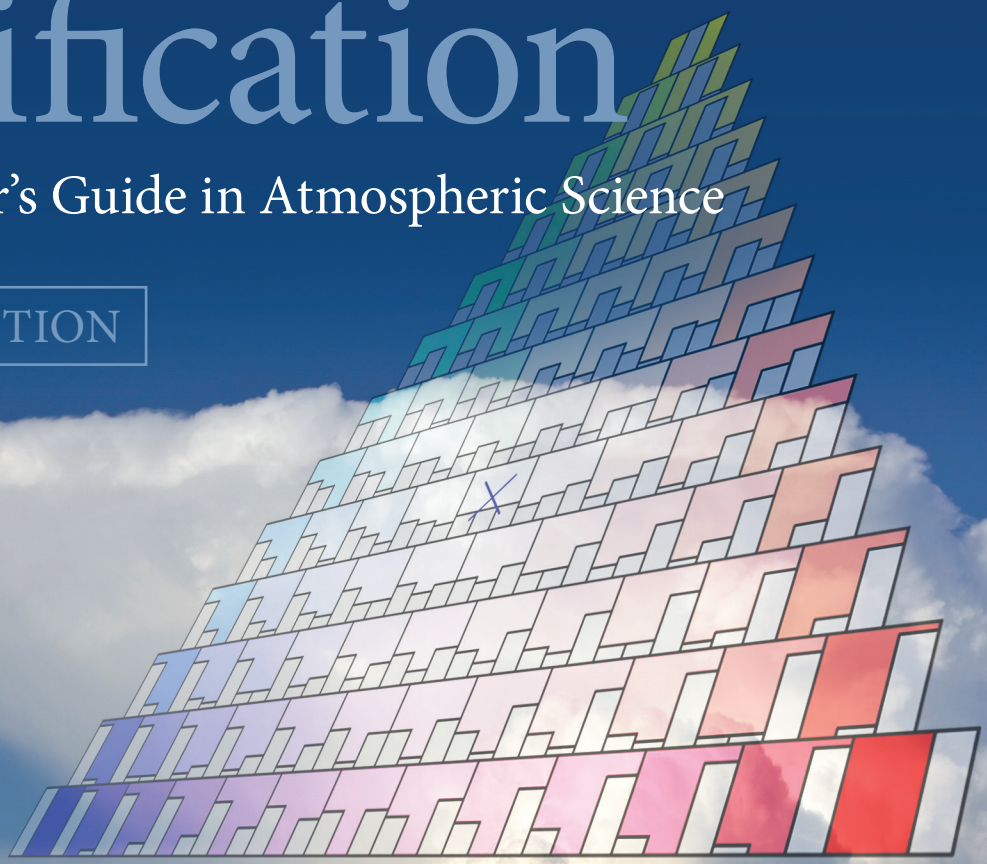
Editors
Ian T. Jolliffe | David B. Stephenson

Forecast Verification

A Practitioner's Guide in Atmospheric Science

SECOND EDITION



WILEY-BLACKWELL

Forecast Verification

Forecast Verification

A Practitioner's Guide in Atmospheric Science

SECOND EDITION

Edited by
Ian T. Jolliffe
David B. Stephenson
University of Exeter, UK

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

This edition first published 2012

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

Registered office
John Wiley & Sons, Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

Editorial offices
9600 Garsington Road, Oxford, OX4 2DQ, UK
The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK
111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell.

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. – 2nd ed.
p. cm.
Includes index.
ISBN 978-0-470-66071-3 (cloth)
1. Weather forecasting–Statistical methods–Evaluation. I. Jolliffe, I. T. II. Stephenson, David B.
QC996.5.F67 2011
551.63–dc23

2011035808

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

This book is published in the following electronic formats: ePDF 9781119960010; Wiley Online Library 9781119960003; ePub 9781119961079; Mobi 9781119961086

Set in 10/12 pt Times by Aptara Inc., New Delhi, India

First Impression 2012

Contents

List of contributors

Dr Jochen Broecker
Max-Planck-Institute for the Physics of Complex Systems, Noethnitzer Str. 38, 01187 Dresden, Germany
broecker@pks.mpg.de

Dr Barbara G. Brown
Research Applications Laboratory, National Center for Atmospheric Research, P.O. Box 3000, Boulder CO 80307-3000, USA
bgb@ucar.edu

Michel Déqué
Météo-France CNRM,CNRS/GAME, 42 Avenue Coriolis, 31057 Toulouse Cedex 01, France
deque@meteo.fr

Dr Elizabeth E. Ebert
Centre for Australian Weather and Climate Research (CAWCR), Bureau of Meteorology, GPO Box 1289, Melbourne, Victoria 3001, Australia
e.ebert@bom.gov.au

Dr Christopher A.T. Ferro
Mathematics Research Institute, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QF, UK
c.a.t.ferro@exeter.ac.uk

Dr Eric Gilleland
Research Applications Laboratory, National Center for Atmospheric Research, P.O. Box 3000, Boulder CO 80307-3000, USA
ericg@ucar.edu

Professor Robin Hogan
Department of Meteorology, University of Reading, P.O. Box 243, Reading RG6 6BB, UK
r.j.hogan@reading.ac.uk

Professor Ian Jolliffe
30 Woodvale Road, Gurnard, Cowes, Isle of Wight, PO31 8EG, UK
ian@sandloch.fsnet.co.uk

Dr Robert E. Livezey
5112 Lawton Drive, Bethesda, MD 20816, USA
bobbilbo@msn.com

Dr Ian B. Mason
32 Hensman St., Latham, ACT, Australia, 2615
ibmason@grapevine.com.au

Dr Simon J. Mason
International Research Institute for Climate and Society (IRI), Columbia University, 61 Route 9W, P.O. Box 1000, Palisades, NY 10964-8000, USA
simon@iri.columbia.edu

Dr Matt Pocernich
Research Applications Laboratory, National Center for Atmospheric Research, P.O. Box 3000, Boulder CO 80307-3000, USA
e-mail: matt_pocernich@hotmail.com

Dr Jacqueline M. Potts
Biomathematics and Statistics Scotland, Craigiebuckler, Aberdeen AB15 8QH, UK
jackie@bioss.ac.uk

David S. Richardson
European Centre for Medium-Range Weather Forecasts (ECMWF), Shinfield Park, Reading, RG2 9AX, UK
david.richardson@ecmwf.int

Professor David B. Stephenson
Mathematics Research Institute, College of Engineering, Mathematics and Physical Sciences, University of Exeter, Harrison Building, North Park Road, Exeter EX4 4QF, UK
d.b.stephenson@exeter.ac.uk

Dr Andreas P. Weigel
Federal Office of Meteorology and Climatology MeteoSwiss, Kraehbuehlstr. 58, P.O. Box 514, CH-8044 Zurich, Switzerland
andreas.weigel@alumni.ethz.ch

Preface

In the eight years since the first edition was published, there has been considerable expansion of the literature on forecast verification, and the time is ripe for a new edition. This second edition has three more chapters than the first, as well as a new Appendix and substantially more references. Developments in forecast verification have not been confined to the atmospheric science literature but, as with the first edition, we concentrate mainly on this area.

As far as we are aware, there is still no other book that gives a comparable coverage of forecast verification, although at least two related books have appeared outside the atmospheric science area. Pepe (2003) is concerned with evaluation of medical diagnostic tests, which, although essentially concerned with 'forecast verification', has a very different emphasis, whilst Krzanowski and Hand (2009) is more narrowly focused on ROC curves.

We have retained many of the authors from the first edition, as well as bringing in a number of other experts, mainly for the new chapters. All are well-regarded researchers and practitioners in their fields. Shortly after the first edition was published, an extended and constructive review appeared (Glahn, 2004; Jolliffe and Stephenson, 2005). In this new edition we and our authors have attempted to address some of the issues raised by Glahn.

Compared with the first edition, the introductory and scene-setting Chapters 1 and 2 have only minor changes. Chapter 3 on 'Deterministic forecasts of binary events' has gained an additional author and has been rewritten. Much material from the first edition has been retained but has been restructured, and a non-trivial amount of new material, reflecting recent developments, has been added. Chapters 4 and 5 on, respectively, 'Deterministic forecasts of multi-category events' and 'Deterministic forecasts of continuous variables' have only minor improvements.

One of the biggest areas of development in forecast verification in recent years has been for spatial forecasts. This reflected by a much-expanded Chapter 6 on the topic, with three new authors, all of whom are leaders in the field.

In the first edition, probability forecasts and ensemble forecasts shared a chapter. This is another area of active development and, as suggested by Glahn (2004) and others, the two topics have been separated into Chapters 7 and 8 respectively, with two new authors. Chapter 9 on 'Economic value and skill' has only minor changes compared to the first edition.

Chapters 10 and 11 are both new, covering areas that have seen much recent research and are likely to continue to do so. Chapter 10 covers the related topics of verification of forecasts for rare and extreme events, and verification of weather warnings. By their nature the latter are often extreme, though many types of warnings are issued for events that are not especially rare. Impact rather than rarity is what warrants a warning. One context in which extremes are of particular interest is that of climate change. Because of the lack of verifying observations, the topic of verification of climate projections is still in its infancy, though likely to develop. There

has been more activity on verification of seasonal and decadal forecasts, and these together with verification of climate projections, are the subject of Chapter 11.

The concluding Chapter 12 reviews some key concepts, summarizes some of the verification/evaluation activity in disciplines other than atmospheric sciences, and discusses some of the main developments since the first edition. As with the first edition, a Glossary is provided, and in addition there is an Appendix on available software. Although such an Appendix inevitably becomes out of date more quickly than other parts of the text, it is arguably the most useful part of the book to practitioners for the first few years after publication. To supplement the Appendix, software and data sets used in the book will be provided via our book website: http://emps.exeter.ac.uk/fvb. We also intend to use this website to record errata and suggestions for future additions.

We hope you enjoy this second edition and find it useful. If you have any comments or suggestions for future editions, we would be happy to hear from you.

Ian T. Jolliffe
David B. Stephenson

Preface to the first edition

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 (2006b [formerly 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. The current book provides a broad coverage, although it does not attempt to be encyclopedic, 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 squared 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 some 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. However, 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 summarises the 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 we would like

to thank the following colleagues for their particularly helpful comments and contributions: 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 out while one us (I.T.J.) 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 and David B. Stephenson
Mathematics Research Institute, University of Exeter

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). The third section 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 current 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 (1996a) notes that although 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

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

Table 1.1 Finley's tornado forecasts

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

10 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 (1996a).

Forecast verification was not a very active branch of research in the first half of the twentieth century. A three-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 (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 (1991a) 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) that gave a comprehensive survey of forecast verification methods in use in the late 1980s. Other WMO documentation is noted in the next subsection.

1.1.2 Current practice

The WMO provides a Standard Verification System for Long-Range Forecasts. At the time of writing versions of this are available at a number of websites. The most up-to-date version is likely to be found through the link to the User's Guide on the website of the Lead Centre for the Long Range Forecast Verification System (http://www.bom.gov.au/wmo/lrfvs/users.shtml). 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. An earlier version is also available as attachments II-8 and II-9 in the WMO *Manual on the Global Data-Processing System* (http://www.wmo.int/pages/prog/www/DPS/Manual/WMO485.pdf). Attachment II-7 in the same document discusses methods used in standardized verification of NWP (Numerical Weather Prediction) products. Two further WMO documents can be found at http://www.wmo.int/pages/prog/amp/pwsp/pdf/TD-1023.pdf and http://www.wmo.int/pages/prog/amp/pwsp/pdf/TD-1103.pdf. These are respectively Guidelines (and Supplementary Guidelines) on Performance Assessment of Public Weather Services. The

latter is discursive in nature, whilst the guidelines in the former are more technical in nature.

European member states report annually on verification of ECMWF (European Centre for Medium Range Weather Forecasts) forecasts in their national weather services, and guidance on such verification is given in ECMWF Technical Memorandum 430 by Pertti Nurmi (http://www.ecmwf.int/publications/library/ecpublications/_pdf/tm/401-500/tm430.pdf).

At a national level, verification practices vary between different National Services, and most use a range of different verification strategies for different purposes. For example, verification scores used at the time of writing by the National Climate Centre at the Bureau of Meteorology in Australia range through many of the chapters that follow, for example proportion correct (Chapter 3), LEPS scores (Chapter 4), root mean square error (Chapter 5), anomaly correlation (Chapter 6), Brier skill score (Chapter 7) and so on (Robert Fawcett, personal communication).

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 30 years ago, are now commonplace (Chapter 8), and the verification of spatial forecasts has advanced significantly (Chapter 6). 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 increasing use of remote sensing by satellite and radar, for example.

An important part of any operational verification system is to have software to implement the system. As well as the widely available software described in Appendix, national weather services often have their own systems. For example, the Finnish Meteorological Institute has a comprehensive operational verification package, which is regularly updated (Pertti Nurmi, personal communication).

A very useful resource is the webpage of the Joint Working Group on Forecast Verification Research (http://www.cawcr.gov.au/projects/verification/). It gives a good up-to-date overview of verification methods and issues associated with them, together with information on workshops and other events related to verification.

1.2 Reasons for forecast verification and its benefits

There are three main reasons for verification, whose description 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 are 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 that 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. In contrast, 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' or 'user-oriented' verification, as it provides information in terms more likely to be understood by the 'customer' or 'user' than a purely 'scientific' approach. Forecast verification using economic value is discussed in detail in Chapter 9. Another aspect of forecasting for specific users is the extent to which users prefer a simple, less informative forecast to one that is more informative (e.g. a probability forecast) but less easy to interpret. Some users may be uncomfortable with probability forecasts, but there is evidence (Harold 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}. User-oriented verification should attempt to ascertain such preferences for the user or 'customer' at hand.

A 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.

This section has been written from the viewpoint of verification of forecasts issued by National Meteorological Services. Virtually all the points made are highly relevant for forecasts issued by private companies, and in other subject domains, but it appears that they may not always be appreciated. Although most National Weather Services verify their forecasts, the position for commercially provided forecasts is more patchy. Mailier *et al*. (2008) reported the findings of a survey of providers and users of commercial weather forecasts in the UK. The survey and related consultations revealed that there were 'significant deficiencies in the methodologies and in the communication of forecast quality assessments' and that 'some users may be indifferent to forecast quality'.

1.3 Types of forecast 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 subdivisions 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 subdivision) 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 a 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 redefinitions 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' that 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.

The chapters of this book cover all the main types of forecasts that require verification, but less common types are not covered in detail. For example, forecasts of wind direction lie on a circle rather than being linearly ordered and hence need different treatment. Bao *et al*. (2010) discuss verification of directional forecasts when the variable being forecast is continuous, and there are also measures that modify those of Chapter 4 when forecasts fall in a small number of categories (Charles Kluepfel, personal communication)

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 later chapters, in particular Chapter 2. 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' of forecasts has many facets: Murphy (1993) identifies three types of goodness:

- Consistency (the correspondence between forecasters' judgements and their forecasts).
- Quality (the correspondence between the forecasts and matching observations).
- Value (the incremental economic and/or other benefits realized by decision-makers through the use of the forecasts).

It seems desirable that the forecaster's best judgement and the forecast actually issued coincide. Murphy (1993) describes this as 'consistency', though confusingly the same word has a narrower definition in Murphy and Daan (1985) – see Chapter 2. 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) from his or her best judgement of what will occur. In that case, the forecaster will be tempted to *hedge,* that is, to forecast something other than his or her best judgement (Murphy, 1978), especially if the forecaster's pay depends on the score. Thus administrators should avoid measuring or rewarding forecasters' performance on the basis of such scoring schemes, as this is likely to lead to biases in the forecasts.

The emphasis in this book is on quality – the correspondence between forecast and observations. Value is concerned with economic worth to the user. Chapter 9 discusses value and its relationship to quality.

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, e.g., Sections 2.7, 3.4, 4.3 and 11.3.1).

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 naive 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, e.g., 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' is in fairly common use, but can have different meanings to different authors. 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. A wider definition includes any prediction made that is not a genuine forecast of a *future* event. With this usage, a prediction for the year 2010 must be a hindcast, even if it is only based on data up to 2009, because year 2010 is now over. The term *retroactive forecasting* is used by Mason and Mimmack (2002) to denote the form of hindcasting in which forecasts are made for past years (e.g. 2006–2010) using data prior to those years (perhaps 1970–2005).

The terminology *ex ante* and *ex post* is used in business 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 for a number of measures or scores are described in Section 3.5.2, and several other chapters 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, for example, Section 4.4.

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 (e.g., protect or don't protect a crop from frost), the baseline can be the best from the list of decision choices ignoring any forecast (e.g., 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.

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. Sometimes the 'verification observations' are not observations at all, but are based on analyses from very-short-range forecast models. This may be necessary if genuine observations are sparse and not conveniently spaced geographically in relation to the forecasts. A common problem is that forecasts may be spatially continuous or on a grid, but observations are available only for an irregular set of discrete spatial points. This is discussed further in Section 6.2.

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. There has been some work, too, on the effect of observation errors on verification scores in a meteorological context. For example, Bowler (2008) shows that the apparent skill of a forecasting system can be reduced by the equivalent of one day in forecast lead time.

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. However, 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. Alternatively, measures may be devised specifically for verification of forecasts or warnings of extreme events – see Chapter 10.

A final practical consideration is that there can be confusion over terminology. This is partly due to the development of verification in several different disciplines, but even within atmospheric science different terms can be used for the same thing, or the same term (or very similar terms) used for different things. For example, *false alarm rate* and *false alarm ratio* are different measures for binary deterministic forecasts (see Chapter 3), but are easily confused. Barnes *et al*. (2009) found that of 26 peer-reviewed articles published in American Meteorological Society journals between 2001 and 2007 that used one or both of the measures, 10 (38%) defined them inconsistently with the currently accepted definitions. The glossary in this book will help readers to avoid some of the pitfalls of terminology, but care is still needed in reading the verification literature.

Even the word 'verification' itself is almost unknown outside of atmospheric science. In other disciplines 'evaluation' and 'assessment' are more common. It seems likely that Finley's use of the phrase 'verification of predictions' in 1884 is the historical accident that led to its adoption in atmospheric science, but not elsewhere.

1.6 Summary

As described in Section 1.2, verification has three main uses:

- Administrative: to monitor performance over time and compare the forecast quality of different prediction systems.
- Scientific: to diagnose the drivers of performance and inform improvements in prediction systems.
- Economic: to build credibility and customer confidence in forecast products by demonstrating that predictions have economic value to users.

Verification is therefore an indispensible part of the development cycle of prediction systems. With increasing complexity and sophistication of forecasts, verification is an active area of scientific research – see, e.g., the review by Casati *et al*. (2008), which is part of a special issue of *Meteorological Applications* on forecast verification. Subsequent chapters of the book give an introduction to some of the exciting developments in the subject, as well as giving a clear grounding in the more established methodology.