cases. It therefore makes sense to develop parametric models of the joint distribution, yet surprisingly few studies have attempted to do this (Katz *et al.* 1982). A parametric modelling approach would also have the advantage of providing estimates of sampling uncertainties on scores. Another promising approach is to model the likelihood of observations given forecasts using suitable regressions (Krzysztofowicz and Long 1991b). Likelihood modelling would allow Bayesian approaches to be used that are good at merging information provided by different forecasting systems (Berliner *et al.* 2000). However, it should be noted that such modelling approaches depend on the validity of the underlying model assumptions and structural uncertainty therefore also needs to be taken into account (Draper 1995).

We speculate here on what are likely to be the most important developments in forecast verification over the next few years. Many forecasts in meteorology and other environmental sciences are not merely forecasts of one number or a vector of numbers but are forecasts of highly structured quantities containing rich inter-relationships (e.g. spatial fields). The verification of such forecasts is confounded by the complex dependency within the data objects. For example, not all pixels in a spatial precipitation forecast are independent but are related to one another due to clustering caused by larger scale coherent structures such as rain bands, storm cells, etc. Methods need to be developed to deal with these complex types of predictand. In particular, techniques such as hierarchical modelling, wavelet analysis and neural networks may all be of use in verifying such complex forecasts. One point that has already been mentioned in Chapter 6 concerns the difficulties posed by the different nature (regular grids, irregularly-spaced stations, continuous maps) and scales of observations and forecasts. At the time of writing, tackling these problems is an active area of research. Another development is to focus on the *shape* of areas of rainfall or of convective activity and verify, among other things, whether shape is correctly forecast. Spatial forecasting of convective activity is of special interest to aviation, and particular aspects of the convective areas are important in this context for routing and landing.

The nature of forecasts and of observations is also likely to change, with a further increase in the prevalence of ensemble forecasts, and an increasing reliance on remotely sensed data, e.g. from satellites. Although there is some methodology for these topics there is still much to do (e.g. for ensemble

forecasts see Wilson *et al*. 1999). The substantial interest in climate change, and the resulting modifications to the distributions of observations, suggests increasing activity in verification of non-stationary data and extremes.

On a final note, it is worth reiterating that the subject of forecast verification could benefit enormously from a wider and more open access to previous forecasts and observations. Sadly, the vast majority of operational centres provide forecast products without also providing clear documentation on previous forecast performance (Thornes and Stephenson 2001). It should be a rule of good practice that forecast providers make such information about their forecasting systems publicly available and should also provide easy access to samples of their past forecasts and verifying observations (e.g. via the internet). Such practice would then enable third parties to examine for themselves the past performance of the forecasting system and ultimately would provide much useful feedback for future forecasting improvements.

In conclusion, this book on forecast verification has covered many aspects in this amazingly rich and important subject. We hope that the book will enable a wider community of people to understand the complexities and issues involved in forecast verification, and thereby encourage the development and application of improved verification methods. We also hope that this book has opened up some exciting new areas for future research in the subject of forecast verification – only time will really tell!

Glossary

DAVID B. STEPHENSON
Department of Meteorology, University of Reading, UK

Many common words and phrases have developed very specific meanings in forecast verification studies. This glossary aims to provide clear explanations and consistent mathematical definitions for the more commonly used expressions. The population mean of a quantity B over all possible values of A for cases where condition C holds true is denoted by $E_A(B|C)$ (the conditional expectation of B over all values of A conditioned on C). The population variance of a quantity B over all values of A is denoted by $\mathrm{var}_A(B)$. Sample estimates of these population quantities are obtained by calculating the sample mean and variance over the appropriate subsample of cases where the condition is valid (stratified/composite means and variances). More comprehensive descriptions of the following can be found in indexed entries discussed in preceding chapters of this book.

Accuracy The average distance/error between forecasts and observations that depends on **Bias, Resolution** and **Uncertainty** attributes. Often estimated using **Mean square error** but can be estimated more robustly using statistics such as **Mean absolute error** that are less sensitive (more resistant) to large outlier errors.

Artificial skill An overestimate of the real skill of a forecasting system caused by including the same data to evaluate the forecast skill as was used to develop/train the forecasting system. Artificial skill can be avoided by using independent training and assessment data sets. Artificial skill often occurs in practice due to the presence of long-term trends in the data set.

Association Overall strength of the relationship/dependency between the forecasts and observations that is independent of the marginal distributions. Linear association is often estimated using the product moment **Correlation coefficient**.

Attributes **Forecast quality** is a multi-dimensional concept described by several different scalar attributes such as overall **Bias, Reliability/Calibration**

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

(Type 1 **Conditional bias), Uncertainty, Sharpness/Refinement, Accuracy, Association, Resolution** and **Discrimination**. All of these attributes provide useful information about the performance of a forecasting system – no single measure is sufficient for judging and comparing forecast quality.

Base rate The marginal probability distribution, $p(x)$, of the observations. In other words, the **Sample climatology** of the event independent of any forecasts.

Bias The difference between the central locations of the forecasts and the observations (also known as **Overall bias, Systematic bias**, or **Unconditional bias**). Most easily quantified using the **Mean error**, $E(\hat{X}) - E(X)$, i.e. the difference between the means of the forecasts and the observations. For **Categorical forecasts**, bias in marginal probabilities is estimated by the ratio of the total number of events forecast to the total number of events observed (i.e. $(a + b)/(a + c)$ for binary categorical forecasts – see **Contingency table**).

Brier probability score The Brier score is the mean square error of probability forecasts for a binary event $X = 0,1$. It is defined as $B = E[(\hat{p} - X)^2]$ and is zero for perfect (deterministic) forecasts and equals 1 for forecasts that are always incorrect.

Calibration (Perfectly Calibrated) See **Reliability** and **Sharpness**.

Categorical forecast A forecast in which a discrete number of K categories of events are forecast. Categories can be either **nominal** (no natural ordering – e.g. clear, cloudy, rain) or **ordinal** (the order matters – e.g. cold, normal, warm). Categorical forecasts can be either **deterministic** (a particular category – e.g. rain or no-rain tomorrow) or **probabilistic** (probabilities for each category – e.g. probability of 0.3 of rain and 0.7 for no-rain tomorrow).

Conditional bias Conditional bias is the difference $E_A(A|B) - B$ between the conditional mean $E_A(A|B)$ (the average over all possible values of A for a given value of B) of a random variable A and the conditioning variable B. The conditional bias is zero when the linear regression of A on B has a slope equal to one and an intercept of zero. Type 1 conditional bias $E_X(X|\hat{X}) - \hat{X}$ is obtained by calculating the mean of the observations for particular values of the forecast, whereas Type 2 conditional bias $E_{\hat{X}}(\hat{X}|X) - X$ is obtained by conditioning on the observed values. Measures of overall conditional bias can be obtained by averaging the mean squared bias over all possible values of the conditioning variable, e.g. $E_{\hat{X}}[(E_X(X|\hat{X}) - \hat{X})^2]$. Type 1 conditional bias is also known as **reliability** or **calibration** and is 0 for all values of $\hat{X}$ for a *perfectly reliable* (well calibrated) forecasting system. A **reliability diagram** can be made by plotting $E_X(X|\hat{X})$ against $\hat{X}$.

Conditional distribution The probability distribution of a variable, given that a related variable is restricted to a certain value. The conditional distribution of the forecasts given the observations, $p(\hat{x}|x)$, determines the **discrimination** or **likelihood**. The conditional distribution of the observations given the forecasts, $p(x|\hat{x})$, determines the **calibration** or **reliability**. These two conditional distributions are related to each via Bayes' theorem $p(x|\hat{x})p(\hat{x}) = p(\hat{x}|x)p(x)$.

Contingency table A two-way contingency table is a two-dimensional table that gives the discrete joint sample distribution of forecasts and observations in terms of cell counts. For dichotomous categorical forecasts, having only two possible outcomes (Yes or No), the following (2×2) contingency table can be defined:

Event forecast	Event observed		
	Yes	No	Total observed
Yes	a (hits)	b (false alarms)	$a+b$
No	c (misses)	d (correct rejections)	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d=n$

Cell count a is the number of event forecasts that correspond to event observations, or the number of **hits**; cell count b is the number of event forecasts that do not correspond to observed events, or the number of **false alarms**; cell count c is the number of no-event forecasts corresponding to observed events, or the number of **misses**; and cell count d is the number of no-event forecasts corresponding to no events observed, or the number of **correct rejections**.

Forecast quality for this (2×2) binary situation can be assessed using a surprisingly large number of different measures, e.g. **Percent correct** (PC), **Probability of detection** (POD), **False alarm ratio** (FAR), **Success ratio** (SR), **Threat score** (TS) or **Critical success index** (CSI), **Heidke skill score** (HSS), and a categorical measure of **Bias**, etc.

Correct rejection In a categorical verification problem, a no-event forecast that is associated with no event observed. See **Contingency table**.

Correlation coefficient A measure of the **Association** between the forecasts and observations independent of the mean and variance of the marginal distributions. The Pearson product moment correlation coefficient is a measure of linear association and is invariant under any shifts or rescalings of the forecast or observed variables. The Spearman rank correlation coefficient measures monotonicity in the relationship and is invariant

under any monotonic transformations of either the forecast or observed variables.

Critical success index (CSI) Also called the **Threat score** (TS) and the **Gilbert score (GS)**, the CSI is a verification measure of categorical forecast performance equal to $a/(a+b+c)$, i.e. the total number of correct event forecasts (hits) divided by the total number of event forecasts plus the number of misses (hits + false alarms + misses). The CSI is not affected by the number of no-event forecasts that are not observed (correct rejections) and is therefore strongly dependent upon the **base rate**.

Deterministic forecasts Non-probabilistic forecasts of either a specific category or particular value for either a discrete or continuous variable. Deterministic forecasts of continuous variables are also known as **Point forecasts**. Deterministic forecasts fail to provide any estimates of possible uncertainty, and this leads to less optimal decision making than can be obtained using **Probabilistic forecasts**. Deterministic forecasts are often interpreted as probabilistic forecasts having only probabilities of 0 and 1 (i.e. no uncertainty), yet it is more realistic to interpret them as probabilistic forecasts in which the uncertainty is not provided (i.e. unknown uncertainty). Sometimes (confusingly) referred to as categorical forecasts in the earlier literature.

Discrimination The sensitivity of the likelihood $p(\hat{x}|x)$ to different observed values of x. It can be measured for a particular forecast value $\hat{x}$ by the **Likelihood-ratio** $p(\hat{x}|x_1)/p(\hat{x}|x_2)$. A single overall summary measure is provided by the variance $\mathrm{var}_X[E_{\hat{X}}(\hat{X}|X)]$ of the means of the forecasts conditioned (stratified) on the observations.

Equitable threat score (ETS) This skill score is commonly used for the verification of deterministic forecasts of rare events (e.g. precipitation amounts above a large threshold). It was developed by Gilbert (1884) as a modification of the **Threat score** to allow for the number of hits that would have been obtained purely by chance – see Schaefer (1990) and Doswell *et al.* (1990). It is sometimes referred to as Gilbert's skill score. In terms of raw cell counts it is defined as

$$\frac{a - a_\mathrm{r}}{a - a_\mathrm{r} + b + c}$$

where $a_\mathrm{r} = (a+b)(a+c)/n$ is the number of hits expected for forecasts independent of observations (pure chance). Note that the appearance of n in the expression for a_r means that the equitable threat score (unlike the threat score) depends explicitly on the number of correct rejections, d.

Equitable/equitability A metaverification property for screening of suitable scores for deterministic categorical forecasts (Gandin and Murphy 1992). An equitable score takes the same, no-skill value for random forecasts and for unvarying forecasts of a constant category. This criterion is based on the principle that random forecasts or constant forecasts of a category should have the same expected no-skill score (Murphy and Daan 1985).

False alarm In a categorical verification problem, an event forecast that is associated with no event observed. See **Contingency table**.

False alarm rate (F) A verification measure of categorical forecast performance equal to the number of false alarms divided by the total number of events observed. For the (2×2) verification problem in the definition of **Contingency table**, $F = b/(b + d)$. Not to be confused with **False alarm ratio**.

False alarm ratio (FAR) A verification measure of categorical forecast performance equal to the number of false alarms divided by the total number of event forecast. For the (2×2) verification problem in the definition of **Contingency table**, $\mathrm{FAR} = b/(a + b)$. Not to be confused with **False alarm *rate*** that is conditioned on observations rather than forecasts.

Forecast quality Statistical description of how well the forecasts match the observations that provides important feedback on the forecasting system. Unlike **Forecast value**, it aims to provide an overall summary of the agreement between forecasts and observations that does not depend on a particular user's requirements. Forecast quality has many different **Attributes** that can all provide useful information on the performance.

Forecast value The economic utility of forecasts for a particular set of forecast users often based on simple cost–loss models. Often strongly dependent on the marginal distributions and forecast bias due to users incurring very different losses for different categories of events.

Forecast verification The process of summarizing and assessing the overall **Forecast quality** of previous sets of forecasts. Although more commonly referred to as *forecast evaluation* in other disciplines, in the meteorological context forecast evaluation implies the study of user-specific **Forecast value** rather than **Forecast quality**. Philosophically, the word *verification* is a misnomer since all forecasts eventually fail and so can only be *falsified* not *verified*.

Gilbert score (GS) Same as **Critical success index** (CSI).

Gilbert's skill score (GSS) Same as **Equitable threat score** (ETS).

Heidke skill score (HSS) A skill corrected verification measure of categorical forecast performance similar to the **Percentage correct** (PC) but which takes into account the number of hits due to chance. Hits due to chance are given as the event relative frequency multiplied by the number of event forecasts.

Hit A forecasted categorical event that is later observed to happen. See **Contingency table**.

Hit rate (H) A categorical forecast score equal to the total number of correct event forecasts (hits) divided by the total number of events observed, i.e. $a/(a+c)$ in the (2×2) contingency table. Also known as the **Probability of detection** (POD) in the older literature.

Joint distribution The probability distribution defined over two or more variables. For serially independent events, the joint distribution of the forecasts and observations, $p(\hat{x}, x)$, contains all of the probabilistic information relevant to the verification problem. The joint distribution can be factored into **Conditional distributions** and **Marginal distributions** in either of two ways:

- The **calibration–refinement** factorization $p(\hat{x}, x) = p(x|\hat{x})\, p(\hat{x})$.
- The **likelihood–base rate** factorization $p(\hat{x}, x) = p(\hat{x}|x)\, p(x)$.

See Murphy and Winkler (1987) for an elegant exposition on this general and powerful framework.

Likelihood The probability $p(\hat{x}|x)$ of a forecast value given a particular observed value. The sensitivity of the likelihood to the observed value determines the **Discrimination** of the system. Note that the concept of likelihood is fundamental in much of statistical inference, but the usage of the term in that context is somewhat different.

Marginal distribution The probability distribution of a single variable, e.g. $p(x)$ or $p(\hat{x})$. The marginal distribution of the observations, $p(x)$, is referred to as the **Base rate**. See also **Uncertainty, Sharpness** and **Refinement**.

Mean absolute error (MAE) The mean of the absolute differences between the forecasts and observations $E(|\hat{X} - X|)$. A more robust measure of forecast accuracy than **Mean square error** that is somewhat more resistant to the presence of large outlier errors. Can be made dimensionless and more

stable by dividing by the mean absolute deviation of the observations $E(|X - E(X)|)$ to yield a **Relative absolute error**.

Mean error (ME) The mean of the differences of the forecasts and observations $E(\hat{X} - X) = E(\hat{x}) - E(X)$. It is an overall measure of the unconditional bias of the forecasts (see **Reliability**).

Mean square error (MSE) The mean of the squares of the differences of the forecasts and observations $E[(\hat{X} - X)^2]$. It is a widely used measure of forecast **Accuracy** that depends on **Bias, Resolution** and **Uncertainty**. Because it is a quadratic loss function, it can be overly sensitive to large outlier forecast errors and is an unreliable and non-resistant measure (see **Mean absolute error**). The MSE can sometimes encourage forecasters to hedge towards forecasting smaller than observed variations in order to reduce the risk of making a large error.

Miss See **Contingency table**.

Non-probabilistic forecast See **Deterministic forecast**.

Percentage correct (PC) The percentage of correct categorical forecasts (hits and correct rejections) equal to $(a + d)/(a + b + c + d)$ for the (2×2) problem (see **Contingency table**).

Point forecasts See **Deterministic forecasts**.

Predictand The observable object x that is to be forecast. In regression, the predictand is known as the **Response variable**, which is predicted using the **Predictor**. Scalar predictands can be nominal categories (e.g. snow, foggy, sunny), ordinal categories (e.g. cold, normal, hot), discrete variables (e.g. number of hurricanes), or continuous variables (e.g. temperature).

Predictor A forecast of either the value $\hat{x}$ of a predictand (deterministic forecasts) or the probability distribution $\hat{p}(x)$ of a predictand (probabilistic forecasts). In regression analysis, the word 'predictor' is used to denote an explanatory variable that is used to predict the predictand.

Probabilistic forecast A forecast that specifies the future probability $\hat{p}(x)$ of one or more events x occurring. The set of events can be discrete (categorical) or continuous. **Deterministic forecasts** can be considered to be the special case of probability forecasts in which the forecast probabilities are always either zero or one – there is never any prediction uncertainty in the predictand. However, it is perhaps more realistic to consider deterministic forecasts to be forecasts in which the prediction uncertainty in the predictand is not supplied as part of the forecast rather than as ones in

which the prediction uncertainty is exactly equal to zero. Subjective probability forecasts can be constructed by eliciting expert advice.

Probability of detection (POD) Same as **Hit rate** (H).

Proper/propriety A metaverification property for screening of suitable scores for probabilistic forecasts. A **Strictly proper** score is one for which the best expected score is only obtained when the forecaster issues probability forecasts that equal the forecasted probabilities consistent with their beliefs (e.g. the forecasting model is correct). Proper scores discourage forecasters from *hedging* their forecasted probabilities towards probabilities that are likely to score more highly. The **Brier score** is **Strictly proper**.

Ranked probability score (RPS) An extension of the **Brier score** to probabilistic categorical forecasts having more than two *ordinal* categories. By using cumulative probabilities, it takes into account the ordering of the categories.

Refinement Refinement is a statistical property of the forecasts that has multiple definitions in the verification literature. It can mean the marginal probability distribution of the forecasts, $p(\hat{x})$, as used in the phrase *calibration–refinement factorization* (see **Joint distribution**). However, often it is more specifically used to refer to the spread of the marginal probability distribution of the forecasts. In addition, refinement is also used synonymously to denote the **Sharpness** of probability forecasts. However, in the Bayesian statistical literature refinement appears to be defined somewhat differently to sharpness (see DeGroot and Fienberg 1983).

Reliability The same as **Calibration**. It is related to Type 1 **Conditional bias** $E_X(X|\hat{X}) - \hat{X}$ of the observations given the forecasts. Systems with zero conditional bias for all $\hat{X}$ are *perfectly reliable* (well calibrated) and so have no need to be recalibrated (bias corrected) before use. Forecasting systems can be made more reliable by posterior recalibration; for example, the transformed forecast quantity $\hat{X}' = E_X(X|\hat{X})$ is perfectly reliable. Similar ideas also apply to probabilistic forecasts where predictors $\hat{X}$ are replaced by forecast probabilities $\hat{p}(x)$.

Reliability diagram A diagram in which the conditional expectation of a predictand for given values of a continuous predictor is plotted against the value of the predictor (forecast). For deterministic forecasts of continuous variables, this is a plot of $f(\hat{x}) = E_X(X|\hat{X} = \hat{x})$ versus $\hat{x}$, i.e. the means of the observations stratified on cases with specific forecast values versus the forecast values. For probabilistic forecasts of binary events, this is a plot of $f(q) = E_X(X \mid \hat{p} = q)$ versus the forecast probability value q. Perfectly

reliable probabilistic forecasts have points that lie on the line $f(\hat{x}) = \hat{x}$ (deterministic forecasts of continuous variables) or $f(q) = q$ (probabilistic forecasts of binary variables). Given enough previous forecasts, recalibrated future forecasts can be obtained by using the reliability curve to non-linearly transform the forecasts: $\hat{x}' = f(\hat{x})$ or $\hat{p}' = f(\hat{p})$.

Resistant measure A verification measure not unduly influenced by the presence of very large or small outlier values in the sample, e.g. mean absolute deviation.

Resolution The sensitivity of the conditional probability $p(x|\hat{x})$ to different forecast values of $\hat{x}$. If a forecasting system leads to identical probability distributions of observed values for different forecast values, i.e. $p(x|\hat{x}_1) = p(x|\hat{x}_2)$, then the system has no resolution. Resolution is essential for a forecasting system to be able to discriminate between observable future events. A single overall summary measure is provided by the variance of the conditional expectation $\text{var}_{\hat{X}}[E_X(X|\hat{X})]$.

Robust measure A verification measure that is not overly sensitive to the form of the probability distribution of the variables.

ROC A relative (or receiver) operating characteristic (ROC) is a signal detection curve for binary forecasts obtained by plotting a graph of the **Hit rate** (y-axis) versus the **False alarm rate** (x-axis) over a range of different thresholds. For deterministic forecasts of a continuous variable, the threshold is a value of the continuous variable used to define the binary event. For probabilistic forecasts of a binary event, the threshold is a probability decision threshold that is used to convert the probabilistic binary forecasts into deterministic binary forecasts.

Root mean square error (RMSE) The square root of the **Mean square error**.

Sample Climatology See **Base rate**.

Sharpness An attribute of the marginal distribution of the forecasts that aims to quantify the ability of the forecasts to 'stick their necks out'. In other words, how much the forecasts deviate from the mean climatological value/category for deterministic forecasts, or from the climatological mean probabilities for probabilistic forecasts. Unvarying climatological forecasts take no great risks and so have zero sharpness; perfect forecasts are as sharp as the time-varying observations. For deterministic forecasts of discrete or continuous variables, sharpness is most simply estimated by the variance $\text{var}(\hat{X})$ of the forecasts. For **Perfectly calibrated** forecasts where $E_X(X|\hat{X}) = \hat{X}$, the sharpness $\text{var}(\hat{X})$ becomes identical to the **Resolution**

$\text{var}_{\hat{X}}[E_X(X|\hat{X})]$ of the forecasts. For probabilistic forecasts, although sharpness can also be defined by the variance $\text{var}(\hat{p})$, it is often defined in terms of the *information content* (negative entropy) $I = E(\hat{p}\log\hat{p})$ of the forecasts. High-risk forecasts in which $\hat{p}$ is either 0 or 1 have maximum information content and are said to be *perfectly sharp*. **Perfectly calibrated** perfectly sharp forecasts correctly predict all events. By interpreting deterministic forecasts as probabilistic forecasts with zero prediction uncertainty in the predictand, deterministic forecasts may be considered to be perfectly sharp probabilistic forecasts. However, it is perhaps more realistic to consider deterministic forecasts to be ones in which the prediction uncertainty in the predictand is not supplied as part of the forecast rather than ones in which the prediction uncertainty is exactly equal to zero. Hence, a deterministic forecast can be considered to be a deterministic forecast with spread/sharpness $\text{var}(\hat{X})$, yet at the same time can also be considered to be a probability forecast with perfect sharpness. The word **Refinement** is also sometimes used to denote sharpness.

Skill score Relative measure of the quality of the forecasting system compared to some (usually 'low-skill') benchmark forecast. Commonly used reference forecasts include mean climatology, persistence (random walk forecast), or output from an earlier version of the forecasting system. There are as many skill scores as there are possible scores and they are usually based on the expression

$$SS = \frac{S - S_0}{S_1 - S_0} \times 100\,\%$$

where S is the forecast score, S_0 is the score for the benchmark forecast, and S_1 is the best possible score. The skill scores generally lie in the range 0 to 1 (0 to 100 %). Compared to raw scores, skill scores have the advantage that they help take account of non-stationarities in the system to be forecast. For example, improved forecast scores often occur during periods when the atmosphere is in a more persistent state.

Success ratio (SR) A categorical binary score equal to the number of hits divided by the total number of events predicted $a/(a+b)$. Conditioned on the forecasts unlike the **Hit rate**, which is conditioned on the observations.

Sufficiency The concept of sufficiency was introduced into forecast evaluation by DeGroot and Fienberg (1983), and developed by Ehrendorfer and Murphy (1988) and Krzysztofowicz and Long (1991b) among others. When it can be demonstrated, sufficiency provides an unequivocal ordering on the quality of forecasts. When two forecasting systems, A and B say, are being

compared, A's forecasts are said to be sufficient for B's if forecasts with the same skill as B's can be obtained from A's by a stochastic transformation. Applying a stochastic transformation to A's forecasts is equivalent to randomizing the forecasts, or passing them through a noisy channel (DeGroot and Fienberg 1983). Note that sufficiency is an important property in much of statistical inference, but the usage of the term is somewhat different in that context.

Threat score (TS) Same as **Critical success index (CSI)**.

Uncertainty The mean spread in the observations related to the width of the marginal probability distribution $p(x)$. Uncertainty is most simply measured by the variance, var(X), of the observations. Important aspect in the performance of a forecasting system, over which the forecaster has no control.

REFERENCES

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. New York: Wiley.

Agresti, A. and Coull, B.A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *Amer. Statist.*, **52**, 119–126.

Altman, D.G. and Royston, P. (2000). What do we mean by validating a prognostic model? *Statist. Med.*, **19**, 453–473.

Anderson, J.L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530.

Anderson, J.L. and van den Dool, H.M. (1994). Skill and return of skill in dynamic extended-range forecasts. *Mon. Weather Rev.*, **122**, 507–516.

Ångström, A. (1922). On the effectivity of weather warnings. *Nordisk Statistisk Tidskrift*, **1**, 394–408.

Armstrong, J.S. and Collopy, F. (1992). Error measures for generalizing about forecasting methods: empirical comparisons. *Int. J. Forecasting*, **8**, 69–80.

Armstrong, J.S. and Fildes, R. (1995). On the selection of error measures for comparisons among forecasting methods. *J. Forecasting*, **14**, 67–71.

Armstrong, J.S. (2001). Evaluating forecasting methods. In: *Principles of Forecasting: A Handbook for Researchers and Practitioners* (ed. J. Scott Armstrong). Norwell, MA: Kluwer, 443–472.

Atger, F. (2001). Verification of intense precipitation forecasts from single models and ensemble prediction. *Nonlinear Processes Geophys.*, **8**, 401–417.

Atger, F. (2002). Spatial and interrannual variability of the reliability of ensemble based probabilistic forecasts. Consequences for calibration, *Mon. Weather. Rev.* **131**, to appear.

Baker, S.G. (2000). Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics*, **56**, 1082–1087.

Barnston, A.G. (1992). Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *J. Climate*, **7**, 699–709.

Barnston, A.G. (1994). Linear statistical short-term climate predictive skill in the Northern Hemisphere. *J. Climate*, **7**, 1513–1564.

Berliner, L.M., Wikle, C.K. and Cressie, N. (2000). Long-lead prediction of Pacific SSTs via Bayesian dynamic modeling. *J. Climate*, **13**, 3953–3968.

Birdsall, T.G. (1966). The theory of signal detectability: ROC curves and their character. *Dissert. Abstracts Int.*, **28**, 1B.

Bland, M. (1995). *An Introduction to Medical Statistics*. Oxford: Oxford University Press.

Forecast Verification: A Practitioner's Guide in Atmospheric Science. Edited by I.T. Jolliffe and D.B. Stephenson © 2003 John Wiley & Sons, Ltd ISBN: 0-471-49759-2

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

Brier, G.W. and Allen, R.A. (1951). Verification of weather forecasts. In: *Compendium of Meteorology* (ed. T.F. Malone). Boston: American Meteorological Society, 841–848.

Briggs, W.M. and Levine, R.A. (1997). Wavelets and field forecast verification. *Mon. Weather. Rev.*, **125**, 1329–1341.

Brooks, H.E. and Doswell, C.A. (1996). A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Weather Forecasting*, **11**, 288–303.

Bross, I.D.J. (1953). *Design for Decision*. New York: Macmillan.

Bryan, J.G. and Enger, I. (1967). Use of probability forecasts to maximize various skill scores. *J. Appl. Met.*, **6**, 762–769.

Buizza, R. (1997). Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Weather Rev.*, **125**, 99–119.

Buizza, R., Barkmeijer, J., Palmer, T.N. and Richardson, D.S. (2000). Current status and future developments of the ECMWF Ensemble Prediction System. *Met. Apps.*, **7**, 163–175.

Buizza, R., Hollingsworth, A., LaLaurette, F. and Ghelli, A. (1999). Probabilistic predictions using the ECMWF ensemble prediction system. *Weather Forecasting*, **14**, 168–189.

Centor, R.M. (1991). Signal detectability: the use of ROC curves and their analyses. *Med. Decis. Making*, **11**, 102–106.

Chatfield, C. (2001). *Time Series Forecasting*. Boca Raton: Chapman & Hall/CRC Press.

Christoffersen, P.F. (1998). Evaluating interval forecasts. *Int. Econ. Rev.*, **39**, 841–862.

Daley, R. (1991). *Atmospheric Data Analysis*. Cambridge: Cambridge University Press.

Daan, H. (1984). *Scoring Rules in Forecast Verification*. WMO Short- and Medium-Range Weather Prediction Research Publication Series No. 4.

Daw, F.A. and Mason, I.B. (1981). *Expressing and Verifying Uncertainty in Public Weather Forecasts*. Meteorological Note 125. Australian Bureau of Meteorology.

Dawid, A.P. (1982). The well-calibrated Bayesian. *J. Amer. Statist. Assoc.*, **77**, 605–610.

Dawid, A.P. (1986). Probability forecasting. In: *Encyclopedia of Statistical Sciences*, vol. 7 (eds. S. Kotz, N.L. Johnson and C.B. Read). New York: Wiley, 210–218.

DeGroot, M.H. (1986). *Probability and Statistics*, 2nd edition. Reading, MA: Addison-Wesley.

DeGroot, M.H. and Fienberg, S.E. (1982). Assessing probability assessors: calibration and refinement. In: *Statistical Decision Theory and Related Topics III*, vol. 1 (eds. S.S. Gupta and J.O. Berger). New York: Academic Press, 291–314.

DeGroot, M.H. and Fienberg S.E. (1983). The comparison and evaluation of forecasters. *Statistician*, **32**, 12–22.

Déqué, M. (1991). Removing the model systematic error in extended range forecasting. *Ann. Geophys.*, **9**, 242–251.

Déqué, M. (1997). Ensemble size for numerical seasonal forecasts. *Tellus*, **49A**, 74–86.

Déqué, M. and Royer, J.F. (1992). The skill of extended-range extratropical winter dynamical forecasts. *J. Climate*, **5**, 1346–1356.

Déqué, M., Royer, J.F. and Stroe, R. (1994). Formulation of Gaussian probability forecasts based on model extended-range integrations. *Tellus*, **46A**, 52–65.

Diebold, F.X. and Lopez, J. (1996). Forecast evaluation and combination. In: *Handbook of Statistics* (eds. G.S. Maddala and C.R. Rao). Amsterdam: North-Holland, 241–268.

Diebold, F.X., Hahn, J. and Tay, A. (1999). Multivariate density forecast evaluation and calibration in financial risk management: high-frequency returns on foreign exchange. *Rev. Econ. Statist.*, **81**, 661–673.

Donaldson, R.J., Dyer, R.M. and Kraus, M.J. (1975). An objective evaluator of techniques for predicting severe weather events. In: *Preprints, Ninth Conference on Severe Local Storms*, Norman, Oklahoma. American Meteorological Society, 321–326.

Dong, B.-W., Sutton, R.T., Jewson, S.P., O'Neill, A. and Slingo, J.M. (2000). Predictable winter climate in the North Atlantic sector during the 1997–1999 ENSO cycle. *Geophys. Res. Lett.*, **27**, 985–988.

Doolittle, M.H. (1885). The verification of predictions. *Bull. Philosophical Soc. Washington*, **7**, 122–127.

Doolittle, M.H. (1888). Association ratios. *Bull. Philosophical Soc. Washington*, **10**, 83–87, 94–96.

Doswell, C.A. III, Davies-Jones, R. and Keller, D.L. (1990). On summary measures of skill in rare event forecasting based on contingency tables. *Weather Forecasting*, **5**, 576–586.

Dowd, K. (1998). *Beyond Value at Risk: The New Science of Risk Management*. Chichester: Wiley.

Draper, D. (1995). Assessment and propagation of model uncertainty. *J. Roy. Statist. Soc. B*, **57**, 45–97.

Draper, N.R. and Smith, H. (1998). *Applied Regression Analysis*, 3rd edition. New York: Wiley.

Drosdowsky, W. and Chambers, L.E. (2001). Near global scale sea surface temperature anomalies as predictors of Australian seasonal rainfall. *J. Climate*, **14**, 1677–1687.

Ebert, E. (2001). Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Weather Rev.*, **129**, 2461–2480.

Ebert, E. and McBride, J.L. (2000). Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrol.*, **239**, 179–202.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. London: Chapman & Hall.

Ehrendorfer, M. (1997). Predicting the uncertainty of numerical weather forecasts: a review. *Meteorol. Z., Neue Folge*, **6**, 147–183.

Ehrendorfer, M. and Murphy, A.H. (1988). Comparative evaluation of weather forecasting systems: sufficiency, quality and accuracy. *Mon. Weather Rev.*, **116**, 1757–1770.

Emery, W.J. and Thomson, R.E. (1998). *Data Analysis Methods in Physical Oceanography*. Amsterdam: Elsevier.

Everitt, B.S. (1994). *The Analysis of Contingency Tables*. London: Chapman & Hall.

Finley, J.P. (1884). Tornado predictions. *Amer. Met. J.*, **1**, 85–88.

Flueck, J.A. (1987). A study of some measures of forecast verification. In: *Preprints, 10th Conference on Probability and Statistics in Atmospheric Science*. Edmonton, AB, Canada. American Meteorological Society, 69–73.

Frederiksen, C.S., Zhang, H., Balgovind, R.C., Nicholls, N., Drosdowsky W. and Chambers, L. (2001). Dynamical seasonal forecasts during the 1997/98 ENSO using persisted SST anomalies, *J. Climate*, **14**, 2675–2695.

Gandin, L.S. and Murphy A.H. (1992). Equitable scores for categorical forecasts. *Mon. Weather Rev.*, **120**, 361–370.

Garthwaite, P.H., Jolliffe, I.T. and Jones, B. (2002). *Statistical Inference*. Oxford: Oxford University Press.

Gerrity, J.P. Jr (1992). A note on Gandin and Murphy's equitable skill score. *Mon. Weather Rev.*, **120**, 2707–2712.

Gibson, J.K., Kållberg, P., Uppala, S., Hernandez, A. and Serano E. (1997). *ERA Description*. ECMWF Reanalysis project report series. ECMWF, Reading, UK.

Gilbert, G.K. (1884). Finley's tornado predictions. *Amer. Met. J.*, **1**, 166–172.

Golding, B.W. (1998). Nimrod: a system for generating automated very short range forecasts. *Met. Appl.*, **5**, 1–16.

Goodman, L.A. and Kruskal, W.H. (1959). Measures of association for cross classifications. II: Further discussion and references. *J. Amer. Statist. Assoc.*, **54**, 123–163.

Goodman, L.A. and Kruskal, W.H. (1979). *Measures of Association for Cross Classifications*. New York: Springer.

Green, D.M. and Swets, J.A. (1966). *Signal Detection Theory and Psychophysics*. Reprinted in 1974. New York: Robert E. Kreiger.

Gringorten, I.I. (1951). The verification and scoring of weather forecasts. *J. Amer. Statist. Assoc.*, **46**, 279–296.

Hamill, T.M. (1999). Hypothesis tests for evaluating numerical precipitation forecasts. *Weather Forecasting*, **14**, 155–167.

Hanssen, A.W. and Kuipers, W.J.A. (1965). On the relationship between the frequency of rain and various meteorological parameters. *Mededeelingen en Verhandelingen*, Royal Netherlands Meteorological Institute, **81**.

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

Hasselmann, K. (1997). Multi-pattern fingerprint method for detection and attribution of climate change, *Climate Dynam.*, **13**, 601–611.

Heidke, P. (1926). Berechnung der erfolges und der gute der windstarkevorhersagen im sturmwarnungdienst. *Geogr. Ann.*, **8**, 301–349.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecasting*, **15**, 559–570.

Hoffman, R.N. and Grassotti, C. (1996). A technique for assimilating SSM/I observations of marine atmospheric storms: tests with ECMWF analyses. *J. Appl. Met.*, **35**, 1177–1188.

Hoffman, R.N., Lui, Z., Louis, J.-F. and Grassotti, C. (1995). Distortion representation of forecast errors. *Mon. Weather. Rev.*, **123**, 2758–2770.

Hoffrage, U., Lindsey, S., Hertwig, R. and Gigerenzer, G. (2000). Communicating statistical information. *Science*, **290**, 2261–2262.

Hoffschildt, M., Bidlot, J.-R., Hansen, B. and Janssen, P.A.E.M. (1999). *Potential Benefit of Ensemble Forecasts for Ship Routing*, ECMWF Technical Memorandum 287, ECMWF, Reading, UK.

Hogg, R.V. and Tanis, E.A. (1997). *Probability and Statistical Inference*, 5th edition. Upper Saddle River, NJ: Prentice Hall.

Hollingsworth, A., Apre, K., Tiedke, M., Capaldo, M. and Savijarvi, H. (1980). The performance of a medium range forecast model in winter: impact of physical parameterizations. *Mon. Weather. Rev.*, **108**, 1736–1773.

Holloway, J.L. and Woodbury, M.A. (1955). *Application of Information Theory and Discriminant Function to Weather Forecasting and Forecast Verification*. Philadelphia: Institute for Cooperative Research, University of Pennsylvania.

Houtekamer, P.L., Lefaivre, L., Derome, J., Ritchie, H. and Mitchell, H.L. (1996). A system simulation approach to ensemble prediction. *Mon. Weather. Rev.*, **124**, 1225–1242.

Huberty, C.J. (1994). *Applied Discriminant Analysis*. New York: Wiley.

Hunt, B. and Hirst, A.C. (2000). Global climate models and their potential for seasonal climate forecasting. In: *Application of Seasonal Climate Forecasting in Agriculture and Natural Ecosystems: The Australian Experience* (eds: G.L. Hammer, N. Nicholls and C. Mitchell). Dordrecht: Kluwer, 89–107.

Huth, R. (1996). An intercomparison of computer-assisted circulation classification methods. *Int. J. Climatol.*, **16**, 893–922.

Johansson, A. and Saha, S. (1989). Simulation of systematic error effects and their reduction in a simple model of the atmosphere. *Mon. Weather. Rev.*, **117**, 1658–1675.

Jolliffe, I.T. (1999). An example of instability in the sample median. *Teaching Statistics*, **21**, 29.

Jolliffe, I.T. (2002). *Principal Component Analysis*, 2nd edition. New York: Springer.

Jolliffe, I.T. and Foord, J.F. (1975). Assessment of long-range forecasts. *Weather*, **30**, 172–181.

Jolliffe, I.T. and Jolliffe, N. (1997). Assessment of descriptive weather forecasts. *Weather*, **52**, 391–396.

Jones, D.A. (1998). *The Prediction of Australian Land Surface Temperatures Using Near Global Sea Surface Temperature Patterns*. BMRC Research Report No. 70.

Jones, R.H. (1985). Time series analysis – time domain. In: *Probability, Statistics and Decision Making in the Atmospheric Sciences* (eds. A.H. Murphy and R.W. Katz). Boulder, Colorado: Westview Press, 223–259.

Kaas, E., Guldberg, A., May, W. and Déqué, M. (1999). Using tendency errors to tune the parameterisation of unresolved dynamical scale interactions in atmospheric general circulation models. *Tellus*, **51A**, 612–629.

Katz, R.W. and Murphy, A.H. (eds.) (1997a). *Economic Value of Weather and Climate Forecasts*. Cambridge: Cambridge University Press.

Katz, R.W. and Murphy, A.H. (1997). Forecast value: prototype decision-making models. In: *Economic Value of Weather and Climate Forecasts* (eds. R.W. Katz and A.H. Murphy). Cambridge: Cambridge University Press, 183–217.

Katz, R.W., Murphy, A.H. and Winkler, R.L. (1982). Assessing the value of frost forecasts to orchardists: a dynamic decision-making approach. *J. Appl. Met.*, **21**, 518–531.

Köppen, W. (1884). Eine rationelle method zur prüfung der wetterprognosen. *Met. Z.*, **1**, 39–41.

Krzysztofowicz, R. (1995). Recent advances associated with flood forecast and warning systems. *Rev. Geophys.*, **33**, 1139–1147.

Krzysztofowicz, R. (2001). The case for probabilistic forecasting in hydrology. *J. Hydrol.*, **249**, 1–4.

Krzysztofowicz, R. and Long, D. (1991a). Forecast sufficiency characteristic: construction and application. *Int. J. Forecasting*, **7**, 39–45.

Krzysztofowicz, R. and Long, D. (1991b). Beta likelihood models of probabilistic forecasts. *Int. J. Forecasting*, **7**, 47–55.

Kumar, A., Hoerling, M. P., Ji, M., Leetmaa, A. and Sardeshmukh, P.D. (1996). Assessing a GCM's suitability for making seasonal prediction. *J. Climate*, **9**, 115–129.

Lee, P.M. (1997). *Bayesian Statistics: An Introduction*, 2nd edition. London: Arnold.

Leigh, R.J. (1995). Economic benefits of Terminal Aerodrome Forecasts (TAFs) for Sydney Airport, Australia. *Met. Apps.*, **2**, 239–247.

Leith, C.E. (1974). Theoretical skill of Monte-Carlo forecasts. *Mon. Weather Rev.*, **102**, 409–418.

Levi, K. (1985). A signal detection framework for the evaluation of probabilistic forecasts. *Organ. Behav. Hum. Decis. Processes*, **36**, 143–166.

Liljas, E. and Murphy, A.H. (1994). Anders Angstrom and his early papers on probability forecasting and the use/value of weather forecasts. *Bull. Amer. Met. Soc.*, **75**, 1227–1236.

Livezey, R.E. (1999). Field intercomparison. In: *Analysis of Climate Variability: Applications of Statistical Techniques*, updated and extended edition (eds. H. von Storch and A. Navarra). Berlin: Springer, 161–178.

Livezey, R.E. and Chen, W.Y. (1983). Statistical field significance and its determination by Monte Carlo techniques. *Mon. Weather Rev.*, **111**, 46–59.

Livezey, R.E., Hoopingarner, J.D. and Huang, J. (1995). Verification of official monthly mean 700-hPa height forecasts: an update. *Weather Forecasting*, **10**, 512–527.

Luce, R.D. (1963). Detection and recognition. In: *Handbook of Mathematical Psychology* (eds. R.D. Luce, R.R. Bush and E. Galanter). New York: Wiley, 103–189.

Mason, I.B. (1979). On reducing probability forecasts to yes/no forecasts. *Mon. Weather Rev.*, **107**, 207–211.

Mason, I.B. (1980). Decision-theoretic evaluation of probabilistic predictions. In: *WMO Symposium on Probabilistic and Statistical Methods in Weather Forecasting*, Nice, 8–12 September 1980, 219–228.

Mason, I.B. (1982a). A model for assessment of weather forecasts. *Austral. Met. Mag.*, **30**, 291–303.

Mason, I.B. (1982b). On scores for yes/no forecasts. In: *Preprints of papers delivered at the Ninth AMS Conference on Weather Forecasting and Analysis*, Seattle, Washington, 169–174.

Mason, I.B. (1989). Dependence of the Critical Success Index on sample climate and threshold probability. *Austral. Met. Mag.*, **37**, 75–81.

Mason, S.J. and Graham, N.E. (1999). Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather Forecasting*, **14**, 713–725.

Mason, S.J. and Mimmack, G.M. (2002). Comparison of some statistical methods of probabilistic forecasting of ENSO. *J. Climate*, **15**, 8–29.

McBride, J.L. and Ebert, E.E. (2000). Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. *Weather Forecasting*, **15**, 103–121.

McCoy, M.C. (1986). Severe-storm-forecast results from the PROFS 1983 forecast experiment. *Bull. Amer. Met. Soc.*, **67**, 155–164.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York: Wiley.

Metz, C.E. (1978). Basic principles of ROC analysis. *Seminars Nucl. Med.*, **8**, 283–298.

Miyakoda, K., Hembree, G.D., Strickler, R.F. and Shulman, I. (1972). Cumulative results of extended forecast experiments. I: Model performance for winter cases. *Mon. Weather. Rev.*, **100**, 836–855.

Miyakoda, K., Situtis, J. and Ploshay, J. (1986). One month forecast experiments – without anomaly boundary forcings. *Mon. Weather. Rev.*, **114**, 2363–2401.

Molteni, F. and Buizza, R. (1999). Validation of the ECMWF ensemble prediction system using empirical orthogonal functions. *Mon. Weather. Rev.*, **127**, 2346–2358.

Molteni, F., Buizza, R., Palmer, T.N. and Petroliagis, T. (1996). The ECMWF ensemble prediction system: methodology and validation. *Q. J. Roy. Met. Soc.*, **122**, 73–119.

Mullenmeister, P. and Hart, T. (1994). *The UNIX Verification Programs*. BMRC Research Report No. 40.

Muller, R.H. (1944). Verification of short-range weather forecasts (a survey of the literature). *Bull. Amer. Met. Soc.*, **25**, 18–27, 47–53, 88–95.

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

Murphy, A.H. (1973). A new vector partition of the probability score. *J. Appl. Met.*, **12**, 534–537.

Murphy, A.H. (1977). The value of climatological, categorical and probabilistic forecasts in the cost-loss ratio situation. *Mon. Weather Rev.*, **105**, 803–816.

Murphy, A.H. (1988). Skill scores based on the mean squared error and their relationships to the correlation coefficient. *Mon. Weather Rev.*, **116**, 2417–2424.

Murphy, A.H. (1991). Forecast verification: its complexity and dimensionality. *Mon. Weather Rev.*, **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.

Murphy, A.H. (1996). The Finley affair: a signal event in the history of forecast verification. *Weather Forecasting*, **11**, 3–20.

Murphy, A.H. (1997). Forecast verification. In: *The Economic Value of Weather and Climate Forecasts* (eds. R.W. Katz and A.H. Murphy). Cambridge: Cambridge University Press, 19–74.

Murphy A.H., Brown B.G. and Chen Y.-S. (1989). Diagnostic verification of temperature forecasts. *Weather Forecasting*, **4**, 485–501.

Murphy, A.H. and Daan, H. (1985). Forecast evaluation. In: *Probability, Statistics and Decision Making in the Atmospheric Sciences* (eds. A.H. Murphy and R.W. Katz). Boulder, Colorado: Westview Press, 379–437.

Murphy, A.H. and Ehrendorfer, M. (1987). On the relationship between the accuracy and value of forecasts in the cost–loss ratio situation. *Weather Forecasting*, **2**, 243–251.

Murphy A.H. and Epstein E.S. (1967a). Verification of probability predictions: a brief review. *J. Appl. Met.*, **6**, 748–755.

Murphy A.H. and Epstein E.S. (1967b). A note on probability forecasts and 'hedging'. *J. Appl. Met.*, **6**, 1002–1004.

Murphy, A.H. and Epstein, E.S. (1989). Skill scores and correlation coefficients in model verification. *Mon. Weather. Rev.*, **117**, 572–581.

Murphy, A.H. and Winkler, R.L. (1974). Credible interval temperature forecasting: Some experimental results. *Mon. Weather Rev.*, **102**, 784–794.

Murphy, A.H. and Winkler, R.L. (1977). Reliability of subjective probability forecasts of precipitation and temperature. *Appl. Statist.*, **26**, 41–47.

Murphy, A.H. and Winkler, R.L. (1984). Probability forecasting in meteorology. *J. Amer. Statist. Assoc.*, **79**, 489–500.

Murphy, A.H. and Winkler, R.L. (1987). A general framework for forecast verification. *Mon. Weather Rev.*, **115**, 1330–1338.

Murphy A.H. and Winkler R.L. (1992). Diagnostic verification of probability forecasts. *Int. J. Forecasting*, **7**, 435–455.

Nicholls, N. (1989). Sea-surface temperatures and Australian winter rainfall, *J. Climate*, **2**, 965–973.

Nicholls, N. (1992). Recent performance of a method for forecasting Australian seasonal tropical cyclone activity. *Austral. Met. Mag.*, **40**, 105–110.

North, G.R., Kim, K.Y., Chen, S.P. and Hardin, J.W. (1995). Detection of forced climate signals. Part I: Filter theory. *J. Climate*, **8**, 401–408.

Palmer, T.N., Brankovic, C. and Richardson, D.S. (2000). A probability and decision-model analysis of PROVOST seasonal multi-model integrations. *Q. J. Roy. Met. Soc.*, **126**, 2013–2033.

Palmer, T.N., Molteni, F., Mureau, R. and Buizza, R. (1993). Ensemble prediction. In: *ECMWF Seminar Proceedings, Validation of Models Over Europe*, vol. 1, ECMWF, Reading, UK.

Palmer, W.C. and Allen, R.A. (1949). *Note on the Accuracy of Forecasts Concerning the Rain Problem*. U.S. Weather Bureau manuscript, Washington, D.C.

Peirce, C.S. (1884). The numerical measure of the success of predictions. *Science*, **4**, 453–454.

Peng, P., Kumar, A., Barnston, A.G. and Goddard, L. (2000). Simulation skills of the SST-forced global climate variability of the NCEP-MRF9 and the Scripps-MPI ECHAM3 models, *J. Climate*, **13**, 3657–3679.

Pickup, M.N. (1982). A consideration of the effect of 500 mb cyclonicity on the success of some thunderstorm forecasting techniques. *Met. Mag.*, **111**, 87–97.

Potts, J.M. (1991). *Statistical Methods for the Comparison of Spatial Patterns in Meteorological Variables*. Unpublished Ph.D. thesis, University of Kent at Canterbury.

Potts, J.M., Folland, C.K., Jolliffe, I.T. and Sexton, D. (1996). Revised 'LEPS' scores for assessing climate model simulations and long-range forecasts. *J. Climate*, **9**, 34–53.

Ramage, C.S. (1982). Have precipitation forecasts improved? *Bull. Amer. Met. Soc.*, **63**, 739–743.

Renshaw, A.C., Rowell, D.P. and Folland, C.K. (1998). Wintertime low-frequency weather variability in the North Pacific-American sector 1949–93. *J. Climate*, **11**, 1073–1092.

Richardson, D.S. (2000). Skill and relative economic value of the ECMWF Ensemble Prediction System. *Q. J. Roy. Met. Soc.*, **126**, 649–668.

Richardson, D.S. (2001). Measures of skill and value of Ensemble Prediction Systems, their interrelationship and the effect of ensemble size. *Q. J. Roy. Met. Soc.*, **127**, 2473–2489.

Rodenberg, C. and Zhou, X.-H. (2000). ROC curve estimation when covariates affect the verification process. *Biometrics*, **56**, 1256–1262.

Roebber, P.J. and Bosart, L.F. (1996). The complex relationship between forecast skill and forecast value: a real-world analysis. *Weather Forecasting*, **11**, 544–559.

Ropelewski, C.F. and Halpert, M.S. (1986). North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). *Mon. Weather Rev.*, **114**, 2352–2362.

Ropelewski, C.F. and Halpert, M.S. (1987). Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation (ENSO). *Mon. Weather Rev.*, **115**, 1606–1626.

Roulston, M.S. and Smith, L. (2002). Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.*, **130**, 1653–1660.

Rowell, D.P. (1998). Assessing potential seasonal predictability with an ensemble of multidecadal GCM simulations. *J. Climate*, **11**, 109–120.

Rowell, D.P. and Zwiers, F.W. (1999). The global distribution of sources of atmospheric decadal variability and mechanisms over the tropical Pacific and southern North America. *Climate Dynam.*, **15**, 751–772.

Saha, S. and van den Dool, H.M. (1988). A measure of the practical limit of predictability. *Mon. Weather Rev.*, **116**, 2522–2526.

Sanders, F. (1958). *The Evaluation of Subjective Probability Forecasts*. Scientific Report No. 5, Department of Meteorology, Massachusetts Institute of Technology.

Sanders, F. (1963). On subjective probability forecasting. *J. Appl. Met.*, **2**, 191–201.

Santer, B.D., Wigley, T.M.L., Barnett, T.P., Anyamba, E. (1995a). Detection of climate change and attribution of causes. In: *Climate Change, The IPCC Second Scientific Assessment* (ed. J.T. Houghton). Cambridge: Cambridge University Press, 407–444.

Santer, B.D., Taylor, K.E., Wigley, T.M.L., Penner, J.E., Jones, P.D. and Cubash, U. (1995b). Towards the detection and attribution of an anthropogenic effect on climate. *Climate Dynam.*, **12**, 77–100.

Saporta, G. (1990). *Probabilités, Analyse de Données et Statistiques*. Paris: Technip.

Schaefer, J.T. (1990). The critical success index as an indicator of forecasting skill. *Weather Forecasting*, **5**, 570–575.

Schervish, M.J. (1989). A general method for comparing probability assessors. *Ann. Statist.*, **17**, 1856–1879.

Scott, D.W. (1979). On optimal and data-based histograms. *Biometrika*, **66**, 605–610.

Seaman, R., Mason, I. and Woodcock, F. (1996). Confidence intervals for some performance measures of yes/no forecasts. *Austral. Met. Mag.*, **45**, 49–53.

Seillier-Moiseiwitsch, F. and Dawid, A.P. (1993). On testing the validity of sequential probability forecasts. *J. Amer. Statist. Assoc.*, **88**, 355–359.

Slonim, M.J. (1958). The trentile deviation method of weather forecast evaluation. *J. Amer. Statist. Assoc.*, **53**, 398–407.

Smith, L.A. (2000). Disentangling uncertainty and error: on the predictability of nonlinear systems. In: *Nonlinear Dynamics and Statistics* (ed. A.I. Mees). Boston: Birkhauser, 31–64.

Stanski, H.R., Wilson, L.J. and Burrows, W.R. (1989). *Survey of Common Verification Methods in Meteorology*. World Weather Watch Technical Report No. 8. World Meteorological Organisation, Geneva.

Stephenson, D.B. (1997). Correlation of spatial climate/weather maps and the advantages of using the Mahalanobis metric in predictions. *Tellus*, **49A**, 513–527.

Stephenson, D.B. (2000). Use of the 'odds ratio' for diagnosing forecast skill. *Weather Forecasting*, **15**, 221–232.

Stephenson, D.B. and Doblas-Reyes, F.J. (2000). Statistical methods for interpreting Monte Carlo forecasts. *Tellus*, **52A**, 300–322.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. B*, **36**, 111–147.

Sturges, H.A. (1926). The choice of a class interval. *J. Amer. Statist. Assoc.*, **21**, 65–66.

Swets, J.A. (1973). The relative operating characteristic in psychology. *Science*, **182**, 990–1000.

Swets, J.A. (1986a). Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychol. Bull.*, **99**, 100–117.

Swets, J.A. (1986b). Form of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychol. Bull.*, **99**, 181–198.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Swets, J.A. (1996). *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. Mawah, New Jersey: Lawrence Erlbaum.

Swets, J.A. and Pickett, R.M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. New York: Academic Press.

Talagrand, O., Vautard, R. and Strauss, B. (1998). Evaluation of probabilistic prediction systems. In: *Proceedings of ECMWF Workshop on Predictability*, 20–22 October 1997, 1–25.

Tanner, W.P. Jr and Birdsall, T.G. (1958). Definitions of d' and η as psychophysical measures. *J. Acoust. Soc. Amer.*, **30**, 922–928.

Tay, A.S. and Wallis, K.F. (2000). Density forecasting: a survey. *J. Forecasting*, **19**, 235–254.

Taylor, J.W. (1999). Evaluating volatility and interval forecasts. *J. Forecasting*, **18**, 111–128.

Taylor, J.W. and Buizza, R. (2003). Using weather ensemble predictions in electricity demand forecasting. *Int. J. Forecasting*, **19**, to appear.

Teweles, S. and Wobus, H.B. (1954). Verification of prognostic charts. *Bull. Amer. Met. Soc.*, **35**, 455–463.

Thompson, J.C. (1952). On the operational deficiencies in categorical weather forecasts. *Bull. Amer. Met. Soc.*, **33**, 223–226.

Thornes, J.E. and Stephenson, D.B. (2001). How to judge the quality and value of weather forecast products. *Met. Apps.*, **8**, 307–314.

Toth, Z. (1991). Intercomparison of circulation similarity measures. *Mon. Weather Rev.*, **119**, 55–64.

Toth, Z. and Kalnay, E. (1997). Ensemble forecasting at NCEP and the breeding method. *Mon. Weather Rev.*, **125**, 3297–3319.

Toth, Z., Zhu, Y., Marchok, T., Tracton, S. and Kalnay, E. (1998). Verification of the NCEP global ensemble forecasts. In: *Preprints of the 12th Conference on Numerical Weather Prediction*, 11–16 January 1998, Phoenix, Arizona, 286–289.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison Wesley.

Venkatraman, E.S. (2000). A permutation test to compare Receiver Operating Characteristic curves. *Biometrics*, **56**, 1134–1138.

von Storch, H. and Zwiers, F.W. (1999). *Statistical Analysis in Climate Research*. Cambridge: Cambridge University Press.

Wallis, K.F. (1995). Large-scale macroeconomic modelling. In: *Handbook of Applied Econometrics* (eds M.H. Pesaran and M.R. Wickens). Oxford: Blackwell, 312–355.

Wang, X.L. and Rui, H.L. (1996). A methodology for assessing ensemble experiments, *J. Geophy. Res.*, **101-D23**, 29 591–29 597.

Ward, M.N. and Folland, C.K. (1991). Prediction of seasonal rainfall in the north Nordeste of Brazil using eigenvectors of sea-surface temperature. *Int. J. Climatol.*, **11**, 711–743.

Wei, M. and Toth, Z. (2002). Ensemble perturbations and forecast errors. Submitted for publication.

Weymouth, G., Mills, G.A., Jones, D., Ebert, E.E. and Manton, M.J. (1999). A continental-scale daily rainfall analysis system. *Aust. Meteor. Mag.*, **48**, 169–179.

Wigley, T.M.L. and Santer, B.D. (1990). Statistical comparison of spatial fields in model validation, perturbation, and predictability experiments. *J. Geophys. Res.*, **95**, 851–865.

Wilks, D.S. (1995). *Statistical Methods in the Atmospheric Sciences: An Introduction*. San Diego: Academic Press.

Wilks, D S. and Hamill, T.M. (1995). Potential economic value of ensemble forecasts. *Mon. Weather Rev.*, **123**, 3565–3575.

Wilson, E.B. (1927). Probable inference, the law of succession, and statistical inference. *J. Amer. Statist. Assoc.*, **22**, 209–212.

Wilson, L. (2000). Comments on 'Probabilistic predictions of precipitation using the ECMWF ensemble prediction system'. *Weather Forecasting*, **15**, 361–364.

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

Winkler, R.L. (1969). Scoring rules and the evaluation of probability assessors. *J. Amer. Statist. Assoc.*, **64**, 1073–1078.

Winkler, R.L. and Murphy, A.H. (1968). 'Good' probability assessors. *J. Appl. Met.*, **7**, 751–758.

Woodworth, R.S. (1938). *Experimental Psychology*. New York: Holt, Rinehart and Winston.

Xie, P. and Arkin, P.A. (1996). Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *J. Climate*, **9**, 840–858.

Youden, W.J. (1950). Index for rating diagnostic tests. *Cancer*, **3**, 32–35.

Yule, G.U. (1900). On the association of attributes in statistics. *Philos. Trans. Roy. Soc. London*, **194A**, 257–319.

Zheng, X. and Frederiksen, C.S. (1999). Validating interannual variability in an ensemble of AGCM simulations. *J. Climate*, **12**, 2386–2396.

Zhu, Y., Iyengar, G., Toth, Z., Tracton, M.S. and Marchok, T. (1996). Objective evaluation of the NCEP global ensemble forecasting system. In: *Preprints of the 15th AMS conference on Weather Analysis and Forecasting*, 19–23 August 1996, Norfolk, Virginia, J79–J82.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D. and Mylne, K. (2002). The economic value of ensemble-based weather forecasts. *Bull. Amer. Met. Soc.*, **83**, 73–83.

Zwiers, F.W. (1999). The detection of climate change. In: *Anthropogenic Climate Change* (eds H. van Storch and G. Flöser). Heidelberg: Springer, 161–206.

Author Index

Subject Index