Quality of Weather Forecasts

Review and recommendations

Authors:

Pascal J. Mailier
Ian T. Jolliffe
David B. Stephenson

January 2006

The project contractors are: Pascal J. Mailier (project manager and principal investigator), Ian T. Jolliffe (first co-investigator) and David B. Stephenson (second co-investigator).

For reasons of confidentiality, none of the names of respondents to the survey, interviewees and providers of the weather forecast data used for the case studies have been revealed in this report. The locations for which the forecasts in the case studies were made have not been disclosed.

SUMMARY

This study has examined current issues regarding the quality (fitness for purpose) of commercial weather forecasts in the United Kingdom, with particular attention given to problems arising from inadequate quality assessment and from the lack of generally agreed standards. Forecast providers and forecast users have been consulted by means of on-line questionnaires, interviews, visits and an open workshop discussion. Results from this consultation have highlighted significant deficiencies in the methodologies and in the communication of forecast quality assessments. The consultation has also revealed that the open dialogue and transparency required to establish commonly agreed standards in the industry are lacking. Moreover, there is evidence that some users are indifferent to forecast quality. A comprehensive review of existing forecast verification methodologies and metrics has been conducted. This review has shown that suitable quality assessment methods are available for nearly all types of quantitative forecasts identified in the consultation. Descriptive or worded forecasts, however, cannot be assessed objectively. A very important finding is that apparently simple and easy-to-understand metrics may have statistical properties that can result in misleading assessments of forecast quality. Furthermore, not enough care is taken to estimate the uncertainty (statistical significance) of quality assessment results. The crucial importance of choosing proper metrics, the impact of their statistical properties on results and the need to estimate statistical significance have been exemplified in four case studies. The findings from the consultation, the literature review, and the lessons learnt from the case studies have led to a set of practical recommendations. These guidelines, which are based on sound scientific principles, aim at establishing the discipline and rigour that are necessary for achieving best practice in the quality assessment of weather forecasts. Specific recommendations have also been made to the Royal Meteorological Society to set up a Special Commission that will promote a sense of community within the industry, and to run an accreditation scheme that will encourage best practice on a voluntary basis.

ACKNOWLEDGEMENTS

We wish to thank the following persons for providing constructive feedback, and for giving us support and encouragement:

- The Members of the Project Steering Group, more especially Dr Peter Ryder (Chairman), Dr Richard Pettifer, Prof Chris Collier, Dr Kirby James, Dr Andrew Eccleston and Dr David Pick;
- The Members of the National Meteorological Service Commissioning Group;
- Dr R. Gary Rasmussen, Director of Enterprise Activity Support, American Meteorological Society;
- Dr Beth Ebert, the Australian Bureau of Meteorology;
- Dr David Forrester and Dr Clive Wilson, the Met Office;
- Prof Brian Hoskins, the University of Reading;
- Mr Chris Blowes, Weather Commerce Ltd;
- Mr Peter Hollingsworth, Metra Information Ltd;
- Mrs Kathy Riviere, the Royal Meteorological Society;
- Dr Pertti Nurmi, the Finnish Meteorological Institute,
- Dr Paul Berrisford, the University of Reading.

We are also grateful to the following contributors:

- Dr Warwick Norton of Weather Informatics Ltd;
- Mr Etienne Moser and Mr Eric Lefèvre of EDF Trading Ltd;
- Mr Dave Parker and Mr Matthew Pollard of EDF Energy Ltd;
- Dr Adam Scaife of the Met Office.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

1.1. Background and motivation

The problem of how to assess the performance of weather forecasts has been a long-standing topic in meteorology. Since 1950, progress in numerical weather prediction (NWP) has prompted meteorologists to develop and apply many different objective verification and evaluation techniques to forecast fields. A comprehensive review and discussion of standard methods was compiled recently by Jolliffe and Stephenson (2003). Over the past two decades, the use of commercial weather forecast products by industrial decision makers has become much more widespread. In parallel the range of competing weather forecast services to choose from has widened. As a result there is a pressing need to include the user's perspective in the verification methodology (referred to as *user-oriented verification* in the recent WMO guidance report by Ebert et al., 2005).

Because the mainstream verification methods were primarily devised to answer the needs of model developers, it is not surprising that a common concern in the literature on the subject is to monitor and improve a forecast system following requirements posed by atmospheric scientists, rather than the needs of specific end-users who are more interested in the performance of a set of forecasts relative to their own specific demands (see e.g. Glahn, 2004).

Another reason for the lack of research in assessing the 'usability' of weather forecasts is the belief that users are only concerned with economic value (Katz and Murphy, 1997). In theory, this would be the case if all users were able to formulate exactly their own utility functions. However, utility functions are often very difficult to determine in practice, and real-world decision-making processes are often a lot more complicated than the idealised cost-loss models used in the meteorological literature. For these reasons, many users often find it simpler to look at the forecast 'fitness for purpose', which in this work will be referred to as *forecast quality*. It goes without saying that good meteorological performance constitutes an essential 'must-have' characteristic of weather forecasts, so that a proper quality assessment strategy cannot be envisaged without due consideration of relevant meteorological parameters. Of course, principles that constitute best practice for meteorological quality control can also be extended to a wider range of verification techniques, including the assessment of economic value.

Nowadays forecast users are able to receive weather forecasts from many different sources. The spread of quality between all available products is considerable, and in some cases the origin of the forecasts is obscure. Even when performance statistics are available from competing providers, in general those measures of quality are not presented in forms that allow immediate comparison, and they do not relate directly to the practical applications for which the forecasts are supplied.

Typically, it is the suppliers rather than the users who assess the quality of weather forecasts. This situation has the potential to lead to conflicts of interests that may affect the credibility of the whole industry. There is therefore an

obvious need to develop independent standards of practice that are scientifically sound, adhered to by forecast suppliers, and trusted by their customers.

In contrast with the developer's viewpoint, the wide range of weather forecast products, the diversity of applications and customer requirements, and the differing nature of delivery systems complicate assessing the quality of weather forecasts from a single user's perspective. Although research in forecast verification is continually developing new methodology for new products, the complexity of this problem goes well beyond what has been previously addressed by the classical forecaster-oriented verification methodology.

1.2. Aim and objectives of this project

This project examines fundamental issues raised by current verification practice and by the lack of generally agreed standards. Its aim is to present practical solutions and make feasible proposals that will allow the industry to tackle these problems in a way that is beneficial to both forecast users and providers.

The project objectives can be summarised as follows:

1. *Identify and select existing verification methods and metrics;*
2. *Develop new approaches and appropriate metrics where needed;*
3. *Communicate the project findings and make the necessary recommendations to the Royal Meteorological Society;*
4. *Propose standard and recommended practices, and a scheme to monitor and encourage their use within the industry.*

1.3. Outline of the report

The core of this report has been organised in three self-contained units - Parts I, II and III - that can be read independently. Text boxes placed at the beginning of each unit help the reader to navigate through the material. A fourth unit -Part IV- is reserved for ancillary material.

Part I (Section 2) gives a detailed account of the consultation that has been carried out to ascertain the current state of affairs in the UK weather forecasting industry. First, the scene is set with a brief description of the UK market, and definitions are given for what is meant in this study by forecast provider, forecast user and forecast quality assessment. After a short explanation of the preliminary work that was required to set up and run the on-line questionnaires, the results of the survey are presented and analysed in detail. Part I finishes with an account and a critical discussion of statements gathered during visits, interviews and a thematic workshop. A summary of the results from this consultation is provided in Part III, Subsection 5.1.

Part II deals with the technical aspects of forecast quality assessment. It contains two sections (Section 3 and Section 4) that complement each other. Section 3, which is of a theoretical nature, provides a comprehensive literature review of

available methodologies and metrics. Quantitative and descriptive forecasts have been considered. As the verification of quantitative forecasts has become a rapidly expanding area of active research, a few paragraphs have been dedicated to the most recent developments in this domain. In view of the surge of interest in high-impact forecasts and the difficulties experienced in assessing their performance, it has also been deemed relevant to include a subsection dealing with forecasts of rare and extreme events. Section 4 demonstrates and consolidates some of the key concepts highlighted in the review by means of practical applications. It provides the reader with concrete examples of what constitutes good and bad practice. The main findings from Part II are summarised in Part III, Subsection 5.2 (Section 3) and Subsection 5.3 (Section 4).

Part III sums up the important findings from the consultation, review and case studies (Section 5), provides recommendations (Section 6) and concludes with some suggestions for the future (Section 7). The recommendations are subdivided in two sets. The first set (Subsection 6.1 and Subsection 6.2) is a list of recommendations on methodology which is meant for forecast providers and users. It aims to give general and specific guidelines for sound quality assessment practice. The second set of recommendations (Subsection 6.3) is for the Royal Meteorological Society and suggests a course of action to put in place a system that will facilitate and foster adherence to best practice in the quality assessment of weather forecasts.

Part IV contains the bibliography and appendices. The first two appendices show the online survey questionnaires used in the consultation with results for forecast providers (Appendix A) and forecast users (Appendix B). Appendix C provides references to existing guidelines for forecast quality assessment.

PART I: CONSULTATION

In this part of the report:

- Results from the consultation (on-line survey, visits and interviews, workshop) are presented (Section 2).

Note:

- Highlights of the consultation are given in Subsection 5.1.

2. WEATHER FORECASTING IN THE UNITED KINGDOM

2.1. The key market agents: forecast providers and forecast users

The largest supplier of meteorological products in the United Kingdom is the Met Office: the British national weather service funded by Government and through revenue from contracts with the public and private sectors. Besides developing and maintaining their own NWP models as the basis of their services, the Met Office provides services based on the operational products of the European Centre for Medium-Range Weather Forecasts (ECMWF). The Met Office also owns shares in WeatherXchange, a private venture that supplies weather products to customers in the trade of energy and weather derivatives.

In parallel, a growing number of entirely private companies also compete to sell weather forecasts to a wide range of end users. Although a few of these companies are able to some extent to run their own proprietary numerical models, most of them need access to basic numerical model outputs from national weather agencies or ECMWF. For example, operational forecast products from the American Global Forecasing System (GFS) can be acquired free of charge from the US National Centers for Environmental Predictions (NCEP). Private companies can also buy products from the Met Office (forecasts and observations) under conditions laid down by ECOMET, the interest consortium of the national meteorological services of the European Economic Area (http://www.meteo.oma.be/ECOMET/).

As mentioned in Subsection 1.1, most weather forecasting agencies and companies carry out various forms of quality control of the forecasts that they produce. But there is no agreement on common procedures and standards to facilitate fair comparisons between competing forecasts. The creation of such standards on forecast quality is also justified by the fact that in some cases, the source of the forecasts provided by a company is obscure.

In the context of this study, we define *forecast provider* to be an agent (national weather service, private person or company) who supplies weather forecasts in exchange for payment. Suppliers of freely available forecast products are not included in this definition. Similarly, a *forecast user* is an individual or organisation (private or public) that *buys* weather forecasts from one or more forecast providers. Again, we exclude ‘users’ who rely entirely on freely available forecasts.

Routine forecast quality control is usually performed by model developers and providers themselves, typically by comparing actual against predicted variables such as 500-hPa geopotential heights, temperatures at various levels in the atmosphere, accumulated precipitation or mean surface wind speed. This form of appraisal is only concerned with the meteorological aspect of the forecasts, and does not necessarily address its practical usefulness for a decision maker. Forecast performance assessment methods used by model developers and providers are commonly referred to as *forecast verification.* The term ‘verification’ implies comparison with some reference that is assumed to be the

'truth'. The approach is conceptually different for users, who are more interested in assessing how they benefit from using forecast products. In the context of this study we will use the more generic expression of *forecast quality assessment* to include measures that express relevant properties of forecasts that help users to judge the usefulness of such forecasts for their purposes. These measures or *metrics* may include some of, but are not limited to, the traditional statistics used in forecast verification. Furthermore, some aspects of forecast quality cannot be measured simply and objectively with simple metrics. For example, the media may be more interested in the style, graphics and attractiveness of the forecasts than in their accuracy. This facet of forecast quality, which is more of a subjective nature, will not be covered in this study. Moreover, the usefulness of forecast products is also affected by aspects of the quality of service - such as timeliness - that are not specific to meteorology, and these aspects have been intentionally ignored.

Forecast quality assessment is typically done by providers, using a selected sample of forecasts and a selection of metrics, assuming that verification results also measure forecast quality for users. It is not clear, however, that the assessment methods and metrics chosen by a provider are necessarily consistent with the user's needs. In a competitive market, providers may naturally select methods and metrics that show their forecasts in a favourable light. This free choice of methods and metrics makes comparison between providers difficult, and sometimes confusing, for users. Ideally, the assessment of forecast quality should be done by the users themselves, or by an independent assessor, using a standard set of methods and metrics that are relevant to the forecast application.

An independent body could be set up to monitor the industry, encourage sound forecast quality assessment practice, and possibly act as an official watchdog to ensure that acceptable quality standards of forecast products are met by providers. The relationships between forecast providers, users and assessors, and the existence of an independent monitoring body are summarised in the simple schematic of Fig. 1.

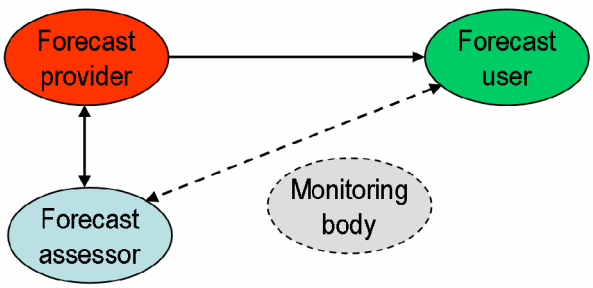

Figure 1 - Key agents in the UK weather forecasting market.

.

2.2. Consultation

Forecast providers and users based in the United Kingdom were widely consulted to examine current quality assessment practice in the industry. This consultation was carried out through an online survey complemented by a dozen of visits and interviews. A thematic workshop followed by a discussion with the participants was also held at the University of Exeter on 15th September 2005.

2.2.1. Survey technical setup

The open-source *PHPSurveyor* package (Code and documentation freely available from http://phpsurveyor.sourceforge.net) was selected for its cost-efficiency, its good reliability record, its advanced features (e.g. powerful administration tools, conditional branching), and the availability of online support. The package was installed and run on the Apache web server of the Royal Meteorological Society.

2.2.2. Legal requirements

A description of the project and a full privacy/confidentiality statement were provided on the survey web page to ensure compliance with the law (Data Protection Act), the ESOMAR Codes and Guidelines, and the UK Market Research Society's Code and Guidelines for Internet Surveys.

2.2.3. Publicity

Announcements were emailed to a large number of possible participants in the month preceding the survey launch (April 2005). Steps were also taken to have the survey listed in popular search engines such as Google and Altavista. In addition, an article in *Nature* (Giles, 2005) contributed to increasing public awareness of the project. Some publicity was also given at the latest ECMWF User Meeting (15-17 June 2005).

2.2.4. *Questionnaire design*

Two online questionnaires – one for forecast providers and the other for forecast users - were designed in consultation with the Project Steering Group and the WMO WWRP/WGNE Joint Working Group on Verification. One anonymous forecast provider and one anonymous forecast user also graciously volunteered to test the questionnaires and gave useful feedback.

2.2.5. Respondents

A total number of 25 UK-based commercial forecast providers were invited to take part in the online survey. Forecast users were invited via various channels. The main difficulty with users is that they are not as easy as providers to contact directly. Providers were asked to forward invitations to their customers and encourage them to participate. As it turned out, this approach to reach forecast users was the survey's 'Achilles' heel' because it required the providers' cooperation and involvement to get their customers to respond, as is discussed further below. We tried to overcome this weakness by using alternative routes. *Nature* agreed to pass on invitations to the users interviewed by Giles (2005). In

addition, invitations were emailed to a dozen major forecast users in the energy and financial sectors. Finally, several organisations representing various sectors of the UK economy were asked to publish announcements and invitations in their newsletters, viz. the Fresh Produce Consortium, the National Farmers' Union, the British Retail Consortium, the British Soft Drink Association and UKinbound. These approaches met with only limited success. If such a survey were repeated in future, additional time and effort would be needed to obtain a larger and more representative sample of users, but this would require more financial resources.

18 responses were received from 12 of the 25 providers contacted. 7 of these responses originated from various specialised departments of the Met Office and each of the remaining 11 responses came from different companies. This large proportion of responses from the Met Office –more than 1/3 - reflects its position as the dominant forecast provider in the UK. However, it may also have introduced some biases in the results. Only 16 responses were received from users. This is unexpectedly low, considering the large number of customers per provider (see the survey results below). The low user response can be explained by a combination of different factors:

- Several large providers with a substantial number of customers declined to take part in the survey. The list of companies reluctant to participate includes Weathernews International, Fugro-Geos, the PA Weather Centre, Metcheck, and Weather Action. Two of these companies declared that they have always been in favour of a coordinated effort to improve standards in the industry, but they have reservations concerning the project. Other reasons stated were:
    - the absence of commercial incentive or reward;
    - the provider's belief that participation was not appropriate;
    - the provider's satisfaction that their own assessment methods are sound, and that their products are of good quality and recognised as such by independent external assessors.
- Providers who responded to the survey have not strongly encouraged their own customers to respond. Interviewed users confirmed that some providers deliberately did not forward the invitations to their customers. In contrast, one of the providers who took part in the survey was particularly successful at getting their customers to respond. This demonstrates that suitable encouragement of users by their providers can be very effective in achieving a good response rate.
- A large number of users do not see any immediate material benefit in taking part. Weather exposure may only be a minor source of risk for some users. Forecast accuracy may also be considered as a secondary issue and the user may prefer to use the cheapest products available. For example, insurance policy clauses may require the use of weather forecasts, but this requirement is not necessarily linked to forecast performance. The media are generally more concerned with forecast communication and presentation than with accuracy.
- A few potential respondents have been unwilling to answer some of the mandatory questions despite guaranteed confidentiality. We found that making key questions mandatory is preferable in order to have a guaranteed

minimum consistency in the number of answers, and to make sure that all respondents go through the questionnaire without leaving out particularly interesting items. We also believe that the questionnaires were designed so as not to be too inquisitive.

2.2.6. Survey Results

Appendices B and C give details of survey Questions and Answers. Some of main features are summarised here.

Figure 2 - Answers to Question 1 in the survey - shows that providers who responded sell forecasts to a wide spectrum of the UK economy. In contrast, the majority of users who responded are in the retail or energy sector while some important forecast consumers (e.g. maritime transport, media, tourism) are absent from the sample. Reported applications of the forecasts (User Question 10) are sales forecasting and stock planning (retail), the prediction of power/gas/oil demand and supply (energy), risk management (energy), the prediction of crop growth and pests (agriculture).

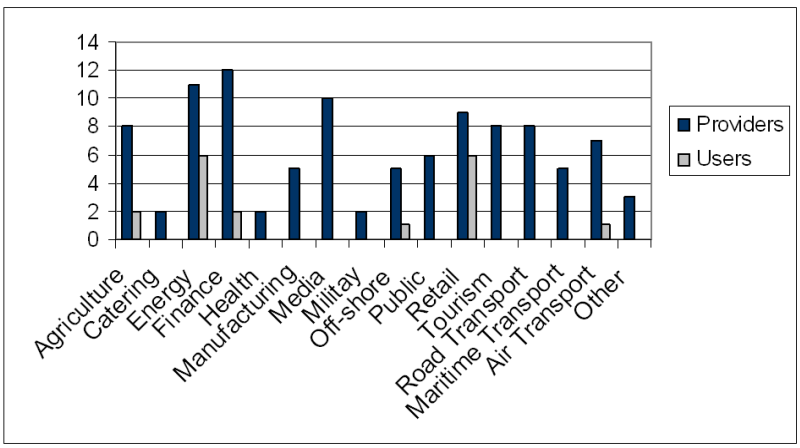

Figure 2 – Sectors of activity of respondents [Question 1].

Answers to Question 2, which are summarised in Fig. 3, indicate that the majority of providers who responded have a fairly large (> 20) number of customers. User responses, however, indicate a preference for buying forecasts from a single provider, though the use of free forecast products from the internet is commonplace (Question 5).

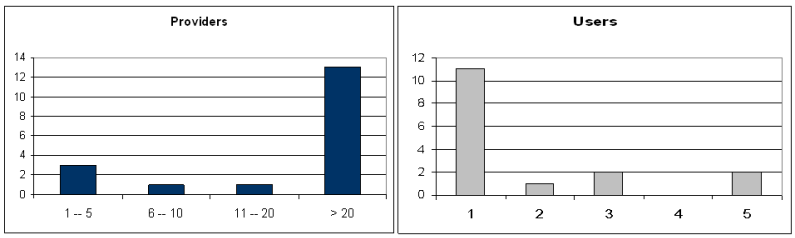

Figure 3 - Respondent's number of customers (left) and providers (right) [Question 2].

Fig. 4 shows the spectrum of forecast product ranges produced/used by respondents (Question 30). Maximum provision and use of products is in the

medium range. The discrepancies between providers and users in the short ranges are linked to the modest representation of users from the transport sector (Fig. 2). The importance of medium to long-range (monthly/seasonal) product use, on the other hand, reflects the dominance of the retail and energy sectors in the sample of users.

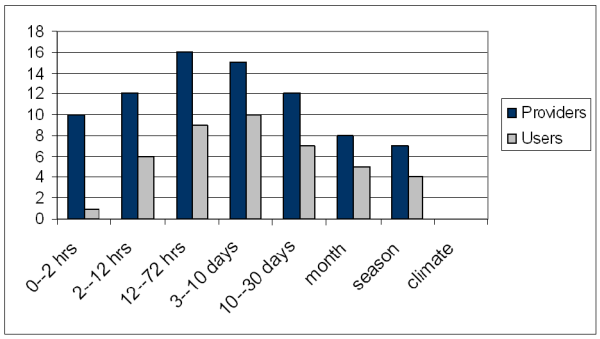

Figure 4 - Ranges of forecast products [Question 30].

In Fig. 5 (Question 40), responses suggest that that the most widely used format of forecast is quantitative (numbers). However, other formats such as symbols and purely descriptive forecasts remain important. Making this distinction between formats is important because qualitative forecasts are much harder to assess than purely quantitative forecasts (See section 3.3 below). However, the value added by qualitative forecast information is implicitly acknowledged by providers/users with a majority of them providing/having access to real-time forecast guidance, e.g. through a dedicated hotline (Question 55).

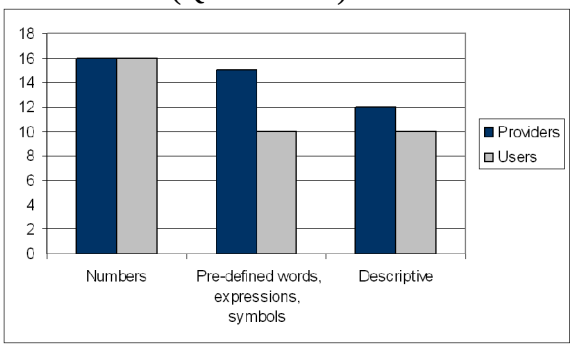

Figure 5 – Forecast formats [Question 40].

The main types of quantitative forecasts available from providers are: point, interval, categorical and probability forecasts (Fig. 6, Question 42). Deterministic (point and categorical) and interval forecasts are the most widely used types in the current user sample. Probability and binary forecasts are much less used. This pattern can be at least partly explained by the characteristic of the sample. Indeed, most users of binary forecasts (e.g. frost/no frost or rain/no rain) would belong to the agricultural, transport and building sectors, and these are not well represented in the sample. Several interviewed providers confirmed that the market for probabilistic forecasts is quite small in general. The main reason given is the difficulties to interpret and apply probabilistic information, despite its higher value for decision makers compared with deterministic forecasts.

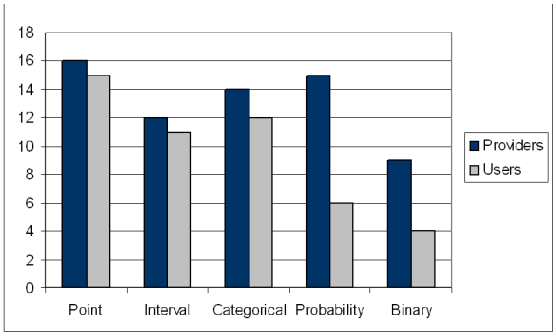

Figure 6 – Types of quantitative forecasts [Question 42].

Information on forecast uncertainty is supplied, or can be supplied, by all responding providers. Conversely, information on forecast uncertainty is provided or available on request to all users in the sample (Question 45).

As regards the methods of forecast delivery, Fig. 7 (Answers to Question 50) confirms the predominance of electronic methods to disseminate forecast information, more particularly methods based on Internet technology. Forecast upload to customers (e.g. via FTP) is the most common method among providers.

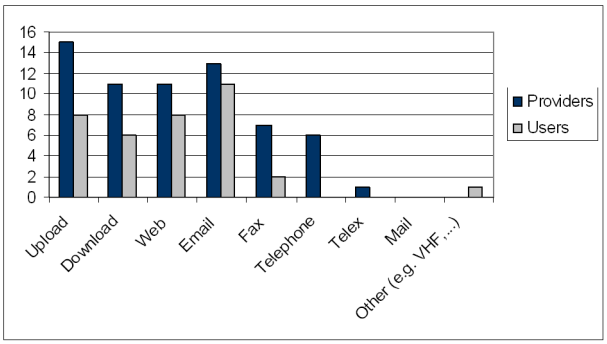

Figure 7 – Methods of forecast delivery [Question 50].

The users' preference for emails may be associated with the sample characteristics. In general, users in the retail sectors receive their forecasts by email, while automatic upload tends to be the preferred method in sectors where time delivery is critical (e.g. energy trading).

The frequency and form of forecast quality assessment was also examined in Question 60. Fig. 8 shows that 16 out of the 18 responding providers issue assessments of forecast quality. However, just over half of the users (9 out of 16) report that they don't receive forecast quality assessments from their providers. Providers also claim a higher frequency of quality assessment than suggested by the sample of users. This divergence between providers and users is increased when the quality assessments are in quantitative form (Question 61): 14 providers out of 18 claim to issue quantitative quality assessments, but only 6 users out of 16 receive quantitative assessments from their providers. Interestingly, 4 of these 6 users are in the energy sector. No respondent from the retail sector receives any quantitative assessment from their providers. This might be an artefact due to each group receiving their forecasts from the same providers, but it may also reflect different attitudes toward quantitative forecast quality

assessment. Interviews have corroborated that the retail sector is generally less concerned than the energy sector about quantitative quality assessment. This point is confirmed by the fact that only 2 out of the 6 respondents from retail make their own quantitative assessment, against 6 out of 6 in energy (Question 71).

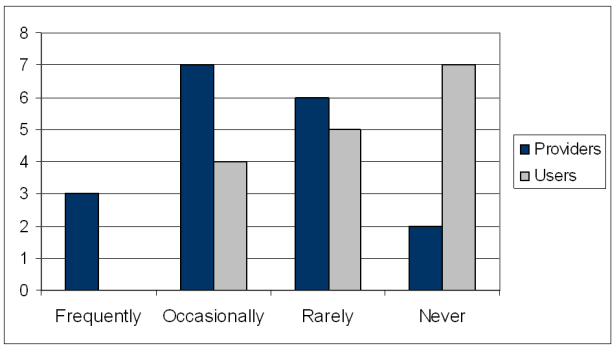

Figure 8 – Frequency of forecast quality assessment [Question 60].

Another striking pattern is that only 6 out of 16 responding users receive quality assessments from their providers that they find easy to understand. The same proportion of users state that they receive useful quality assessments from their providers (Questions 63 & 64).

Nearly ¾ of users (11 out of 16) make their own assessment of the quality of the forecast products they buy (Question 70), and in 7 cases (less than half of responding users) this assessment is discussed at least occasionally with the providers (Question 80). However, only 7 out of 16 users (6 from the energy sector) use their own quantitative quality assessment to decide which forecast product to buy (Question 87). Furthermore, users appear to be much less specific about their assessment methods than providers (Question 75). One user from the energy sector even refused explicitly to reveal their methodology.

Forecast quality is a multi-dimensional concept described by several different attributes. Fig. 9 shows the main forecast attributes used by providers and users to assess forecast quality (Question 73).

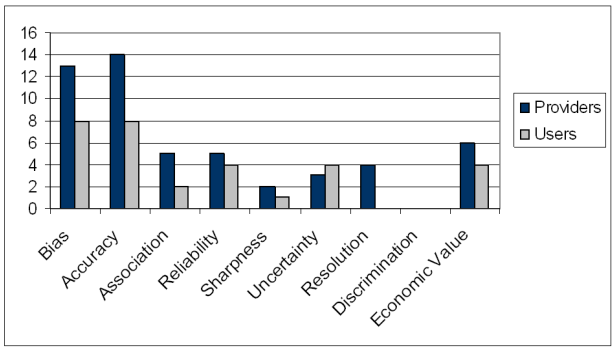

Figure 9 – Forecast attributes assessed by providers and users [Question 73].

The measures reported by respondents for quality assessment are (Question 75):

- Mean Error (ME), Mean Absolute Error (MAE), (Root) Mean Squared Error ((R)MSE) and Skill Scores (SS) based on them;

- Mean Absolute Percentage Error (MAPE);
- Brier Score (B), Brier Skill Score (BSS);
- Binary contingency tables measures, e.g.: Percentage Correct (PC), Hit Rate (H), False Alarm Rate (F), Relative Operating Characteristic (ROC) curve;
- Proportion within tolerance interval;
- Correlation measures, e.g. Anomaly Correlation Coefficient (ACC);
- Gerrity Score (GS);
- Service Quality Index (SQI).

The SQI is a composite measure of performance that is used at the Met Office to assess TAF performance. The other metrics will be briefly reviewed and defined in Sections 3 and 4.

Question 97 probes the desirability of establishing an independent body to monitor the weather forecasting sector and encourage good practice in the assessment of forecast quality. A large majority of respondents believe that although such a body is not necessary, it would be useful (Fig. 10).

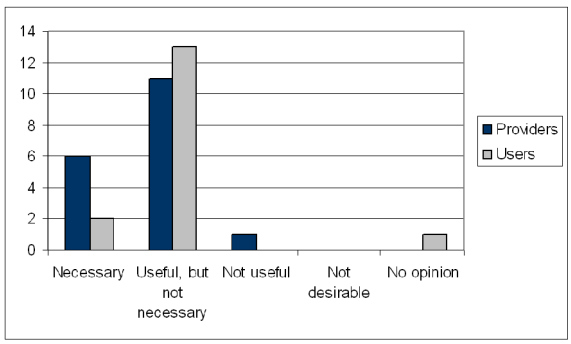

Figure 10 – Desirability of an independent monitoring body [Question 97].

There is also good support in Question 98 for an independent online forum where users and providers could submit their problems concerning forecast quality issues and find/offer practical solutions (Fig. 11). Some respondents were critical about the idea, expressing concerns that without appropriate moderation the forum might not fulfil its objective and even be open to abuse.

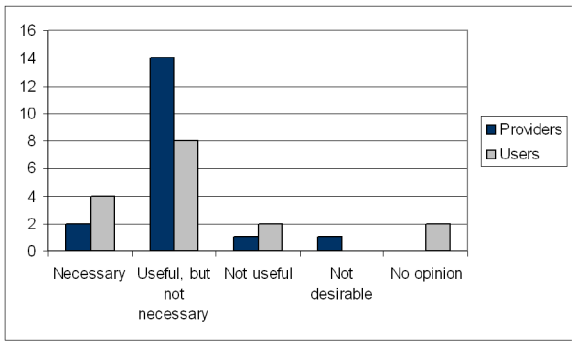

Figure 11 - Desirability of an independent online forum [Question 98].

2.2.7. Visits and interviews

In general, statements gathered through visits and interviews consolidate the main findings of the survey. The important roles of perception and psychology in the notion of quality were emphasised by several providers. Their opinion is that the customer's interest in forecast quality is often selective and subjective, except for some typical users with 'quantitative' profiles (e.g. energy, aviation, marine).

Forecast quality is not necessarily the prime concern of a user as long as poor forecast performance does not affect their business beyond certain limits. Users may be more interested in other aspects, such as price and quality of service. These two factors provide a plausible explanation for the high level of satisfaction expressed by some users in the survey despite the total absence of forecast quality assessment in their case (1 fairly satisfied, 2 very satisfied, all three from the retail sector). In the absence of quantitative quality assessment, cheaper forecast products are easily perceived by users as having greater value most of the time. In addition, forecast quality assessment entails a clear economic cost that some users are not prepared to bear.

There were calls – also echoed in the survey - for quality assessment metrics that are more 'user-friendly'. A couple of providers expressed the need to engage the users more closely in collaborative projects. This would require a more active and open involvement of users, e.g. to compile accurate long-term records that quantify the impact of weather on their business. However, several interviewed users did not feel comfortable with the idea of disclosing that information because they view it as commercially sensitive. Another user pointed out that it is not always feasible to separate clearly the effect of weather from other factors.

The appeal of summarising forecast quality in one single composite metric was also discussed with one provider. There is increased pressure through a target-driven culture to provide simple summary measures. Although this approach has the advantage of simplicity, it has the drawback of being difficult to interpret and potentially misleading, as quality is inherently multidimensional.

One provider who did not want to participate in the survey stated that the performance of his own forecasts has been tested successfully by gambling against statistically fair odds. He believes that the profit made using his forecasts proves that they are useful.

2.2.8. Workshop on the quality of weather forecasts

Preliminary results and recommendations were presented to the public at an open (free) workshop that was held on 15th September 2005 at the biennial Conference of the Royal Meteorological Society, University of Exeter. The workshop started with a 1-hour presentation given by the principal consultant Pascal Mailier. A short preview of the same presentation had been given two days before by Prof Ian Jolliffe at the European Meteorological Society's (EMS) 7th European Conference on Applications of Meteorology (ECAM) in Utrecht. The workshop presentation was followed by a ½-hour discussion with the participants. It is estimated that around 50 people were present. The same reasons as those given

for the modest survey response rate could be invoked to explain – at least partially - the limited turn-out at the workshop. In addition, many potential participants from the industry were drawn to other events on weather prediction which took place elsewhere in Europe and overseas near the time of the workshop (e.g. ECAM).

Most of the interventions made during the discussion repeated or confirmed the findings and statements gathered from the survey, visits and interviews.

Several participants advocated a very radical approach to user-oriented quality assessment. The forecast provider should incorporate the user's decision model into the quality assessment scheme. Furthermore, probability forecasts should be converted into user-specific decisions by the provider, who would then communicate these decisions to the user. This arrangement would absolve the user of any blame in cases where the decision turns out to be wrong. It also implies that the quality assessment metrics may become unique to a user. In its extreme form, this system leads to a reversal of the traditional roles, i.e. a situation where the forecast provider takes responsibility for the decisions and the user assesses the quality of these decisions. These views reflect a genuine desire for a service tailored to the specific user, but in any case a proper quality assessment scheme should always be able to discern the 'meteorological goodness' of the forecasts.

Some comments were made concerning the guesstimates for the number of UK users (~500?) and providers (~25?) on the introductory slide. These numbers are likely to be underestimated, but the current lack of participation and transparency in the industry makes it very difficult to estimate these numbers with accuracy.

The possible effects that performance-related contracts may have on best practice were also considered. For example, the choice of metrics should be such that hedging is not encouraged. However, the fact that some hedging is acceptable in the case of high-impact forecasts (e.g. storm or flood warnings) must be acknowledged.

No one dissented from the desirability of an independent body to monitor quality assessment practices. There was no clear reaction to the suggestion that this might be done by the Royal Meteorological Society. More transparency is required from providers regarding their quality assessment practices.

It was felt that users or providers may not really wish to talk openly to each other because of competition. A provider pointed out that users do not necessarily tell the truth when giving feedback on forecast performance. These arguments justify some reserve as to the feasibility of an on-line forum.

In connection with the slides on the NAO index forecasts, there was some concern that the quality assessment standards applied to seasonal forecasts are often low, and this situation may result in giving the industry a bad name.

PART II: LITERATURE REVIEW AND CASE STUDIES

In this part of the report:

- A comprehensive critical review of existing quality assessment methods and metrics is presented (Section 3);
- Do's and Don'ts of assessment methods and metrics are illustrated by means of several case studies (Section 4).

Note:

This part is very technical by nature. A recap in plain language is written in Section 5, more specifically:

- The important findings of the literature review of Section 3 are summed up in Subsection 5.2;
- The case studies of Section 4 and the lessons learnt from them are summarised in Subsection 5.3.

3. FORECAST VERIFICATION METHODS AND METRICS

This section offers a short, but comprehensive literature review of existing weather forecast verification methods and metrics. This special field is growing rapidly in extent and complexity, and the text of this section is inevitably very technical. Some readers may find it easier to read the summary of this section that is presented in Subsection 5.2.

3.1. Quantitative forecasts

Methods and metrics for all the types of quantitative forecasts recorded in the survey are considered.

3.1.1. Review of the literature until 2003

The book edited by Jolliffe and Stephenson (2003), with contributions by many experts in the field, was largely written from a developer's point of view rather than with the user's perspective. However, this book provides an excellent overview of the forecast verification literature up to 2002 with numerous references. It contains formal definitions and critical discussions of the most commonly used metrics. Jolliffe and Stephenson (2003) use a similar classification of forecasts to that adopted in this project's survey, with chapters on binary forecasts, categorical forecasts, continuous variable forecasts and probability forecasts. All of these types of forecasts are provided by a majority of providers in the survey, and the relevant chapters of the book have played an important role in formulating our own recommendations.

For binary forecasts, there are an amazingly large number of possible measures (see Mason, 2003). With a binary forecast of an event there are two ways of getting the forecast right (hit: the event is forecast and occurs; correct rejection: the event is correctly forecast not to occur), and two possible types of incorrect forecast (a missed event or a false alarm). Two commonly used measures of the quality of a binary forecast are the hit rate H, which is proportion of observed events that were correctly forecast, and the false alarm rate F, the proportion of non-occurrences that were incorrectly forecast. The false alarm rate is often confused with the false alarm ratio, which is the proportion of forecasts of occurrence not followed by an actual occurrence. The percentage correct, PC gives the proportion of hits and correct rejections. PC is not a reliable measure of forecast performance because it is heavily dependent on the underlying frequency, or base rate, of the event of interest. Since forecast skill depends on maximising the number of hits while minimising the number of false alarms, H and F together give a useful summary of the quality of a set of binary forecasts, and when a series of binary forecasts, corresponding to different thresholds of an underlying variable, is issued, a plot of H against F for the different thresholds gives the so-called ROC curve. Two further measures, among the many available, should be mentioned. The equitable threat score seems to be one of the most popular, but suffers from a disadvantage shared by several other measures, namely a strong dependence on the base rate. The odds ratio skill score (Stephenson, 2000) is a simple measure of association in (2x2) contingency tables. It has not yet been

used widely in the atmospheric science literature as a verification measure, though it avoids many of the poor properties of most other measures. Another advantage of the odds ratio is that approximate confidence intervals can be easily calculated. However, it should be stressed that it is hardly ever advisable, for binary or any other type of forecast, to rely on a single verification measure. It is impossible to evaluate fully the quality of a set of forecasts with a single measure.

A disadvantage of the traditional binary "yes/no" forecast assessment is that it does not discriminate between "near misses" and forecasts that are far away from the observations. Similarly, a "hit" in the traditional view has some probability of being a "false alarm" when the observation is uncertain. Fuzzy forecast quality assessment attempts to take account of the uncertainties in the forecasts and/or the observations by giving some credit for getting a forecast partially correct, and by giving a penalty for getting it partially wrong (Ebert, 2002). Forecast and observation are assigned to either side of the "yes/no" threshold using membership functions that are derived from their probability density functions.

Categorical forecasts (Livezey, 2003) can be assessed either by reducing them to a series of binary forecasts, or by a single measure. Both approaches have advantages and disadvantages. The single measure approach is recommended by Livezey (2003), based on a family of scores introduced by Gandin and Murphy (1992) and Gerrity (1992). Such scores take into account the fact that the categories are often ordered and that a forecast that is only one category away from the correct category is better than a forecast that is two or more categories in error. Each combination of forecast and observed categories is assigned a score, measuring how close the forecast is to the observed category and the overall skill score is based on averaging these individual scores according to how frequently each combination occurs. Individual scores are chosen so that the overall skill score has desirable properties. However, using a single measure hides some detail of the quality of a set of forecasts, and users may also find it easier to have the same type of information available for multi-category forecasts as for binary forecasts.

For continuous variables (Déqué, 2003), a number of measures based on bias, mean square error, mean absolute error and correlation are in common use. Bias, or mean error, measures whether, on average, forecasts are consistently too high or low. Mean square error (MSE) is an average of the squared difference between forecasts and observations. Its square root (RMSE) is often used, as it is in the same units as the quantity being forecast. Because they are quadratic rules, in MSE and RMSE large errors are given substantially more weight than small errors. In applications where this sensitivity to large errors is not desirable, the mean absolute error (MAE) can be used instead. The mean absolute percentage error (MAPE) is similar to the MAE, except that each absolute error is divided by the modulus of the corresponding observation. This adjustment helps to offset possible error increases that arise as the observations get larger. For this reason, MAPE may be a useful measure to assess the accuracy of precipitation forecasts. None of these error-based measures is very useful unless it is standardised in some way to convert it to a skill score, because otherwise there is a lack of

comparability on different datasets. A skill score measures the relative quality of a forecasting system compared to another reference forecasting system. Climatology (long-term average) is commonly used as a reference for medium-range or long-range forecasts, whereas persistence may be used in the case of short-term forecasts. However, a reference forecast that is clearly unskilful, using no knowledge at all about climate or weather (e.g. some sort of a random forecast) is required to obtain an absolute measure of forecast skill.

If forecast and observed values of the variable of interest are plotted against each other, the (Pearson) correlation coefficient measures how close to a straight line are the plotted points. If the observations and forecasts are replaced by their ranks, then the correlation calculated between these ranks is Spearman's rank correlation. This is less sensitive to extreme observations than Pearson's correlation and measures how close the plot of forecast vs. observed values is to monotonicity (not necessarily a straight line). It is sometimes useful to calculate correlations using anomalies relative to climatology. Again, no single measure is adequate to describe the quality of a set of forecasts. MSE can be decomposed into the sum of a number of terms, two of which measure bias and correlation respectively (Murphy, 1988).

The Brier score, and its extension the ranked probability score, predominate in the assessment of probability forecasts (Toth et al. 2003). The Brier score is rather like MSE, measuring the average squared difference between the forecast probability and the corresponding observation, which in this case takes only the values 0 or 1. Like the MSE, it can be decomposed into terms measuring different attributes of the forecast system. In this case there are terms measuring reliability and resolution, and a third term related to the base rate of the event of interest. Reliability measures how close is the proportion of occurrences of the event to the forecast probability, conditional on the value of the forecast probability. It is often plotted for a number of values of the forecast probability on a so-called reliability diagram. Resolution measures how different are the distributions of the observations for different forecast probabilities. Simple metrics based on entropy have been proposed to assess resolution of probabilistic forecasts (Stephenson and Doblas-Reyes, 2000; Roulston and Smith, 2002). Sharpness quantifies the ability of the forecasts to 'stick their neck out'. It can be defined simply by the variance of the forecast probabilities or in term of the information content (negative entropy) of the forecasts. For perfectly calibrated forecasts, sharpness is identical to resolution. If a probability forecast is converted into a deterministic one by predicting the event to occur whenever the probability exceeds some threshold, and this procedure is repeated for several thresholds, a ROC curve can be constructed as an alternative way of assessing probability forecasts. A variation of this idea of reducing a probability forecast to a binary forecast by forecasting the event if the probability exceeds some threshold is discussed by Zhang and Casey (2000). They categorise into three groups, rather than two, using the forecast probability. A probability that is sufficiently far above/below the climatological probability for the event of interest leads to a forecast that the event will/will not occur. A probability close to the climatological forecast is deemed to be a 'non-applicable' forecast, because it is unlikely to influence a user's

behaviour compared to having no forecast at all. Zhang and Casey (2000) discuss verification based on this categorisation.

Hartmann et al. (2002) also discuss verification of probability forecasts from the users' perspectives. They concentrate on graphical representations of the quality of forecasts, giving imaginative displays of conventional metrics such as the Brier score, but also presenting descriptive but informative plots of forecasts and observations.

Jolliffe and Stephenson (2003) have separate chapters on verification of spatial forecasts, and verification based on economic value; neither of these was explicitly addressed in the survey. As noted in the introduction, consideration of economic value was deliberately excluded from the project. With respect to spatial forecasts, the emphasis in the project has been on single value forecasts, but many meteorological forecasts take the form of maps. At present the most common ways of assessing the quality of map forecasts use MSE (or RMSE), taking averages over the stations or gridpoints on the map, or some type of correlation or anomaly correlation (Drosdowsky and Zhang, 2003). Neither of these approaches takes into account the spatial nature of the forecasts, and the development of measures that do so is an active area of research (see the next section on recent literature). Verification of map forecasts also needs careful consideration of how to match the forecasts, which are typically made for regularly-spaced gridpoints, and observations, which are often at irregularly-spaced stations.

3.1.2. Review of the recent literature

Jolliffe and Stephenson (2003) was perceptively reviewed by Glahn (2004) who raised several very interesting and important issues. These issues were then addressed by Jolliffe and Stephenson (2005) in a forum article published in *Weather and Forecasting*. Since the publication of Jolliffe and Stephenson (2003), several new papers have been published on verification methods, some of which will be briefly reviewed here together with a handful of slightly earlier papers that were missed by Jolliffe and Stephenson.

Quantitative Precipitation Forecasting (QPF) is one of the areas that have received the most attention in development of new approaches. Verification of spatial precipitation maps is particularly challenging due to precipitation amounts being spatially discontinuous (non-smooth), highly skewed, and containing a mixture of zero and non-zero values. New multi-scale approaches have recently been developed to address such problems (Tustison et al., 2002; Casati et al., 2004; Venugopal et al., 2005). Other studies have developed and tested verification methods for forecasts of precipitation over regional domains (Accadia et al., 2003a,b; Haklander and Van Delden, 2003; Saulo and Ferreira, 2003).

There has also been a growing appreciation for the need for probabilistic rather than deterministic forecasts especially for longer lead-time forecasts (e.g. medium-range or seasonal). There are fundamental issues on the interpretation of probability in such systems (Wilks, 2000; de Elia and Laprise, 2005). The prequential interpretation of probability has recently been illustrated in an

assessment of financial forecasts (Bessler and Ruffley, 2004). An extensive discussion of scoring rules for probability forecasts was given by Wrinkler (1996) and discussants. Later studies have examined in more detail some of the undesirable properties of well-known probability scores such as the Brier score and have proposed variant approaches (Mason, 2004; Muller et al., 2005; Gneiting et al., 2005). All proper probability scores are inequitable (work in progress, Stephenson and Jolliffe) and so different no-skill forecasts yield different scores, which causes unavoidable problems when defining skill scores based on a no-skill reference forecast. Distribution-oriented measures (Bradley et al., 2004) such as the ROC score are now being routinely employed in the verification of probability forecasts (Mason and Graham, 1999; Mason and Graham, 2002; Kharin and Zwiers, 2003). The use of information-theoretic measures such as entropy has been proposed (Roulston and Smith, 2002) and such approaches also appear to be promising for the verification of forecasts of point process events such as earthquakes (Daley and Vere-Jones, 2004; Harte and Vere-Jones, 2005). Such approaches are potentially useful for the point process events that occur in meteorology such as individual weather systems, extreme rainfall events, etc…

One possible approach for generating probability forecasts is to produce an ensemble of forecasts. Ensemble forecasting is now a major activity at many forecasting centres around the world. Various approaches have been developed for evaluating such systems such as economic value (Wilks, 2001), new types of correlation analysis (Wei and Toth, 2003), minimum spanning trees (Smith and Hansen, 2004; Wilks, 2004) and new reliability measures (Atger, 2004). The properties of various scores and their relationships with other scores and value measures have been addressed in several studies (Wandishin and Brooks, 2002; Hiliker, 2004). Several recent studies have discussed the use of value (utility) measures based on decision theory models of how a simple user might operate (e.g. Thornes and Stephenson, 2001; Mylne, 2002; Mason, 2004; Stewart et al., 2004).

There has been much activity in generating and verifying multi-model seasonal forecasts (e.g. Wilks and Godfrey, 2002; Goddard et al., 2003; Potgieter et al., 2003). The skill of such forecasting systems can be improved by using statistical approaches to combine and calibrate the climate model outputs. Bayesian methods of combination and calibration have been demonstrated to work well (Rajagopalan et al., 2002; Coelho et al., 2004). The calibration and combination of forecasts can be shown to be mathematically analogous to data assimilation and hence should be considered to be an integral part of the forecasting process rather than an optional post-processing stage (Stephenson et al., 2005). The likelihood regression of the forecasts on the observations in these procedures provides much valuable verification information on the reliability and resolution of the forecasts. However, calibration is not always a possibility for forecast users who do not have access to all previous forecasts and observations (see Glahn, 2004 and Jolliffe and Stephenson, 2005 for an interesting debate on this important issue). Another important aspect of seasonal forecasting is that the sample size of past

forecasts and observations is small (typically less than 50) and so care needs to be taken in using standard verification approaches (Bradley et al., 2003).

Hamill and Juras (2005) demonstrated that one should be careful in pooling data with different climatologies, because pooling can lead to misleading verification statistics.

3.2. Descriptive forecasts

Descriptive or worded forecasts are particularly difficult to verify. Examples are phrases such as 'scattered showers in the west', 'best of the sunshine in the south east', 'windy in the north west later'. A typical descriptive forecast consists of several such phrases. What is common to all is that the forecast as it stands is not quantitative or even categorical, and different users of the forecast can, and most likely will, interpret it differently. Hence there is inevitable subjectivity in deciding whether or not the forecast was a good one.

Many descriptive forecasts can, in theory, be made more definite. For example, 'windy in the northwest later' could mean 'at some time in a 12-hour period, the mean wind speed at station A in the northwest will exceed a specified threshold'; 'frost is likely' may mean 'the probability of frost exceeds 0.7', and so on. In circumstances where a technical definition underlies a descriptive forecast, the descriptive forecast can be verified by going back to the technical definition. Depending on the nature of that definition (binary, continuous, probabilistic,…) an appropriate verification strategy can be chosen. The forecaster may, or may not, have such a definition in mind, but the forecasts are often aimed at a general audience, for whom the more technical version would be unattractive.

If no underlying technical definition is available, verification is inevitably subjective, Consider again 'windy in the northwest later', and a number of possible outcomes. It may be windy everywhere, not only in the northwest, or it may be windy later in the forecast period, but even windier earlier. It is quite possible to treat these outcomes as corresponding to 'good' forecasts, 'poor' forecasts, or somewhere in between.

Various strategies have been suggested for introducing an element of objectivity. One approach that dates back to Wright and Flood (1973) is to create an 'anti-forecast' that is notionally the opposite of the forecast phrase. By comparing the forecast with a baseline (the anti-forecast) that should be much worse (have negative skill), a more objective way of assessing the forecast is achieved. Another approach is use as a baseline a forecast/observation pair where there should be zero skill. For example the current weather could be compared with the forecast made for today, and with one or more forecasts made for completely different days, say one year ago. The proportion of times that the relevant forecasts appear to be better than the irrelevant ones gives a measure of the forecasts' skill. Jolliffe and Jolliffe (1997) considered two variants of this approach. Neither of these approaches eliminates subjectivity: different assessors may still come to different conclusions.

Our view is that if underlying technical definitions are unavailable and subjectivity therefore remains, descriptive forecasts cannot be verified without the potential for bias. Such forecasts are nevertheless valuable for some users, and we would not completely discourage their provision. However users should be aware that they are impossible to verify with objectivity and claims for the skill of descriptive forecasts therefore need to be treated with scepticism.

3.3. Forecasts of rare and extreme events

Another aspect of forecast quality that has become increasingly important over the past decade is the ability to predict rare and extreme events. The assessment of such forecasts is problematic. The adverse impact of small sample sizes on the validity of results presents a first difficulty. Furthermore, in situations where rare events or the tails of distributions are important to the decision maker, traditional measures such as H, MSE and MAE become virtually useless (Murphy, 1996; Marshall and Oliver, 1995; Matthews, 1997). For vanishingly rare events, most classical quality assessment metrics for binary forecasts tend to 0 or 1 and can be improved without skill by under-forecasting. However, the extreme dependency score (Coles *et al.*, 1999) and, to a lesser extent, the odds ratio give useful information. The assessment of extreme forecasts is the subject of ongoing research work in the field of ensemble prediction. Exploratory assessment strategies based on extreme value theory have been developed (e.g. Mailier, 2001), but these techniques are not practical for operational use.

3.4. Effect of forecast type on perceived usefulness

The potential differences in perceived usefulness of various forecast types were examined quantitatively by Önkal and Bolger (2003) in the context of financial (stock-market) forecasting, from both the provider's and the user's perspective. The metric used by participants to assess the usefulness of the forecasts was a simple 7-point rating scale (1 = 'not useful at all' and 7 = 'extremely useful'). They found that directional (trend) and interval forecasts were perceived as more useful than point (deterministic) forecasts. This confirms the assertion that users need an indication of forecast confidence (Fischhoff, 1994). Interestingly, their results also reveal that the 50% interval forecasts were rated as being less useful than 95% interval forecasts, despite much better reliability scores in the case of 50% intervals (forecasters were found to be overly confident, with 95% confidence intervals being too narrow). Preference for 95% forecast intervals could be due to concerns about the 50% chance of being incorrect. In a study on the perception of probabilistic forecasts, Yates *et al.* (1996) had found earlier that 50% probabilities were seen by forecast users as indications of incompetence, ignorance or even laziness. This perception is consistent with the fact that for a forecast with only two possible outcomes (either within or outside the interval), a 50% probability corresponds to a state with maximum degree of uncertainty (entropy).

Önkal and Bolger's experimental results also showed that usefulness ratings from users were on average better than usefulness ratings from providers, and also that

usefulness ratings were, on average, higher without feedback on forecast performance.

3.5. Special methods for wind direction

Forecasts of wind direction are often made in the form of 8 (or sometimes 16) categories corresponding to compass points. They are therefore similar to multi-category forecasts whose verification is discussed by Livezey (2003), but with one crucial difference. The circularity of categories, with no maximum or minimum category, means that the Gandin and Murphy (1992) family of scores cannot be straightforwardly used. It is possible to use the same underlying reasoning of averaging individual scores for forecast/observed category combinations to give an overall skill score with the desirable properties, but the nature of the properties changes. A promising approach along these lines is given by Kluepfel (personal communication).

4. CASE STUDIES

Various approaches for assessing forecast performance have been discussed in the review of Section 3. The reviewed approaches are still primarily concerned with the needs of forecast developers. The purpose of this section is to:

- demonstrate with real cases that many of the reviewed methods can just as well be applied for user-oriented quality assessment;
- give examples of good and bad practice;
- introduce a new metric to assess the accuracy of interval forecasts and illustrate its functionality by means of a practical example.

The principal fields of user application reported in the survey are in the energy and retail sectors. The examples given in this section pertain to the prediction of weather variables that are particularly relevant to these two classes of users: daily average temperatures in the short and medium range (cases 1, 2 and 3) and seasonal temperature anomalies (case 4).

The contents of this section are of a very technical nature, and some readers may find it easier to read the summary that is presented in Subsection 5.3.

4.1. Simple classical methods for point forecasts

Existing metrics are suitable for many user applications, and in these cases it is not necessary to develop alternative methods. A successful application of classical metrics for users is presented in this Section. Several simple metrics were selected to evaluate and compare the performances of 7 different point forecasts of daily average 2-m temperatures (Fig. 12) at a particular station. Each one of the 5 forecast providers involved was assigned a distinct colour for identification: Brown, Red, Yellow, Green and Blue. The deterministic forecasts (Deterministic 1 to 5) were produced through single runs of 5 different numerical models. The two remaining forecasts (Ensemble 1 and Ensemble 2) were produced by averaging all 51 members of an ensemble of forecasts with equal weights (1/51). The two ensembles were produced by the same model, but were post-processed using different techniques. This benchmark test covered a 6-month period (from October 2003 to March 2004).

Given a series of n forecasts $\{F_i(t), i \le n\}$ at a particular time step t (e.g. 2, 3 or 5 days), and the corresponding observations (actuals) $\{O_i(t), i \le n\}$, the chosen metrics are as follows:

Mean Error: $ME(t) = \frac{1}{n}\sum_{i=1}^{n}[F_i(t) - O_i(t)]$,

The ME highlights systematic biases in the forecast systems. Negative (cold) or positive (warm) biases are relatively simple to correct by statistical post-processing , e.g. using a Kalman filter (Persson, 1991).

Root Mean Squared Error: $RMSE(t) = \sqrt{\frac{1}{n}\sum_{i-1}^{n}[F_i(t) - O_i(t)]^2}$.

This measure evaluates the overall accuracy of the forecasts. Because it is a quadratic rule, the *RMSE* penalises larger errors much more heavily than smaller ones. For a non-biased forecasting system (i.e. $ME = 0$) the *RMSE* is equivalent to the standard deviation of the error.

Skill Score based on the *RMSE* **:** $RSS(t) = \frac{RMSE_{norm}(t) - RMSE(t)}{RMSE_{norm}(t)} x100\%,$

where $RMSE_{norm}$ is the *RMSE* obtained by using the climatological values (i.e. the long-term averages) in lieu of the forecasts. The *RSS* compares the forecasting system with a much cheaper alternative method, in this case the long-term average (climatology). A positive (negative) *RSS* reveals a forecast system that performs better (worse) than a system that merely forecasts climatological values. A 'perfect' forecast system (here, perfect means $RMSE = 0$) scores 100%.

Trend Correlation: $TC = r[(F(t+1) - F(t), O(t+1) - O(t)],$

where r refers to Pearson's product moment correlation coefficient. The *TC* shows how well the forecasting system is able to pick up the day-to-day fluctuations of the weather variable of interest. This measure penalises forecast systems that try to minimise the *RMSE* by systematically predicting values close to the long-term averages. Note that it does not make much sense to try and construct a skill score with the *TC* because, by definition, the systematic use of long-term averages does not allow the detection of daily fluctuations.

One of the contenders (Deterministic 4), which was one of the most expensive forecasts, failed the acceptance test (the actual acceptance criteria have been kept confidential) because it showed a strong systematic negative bias and a poor trend correlation. Despite a relatively small bias throughout the forecast range, Deterministic 1 also failed owing to errors of large magnitude as shown by the RMSE and skill scores. Deterministic 3 is also plagued by fairly large errors, apparently associated with an uncontrolled positive bias beyond day 3. Deterministic 5 has highest accuracy in the first 4 days of the forecast. Ensembles 1 and 2 are the most accurate forecasts beyond day 5, Ensemble 1 gradually losing accuracy compared to Ensemble 2 due to a growing bias. The ensemble means become superior after 4-5 days because they filter out the temperature fluctuations that become less predictable in the medium range. Also note the differences between the skill scores based on climatology and those based on persistence. Persistence is a more/less skilful reference forecast than climatology in the short/long range.

The same metrics were also applied to assess the quality of 10-m wind speed forecasts (Fig. 13) produced by three of the above providers (Red, Yellow and Blue). Yellow (Deterministic 3) systematically underestimates the wind speed

(negative bias), so badly that it performs worse than the climatology (negative skill from day 1!). To have an insight into the possible reason for this dramatic bias, the mean wind speed ratio $\frac{1}{n}\sum_{i=1}^{n}[F_i(t)/O_i(t)]$ was calculated.

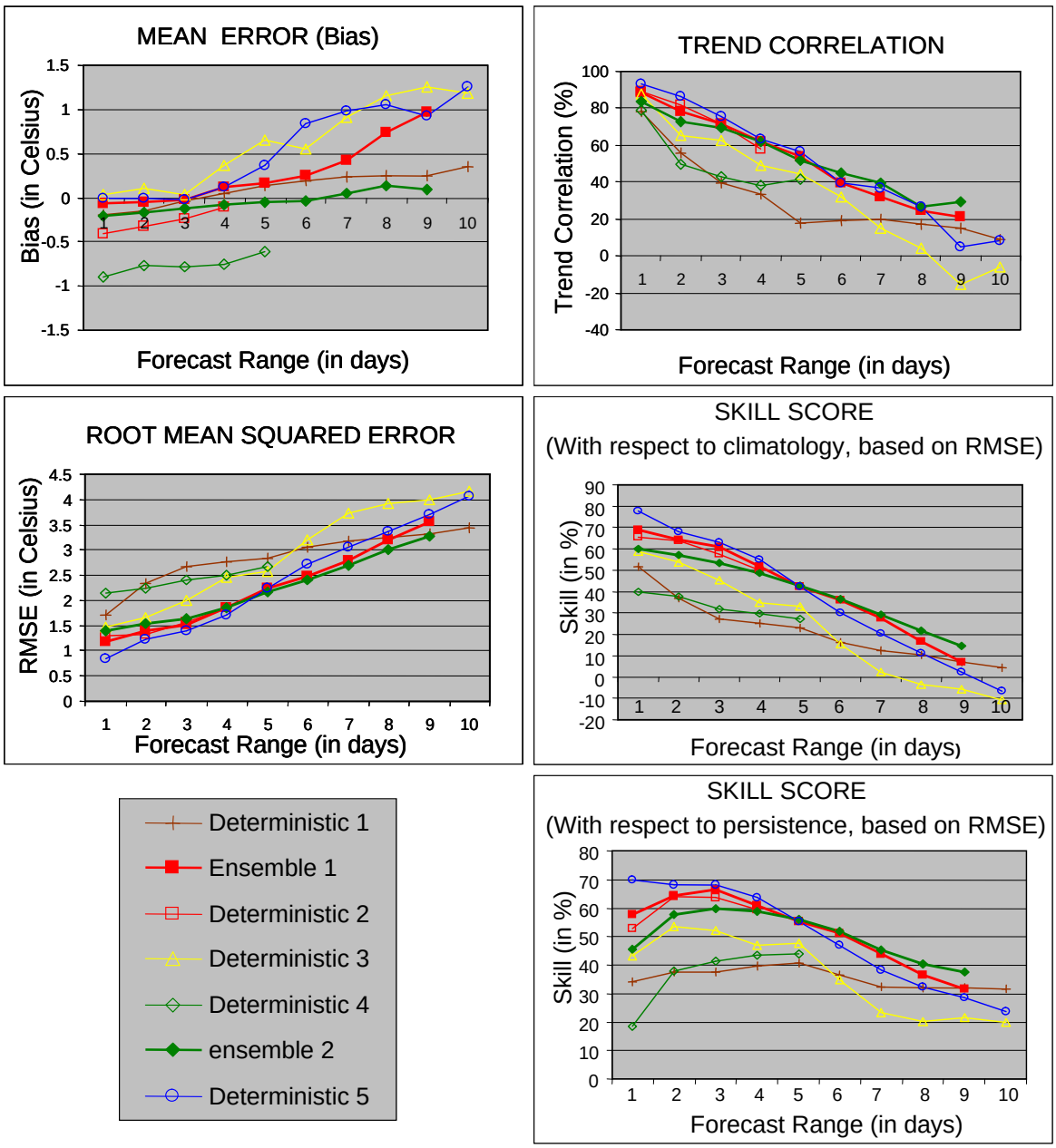

Figure 12 - Results of a simple benchmark test between different forecast providers using the following metrics: Mean Error (top left), Trend Correlation (top right) and Root mean Squared Error (middle left). Skill scores based on RMSE - using climatology (middle right) and persistence (bottom right) as reference forecasts - are also shown. The forecast variable under scrutiny is the daily mean surface temperature.

Values very close to 0.5 pointed to wrong units (confusion between m/s and knots), a suspicion that was confirmed later by the provider in question. Blue (Deterministic 5) was the only provider to issue wind speed forecasts up to day

10. Blue's forecasts appear to be the most accurate in the first four days. However, they do not show any skill with respect to climatology beyond day 4. This suggests that values close to climatology were used by the provider in the latter part of the forecasts to avoid negative skill due to sharp loss of predictability. The last 6 days of Blue's forecasts have in fact no value for a user who knows the climatological wind speeds.

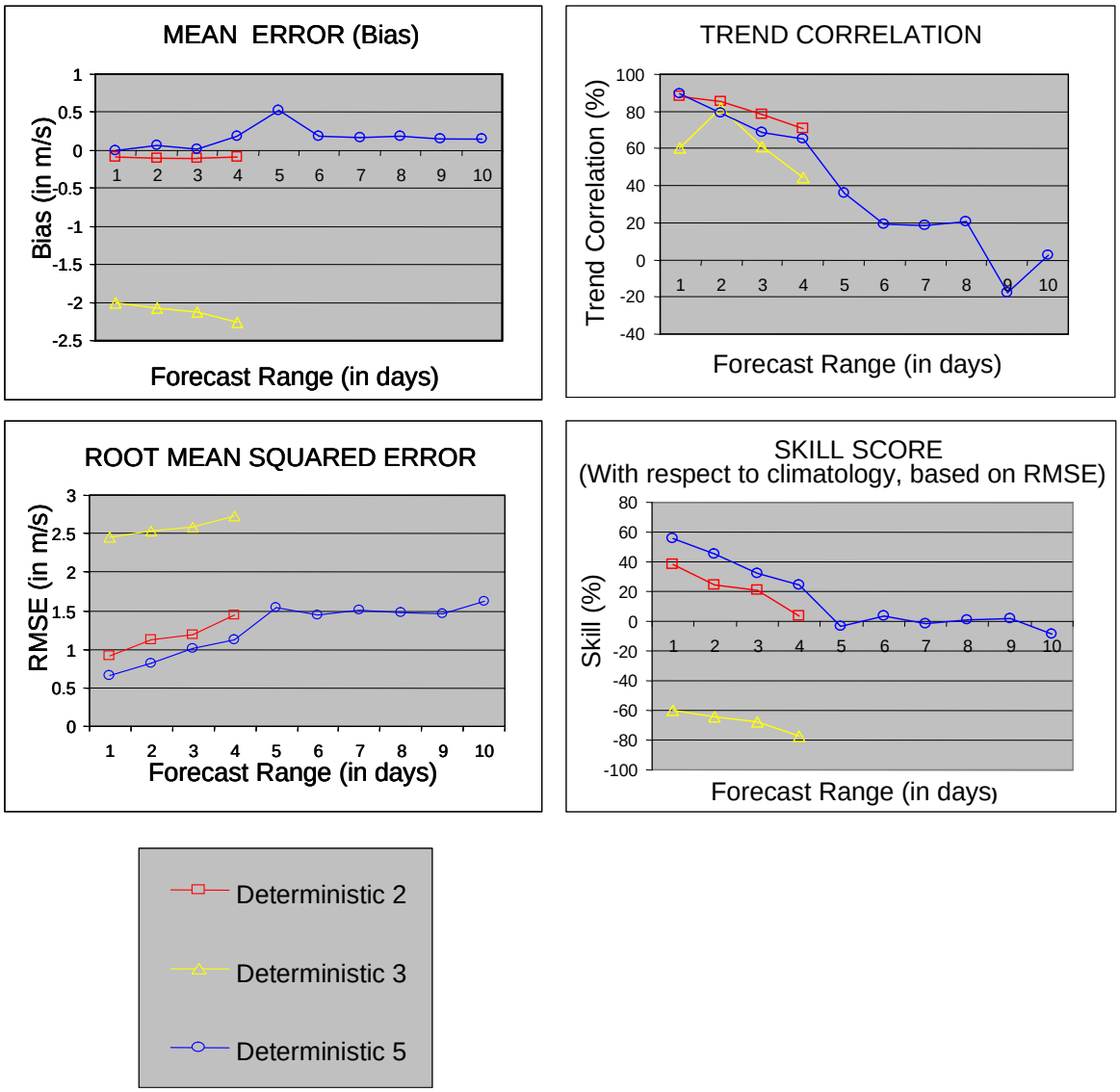

Figure 13 - Benchmark test results for surface wind speed forecasts (skill score using persistence not shown).

It could be argued that the small differences in skill between Deterministic 5, Ensemble 1 and Ensemble 2 in Fig. 12, and between Deterministic 2 and Deterministic 5 in Fig. 13, are not necessarily statistically significant. Ideally, confidence intervals for the metrics should have been plotted, e.g. in the form of error bars. Evaluating the uncertainty of metric estimates is an important component of a proper quality assessment strategy that is often omitted by providers and users, and this was the case in this example. The importance of

assessing the statistical significance of metric estimates is illustrated in the next section.

4.2. Statistical significance of metric estimates

An important aspect that is often currently neglected in forecast quality assessment is the uncertainty or statistical significance of the metric estimates (e.g. Jolliffe, 2005). A particular set of forecasts used for the assessment can be regarded as just one of many possible samples from a population with certain fixed characteristics. The metric estimates are therefore only finite sample estimates of 'true' population values, and as such are subject to sampling uncertainty.

Uncertainty in the measures can be indicated as confidence intervals or error bars calculated using approximations to statistical distributions, or through computer-intensive techniques such as re-sampling (Wilks, 1995, Section 5.3.2). Many assessment metrics used for binary forecasts – e.g. H and F - are sample proportions (probabilities), so confidence intervals for these metrics can be derived from the binomial distribution.

When comparing the performances of two forecast systems, an approach based on hypothesis testing can also be adopted (Mailier, 1997; Hamill, 1999). The null hypothesis is that there is no underlying difference in quality between the two forecasts, and the alternative hypothesis is that one of the forecasts is better than the other for the user application considered. A suitable metric must be chosen, and the values of the metric estimated for each forecast. Then, the significance level (p-value) of the difference between the two estimated values is evaluated. If the p-value is very small the data support the alternative hypothesis. If the p-value is large the data support the null hypothesis.

An example is given of two sets of 182 deterministic 10-day forecasts (1 October 2004 – 31 March 2005). The variable considered is again the daily average 2-m temperature at an undisclosed location. Forecast accuracy has been measured using the Mean Absolute Error (MAE):

$$MAE(t) = \frac{1}{n}.\sum_{i=1}^{n} | F_i(t) - O_i(t) | .$$

The results for the period 1-31 December 2004 (31 paired forecasts) are shown in Fig. 14. The plot indicates that during December 2004, one set of forecasts (blue dots) has errors of smaller magnitude on average than the other (red dots) throughout the forecast range. The differences in accuracy drop after day 6, then seem to increase again at day 10. Does this really mean that the blue forecasts have become more accurate again at day 10? More generally, how certain are these MAE estimates?

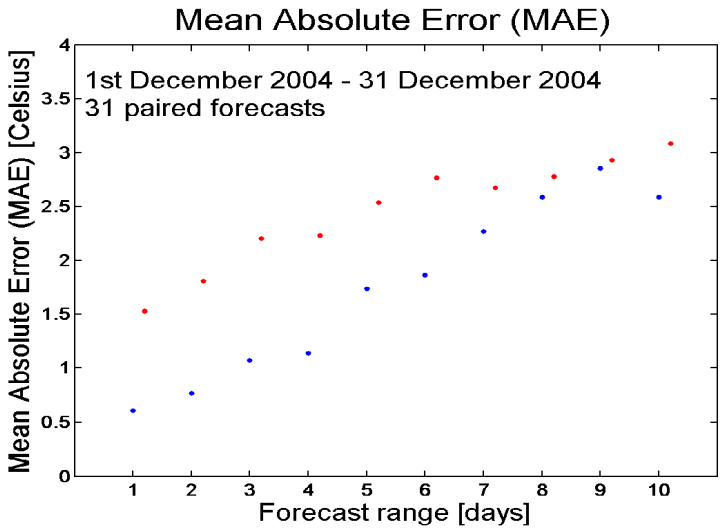

Figure 14 – Point estimates of Mean Absolute Errors for two paired sets of 31 forecasts. Blue and red colours have been used to identify the different sets.

A possible way to quantify this uncertainty is to calculate confidence intervals for the metric. In this case, 95% confidence intervals of the MAE were obtained using a simple bootstrap resampling procedure (Wilks, 1995, Section 5.3.2). The results in Fig. 15 have been computed using 10,000 bootstrap iterations. In the first four days of the forecast range, separate intervals provide evidence that the blue forecasts are more accurate than the red forecasts at the 95% level. In the latter part of the forecast range, the intervals overlap, and the evidence in support of the blue forecasts' higher accuracy weakens. This convergence in accuracy beyond day 5, which is typical with deterministic forecasts, is due to the gradual loss of predictability in the medium range.

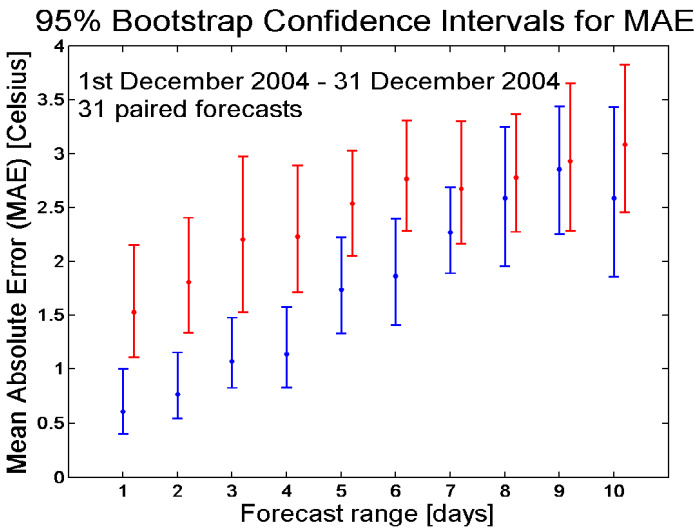

Figure 15 – 95% bootstrap confidence intervals for the Mean Absolute Errors of Fig. 14.

More accurate MAE estimates (narrower confidence intervals) can be achieved by increasing the size of the forecast sample. MAE point estimates and 95% confidence intervals have been calculated for the whole 6-month period. The

results in Fig. 16 below confirm that, at the 95% level, the blue forecasts are more accurate than the red forecasts throughout the first half of the forecast range, and that there is no evidence of a significant difference in accuracy in the latter part of the forecast range. Also note that the MAEs are smaller overall in Fig. 16 than in Fig 15. This difference is at least partly accounted for by the influence of the flow pattern and season on the accuracy of weather forecasts. December 2004 appears to have been a difficult month for forecasting compared to the rest of the 6-month period.

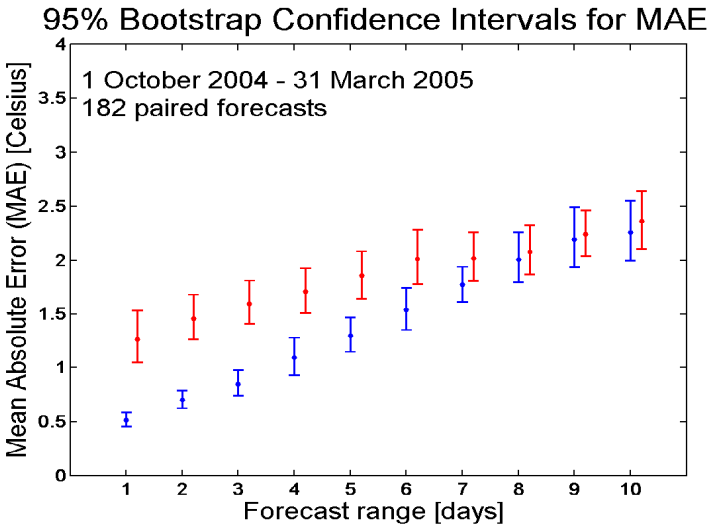

Figure 16 – MAE point estimates and 95% confidence intervals for the two paired sets of 182 forecasts.

The null hypothesis H_0 that the blue forecasts are on average as accurate as the red forecasts has been formally tested against the alternative hypothesis H_1 that the blue forecasts are on average more accurate than the red forecasts. The non-parametric Wilcoxon signed-rank test (Wilks, 1995, Section 5.3.1) is used here because it is more powerful than the sign test and the assumptions necessary to use a t-test are dubious. The p-values of the Wilcoxon test statistic calculated from the observed differences in absolute errors are shown in Fig. 17. The very small p-values in the first seven days of the forecast range (virtually indistinguishable from zero from day 1 to day 6) constitute very strong evidence against H_0 and in favour of H_1. However, the p-values shoot well above the 5% line in the latter part of the forecast range, and H_0 is not rejected at the 5% level from day 8 to 10. Also note that, at the 1% level, H_0 cannot be rejected at day 7.

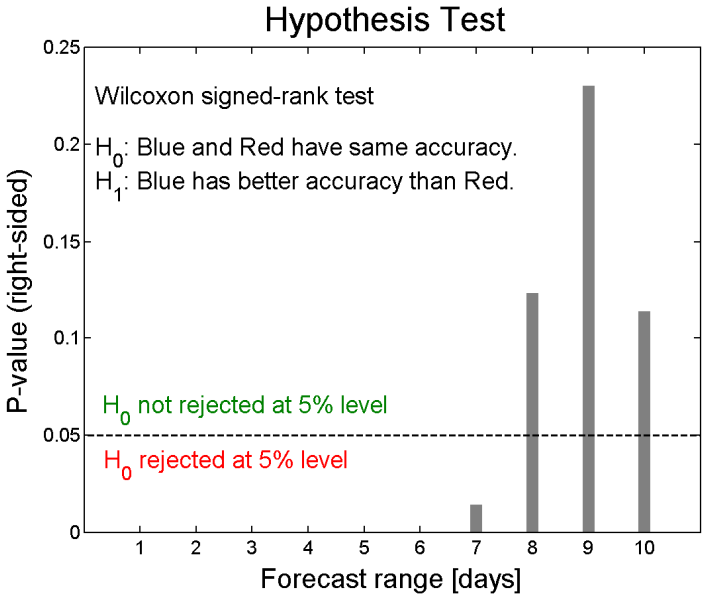

Figure 17 – P-values of the Wilcoxon test statistic calculated from the observed differences in absolute errors. The null hypothesis (H_0) is that both forecasts have equal accuracy, and the alternative hypothesis (H_1) is that the blue forecasts are more accurate than the red forecasts (one-sided test).

The results from this hypothesis test are mostly consistent with the information given by the confidence intervals, but there seems to be an inconsistency at day 7. Figure 17 shows a significant difference at the 5% level, but the confidence intervals in Fig. 16 have considerable overlap. This apparent discrepancy occurs because using overlap of confidence intervals to assess differences is rarely a powerful way of finding such differences. It is preferable to find a single confidence interval for the difference itself (Jolliffe, 2005). Figure 18 shows 95%

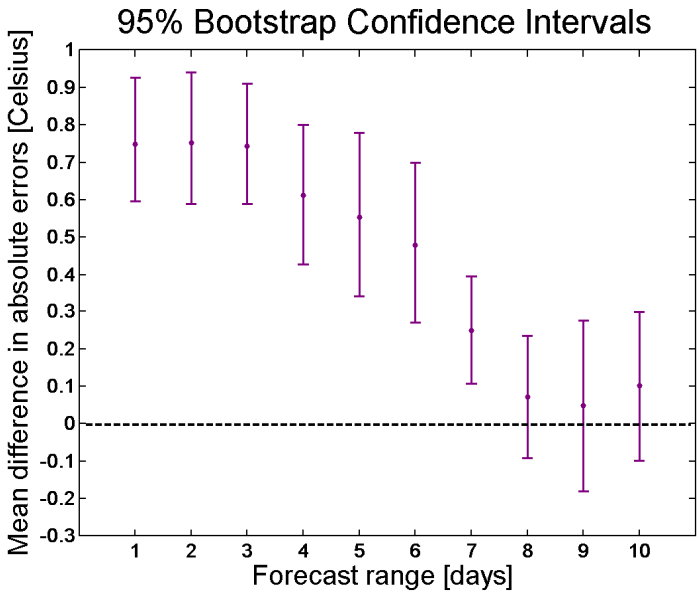

Figure 18 – Point estimates and **95% c**onfidence intervals for the mean difference in absolute errors between the paired forecasts. Values above/below the dashed zero line correspond to the blue forecasts being more/less accurate (smaller absolute errors) than the red forecasts.

confidence intervals for the mean difference in absolute errors between forecasts. Positive/negative values correspond to the blue forecasts being more/less accurate than the red forecasts. The confidence intervals are entirely above the zero line from day 1 to day 7 whereas they lap over it at days 8, 9 and 10. This time, the result given by the confidence intervals is completely in accord with the results from the hypothesis test.

4.3. Interval forecasts

One type of forecast included in the survey, but not in Jolliffe and Stephenson (2003), are interval forecasts. Verification of these has not been much discussed in the literature, but survey results indicate that interval forecasts are commonly produced and used. Although it is never possible to evaluate the quality of an *individual* interval, an obvious way to judge the reliability of a set of interval forecasts is to count the proportion of times that observations fall in the interval, and compare this proportion with the nominal confidence assigned to the interval. For example, perfectly reliable 75% interval forecasts should include the observations 75% of the time. If the observations fall inside the intervals less than 75% of the time (i.e. outside more than 25% of the time), then the forecaster is overconfident (interval too narrow). Conversely, if the observations fall inside the intervals more than 75% of the time, then the forecaster lacks confidence (interval too large). Reliability, however, is only one forecast attribute. Another important attribute that is often neglected with interval forecasts is accuracy. Gneiting and Raftery (2004) have proposed a special score to assess the accuracy of interval forecasts. This metric is based on a cost function that penalizes both wide intervals and intervals that 'miss' an observation. An illustration of the method is given below.

In this example, we examine a set of 182 interval forecasts based on ensembles of 51 members for the period 1 October 2003 – 31 March 2004. The products are 9-day 90% central prediction intervals of daily mean temperatures at one undisclosed location. The question asked is "how good" these intervals are for a user who wants them as narrow and accurate as possible. To be perfectly reliable, the intervals must contain the observations 90% of the time. The solid blue curve in Fig. 19 shows that this is not the case, and that in general the prediction intervals are too narrow (overconfident forecasts). Reliability is best when the observations fall inside the intervals 86 to 88% of the time at days 3, 4 and 5, but it is very poor earlier in the forecast range (days 1 and 2), and drops again beyond day 5. The severe lack of reliability in the short range is due to the nature of the initial perturbations that are used to generate the ensemble members. These perturbations are designed to work optimally in the medium range, i.e. beyond day 2. The deteriorating reliability beyond day 5 is indicative of insufficient spread in the ensemble. The reliability of prediction intervals obtained from ensemble forecasts can nonetheless be substantially improved through calibration. However, a sophisticated and expensive ensemble prediction system is not necessary to deliver very reliable forecast intervals. The dashed red line in Fig.

19 demonstrates that excellent reliability can be easily achieved using prediction intervals based on climatology. The lower and upper bounds of these intervals correspond respectively to the 5th and 95th quantiles of the climate distribution (calculated from a 100-year long de-trended time series) of daily mean temperatures for the location and period considered in this example. In spite of their high reliability, prediction intervals based on climatology are not accurate at all, and therefore useless for users who are interested in the actual fluctuations of daily mean temperatures during the 9-day forecast horizon.

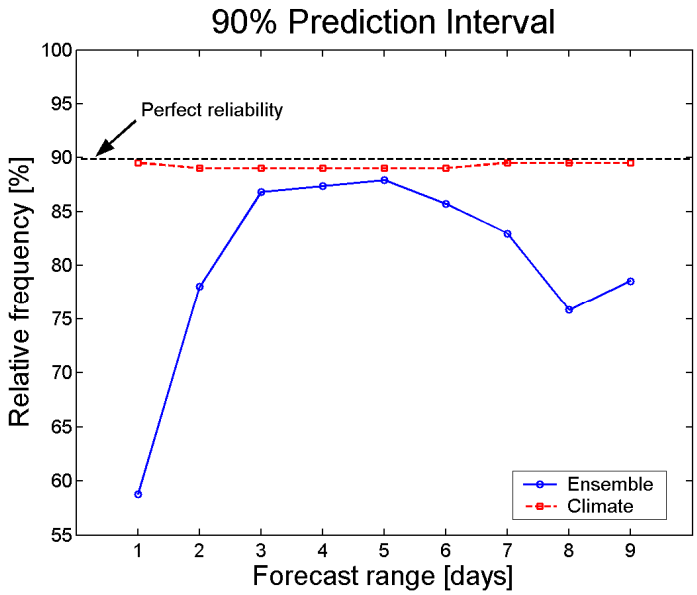

Figure 19 – Estimated probability that the observed daily mean temperatures fall within the prediction intervals for intervals based on ensemble forecasts (blue solid line) and for intervals based on climatological data (red dashed line).

The interval score (Gneiting and Raftery, 2004) is a new metric that has been especially designed to measure the accuracy of prediction intervals. Like the MAE, the MSE and the Brier score, this metric is in essence a cost function which assigns a fixed penalty proportional to the width of the interval, and an additional penalty when the observation falls outside the prediction interval that is proportional to how far the observation is from the interval. For n central $(1-\alpha)\,100\%$ prediction intervals, if the interval forecast is $[l_i, u_i]$ and the observation x_i, then the penalty $P_{\alpha,i}$ is defined by:

$$P_{\alpha,i} = \begin{cases} \alpha(u_i - l_i) + 2(l_i - x_i) & \text{if } x_i < l_i \\ \alpha(u_i - l_i) & \text{if } x_i \in [l_i, u_i] \\ \alpha(u_i - l_i) + 2(x_i - u_i) & \text{if } x_i > u_i \,, \end{cases}$$

and the interval score S_α is simply the average penalty:

$$S_\alpha = \frac{1}{n}\sum_{i=1}^{n} P_{\alpha,i} \,.$$

The interval scores achieved by the two types of prediction intervals (ensemble forecasts and climatology) are shown in Fig. 20. Lower (higher) scores indicate higher (lower) accuracy. Prediction intervals based on climatology achieve a

nearly constant score of 1.5°C because the penalties incurred are due to the fact that climatological intervals are very wide and almost invariable. Despite being much less reliable in the short range, prediction intervals based on ensemble forecasts score much better in accuracy than those based on climatology thanks to their narrowness and ability to stay close to the observations (typically less than 0.5°C off). From day 3 onwards though, they undergo a gradual loss in accuracy and at day 9 both prediction systems are equally accurate.

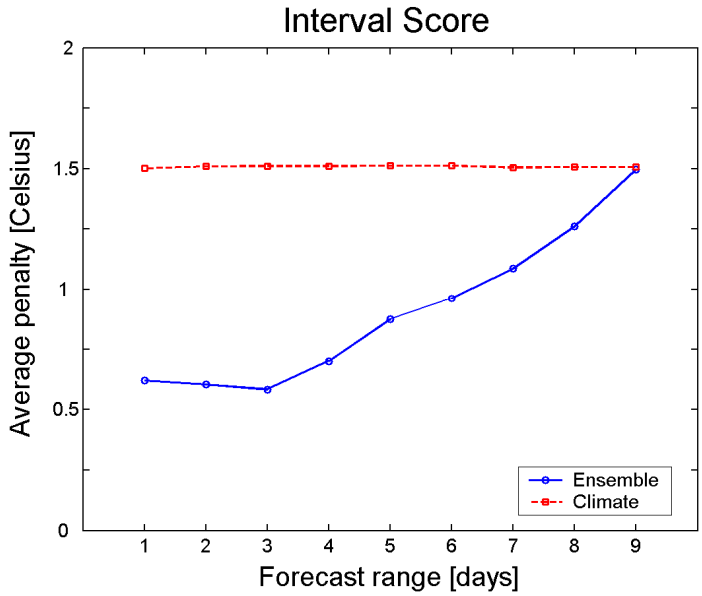

Figure 20 – Interval scores achieved by the prediction intervals based on ensemble forecasts (solid blue line) and climatological data (red dashed line).

The two measures of accuracy can be combined in a single metric that measures the relative performance of the ensemble-based intervals compared to climatological intervals. A simple skill score SS can be easily defined as:

$$SS = \frac{S_\alpha(c) - S_\alpha(f)}{S_\alpha(c)},$$

where $S_\alpha(f)$ and $S_\alpha(c)$ are the interval scores based on ensemble forecasts and climatology, respectively. Perfect ensemble-forecast intervals ($S_\alpha(f) = 0$, corresponding to point forecasts with no errors) would yield $SS = 1$. $SS = 0$ when $S_\alpha(f) = S_\alpha(c)$, i.e. when ensemble-forecast intervals are as accurate as climatological intervals. Positive (negative) values of SS mean that the ensemble-forecast intervals are more (less) accurate, than the climatological intervals. The estimated values of SS expressed in % are plotted in Fig. 21. Approximate 95% bootstrap confidence intervals for SS have been added for completeness.

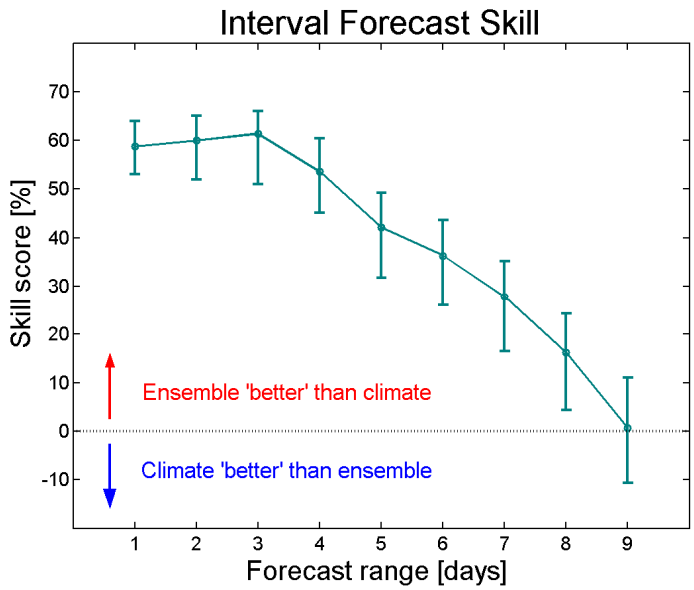

Figure 21 – Estimated skill scores with 95% bootstrap confidence bars for the prediction intervals based on ensemble forecasts compared with climatology.

4.4. Prediction of the winter NAO index

Predominant winter conditions in northern Europe are associated with the average sign of the winter North Atlantic Oscillation (NAO) index. A positive (negative) sign of the winter NAO corresponds to predominantly mild and wet (cold and dry) winter weather. Figure 22 shows the time series of forecast and observed December-to-January (DJF) average NAO indices from 1948/9 to 2004/5. Estimates of several quality assessment metrics for this set of forecasts are given in the second column of Table 1 (assessment period: 1950/1 – 2004/5). Estimates of the same metrics obtained by forecasting no change (persistence from the previous year) are given for comparison in the third column. The proportion of correct forecasts of the sign of the winter NAO index (percent correct PC) is 67%. The p-value for $\hat{p} = 0.67$ given by the binomial distribution with parameter $p = 0.5$ and number of trials $n = 55$ is 0.003, i.e. very small. Therefore, it can be claimed that the forecasts show significant skill compared to purely random forecasts with fair odds (like the tossing of a fair coin). For large enough sample sizes n, 95% confidence intervals (CI) for estimated sample proportions $\hat{p}$ are given by a modified form of the Wald CI (Agresti and Coull, 1998):

$$\tilde{p} \pm 1.96\sqrt{\tilde{p}(1-\tilde{p})/n},$$

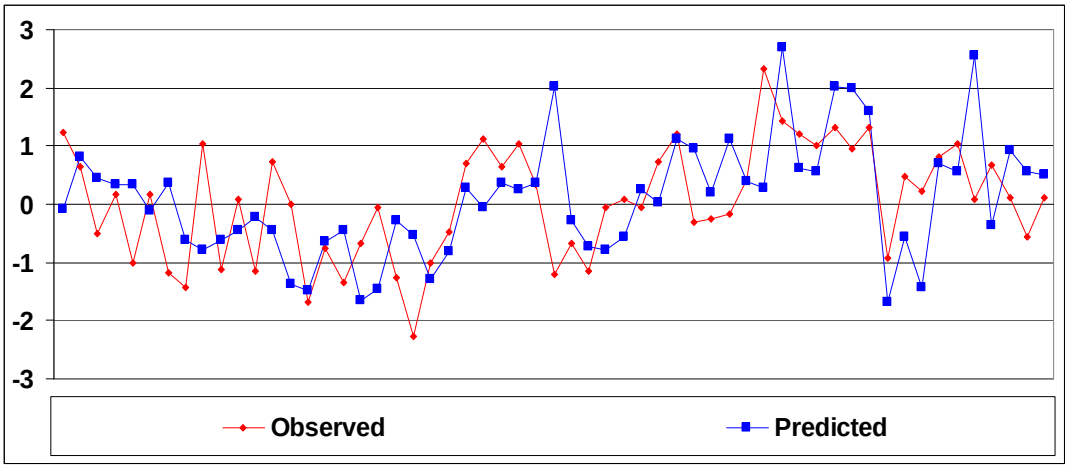

Figure 22 – Time series of forecasts (in blue) and observations (in red) of DJF NAO indices from 1948/9 to 2004/5.

where $\tilde{p}$ is a new proportion obtained by adding 2 successes and 2 failures to the sample. For the NAO forecasts, $\hat{p} = 0.67$, $\tilde{p} = 0.66$, and the CI is [0.54,0.78].

Metric	Forecast	Persistence
Percent Correct for Sign	0.67	0.60
Mean Error	0.08	0.01
Variance Ratio	1.05	1.00
Mean Absolute Error	0.85	0.85
Mean Squared Error	1.12	1.06
Correlation	0.43	0.42

Table 1 – Estimates of various quality assessment metrics for the NAO forecast of Fig. 22 (second column), and for persistence (third column).

In the case of persistence, the CI is [0.47, 0.72]. The considerable CI overlap indicates that the forecasts are only marginally better than persistence when assessed on PC. Estimates of other metrics in Table 1 confirm that there is little difference in performance between the forecasts and persistence.

Another reference forecast, also based on persistence, can be made by simply predicting the average index over the two preceding winters:

$$NAO_i = \frac{1}{2}\left(NAO_{i-1} + NAO_{i-2}\right).$$

The time series for this new benchmark forecast – hereafter labelled as "MA-2 persistence" – is displayed in Fig. 23. MA-2 stands for "2-year moving average". Note that the first forecast in the sample is for the winter 1950/1.

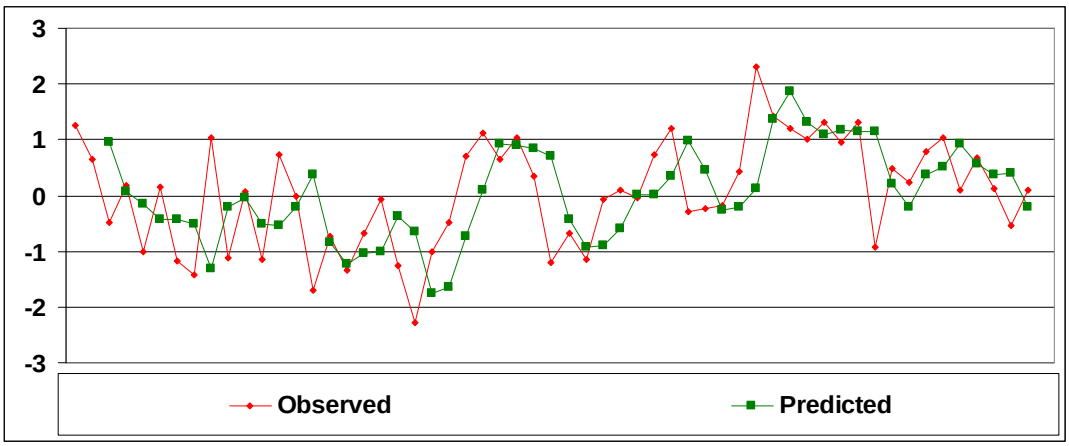

Figure 23 – Time series of MA-2 persistence forecasts (in green) and observations (in red) of DJF NAO indices from 1948/9 to 2004/5 (from 1950/1 to 2004/5 for the observations).

Inspection of the metric estimates in Table 2 reveals that the forecasts do not perform any better than MA-2 persistence. Some metric estimates (PC, ME, correlation, F for "+", H for "-", odds ratio) rather suggest that MA-2 persistence performs marginally better than the forecasts, but the difference is not significant considering the sample size ($n = 55$). The difference of 2% in PC corresponds to just one more correct prediction of the sign of the winter NAO, which may be very

Metric	Forecast	MA-2 Persistence
Percent Correct for Sign	0.67	0.69
Mean Error	0.08	0.03
Variance Ratio	1.05	0.73
Mean Absolute Error	0.85	0.72
Mean Squared Error	1.12	0.87
Correlation	0.43	0.45
Hit Rate for + Sign	0.69	0.69
False Alarm Rate for + Sign	0.35	0.30
Hit Rate for - Sign	0.65	0.69
False Alarm Rate for - Sign	0.31	0.31
Odds Ratio Skill Score	0.62	0.67

Table 2 – Estimates of various quality assessment metrics for the NAO forecast of Fig. 22 (second column), and for MA-2 persistence (third column).

well explained by random effects. MA-2 persistence also shows higher accuracy (smaller errors) than the forecasts. This apparent gain in accuracy is an artefact of the smoothing that results from the two-winter averaging process. More in particular, MA-2 persistence is inherently incapable of predicting extremes, and this inability to reproduce the natural variability of the winter NAO indices is reflected by the small value of the estimated variance ratio (0.73). The variance ratio estimate is obtained by dividing the sample variance of the predicted NAO indices by the sample variance of the observed NAO indices. Note that the p-value given by the F distribution for this estimate is 0.125, therefore the statistical evidence is too weak to reject with reasonable confidence the hypothesis that predicted and observed NAO indices have the same variance. However, we know that the variance of the predicted NAO indices must be smaller in this case

because they are average indices. The lack of statistical significance results from the small sample size (55 cases), a problem that is common in seasonal forecasting.

So, the winter NAO index forecasts do not perform any better than simple alternative forecasts which, because they are based on persistence, have no skill to predict index changes from winter to winter. This point is fundamental, because forecasts that are not capable of predicting change are virtually useless for making decisions. The skill of the forecasts with respect to random forecasts is essentially due to their ability to follow the decadal trends in the winter NAO indices. A metric that measures the forecast aptitude to predict changes in sign/magnitude of the NAO index would better reflect its useful forecasting ability.

Finally, it is important to emphasise that owing to its high sensitivity to the observed frequency of occurrence of positive or negative indices, the use of PC for assessing the performance of the forecasts is not recommended. Figure 24 shows the time series of a December-to-March (DJFM) NAO index based on sea-level pressure differences between Lisbon and Reykjavik. The time series is characterised by decadal trends resulting in a prolonged period when negative indices predominate,

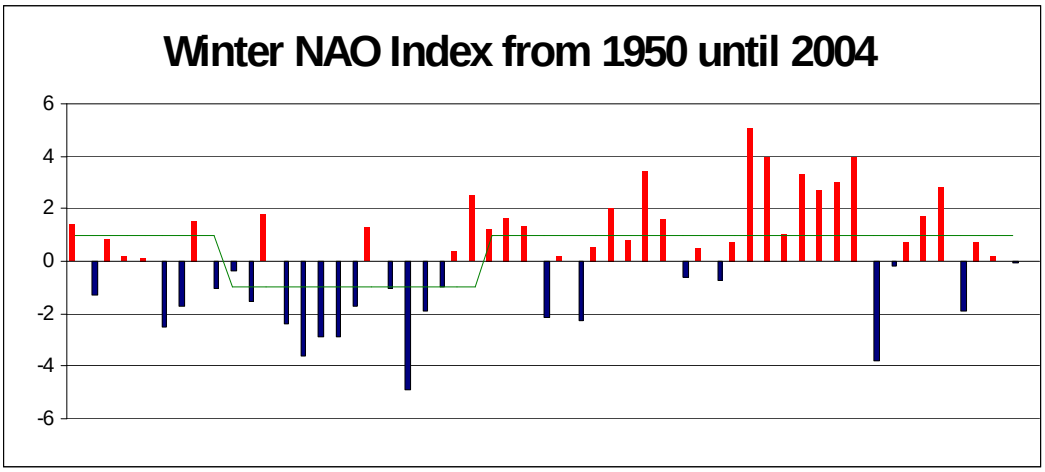

Figure 24 – Time series of DJFM NAO indices from 1950/1 to 2004/5 (Source: J. Hurrell). Positive values are coloured in red, and negative values in blue. The green (+1/-1) line represents the forecasted sign.

followed by another, longer period dominated by positive indices. It is easy to design a forecasting system that identifies the 'mood' (negative or positive) of the NAO index, and always predicts the appropriate sign until it detects a change of mood. By systematically predicting the prevailing sign, it is possible to achieve a high PC while still not being able to predict one single change of sign correctly! An algorithm for such a forecast is given by:

$$sign(NAO_i) = sign\left[\sum_{j=i-4}^{i-1} sign(NAO_j) + 0.5\right],$$

where sign(.) is either +1 or -1 depending on whether the argument is positive or negative. The forecasted sign is represented by the green line in Fig. 24. In this case, the PC achieved by the forecasts is 71%, demonstrating that a relatively high PC does not necessarily mean a skilful forecast.

PART III: CONCLUSIONS AND RECOMMENDATIONS

In this part of the report:

- Section 5 sums up the main results of Sections 2, 3 and 4. More specifically, it recalls :
  - The highlights of the consultation covered in Section 2 (Subsection 5.1);
  - The most important findings from the literature review of Section 3 (Subsection 5.2);
  - The case studies of Section 4 and the lessons learnt from them (Subsection 5.3).

- Section 6 lists the recommendations that follow from the findings summarised in Section 5. It contains:
  - Recommendations to forecast providers and forecast users (Subsections 6.1 and 6.2);
  - Recommendations to the Royal Meteorological Society (Subsection 6.3).

- Finally, Section 7 presents a general conclusion and gives some suggestions for future work.

5. SUMMARY OF MAIN FINDINGS

5.1. Consultation (survey, visits, interviews and workshop)

The consultation was characterised by a surprisingly low participation rate in the on-line survey from both forecast providers and users. Received responses still provide useful information, but for many questions it has been difficult to generalise the results and draw firm conclusions owing to the small sample sizes.

5.1.1. Providers

The majority of consulted providers recognise the overall benefits of setting up quality standards in the industry, but there is currently no agreement on a common strategy. Although a significant number of providers took part in the survey, a large section of the industry – more than 50% – has opted not to do so. This lack of enthusiasm suggests that in fact there may be less appetite for quality standards in the industry than suggested by the survey responses and interviews. Consulted providers acknowledge the need for close collaboration with customers to produce measures of forecast quality that are more meaningful for users than the current ones. However, user feedback indicates that a number of providers have been reluctant to engage their customers in the consultation.

5.1.2. Users

The poor response rate from the user community is due to a combination of various factors:

- Users haven't been strongly encouraged to respond by their providers;
- Forecast quality is not necessarily the user's prime concern;
- Some users have been reluctant to give information deemed commercially sensitive.

A few users voiced the desirability of a better understanding of the customer's needs in the quality assessment of weather forecasts. Despite the fact that information on forecast performance can help the customer to use the forecasts sensibly, a large proportion of consulted users do not receive quality assessments that they find easy to understand and/or useful. However, all responding users declare that they are satisfied with the forecasts they buy. This apparent non-association between customer satisfaction and the availability of clear quality assessment suggests that users may be less concerned about information on forecast quality than initially thought, or that the importance of such information is not well understood.

5.1.3. Classification of forecast products

Forecast products commonly in use have time horizons from the very short up to the seasonal range. The most widely used forecast format is quantitative. This format lends itself well to objective assessment methods. Descriptive forecasts are also in common use, but their quality is much harder to assess. Types of quantitative forecasts reported in the survey are: deterministic, interval, categorical, probabilistic and binary. Probability and binary forecasts do not

appear to be used much, but this may be due to the characteristics of the user sample in the survey. However, providers have confirmed that the market for probability forecasts is indeed quite small. The lukewarm reception of probabilistic products may be at least partly explained by:

- Some reluctance to transfer the "Yes/No" decision stage from the forecaster to the user (no-one to blame when the wrong decision is taken);
- The negative connotation of probability implying ignorance.

5.1.4. Methods of forecast delivery

The internet is the most common means of product dissemination. This technology has allowed fast delivery to an increasing number of users with early availability requirements. For these users, forecast timeliness has become part and parcel of product quality.

5.1.5. Methods of assessment and metrics

Most of the methods and metrics commonly used by consulted providers are well documented in the existing literature (Subsection 5.1.2). Users are less open about their assessment methodologies. Areas that lack coverage in the literature or need special attention are dealt with in Section 4.

5.1.6. Independent monitoring body and online forum

The majority of providers and users who took part in the survey are clearly in favour of both an independent monitoring body and an online forum. However, in view of the limited success of the survey, it is reasonable to suspect that respondents are naturally inclined to support these ideas, but that a significant proportion of forecast providers and users who did not respond may be indifferent, or even hostile to them.

5.2. Literature review

The literature on forecast quality assessment is largely written to cater for the needs of forecast model developers. However, many existing methods and metrics are also relevant to assess the quality of forecasts from a user's viewpoint. A comprehensive list of common assessment metrics with full discussion of their merits and limitations is presented in the book edited by Jolliffe and Stephenson (2003). The use of one single metric may be appealing to convey information on forecast quality in a simple and easily understandable way. However, one metric on its own is inadequate to quantify the overall quality of a set of forecasts. Furthermore, forecast samples must be representative and large enough to achieve statistical significance. Methods to calculate the required sample sizes required are given in the classical statistical literature. The small sample size of past forecasts and observations is particularly problematic in seasonal forecasting. Pooling data with different climatologies may also lead to misleading assessment results. Furthermore, it must be pointed out that the common quality assessment metrics listed below do not work well when dealing with rare or extreme events.

5.2.1. Binary forecasts

Either the hit rate or the false alarm rate alone does not provide a measure of the skill of a binary forecast, but together they do. Despite its simplicity, the use of percent correct is often misleading because of its high sensitivity to the frequency of occurrence of the event being forecasted. The odds ratio skill score offers a good alternative, but it is less easy to explain to users.

5.2.2. Categorical forecasts

The easiest way to assess the quality of categorical forecasts is by reducing them to a series of binary forecasts. There are also metrics that take order and proximity between categories into account, but these metrics are more complex and difficult to explain to users. A similar approach can be used to assess wind direction forecasts, but in this case the problem is complicated by the circularity of the categories.

5.2.3. Point forecasts

Metrics based on errors (forecast minus observed) are commonly used. The mean error measures forecast bias. The mean absolute error measures forecast accuracy. The (root) mean squared error also measures forecast accuracy, but it penalises more large errors than small errors. The mean absolute percentage error is another measure of accuracy that compensates for error increases that are associated with observations of larger amplitude. Correlation measures are commonly used to assess forecast association. Pearson's product-moment correlation is less resistant to outliers than Spearman's rank correlation. When using these metrics, forecast skill can be assessed by comparing the score of the forecast with the corresponding score of a reference forecast that has no skill, e.g. persistence, climatology or random forecast, and hence computing a skill score.

5.2.4. Probabilistic forecasts

The Brier score, which is similar to the mean squared error, is the dominant basic metric. It measures several attributes of the forecasts at the same time, the most important being reliability (calibration) and resolution. Measures of these attributes must be obtained separately either through decomposition of the Brier score, or by direct computation. Sharpness, which is equivalent to resolution for perfectly calibrated forecasts, measures the information content of the forecasts. Another powerful method for assessing probability forecasts is based on the Relative Operating Characteristic curve.

5.2.5. Interval forecasts

Interval forecasts are the most popular type of probability forecasts. This popularity is explained by the fact that intervals provide the user with a combination of point value and a measure of forecast confidence. The quality assessment of interval forecasts is traditionally concerned with their reliability, but their accuracy has been neglected in the literature. A new metric that measures the accuracy of interval forecasts is introduced in Subsection 4.3.

5.2.6. Descriptive forecasts

Although such forecasts are valuable to some users, they are virtually impossible to verify with total objectivity.

5.3. Case studies

Examples involving most types (binary, point, probabilistic-interval) and time ranges (short, medium and seasonal) are considered in these cases.

5.3.1. Simple methods for point forecasts

This example illustrates the use of several simple classical metrics for benchmarking point-value forecasts. It highlights the facts that each aspect of forecast quality in which the user is interested has to be assessed individually using appropriate metrics. It also exemplifies the relative character of forecast skill, showing that forecasts can appear more or less skilful depending on the reference forecast being used. Finally, this case study draws the attention to the fact that a proper quality assessment strategy requires that the uncertainty of metric estimates be evaluated, a point that introduces the next example.

5.3.2. Statistical significance of metric estimates

The performance of weather forecast systems depends on the season and flow pattern, but quality assessment metrics can only be estimated using samples of limited size and characteristics. This case study demonstrates the importance of assessing the uncertainty of metric estimates that results from sampling limitations. The reduction of uncertainty owing to longer, more representative samples is exemplified. Information on metric uncertainty is crucial to quantify and understand the significance of apparent differences in forecast quality. The computation of confidence intervals for metric estimates can be best achieved by means of resampling methods when no standard statistical distribution can be used. Hypothesis tests are strikingly rare in the current assessment methodology although they provide a clear and rigorous decision framework that takes due account of metric uncertainty, and this point is illustrated by an example that corroborates the results from the confidence intervals.

5.3.3. Interval forecasts

Quantitative forecasts expressed in the form of prediction intervals are used commonly, yet the traditional methodology does not offer appropriate metrics to assess the accuracy of such intervals. The most commonly assessed attribute of interval forecasts is reliability. This case study demonstrates that good reliability does not necessarily guarantee prediction intervals that are informative. More generally, it illustrates the incompleteness of one single measure for assessing forecast quality. A new metric for measuring the accuracy of interval forecasts is introduced, and it is used to show that despite being less reliable than climatological intervals, prediction intervals based on a large ensemble of forecasts are more accurate.

5.3.4. Prediction of the winter NAO index

Forecasts of the winter NAO index have been examined using a set of metrics and two reference forecast systems based on persistence. Statistical significance and confidence intervals for metric estimates have been formally calculated where

required. It has been found that the forecasts of the sign of the winter NAO index are significantly more skilful than purely random forecasts with fair odds. However, winter NAO index forecasts are not more skilful than forecasts based on persistence. This point is crucial because forecasts based on persistence have no skill for predicting change, but it is precisely in possible future changes that users have an interest. The skill of the winter NAO index forecasts with respect to random forecasts is essentially due to their ability to follow the decadal trends. A metric that measures the forecast aptitude to predict changes would better reflect its useful forecasting ability.

This example demonstrates that any claim of forecast skill is always relative to some reference forecasts (in this case random and persistence forecasts), and hence that such claim does not make sense to the user if the reference forecasts are unknown. The weaknesses of two common metrics have also been exposed in this case study. The percent correct is not a reliable accuracy metric for binary forecasts because it is very sensitive to the observed frequency of occurrence of the event in the sample. In addition, mean absolute and mean squared errors of point forecasts can be artificially reduced by smoothing.

6. RECOMMENDATIONS

6.1. Recommendations for good quality assessment practice

6.1.1. *Forecast quality assessments on a routine basis.*

The provision of regular quality assessments by providers should be seen as part and parcel of a proper weather forecasting service. The information provided in such assessments should enable the users to know the performance characteristics of the forecast products they buy – more particularly their limitations, and thereby help them to use the forecasts sensibly.

6.1.2. *Assessment methodology*

The assessment procedures should be clearly and fully described with all technical terms and jargon explained. Metrics should be unambiguously defined in plain words and/or by using correct equations. The systematic use of graphics (e.g. plots, histograms, boxplots, bull's eye diagrams, etc.) is encouraged to illustrate the assessment results.

6.1.3. Forecast format

Forecasts should be presented so far as possible in formats that are amenable to objective quality assessment. Qualitative terms should be avoided wherever feasible, and any claim for skill of descriptive forecasts should be treated with scepticism.

6.1.4. Reproducibility of assessment results

The assessment methodology, metrics and documentation should be such that the quality assessment could in principle be repeated by the user or an independent third party.

6.1.5. Relevance of assessments to the user application

The methodology and metrics should be carefully chosen so as to produce information that is meaningful to the user. Providers should accept responsibility for ensuring that this is so, if necessary by education of users. A two-way dialogue is necessary to ensure that the users get what they need.

6.1.6. Completeness of assessments

Assessments should take into account the multi-faceted nature of quality. Methodologies and metrics that attempt to summarise various forecast attributes into one single composite measure are not encouraged. A sufficiently large number of metrics should be presented so as to give an honest and comprehensive summary of the different facets of forecast performance. If required, quantitative assessment results should be illustrated with graphics and complemented with explanations and commentary in plain words.

6.1.7. Use of skill scores

Whenever possible, forecast quality measures should be compared to the ones obtained using reference forecasts for the same assessment period – e.g. persistence, climatology or random forecasts. This helps put the forecast performance in context. Care should be taken to select appropriate reference forecasts so that the measured skill reflects the true usefulness of the forecast. When selecting reference forecasts, one should be aware that random forecasts are generally the least skilful reference, and that persistence is more (less) skilful than climatology in the short (long) ranges. Any claim of forecast skill should always mention the reference forecast that has been used.

6.1.8. Statistical properties of metrics

Metrics may possess statistical properties that sometimes make a forecast system look good when in fact it is poor for a particular application. Users should be made aware of the statistical properties and possible deficiencies of the metrics used for the quality assessment.

6.1.9. Statistical significance of metric estimates

Uncertainty in the metric estimates due to the finite assessment period should be quantified and presented in a simple format that the user can easily understand. Recommended formats are confidence intervals, standard errors (square root of estimated error variance) or p-values.

6.1.10. Sample choice

The choice of sample used for the assessment, more particularly its meteorological and statistical characteristics (weather types, size, homogeneity), should be justified. The chosen period should be long enough to provide stable and representative metric estimates, and the data should be as homogeneous as feasible in space and time. In cases where heterogeneity arises due to missing data, the presence of trends or different flow regimes, the impact of these sample features on the results must be appraised. When testing forecast systems, adequate procedures such as cross-validation (i.e. the data used for the verification are not used in the forecast) should be in place in order to prevent artificial skill. Where feasible, retroactive forecasting (hindcast) should be avoided.

6.2. Specific recommendations concerning quality assessment metrics

6.2.1. Simplicity

Metrics should be as simple as possible so as to provide meaningful and easy to understand quality summaries. However, they should not be overly simple so as to be inappropriate. The purpose of a metric should be to reveal, and not to conceal, one particular aspect of forecast quality. Single composite metrics that combine several aspects of forecast quality should be avoided because unexpected changes in value may be more difficult to interpret.

6.2.2. Robustness

The evaluation of uncertainty on metric estimates should rely on as few assumptions as possible. Care should be taken that the assumptions made are realistic, and that the results are sufficiently stable when departing slightly from them.

6.2.3. Resistance

Metrics should not be unduly dependent on the presence of outlier observations or forecasts in the verification period.

6.2.4. *Consistency*

Metrics should be difficult to improve by "hedging" the forecasts. The best scores should be obtained for forecasting systems that are consistent with the forecaster's true beliefs rather than for systems that have been modified so as to get improved scores.

6.2.5. *Independence*

Metrics should not take account of the means by which the forecasts are produced.

6.2.6. Discrimination

Metrics should separate the net forecast effect on value from the impact of the decision maker's policy.

6.2.7. Specific recommendations on which metrics to use

Formal definitions, discussions, and further references for the metrics below are given in Jolliffe and Stephenson (2003), Wilks (1995), and Gneiting and Raftery (2004).

- Binary forecasts – To measure accuracy, the hit and false alarm rates are appropriate in most situations, but they should always be used together; the proportion correct should be avoided. The base rate (event probability) should always be quoted. The odds ratio is appropriate to measure forecast association. The frequency bias is useful to detect systematic over/underforecasting.

- Categorical forecasts – Forecasts with multiple categories can be reduced to a series of binary forecasts. Gerrity scores may be more appropriate for ordinal categories, but they are not easy to explain and interpret.

- Point (deterministic) forecasts – Forecast bias is measured by the mean error. Good accuracy measures are the mean absolute error and the (root) mean squared error (less resistant to outliers). The mean absolute percentage error may be useful in cases where forecast errors increase as the observations get larger (e.g. quantitative precipitation forecasts). Association is assessed using

simple correlation measures (Pearson's product-moment correlation less resistant to outliers than rank correlation). The variance ratio is useful to show how realistic the forecasts are in reproducing the observed variability.

- Probabilistic – The use of the Brier score alone without decomposition is not recommended. Reliability together with resolution and/or sharpness provide useful summaries of forecast performance. The ROC curve and the area under it are also powerful assessment tools that are closely linked with economic value and other quality assessment metrics for binary forecasts.

- Interval - Reliability is the best measure to assess the probabilistic fitness of the intervals, but is inadequate on its own. The interval score is recommended to determine accuracy.

- Forecast skill – The use of skill scores is strongly encouraged. Any claim for skill should always specify the 'no-skill' reference used against the forecasts (e.g. random guess, persistence, climatology).

6.3. Recommendations to the Royal Meteorological Society

The low level of participation in the consultation has revealed that it is extremely difficult to engage the whole marketplace in a comprehensive and open debate around the issue of the quality of weather forecasts. Findings from the consultation summarised in Subsection 5.1 point to several behavioural and market-related factors that may account for these difficulties. A large number of users may not be motivated simply because they are not interested, or because they do not understand the importance of the issue. Some providers have been hostile to the project. Others may be satisfied with the current situation and feel uncomfortable at the prospect of seeing their customers becoming more aware of forecast quality matters. The recommendations that follow aim at raising the profile of the issue of weather forecast quality, increasing user awareness, and promoting a more progressive and open culture in the industry that favours and maintains high quality standards for the benefit of the whole community.

6.3.1. Establish a Special Commission on the weather forecasting industry

There is clear support from consulted forecast providers and users for establishing an independent regulatory and monitoring body, but this does not include input from some of the key players in the private sector and from many users who may be indifferent, or even hostile, to this idea. An official watchdog would have to be funded by the industry, and this is unlikely to happen in the current situation. As an alternative, we propose a less coercive scheme where participation and voluntary adherence to a code of practice are encouraged.

It is recommended that the Royal Meteorological Society first establishes a specialised commission that would deal with matters specific to the weather forecasting industry. Its main mission should be essentially to facilitate communication and openness, to inform and educate forecast users, and to promote the benefits of adopting common quality assessment standards and practices. The proposed commission should play a role similar to the US

Commission on the Weather and Climate Enterprise (CWCE[1]), which was recently set up by the Council of the American Meteorological Society (AMS). The CWCE is charged with the following responsibilities (as quoted from their web site):

- Develop and implement programs that address the needs and concerns of all sectors of the weather and climate enterprise;
- Promote a sense of community among government entities, private sector organizations, and universities;
- Foster synergistic linkages between and among the sectors;
- Entrain and educate user communities on the value of weather and climate information;
- Provide appropriate venues and opportunities for communications that foster frank, open, and balanced discussions of points of contention and concern.

6.3.2. Set up a committee on weather forecast quality standards

The proposed commission should appoint an ad-hoc committee to run a certification scheme for providers who adhere to a prescribed code of practice. This code of practice should specify the professional, scientific and technical standards to be met for accreditation. Standards and recommended practices in the field of forecast quality assessment should be based on the recommendations made above in Subsections 6.1 and 6.2. Companies applying for accreditation should agree to submit themselves to independent, regular audits.

6.3.3. Develop and maintain dedicated web pages

The creation of an open online forum where users and providers would be able to submit their problems on forecast quality issues has been found desirable by a majority of respondents to the survey. However, the difficulties experienced to get providers to mobilise their customers and the surprisingly low user response rate suggest that an online forum may not have the success than one might assume from the survey responses. Furthermore, it is probable that without appropriate moderation, the forum will not fulfil its objective and even be open to abuse.

However, in order to facilitate information and education, the proposed commission should endeavour to develop and maintain dedicated pages on the Society's web site. These web pages should be adequately advertised and made publicly available. They should include the code of practice for providers, and dispense information, education and basic guidelines on matters regarding weather forecast quality.

[1] http://www.ametsoc.org/boardpges/cwce/

6.3.4. *Raise public awareness through publicity at high-profile events and in the media.*

It is possible that a better response to the survey would have been achieved if some resources could have been allocated to a preliminary marketing and advertising campaign. The proposed commission should use every opportunity to raise the profile and awareness of the issue of weather forecast quality through appropriate communication channels, in particular:

- Encourage the publication of letters and articles on weather forecast quality topics in the non-meteorological literature;
- Publish information leaflets to be freely distributed at conferences and workshops;
- Organise high-profile events – e.g. forecasting contests similar to those held under the auspices of the AMS[2] - that demonstrate the importance of good practice and quality standards.

[2] See http://www.meteor.iastate.edu/~miraje/AFC/ for the 2005/6 competition.

7. CONCLUSION AND FUTURE DIRECTIONS

An important lesson learned from this project is the absence of a sense of community between weather forecast providers and between forecast users. This fragmented state - and the lack of constructive dialogue that results - constitute a major obstacle to establishing a commonly agreed strategy for better quality standards in the industry. Fundamental changes of disposition and attitude are required. The role of forecast providers should reach from the mere distribution of products to the delivery of a genuine service that includes the provision of user-oriented forecast quality assessments and the necessary user education. Information on forecast performance should be seen as an essential part of a 'User Guide' that helps users to make sensible use of the products they buy. Uncertainty in the forecasts and in the metric estimates should be treated as valuable information instead of ignorance. Unfortunately, the current background of increasingly aggressive competition in the marketplace does not favour openness on forecast performance at a time when more transparency is needed. It is hoped that the commission proposed to the Society in Subsection 6.3 will foster a more cooperative and participative culture within the industry.

The problem of assessing the quality of weather forecasts from a user standpoint is far more complex than the traditional forecaster-oriented verification because it must take the user's own requirements into account. Many of the already existing techniques can be easily applied to assess forecast quality for users. If needed, new, simple methods and metrics can also be designed to answer specific questions from a user on forecast performance. However, there are important aspects of the quality of service offered by weather forecast providers that cannot be assessed by simple objective metrics, for example the way the forecasts are presented to the user, or the provision of subjective forecast guidance by a meteorologist. Moreover, the case studies in this project have looked at forecast quality from the perspective of industrial, agricultural or financial decision makers who use weather forecasts to mitigate (optimise) weather-related losses (profits). The principal reason for this selection is simply that it has been mainly users from this category who have responded to the survey. A definition of forecast quality for the media gives probably more weight to the efficacy of the graphics and attention getters while giving less weight to accuracy. Nevertheless, a standard checklist containing the important basic questions that providers should be asked could be a useful aid for many users whatever their profile, and the drawing up of such a checklist could be a future task for the committee proposed in Subsection 6.3.

Specific recommendations on which metrics to use have been made in Subsection 6.2. These recommendations do no purport to confine forecast quality assessment to a rigid set of prescribed metrics. Considering the increasing variety of weather forecast products and the growing number of applications, quality assessment techniques are bound to become more complex and diversified. In Subsection 2.2, some consulted stakeholders expressed the wish to see more collaborative work involving providers and users. There is no doubt that the successful

development of future user-specific quality assessment methods and metrics will require more synergy between both ends of the forecasting line.

An approach to forecast quality assessment based on fuzzy logic was briefly discussed in Subsection 3.1. When adopting a 'fuzzy quality assessment' strategy, forecasts that are 'close' to the observations are not so bad as forecasts that are 'far' off, and therefore they are potentially more useful. This approach is very appealing for practical applications because it gives the user considerable flexibility to specify – objectively or subjectively – the structure of the membership functions that define the 'goodness' or 'badness' of the forecasts. This avenue of research may be worth pursuing, but ultimately it is the demand arising from practical user applications that must give the directions for future advances in the development of user-oriented methodologies and metrics.

PART IV: BIBLIOGRAPHY AND APPENDICES

8. REFERENCES

Accadia, C., M. Casaioli, S. Mariani et al. , 2003a: Application of a statistical methodology for limited area model intercomparison using a bootstrap technique. *Nuovo Cimento C*, **26**(1), 61-77.

Accadia, C., S. Mariani, M. Casaioli et al., 2003b: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18** (5), 918-932.

Agresti, A. and B.A. Coull, 1998: Approximate is better than 'exact' for interval estimation of binomial proportions. *Amer. Statist.*, **52**, 119-126.

Atger, F., 2004: Estimation of the reliability of ensemble-based probabilistic forecasts. *Q.J.R. Meteorol. Soc.*, **130** (597), 627-646 Part B.

Bessler, D.A and R. Ruffley, 2004: Prequential analysis of stock market returns. *Appl. Econ.*, **36**(5), 399-412.

Bradley, A.A. and T. Hashino, S.S. Schwartz, 2003: Distributions-oriented verification of probability forecasts for small data samples. *Wea. Forecasting*, **18**(5), 903-917.

Bradley, A.A., S. S. Schwartz, and T. Hashino, 2004: Distributions-oriented verification of ensemble streamflow predictions. *J. Hydrometeorol.*, **5**(3), 532-545.

Casati, B., G. Ross, and D.B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Met. Applications*, **11**, 141-154.

Coelho, C.A.S., S. Pezzulli, M. Balmaseda, F. J. Doblas-Reyes and D.B. Stephenson, 2004: Forecast calibration and combination: A simple Bayesian approach for ENSO. *J. Climate*, **17**(7), 1504-1516.

Coelho, C.A.S., S. Pezzulli, M. Balmaseda, F. J. Doblas-Reyes and D.B. Stephenson, 2004: Skill of coupled model seasonal forecasts: A Bayesian assessment of ECMWF ENSO forecasts. *ECMWF Technical Memorandum* No. 426, 16pp.

Coles, S., J. Heffernan, and J. Tawn, 1999: Dependence measures for extreme value analyses, Extremes, 2, 339-365.

Daley, D.J. and D. Vere-Jones, 2004: Scoring probability forecasts for point processes: The entropy score and information gain. *J. Appl. Prob.*, **41A**, 297-312, Sp. Iss. SI 2004.

Déqué, M., 2003: Continuous variables. In: Jolliffe, I.T. and D.B. Stephenson, (Eds.), *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* John Wiley and Sons, 97-119.

Drosdowsky, W. and H. Zhang, 2003: Verification of spatial fields. In: Jolliffe, I.T. and D.B. Stephenson, (Eds.), *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* John Wiley and Sons, 121-163.

Ebert, B., 2002: Fuzzy verification: Giving partial credit to erroneous forecasts. *Presentation given at the NCAR/FAA Workshop on "Making Verification More Meaningful", 30 July – 1 August 2002, NCAR Foothills Laboratory.*

Ebert, B., 2004: Forecast Verification - Issues, Methods and FAQ. *http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html.*

de Elia, R. and R. Laprise, 2005: Diversity in interpretations of probability: Implications for weather forecasting. *Mon. Weather Rev.*, **133**(5), 1129-1143.

Fischoff, B., 1994: What forecast (seem to) mean, *Int. J. Forecasting*, **10**, 387-403.

Fuller, S.R., 2004: Book reviews. *Weather,* **59**(5), 132.

Gandin, L.S. and A.H. Murphy, 1992: Equitable scores for categorical forecasts. *Mon. Wea. Rev*., **120**, 361-370.

Gerrity, J.P.Jr,1992: A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev.*, **120**, 2707-2712.

Giles, G., 2005: We'll rain on your parade, forecasters tell rogue pundits. Nature, **435**, 396-397.

Glahn, H.R., 2004: Discussion of verification concepts in *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. *Wea. Forecasting*. **19**, 769-775.

Gneiting, T. and A. E. Raftery, 2004. Strictly proper scoring rules, prediction and estimation. Technical Report 463, Department of Statistics, University of Washington.

Gneiting, T., A.E. Raftery, A. H. Westveld, et al, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.*, **133**(5), 1098-1118.

Göber, M., C.A. Wilson, S.F. Milton and D.B. Stephenson, 2004: Fairplay in the verification of operational quantitative precipitation forecasts. *J. Hydrology*, **288**, 225-236.

Goddard, L., A. G. Barnston and S. J. Mason, 2003: Evaluation of the IRI's "net assessment" seasonal climate forecasts 1997-2001. *B. Am. Meteorol. Soc.*, **84** (12), 1761-1781.

Haklander, A.J. and A. Van Delden, 2003: Thunderstorm predictors and their forecast skill for the Netherlands. *Atmos. Res.***67**(8*),* 273-299 Sp. Iss.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, 14, 155-167.

Hamill, T. M. and J. Juras, 2005: Measuring forecast skill: is it real skill, or is it the varying climatology? Submitted to *Mon. Wea. Rev.*

Harte, D. And D. Vere-Jones, 2005: The entropy score and its uses in earthquake forecasting. *Pure Appl. Geophys.*, **162**(6-7), 1229-1253.

Hartmann, H. C., T.C. Pagano, S. Sorooshanian and R. Bales, 2002:Confidence builders. Evaluating seasonal climate forecasts from user perspectives. *Bull. Amer. Meteor. Soc.*, **83**, 683-698.

Hilliker, J.L., 2004: The sensitivity of the number of correctly forecasted events to the threat score: A practical application. *Wea. Forecasting*, **19**(3), 646-650.

Jolliffe, I.T., 1979: How accurate are long-range weather forecasts? *New Scientist*, **84**, 192-194.

Jolliffe, I. T., 2005: Uncertainty and inference for verification measures. Submitted to *Wea. Forecasting.*

Jolliffe, I.T. and J.F. Foord, 1975: Assessment of long-range forecasts. *Weather,* **30**, 172-181.

Jolliffe I. T. and N. M. N. Jolliffe, 1997: Assessment of descriptive weather forecasts. *Weather*, **52**, 391-396.

Jolliffe, I.T. and J.M. Potts, 1992: Skill scores based on mean square error and correlations: some further relationships and insights. *Fifth International Meeting on Statistical Climatology*, 343-346.

Jolliffe, I.T. and D.B. Stephenson, Eds., 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, 254 pp.

Jolliffe, I.T. and D.B. Stephenson, 2005: Comment on Discussion of verification concepts in Forecast Verification: A Practitioner's Guide in Atmospheric Science. *Wea. Forecasting*, (in press).

Katz, R.W. and A.H. Murphy, Eds., 1997: *Economic value of weather and climate forecasts*. Cambridge University Press, 222 pp.

Kharin, V.V. and F.W. Zwiers, 2003: On the ROC score of probability forecasts. *J. Climate*, **16** (24): 4145-4150.

Livezey, R. E., 2003: Categorical events. In: Jolliffe, I.T. and D.B. Stephenson, (Eds.), *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* John Wiley and Sons, 77-96.

Lopez, J..A., 2001: Evaluating the predictive accuracy of volatility models. *J. Forecasting*, **20**(2), 87-109.

Mailier, P., 1997: Experimental 4-D VAR evaluation. *ECMWF Daily Meteorological Operations Summary*, 04/10/1997.

Mailier, P., 2001: Ensemble forecasting of extreme mid-latitude cyclones. *MSc dissertation*, U. of Reading, 74 pp.

Marshall, K.T. and R. M. Oliver, R. M., 1995: *Decision Making and Forecasting. M*cGraw-Hill, 427 pp.

Mason, I.B., 2003: Binary events. In Jolliffe and Stephenson (2003), 37-76.

Mason, I., 2004: The cost of uncertainty in weather prediction: Modelling quality-value relationships for yes/no forecasts. *Aust. Meteorol. Mag.*, **53**(2), 111-122.

Mason, S.J., 2004: On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Mon. Weather Rev.*, **132**(7), 1891-1895.

Mason, S.J. and N. E. Graham, 1999: Conditional probabilities, relative operating characteristics, and relative operating levels. *Wea. Forecasting*, **14**(5), 713-725.

Mason, S.J. and N.E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Q. J. R. Meteorol. Soc.*, **128**(584), 2145-2166.

Matthews, R., 1997: 'How right can you be?' *New Scientist*, **2072**, 28-31.

Mozer, J.B. and W.M. Briggs, 2003: Skill in real-time solar wind shock forecasts. *J. Geophys. Res.-Space*, **108**(A6), Art. No. 1262.

Muller, W.A., C. Appenzeller, F.J. Doblas-Reyes et al., 2005.: A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *J. Climate*, **18**(10), 1513-1523.

Murphy, A.H., 1988: Skill scores based on the mean squared error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417-2424.

Murphy, A. H., 1996: 'The Finley Affair: A signal event in the history of forecast verification'. *Wea. Forecasting*, **11**, 236-241.

Mylne, K.R., 2002: Decision-making from probability forecasts based on forecast value. *Meteorol. Appl.* , **9**(3), 307-315.

Önkal, D. and F. Bolger, 2003: Provider-user differences in perceived usefulness of forecasting formats, *Omega*, **32**, 31-39.

Persson, A., 1991: Kalman filtering – a new approach to adaptive statistical interpretation of numerical meteorological forecasts. Lecture presented at the WMO Training Course, Wageningen, The Netherlands, 29 July-9 August 1991. WMO/TD No 421.

Potgieter, A.B., Y. L. Everingham and G.L. Hammer, 2003: On measuring quality of a probabilistic commodity forecast for a system that incorporates seasonal climate forecasts. *Int. J. Climatol.*, **23**(10), 1195-1210.

Potts, J.M., C.K. Folland, I.T. Jolliffe and D.Sexton, 1996: LEPS scores for assessing climate model simulations and long-range forecasts. *J.Climate*, **9**, 34-53. See also letter 1976, **31**, 101.

Rajagopalan, B., U. Lall, S.E. Zebiak, 2002: Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Weather Rev.*, **130**(7), 1792-1811.

Roulston, M.S. and L.A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.*, **130**(6), 1653-1660.

Saulo, A.C. and L. Ferreira, 2003: Evaluation of quantitative precipitation forecasts over southern South America. *Aust. Meteorol. Mag.*, **52**(2), 81-93.

Smith, L.A. and J.A. Hansen, 2004: Extending the limits of ensemble forecast verification with the minimum spanning tree. *Mon. Weather Rev.*, **132**(6), 1522-1528.

Stephenson, D.B., 1997: Correlation of spatial climate/weather maps and the advantages of using the Mahalanobis metric in predictions. *Tellus*, **49A**(5), 513-527.

Stephenson, D.B., 2000: Use of the "odds ratio" for diagnosing forecast skill. *Wea. Forecasting*, **15**(2), 221-232.

Stephenson, D.B., C.A.S. Coelho, M. Balmaseda and F.J. Doblas-Reyes, 2005: Forecast Assimilation: A unified framework for the combination of multi-model weather and climate predictions, *Tellus*, **57A**, 253-264.

Stephenson, D.B. and F.J. Doblas-Reyes, 2000: Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus*, **52**(3), 300-322.

Stewart, T.R., R. Pielke and R. Nath: Understanding user decision making and the value of improved precipitation forecasts - Lessons from a case study. *B. Am. Meteorol. Soc.*, **85**(2), 223-+ .

Thornes, J.E. and D.B. Stephenson, 2001: How to judge the quality and value of weather forecast products, *Meteorological Applications*. 8, 307-314.

Thornes, J.E. and D.B. Stephenson, 2001: How to judge the quality and value of weather forecast products. *Meteorol. Appl.*, **8**(3), 307-314.

Toth, Z., O. Talagrand, G. Candille and Y. Zu, 2003: Probability and ensemble forecasts. In: Jolliffe, I.T. and D.B. Stephenson, (Eds.), *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* John Wiley and Sons, 137-163.

Tustison, B., E. Foufoula-Georgiou and D. Harris, 2002: Scale-recursive estimation for multisensor Quantitative Precipitation Forecast verification: A preliminary assessment. *J. Geophys. Res.-Atmos.*,**108**(D8): Art. No. 8377.

Venugopal, V., S. Basu and E. Foufoula-Georgiou, 2005: A new metric for comparing precipitation patterns with an application to ensemble forecasts. *J. Geophys. Res.-Atmos.*, **110**(D8): Art. No. D08111.

Wandishin, M.S. and H.E. Brooks, 2002: On the relationship between Clayton's skill score and expected value for forecasts of binary events. *Meteorol. Appl.*, **9**(4), 455-459.

Wei, M.Z. and Z.A. Toth, 2003: New measure of ensemble performance: Perturbation versus error correlation analysis (PECA). *Mon. Weather Rev.*, **131**(8), 1549-1565.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.

Wilks, D.S., 2000: On interpretation of probabilistic climate forecasts. *J. Climate*, **13**, 1965-1971.

Wilks, D.S., 2001: A skill score based on economic value for probability forecasts. *Meteorological Applications*, **8**, 209-219.

Wilks, D.S., 2002: Smoothing forecast ensembles with fitted probability distributions. *Q. J.R. Meteorol. Soc.*, **128,** 2821-2136.

Wilks, D.S., 2004: The Minimum Spanning Tree (MST) histogram as a verification tool for multidimensional ensemble forecasts. *Mon. Weather Rev.*, **132**, 1329-1340.

Wilks, D..S. and C.M. Godfrey, 2002: Diagnostic Verification of the IRI Net Assessment Forecasts 1997-2000. *J. of Climate*, **15**, 1369-1377.

Winkler, R. L., 1996: Scoring rules and the evaluation of probabilities (with comments). *Test*, **5**, 1-60.

Wright P B and C. R. Flood, 1973: A method of assessing long-range weather forecasts. *Weather*, **28**, 178-187.

Yates, J.F., P.C. Price, J.W. Lee and J. Ramirez, 1996: Good probabilistic forecasters: the 'consumer's' perspective. *Int. J. Forecasting*, **12**, 41-56.

Yuan, H., S.L. Mullen, X. Gao et al., 2005: Verification of probabilistic quantitative precipitation forecasts over the southwest United States during winter 2002/03 by the RSM ensemble system. *Mon. Weather Rev.*, **133**(1), 279-294.

Zhang, H. and T. Casey, 2000: Verification of categorical probability forecasts. *Wea. Forecasting*, **15**, 80-89.

APPENDIX A: PROVIDER SURVEY WITH RESULTS

Q1: Which areas below best correspond to the sector(s) of activity of your customers?

Please choose all that apply

Agriculture	8
Catering	2
Energy	11
Financial	12
Health	2
Manufacturing	5
Media (Radio/Television/Press/Film industry)	6
Military (Army/Navy/RAF)	2
Offshore oil/gas industry	5
Public Sector	6
Retail	9
Tourism/Entertainment	8
Transport (road)	8
Transport (maritime)	5
Transport (air)	7
Other:	3

Q2: How many paying customers do you have?

Please choose only one of the following:

1-5	3
6-10	1
11-20	1
More than 20	13

Q10: How do you produce the forecasts you sell?

Please choose all that apply

We run our own proprietary forecasting model(s) - e.g MM5	11
We receive and process gridded values (e.g. GRIB) given by numerical prediction models from other agencies (Met Office, ECMWF, NCEP,...)	11
We receive forecast products from other providers and re-interpret them for the specific needs of our own customers	5
Other:	0

Q15: Do you apply statistical corrections (e.g. MOS, Kalman filtering) or other techniques to fine-tune the forecasts you sell? If this is the case, please list the methods you use.

Please choose only one of the following:

Yes	12
No	5
No answer	1

Please enter your comment here:

The next questions deal exclusively with commercial weather forecast products you sell to paying customers.
The emphasis is on forecasts applicable to the UK and its marine environs at and near to the earth's surface.

Q30: Please indicate the range(s) of all the forecast products you sell:

Please choose all that apply

0-2 hours (nowcast)	10
2-12 hours (very short range)	12
12-72 hours (short range)	16
3-10 days (medium range)	15
10-30 days (extended range)	12
1-3 months (long range/monthly)	8
3 months-2 years (long range/seasonal)	7
beyond 2 years (long range/climate)	0

Q31: What is the MAXIMUM forecast range for each of the predicted weather variables you sell?

Please choose the appropriate response for each item

	0-2 hrs	2-12 hrs	12-72 hrs	3-10 days	10-30 days	1-3 mths	3mths -2yrs	>2yrs
Temperature	0	0	3	4	3	2	6	0
Precipitation	0	0	3	6	1	2	5	0
Wind speed	0	0	4	9	1	1	1	0
Wind direction	0	0	3	9	1	1	1	0
Humidity	0	0	6	6	0	1	0	0
Pressure	0	0	3	8	2	1	0	0
Cloud cover	0	0	8	5	0	1	0	0
Visibility	0	0	8	3	1	1	0	0
Solar radiation	0	0	4	2	0	1	0	0
Wave height	0	0	3	6	0	0	0	0
Significant weather	0	0	6	7	0	1	1	0
Extreme weather	0	0	9	3	0	1	1	0
Other	0	0	2	6	0	0	1	9

Q40: Information in weather forecasts can be conveyed to end users using quantitative and/or qualitative formats: series of numbers, symbols, predefined words or expressions, purely descriptive texts and maps. Please indicate to what extent each type of format is used in the forecast products you sell to your customers.
Data in GRIB (GRIdded Binary) format count as numbers.
Please choose the appropriate response for each item

	Not used	Not used much	Used moderately	Used a lot
Numbers (predicted values, probabilities,…)	0	2	0	16
Pre-defined word, expressions or symbols (e.g. significant weather)	1	2	6	9
Purely descriptive text or pictures, no pre-agreed definition (e.g.'quite cold with some rain')	2	4	5	7

Q42: For products with quantitative content (numbers), please indicate the forecast type(s).
Please choose all that apply

Simple point-value forecasts (e.g. 'maximum temperature of 17 Celsius')	16
Interval forecasts (e.g. 'wind speed between 5 and 10 knots')	12
Categorical forecasts (e.g. 'above normal, normal, below normal temperatures')	14
Probability forecasts (e.g. probability that rainfall exceeds 5 mm is 30%')	15
Binary forecasts (e.g. 'frost/no frost')	9
Other:	0

Q45: Do you provide your customers with estimates of forecast uncertainty?
Forecast uncertainty can be expressed e.g. by means of confidence intervals, ranges of values, probability, PDFs.
Please choose only one of the following:

Yes, forecasts and uncertainty estimates are provided together.	16
Yes, uncertainty estimates are provided as separate products.	0
No, but uncertainty estimates can be provided on request.	1
No, estimates of forecast uncertainty are not available.	0
Other Some customers have uncertainty provided	1

[Only answer this question if you answered 'No, estimates of forecast uncertainty are not available.' to question 'Q45 ']
Q46: Do you think estimates of forecast uncertainty could improve the usefulness of the forecast products you sell?
Please choose only one of the following:

Yes
No
Don't know

Please enter your comment here:

[Only answer this question if you answered 'Yes, uncertainty estimates are provided as separate products.' or 'Yes, forecasts and uncertainty estimates are provided together.' to question 'Q45 ']
Q47: Please indicate how forecast uncertainty is conveyed to users in the products you sell.
Please choose all that apply

Probabilities (e.g. 'the probability of frost tonight is 70%')	10
Confidence intervals (e.g. 'a 90% confidence interval for the maximum temperature is [7-11]')	9
Confidence indices (e.g. 'the confidence in the warm forecast has risen from 1 to 3')	2
Various forecast scenarios (ensemble forecasts)	6
Pre-agreed expressions/symbols (e.g. 'the uncertainty is high')	6
Freely chosen words (e.g. 'the latest forecast runs are inconsistent with the previous runs')	8
Other:	0

Q50: Please indicate how forecast products are delivered to your customers.

Please choose the appropriate response for each item

	Not used	Not used much	Used moderately	Used a lot
You upload forecasts to customers	2	1	3	12
Customers download forecasts from you	4	3	4	7
Customers read forecasts in your web pages	6	1	4	7
Forecasts are sent to customers by e-mail	2	3	5	8
Forecasts are sent to customers by Fax	5	6	3	4
Forecasts are sent to customers by telephone	6	6	4	2
Forecasts are sent to customers by telex	14	3	1	0
Forecasts are sent to customers by mail/courier	13	5	0	0
Other (e.g. VHF,…)	16	2	0	0

Q55: Do you give your customers the possibility to consult forecasters (e.g. through a dedicated hotline) whenever they require additional forecast guidance?

Please choose only one of the following:

Yes	13
No	5
No answer	0

Please enter your comment here:

Q60: How often do you issue forecast quality assessments to your customers?

Please choose only one of the following:

Frequently (at least once a month)	3
Occasionally (several times a year)	7
Rarely (once a year or less)	6
Never	2

[Only answer this question if you answered 'Frequently (at least once a month)' or 'Occasionally (several times a year)' or 'Rarely (once a year or less)' to question 'Q60 ']

Q61: In what form do you present the quality assessments to your customers?

Please choose all that apply

Quantitative assessment (statistics, e.g. summary of recent forecast errors)	14
Qualitative assessment (e.g. 'The cold wave was well predicted')	9
Other:	0

[Only answer this question if you answered 'Frequently (at least once a month)' or 'Occasionally (several times a year)' or 'Rarely (once a year or less)' to question 'Q60 ']

Q63: Do you think your customers find the quality assessment information easy to understand?

Please choose only one of the following:

Yes	8
No	3
No opinion	5

Please enter your comment here:

[Only answer this question if you answered 'Rarely (once a year or less)' or 'Frequently (at least once a month)' or 'Occasionally (several times a year)' to question 'Q60 ']

Q64: Do you think your customers find the forecast quality assessment information useful?

Please choose only one of the following:

Yes	12
No	0
Uncertain	4

Please enter your comment here:

[Only answer this question if you answered 'Never' to question 'Q60']

Q66: Do you believe that providing your customers with quality assessment information would benefit them?

Please choose only one of the following:

Yes	1
No	1
Don't know	0

Please enter your comment here:

Q70: Do you receive feedback from your customers on the quality of the products that you sell to them?
Please choose only one of the following:

Yes	17
No	1

[Only answer this question if you answered 'Yes' to question 'Q70 ']
Q71: What form of feedback on forecast quality do you receive from your customers?
Please choose all that apply

Quantitative assessment (e.g. look at forecast errors)	8
Qualitative assessment	15
Other:	2

[Only answer this question if you answered 'Yes' to question 'Q70 ']
Q72: On what sample is their quality assessment based?
Please choose all that apply

ALL forecasts in a recent period (within one year or less)	10
A REPRESENTATIVE SAMPLE of recent forecasts	3
A set of SELECTED EVENTS	3
Don't know	3
Other:	2

Q73: Which of the quantities below do you use to assess the quality of the forecasts you sell?
Please choose all that apply

Bias (mean error)	13
Accuracy (mean squared error, mean absolute error, …)	14
Association (e.g. correlation, odds ratio, …)	5
Reliability/Calibration (conditional bias)	5
Sharpness (spread -or information content- of the forecasts)	2
Uncertainty (spread of the observations)	3
Resolution (forecast ability to distinguish between distinct observed events)	4
Discrimination (sensitivity of forecast likelihood to observed values)	0
Economic value (financial benefit from using the forecasts)	6
Other:	0

Q75: Please explain/list the methods/measures (scores) you use to assess forecast quality.
This question is not mandatory.
Please write your answer here:

Q77: Are there any aspects of forecast quality important to users that the available methodologies and measures do not assess sufficiently well? Please explain your answer.
Please choose only one of the following:

Yes	9
No	2
No opinion	7

Please enter your comment here:

Q80: How often do you discuss the quality of your forecasts with your customers?
Please choose only one of the following:

Frequently (at least once a month)	7
Occasionally (several times a year)	9
Rarely (once a year or less)	2
Never	0

Q87: Do you use objective quality assessment to measure the performance of your products against those of other forecast providers?
This question is not mandatory.
Please choose only one of the following:

Yes	6
No	12

Please enter your comment here:

Q90: How satisfied are you with the quality of the forecast products that you sell?
Please choose only one of the following:

Very satisfied	10
Fairly satisfied	8
Dissatisfied	0

Please enter your comment here:

Q95: Please feel free to write any suggestion you may have to improve forecast quality assessment for users.
This question is not mandatory.
Please write your answer here:

Q97: It has been suggested that an independent body might be established that would monitor the weather forecasting sector and encourage good practice in the assessment of forecast quality. How would you view such a body?
Please choose only one of the following:

Necessary	6
Useful, but not necessary	11
Not useful	1
Not desirable	0
No opinion	0

Please enter your comment here:

Q98: It has also been suggested to set up an independent on-line forum where forecast users and providers can submit their problems concerning quality issues and find/offer practical solutions. What do you think of this idea?
Please choose only one of the following:

Necessary	2
Useful, but not necessary	14
Not useful	1
Not desirable	1
No opinion	0

Please enter your comment here:

Q99: We would be grateful if you agreed to leave your email address and/or other contact details in the box below. This information will enable us to get in touch in case there is a problem with your entry.
Optional!
Please write your answer here:

APPENDIX B: USER SURVEY WITH RESULTS

Q1: Which areas below best correspond to your sector(s) of activity?
More than one category may be selected, e.g. for 'Energy trading' check 'Financial' and 'Energy'.
Please choose all that apply

Agriculture	2
Catering	0
Energy	6
Financial	2
Health	0
Manufacturing	0
Media (Radio/Television/Press/Film industry)	0
Military (Army/Navy/RAF)	0
Offshore oil/gas industry	1
Public Sector	0
Retail	6
Tourism/Entertainment	0
Transport (road)	0
Transport (maritime)	0
Transport (air)	1
Other:	0

Q2: How many forecast providers do you BUY products from?
Please do not include forecasts from the media (radio/television/newspapers) or from freely accessible web sites.
Please write your answer here:

1	2	3	4	5	>5
11	1	2	0	2	0

Q5: Do you also use free forecasts for business-related decisions?
Please do not count the use of free forecasts for private decisions like: "Shall I take my umbrella today?"
Please choose only one of the following:

Yes	10
No	6
No answer	0

[Only answer this question if you answered 'Yes' to question 'Q5 ']
Q7: What source(s) of free forecast information do you use?
This question is not mandatory.
Please choose all that apply

Radio/Television	4
Written press (e.g. newspapers)	2
The Internet	10
Other:	1

Q10: What do you use forecast products for?
This question is not mandatory.
Please write your answer here:

The next questions deal exclusively with commercial weather forecast products you BUY from one or several providers.
The emphasis is on forecasts applicable to the UK and its marine environs at and near to the earth's surface.

Q30: Please indicate the range(s) of all the forecast products you buy:
The range (maximum lead time/horizon) is how far ahead the forecast goes.
Please choose all that apply

0-2 hours (nowcast)	1
2-12 hours (very short range)	6
12-72 hours (short range)	9
3-10 days (medium range)	10
10-30 days (extended range)	7
1-3 months (long range/monthly)	5
3 months-2 years (long range/seasonal)	4
beyond 2 years (long range/climate)	0

Q31: What is the MAXIMUM forecast range for each of the predicted weather variables you use?
Please choose the appropriate response for each item

	0-2 hrs	2-12 hrs	12-72 hrs	3-10 days	10-30 days	1-3 mths	3mths -2yrs	>2yrs
Temperature	1	0	1	7	2	3	2	0
Precipitation	0	0	4	6	2	1	1	0
Wind speed	0	0	4	6	1	1	1	0
Wind direction	0	0	5	6	1	1	0	0
Humidity	0	0	3	3	1	0	0	0
Pressure	1	0	1	3	1	0	0	0
Cloud cover	0	0	5	6	1	1	0	0
Visibility	1	0	2	2	1	1	0	0
Solar radiation	0	1	1	1	0	0	0	0
Wave height	0	1	0	0	0	0	0	0
Significant weather	0	1	2	4	2	1	1	0
Extreme weather	0	1	1	8	1	2	0	0
Other	0	0	0	0	0	0	0	0

Q40: Information in weather forecasts can be conveyed to end users using quantitative and/or qualitative formats: series of numbers, symbols, predefined words or expressions, purely descriptive texts and maps. Please indicate to what extent each type of format is used in the forecast products you purchase from your provider(s).
Data in GRIB (GRIdded Binary) format count as numbers.
Please choose the appropriate response for each item

	Not used	Not used much	Used moderately	Used a lot
Numbers (predicted values, probabilities,…)	0	0	3	13
Pre-defined word, expressions or symbols (e.g. significant weather)	2	5	5	5
Purely descriptive text or pictures, no pre-agreed definition (e.g.'quite cold with some rain')	4	2	4	6

Q42: For products with quantitative content (numbers), please indicate the forecast type(s) you use.
Please choose all that apply

Simple point-value forecasts (e.g. 'maximum temperature of 17 Celsius')	15
Interval forecasts (e.g. 'wind speed between 5 and 10 knots')	11
Categorical forecasts (e.g. 'above normal, normal, below normal temperatures')	12
Probability forecasts (e.g. probability that rainfall exceeds 5 mm is 30%')	6
Binary forecasts (e.g. 'frost/no frost')	4
Other:	0

Q45: Are estimates of forecast uncertainty provided to you?
Forecast uncertainty can be expressed e.g. by means of confidence intervals, ranges of values, probability, PDFs.
Please choose only one of the following:

Yes, forecasts and uncertainty estimates are provided together.	14
Yes, uncertainty estimates are provided as separate products.	0
No, but uncertainty estimates can be provided on request.	2
No, estimates of forecast uncertainty are not available.	0
Other	0

[Only answer this question if you answered 'No, estimates of forecast uncertainty are not available.' to question 'Q45 ']
Q46: Do you think estimates of forecast uncertainty could improve the usefulness of the forecast products you buy?
Please choose only one of the following:

Yes
No
Don't know

Please enter your comment here:

[Only answer this question if you answered 'Yes, uncertainty estimates are provided as separate products.' or 'Yes, forecasts and uncertainty estimates are provided together.' to question 'Q45 ']
Q47: Please indicate how forecast uncertainty is conveyed to you in the products you buy.
Please choose all that apply

Probabilities (e.g. 'the probability of frost tonight is 70%')	6
Confidence intervals (e.g. 'a 90% confidence interval for the maximum temperature is [7-11]')	7
Confidence indices (e.g. 'the confidence in the warm forecast has risen from 1 to 3')	1
Various forecast scenarios (ensemble forecasts)	4
Pre-agreed expressions/symbols (e.g. 'the uncertainty is high')	4
Freely chosen words (e.g. 'the latest forecast runs are inconsistent with the previous runs')	1
Other:	0

Q50: Please indicate how you receive forecast products from your providers.

Please choose the appropriate response for each item

	Not used	Not used much	Used moderately	Used a lot
Provider(s) upload(s) forecasts to you	7	1	0	8
You download forecasts from your provider(s)	7	3	3	3
You browse providers' web pages	6	1	3	6
Forecasts are sent to you by e-mail	1	3	3	9
Forecasts are sent to you by Fax	11	3	2	0
Forecasts are sent to you by telephone	13	3	0	0
Forecasts are sent to you by telex	15	1	0	0
Forecasts are sent to you by mail/courier	16	0	0	0
Other (e.g. VHF,…)	15	0	1	0

Q55: Do/does your provider(s) give you the possibility to consult forecasters (e.g. through a dedicated hotline) whenever you require additional forecast guidance?

Please choose only one of the following:

Yes	13
No	3
No answer	0

Please enter your comment here:

Q60: How often do you receive forecast quality assessments from your provider(s)?

Please choose only one of the following:

Frequently (at least once a month)	0
Occasionally (several times a year)	4
Rarely (once a year or less)	5
Never	7

[Only answer this question if you answered 'Frequently (at least once a month)' or 'Occasionally (several times a year)' or 'Rarely (once a year or less)' to question 'Q60 ']

Q61: In what form is the quality assessment presented?

Please choose all that apply

Quantitative assessment (statistics, e.g.summary of recent forecast errors)	6
Qualitative assessment (e.g. 'The cold wave was well predicted')	4
Other:	1

Q62: On what sample is the quality assessment based?

Please choose all that apply

ALL forecasts in a recent period (within one year or less)	3
A REPRESENTATIVE SAMPLE of recent forecasts	2
A set of SELECTED EVENTS	2
Don't know	2
Other:	0

[Only answer this question if you answered 'Frequently (at least once a month)' or 'Occasionally (several times a year)' or 'Rarely (once a year or less)' to question 'Q60 ']

Q63: Do you find the quality assessment information easy to understand??

Please choose only one of the following:

Yes	6
No	1
No opinion	2

Please enter your comment here:

[Only answer this question if you answered 'Rarely (once a year or less)' or 'Frequently (at least once a month)' or 'Occasionally (several times a year)' to question 'Q60 ']

Q64: Do you find the forecast quality assessment information useful?

Please choose only one of the following:

Yes	6
No	0
Uncertain	3

Please enter your comment here:

[Only answer this question if you answered 'Never' to question 'Q60']
Q66: Do you believe that receiving quality assessment information from your provider(s) would benefit you as a user?
Please choose only one of the following:

Yes	4
No	1
Don't know	2

Please enter your comment here:

Q70: Do you make your own assessment of the quality of the forecast products that you buy from providers?
Please choose only one of the following:

Yes	11
No	5

[Only answer this question if you answered 'Yes' to question 'Q70 ']
Q71: How do you assess the quality of the forecast products that you buy?
Please choose all that apply

Quantitative assessment (e.g. look at forecast errors)	9
Qualitative assessment	3
Other:	1

[Only answer this question if you answered 'Yes' to question 'Q70 ']
Q72: On what sample is your quality assessment based?
Please choose all that apply

ALL forecasts in a recent period (within one year or less)	5
A REPRESENTATIVE SAMPLE of recent forecasts	5
A set of SELECTED EVENTS	1
Other:	0

Q73: Which of the quantities below do you use to assess forecast quality?

Please choose all that apply

Bias (mean error)	8
Accuracy (mean squared error, mean absolute error, …)	8
Association (e.g. correlation, odds ratio, …)	2
Reliability/Calibration (conditional bias)	4
Sharpness (spread -or information content- of the forecasts)	1
Uncertainty (spread of the observations)	4
Resolution (forecast ability to distinguish between distinct observed events)	0
Discrimination (sensitivity of forecast likelihood to observed values)	0
Economic value (financial benefit from using the forecasts)	4
Other:	0

Q75: Please explain/list the methods/measures (scores) you use to assess forecast quality.

This question is not mandatory.

Please write your answer here:

Q76: Do you use your own quantitative quality assessment to statistically fine-tune (e.g. calibrate) the forecasts yourself?

Please choose only one of the following:

Yes	1
No	8
No answer	0

Make a comment on your choice here:

Q77: Are there any aspects of forecast quality important to you that the available methodologies and measures do not assess sufficiently well? Please explain your answer.

Please choose only one of the following:

Yes	5
No	5
No opinion	6

Please enter your comment here:

Q80: How often do you discuss your own assessment of forecast quality with your provider(s)?

Please choose only one of the following:

Frequently (at least once a month)	2
Occasionally (several times a year)	5
Rarely (once a year or less)	3
Never	1

Q85: Do you use your own quantitative quality assessment to decide which forecast product to buy -e.g. by testing forecasts from different providers?

Please choose only one of the following:

Yes	7
No	2

Please enter your comment here:

[Only answer this question if you answered 'No' to question 'Q85 ']

Q87: Please explain how you decide which provider to buy products from.

This question is not mandatory.

Please write your answer here:

Q90: How satisfied are you with the quality of the forecast products that you buy?

Please choose only one of the following:

Very satisfied	5
Fairly satisfied	11
Dissatisfied	0

Please enter your comment here:

Q95: Please feel free to write here any suggestion you may have to improve forecast quality assessment regarding your needs as a user.

This question is not mandatory.

Please write your answer here:

Q97: It has been suggested that an independent body might be established that would monitor the weather forecasting sector and encourage good practice in the assessment of forecast quality. How would you view such a body?

Please choose only one of the following:

Necessary	2
Useful, but not necessary	13
Not useful	0
Not desirable	0
No opinion	1

Please enter your comment here:

Q98: It has also been suggested to set up an independent on-line forum where forecast users and providers can submit their problems concerning quality issues and find/offer practical solutions. What do you think of this idea?

Please choose only one of the following:

Necessary	4
Useful, but not necessary	8
Not useful	2
Not desirable	0
No opinion	2

Please enter your comment here:

Q99: We would be grateful if you agreed to leave your email address and/or other contact details in the box below. This information will enable us to get in touch in case there is a problem with your entry.
Optional!

Please write your answer here:

APPENDIX C: EXISTING GUIDELINES FOR FORECAST QUALITY ASSESSMENT

Several guidelines on forecast quality assessment have been compiled. For example, the World Meteorological Organisation (WMO) has commissioned and compiled several reports. Online guidance on verification can be found on these sites:

- **WMO Public Weather Services (PWS) main page on forecast verification.** General guidance regarding verification (go to http://www.wmo.ch and find Forecast Verification under Search by Alphabetical Topics). Currently available directly from http://www.wmo.ch/web/aom/pwsp/qualityassuranceverification.htm

- **WMO Guidelines on Performance Assessment of Public Weather Services, WMO/TD No. 1023.**
  More detailed discussion on forecast verification for public weather services.
  http://www.wmo.ch/web/aom/pwsp/downloads/guidelines/TD-1023.pdf

- **WMO Manual on the Global Data-Processing and Forecasting System (GDPFS), Vol. 1, WMO - No. 485.**
  Standardised verification methods and metrics for NWP products are documented in Annex II.7, Table F. The standardised verification system for long-range forecasts is described in Annex II.9.
  http://www.wmo.int/web/www/DPS/Publications/WMO485.pdf

- **WMO WGNE survey of verification methods for numerical prediction of weather elements and severe weather events.** P. Bougeault, January 2003. Appendix C of Report of the 18th session of the CAS/JSC WGNE.
  http://www.wmo.ch/web/wcrp/wgnepublications.htm

- **WMO WWRP/WGNE Joint Working Group on Verification report on Recommendations for the verification and intercomparison of QPFs from operational NWP models.** December 2004.
  Recommendations on best practice for the quality assessment of quantitative precipitation forecasts (QPF).
  http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/WGNE/QPF_verif_recomm.pdf

- Dr Beth Ebert's web site on forecast verification
  http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html

Unfortunately, there is much duplication of information on verification and no coordinated body within WMO that is solely concerned with the important aspect of forecast quality assessment for users.