METEOROLOGICAL APPLICATIONS
Meteorol. Appl. (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/met.85

Review

Assessing and reporting the quality of commercial weather forecasts

Pascal J. Mailier,* Ian T. Jolliffe and David B. Stephenson
Royal Meteorological Society, Reading, UK

ABSTRACT: In 2005 the Royal Meteorological Society commissioned a study to examine current issues regarding the quality (fitness for purpose) of commercial weather forecasts in the United Kingdom. UK commercial weather forecast providers and users were consulted by means of on-line questionnaires, interviews, visits and an open workshop discussion. Results from this consultation uncovered significant deficiencies in the methodologies and in the communication of forecast quality assessments, a lack of open dialogue and transparency in the industry, and that some users may be indifferent to forecast quality. Descriptive or worded forecasts cannot be assessed objectively. However, suitable quality assessment methods are available for nearly all types of quantitative forecasts identified in the consultation. The crucial importance of choosing proper quality assessment metrics, the impact of their statistical properties on results and the need to estimate the statistical significance of quality assessment results were exemplified by means of four case studies, one of which is presented in this paper. The findings from this study have led to a set of practical recommendations aiming to establish the discipline and rigour that are necessary for achieving best practice in the quality assessment of weather forecasts. Specific recommendations were also made to the Royal Meteorological Society to set up a special commission that would promote a sense of community within the industry, and to run an accreditation scheme that would encourage best practice on a voluntary basis.

KEY WORDS forecast quality; user-oriented verification; forecast value; commercial weather forecasts

Received 3 December 2007; Revised 28 April 2008; Accepted 9 June 2008

1. Introduction

In 2005 the Royal Meteorological Society commissioned a study to examine 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. This paper presents a concise review of the work undertaken and summarizes the results laid out in the original report by Mailier *et al*. (2006). The full report, which also contains numerous references to the literature, can be downloaded freely from the Society's web site.

An open consultation of UK-based commercial weather forecast providers and users was carried out in order to probe current quality-assessment practices in the industry and identify possible issues. The main findings of this consultation are outlined in Section 2. Section 3 focusses on the existing quality assessment methodology and provides a concrete example of application to real forecast data. Recommendations for best practice are made in Section 4. Finally, some concluding remarks and possible future directions are given in Section 5.

* Correspondence to: Pascal J. Mailier, Royal Meteorological Society, 104 Oxford Road, Reading, RG1 7LL, UK. E-mail: survey@rmets.org

2. Consultation

UK commercial weather forecast providers and users were consulted by means of on-line questionnaires, interviews, visits and an open workshop discussion. The consultation was characterized by a surprisingly low participation rate in the on-line survey from both forecast providers and users. Received responses provided useful information, but for many questions it has been difficult to generalize the results and draw firm conclusions owing to the small sample sizes.

The majority of consulted providers recognize the overall benefits of setting up quality standards in the industry, but do not agree on a common strategy. Although a significant number of providers took part in the survey, a large section of the industry – more than 50% – 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.

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 declared 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.

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.

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.

Most of the methods and metrics commonly used by the consulted providers are well documented in the existing literature. Users are less open about their assessment methodologies. The quality assessment of some types of forecasts lacks coverage in the literature and needs special attention (e.g. interval forecasts, extreme events).

The majority of providers and users who took part in the survey are clearly in favour of an independent body to monitor the weather forecasting industry and encourage good practice in the assessment of forecast quality. There is also good support for an independent online forum where users and providers could submit their problems concerning forecast quality issues and find/offer practical solutions. However, in view of the limited success of the survey it is reasonable to suspect that the 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.

3. Methods and metrics for quality assessment

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 suitable to assess nearly all types of quantitative forecasts identified in the consultation. Descriptive or worded forecasts, however, cannot be assessed objectively.

A comprehensive list of common quality assessment metrics with discussion of their merits and limitations is presented in the book edited by Jolliffe and Stephenson (2003) as well as on the web site of the WWRP/WGNE Joint Working Group on Verification (http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html). 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 generally inadequate to quantify the overall quality of a set of forecasts. Furthermore, forecast samples must be representative and large enough to achieve statistical significance. The small sample size of past forecasts and observations is particularly problematic in seasonal forecasting. Furthermore, the common quality assessment metrics do not work well when dealing with rare or extreme events.

The crucial importance of choosing proper metrics, the impact of their statistical properties on results and the need to estimate statistical significance were exemplified by means of four case studies. In order to keep the length of this paper within reasonable limits, only one case study is presented here and the reader is referred to Mailier *et al*. (2006) for more.

The verification of interval forecasts has not been much discussed in the literature, but survey results indicate that this type of forecast is 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 within 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

DOI: 10.1002/met

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 (2007) 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, a set of 182 interval forecasts based on ensembles of 51 members for the period 1 October 2003 to 31 March 2004 are examined. 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 curve in Figure 1 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 86–88% 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 line in Figure 1 demonstrates that excellent reliability can easily be 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 detrended 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.

The interval score (Gneiting and Raftery, 2007) is a metric that has been especially designed to measure the accuracy of prediction intervals. Like the mean absolute error, the mean squared error 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:

$$\alpha(u_i - l_i) + 2(l_i - x_i) \text{ if } x_i < l_i$$

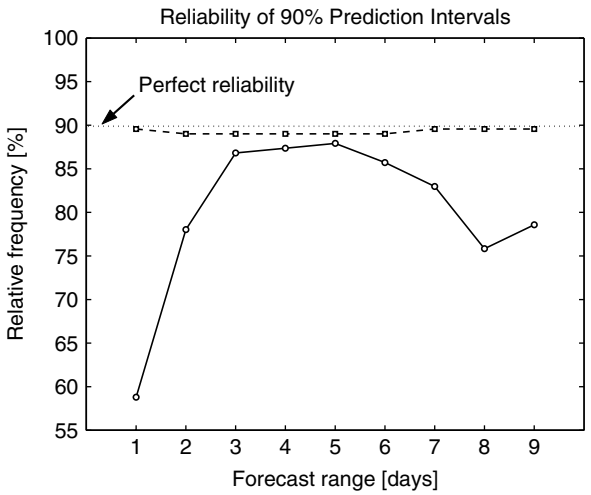

Figure 1. Estimated probability that the observed daily mean temperatures fall within the 90% prediction intervals for intervals based on ensemble forecasts (solid line) and for intervals based on climatological data (dashed line). The dotted horizontal line represents perfect reliability.

$$P_{\alpha,i} = \alpha(u_i - l_i) \text{ if } x_i \in [l_i, u_i]$$
$$\alpha(u_i - l_i) + 2(x_i - u_i) \text{ if } x_i > u_i, \quad (1)$$

and the interval score S_α is simply the average penalty:

$$S_\alpha = \frac{1}{n}\sum_{i=1}^{n} P_{\alpha,i}. \quad (2)$$

The interval scores achieved by the two types of prediction intervals (ensemble forecasts and climatology) are shown in Figure 2. 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.

The two measures of accuracy can be combined in one 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:

$$SS = \frac{S_\alpha(c) - S_\alpha(f)}{S_\alpha(c)} \quad (3)$$

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

DOI: 10.1002/met

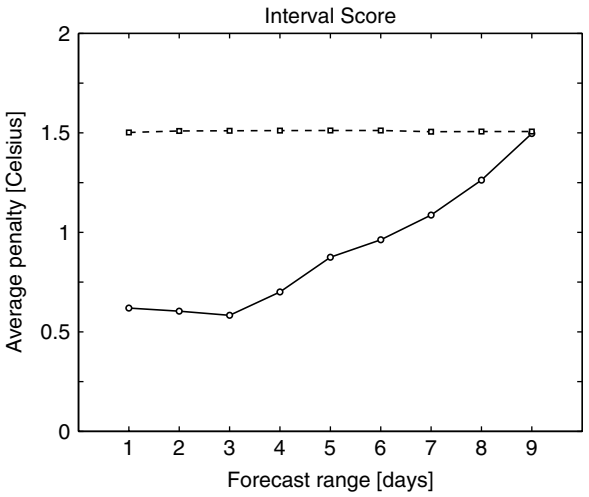

Figure 2. Interval scores achieved by the 90% prediction intervals based on ensemble forecasts (solid line) and climatological data (dashed line).

ensemble-forecast intervals are more (less) accurate, than the climatological intervals. The estimated values of *SS* expressed in percentage are plotted in Figure 3. Approximate 95% bootstrap confidence intervals (Wilks, 1995) *SS* for have been added to quantify the uncertainty (statistical significance) of the results.

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. In this example, a metric for measuring the accuracy of interval forecasts has been introduced and used to show that despite being less reliable than climatological intervals, prediction intervals based on a large ensemble of forecasts can be much more accurate. The use of a skill score here highlights the fact that skill is always measured relative to some other reference forecast that is usually cheap and easy to generate. Climatological intervals have been used in this case, but alternatives based on persistence or randomness are also possible. Therefore, any claim of forecast skill should always clarify which reference forecast was used to calculate the skill statistic. In other words, simply claiming that a forecasting system has skill doesn't mean anything. An indication of statistical significance is also important to ascertain that the quality assessment results are due to the properties of the forecasting system being evaluated rather than to chance.

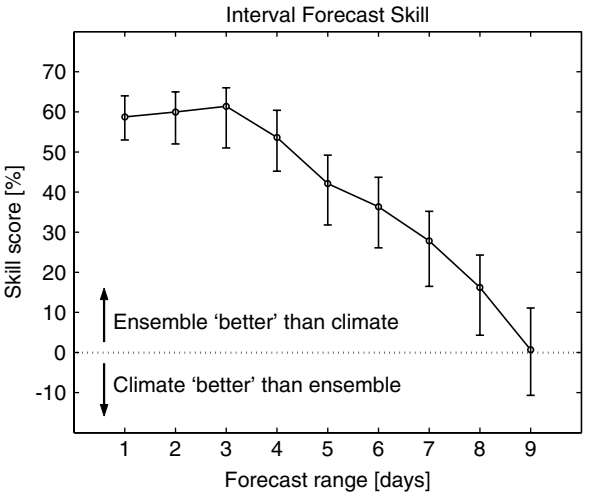

Figure 3. Estimated skill scores with 95% bootstrap confidence bars for the prediction intervals based on ensemble forecasts compared with climatology.

4. Recommendations

4.1. Recommendations for good forecast quality assessment practice

4.1.1. Routine forecast quality assessments

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.

4.1.2. Clear and documented quality assessment methodology

The quality 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) is encouraged to illustrate the assessment results.

4.1.3. Forecast format suitable for objective quality assessment

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.

4.1.4. Reproducibility of quality 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.

4.1.5. Relevance of quality 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.

DOI: 10.1002/met

4.1.6. Completeness of quality assessments

Assessments should take into account the multi-faceted nature of forecast quality. Methodologies and metrics that attempt to summarize 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.

4.1.7. 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 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) skillful than climatology in the short (long) ranges. Any claim of forecast skill should always mention the reference forecast that has been used.

4.1.8. 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.

4.1.9. 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.

4.1.10. 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. Retroactive forecasting (hindcast) should be avoided where feasible.

4.2. Specific recommendations concerning quality assessment metrics

4.2.1. 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.

4.2.2. 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.

4.2.3. Resistance

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

4.2.4. 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.

4.2.5. Independence

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

4.2.6. Discrimination

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

4.2.7. Specific recommendations on which metrics to use

Formal definitions, discussions and further references for the metrics below are given in Wilks (1995); Jolliffe and Stephenson (2003) and Gneiting and Raftery (2007).

- 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.

DOI: 10.1002/met

- 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, the latter measure being 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 is 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 forecasts: 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 Relative Operating Characteristic (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 forecasts: reliability is the best measure to assess the probabilistic fitness of the intervals. 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).

4.3. 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 summarized in Section 2 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.

4.3.1. 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, less coercive scheme is proposed where participation and voluntary adherence to a code of practice are encouraged.

It is recommended that the Royal Meteorological Society first establishes a specialized 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), which was 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.

4.3.2. 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 Sections 4.1 and 4.2. Companies applying for accreditation should agree to submit themselves to independent, regular audits.

4.3.3. 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 was found desirable by a majority of respondents to the survey. However, the difficulties experienced to get providers to mobilize 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

DOI: 10.1002/met

probable that without appropriate moderation, the forum will not fulfill 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.

4.3.4. 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;
- organize high-profile events – e.g. forecasting contests similar to those held under the auspices of the AMS – that demonstrate the importance of good practice and quality standards.

5. 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 a body such as the commission proposed to the Society in Section 4.3 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. 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 Section 4.3.

Specific recommendations on which metrics to use have been made in Section 4.2. These recommendations do not 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. 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.

Acknowledgements

The authors and the Royal Meteorological Society wish to thank the National Meteorological Service Commissioning Group (NMSCG) for agreeing to fund this study. We are particularly grateful to Dr Peter Ryder, Dr Richard Pettifer, Prof Chris Collier, Dr Kirby James, Dr Andrew Eccleston and Dr David Pick for kindly providing guidance and support throughout this project. The Editor Dr Peter Burt and two anonymous reviewers also deserve credit for their constructive comments.

References

Gneiting T, Raftery AE. 2007. Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association* **102**: 359–378.

Jolliffe IT, Stephenson DB (eds). 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons: Chichester; 254.

Mailier PJ, Jolliffe IT, Stephenson DB. 2006. Quality of weather forecasts – review and recommendations. Royal Meteorological Society Report Available from http://www.subscriptions.rmets.org/pdf/fqp_report.pdf.

Wilks DS. 1995. *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press: London; 467.

DOI: 10.1002/met