METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **15**: 3–18 (2008)
Published online in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/met.52

Review

Forecast verification: current status and future directions

B. Casati,[a]* L. J. Wilson,[a] D. B. Stephenson,[b] P. Nurmi,[c] A. Ghelli,[d] M. Pocernich,[e] U. Damrath,[f] E. E. Ebert,[g] B. G. Brown[e] and S. Mason[h]

[a] *Meteorological Research Division, Environment Canada, Dorval, QC, Canada*
[b] *University of Exeter, Exeter, UK*
[c] *Finnish Meteorological Institute, Helsinki, Finland*
[d] *European Centre for Medium-Range Weather Forecasts, Reading, UK*
[e] *National Center of Atmospheric Research, Boulder, CO, USA*
[f] *Deutscher Wetterdienst, Offenbach, Germany*
[g] *Bureau of Meteorology, Melbourne, Australia*
[h] *Columbia University, New York, USA*

ABSTRACT: Research and development of new verification strategies and reassessment of traditional forecast verification methods has received a great deal of attention from the scientific community in the last decade. This scientific effort has arisen from the need to respond to changes encompassing several aspects of the verification process, such as the evolution of forecasting systems, or the desire for more meaningful verification approaches that address specific forecast user requirements. Verification techniques that account for the spatial structure and the presence of features in forecast fields, and which are designed specifically for high-resolution forecasts have been developed. The advent of ensemble forecasts has motivated the re-evaluation of some of the traditional scores and the development of new verification methods for probability forecasts. The expected climatological increase of extreme events and their potential socio-economical impacts have revitalized research studies addressing the challenges concerning extreme event verification. Verification issues encountered in the operational forecasting environment have been widely discussed, verification needs for different user communities have been identified, and models to assess the forecast value for specific users have been proposed. Proper verification practice and correct interpretation of verification statistics has been extensively promoted with recent publications and books, tutorials and workshops, and the development of open-source software and verification tools. This paper addresses some of the current issues in forecast verification, reviews some of the most recently developed verification techniques, and provides recommendations for future research. Copyright © 2008 Royal Meteorological Society and Crown in the right of Canada.

KEY WORDS spatial verification approaches; probability forecasts and ensemble verification; extreme events verification; operational verification; verification packages; user-oriented verification; value

Received 17 September 2007; Revised 7 December 2007; Accepted 2 January 2008

1. Introduction

Verification is an indispensable part of meteorological research and operational forecasting activities. If the methodology is properly designed, verification results can effectively meet the needs of many diverse groups, including modellers, forecasters, and users of forecast information. It can be used to direct research, to help determine where research funding is most needed, to check that forecasts are improving with time, to help operational modelling centres select model upgrades, or to help power companies make decisions on the purchase and distribution of power to their customers, to name just a few.

In general, the vast majority of verification efforts over the past decades have focused on the calculation of one or more verification scores over a forecast-observation dataset, where the observations usually consist of surface or upper air point observations or analyses onto grids. These methods are sometimes referred to as 'traditional verification' to contrast them with more recent developments in verification methodology (see Stanski *et al.*, 1989; Jolliffe and Stephenson, 2003; and Wilks, 2006, for reviews of traditional verification methods). Research and development of new approaches to verification has increased greatly over the last 10 years or so, and has been motivated by several factors, including the availability of new sources of data such as satellite and radar, the desire to generate verification results which are more meaningful to specific users or user groups, the advent of new modelling strategies such as ensembles, and the

* Correspondence to: Dr B. Casati, Visiting Fellow, 2121 Trans-Canada Highway, 5th floor, Dorval, H9P 1J3, QC, Canada.
E-mail: barbara.casati@ec.gc.ca

evolution of models and forecasts to higher spatial and temporal resolution.

The foci of recent research projects in verification methodology are many and varied. For example, much work has been done on spatial techniques, which are designed to account for the spatial structures and features characterizing weather maps, both for large scale and for high-resolution regional models. Verification methods for ensemble forecasts have also received considerable attention, leading to new methods for the evaluation of forecast probability distributions, and further investigation into the properties of traditional verification measures for probability forecasts. The need for estimates of confidence in verification statistics, long known in the research community, is finally being addressed, spurred on by increased interest in verification of extreme and rare events, where sample sizes are often too small to permit a high degree of confidence in the results obtained. Verification research and development is also beginning to focus more toward users, to provide the information they would need to make optimal decisions and to assess the value of the forecast for their specific weather-sensitive operation.

Various international research projects such as the Sydney 2000 and Beijing 2008 Olympic Forecast and Research Demonstration Projects (FDP/RDP) (Keenan *et al.*, 2002; Ebert *et al.*, 2004; Yu, 2005), and later the Mesoscale Alpine Programme (MAP) project (Volkert, 2005), and other test beds, provide ideal frameworks to analyse and address some of the current issues in verification, and develop and test new verification strategies. The scientific verification community has been very active in trying to respond to verification user needs with new techniques, in spreading the knowledge of verification methods and trying to unify the verification terminology (e.g. Jolliffe and Stephenson, 2003, glossary). A Joint Working Group on Verification (JWGV) under the World Meteorological Organization (WMO)/World Weather Research Program (WWRP) and the WMO Working Group on Numerical Experimentation (WGNE) was constituted in January 2003. In addition to promoting verification practice and research, the JWGV maintains a verification web-page (http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html) which outlines basic verification score and techniques, reviews the most recent research, runs a discussion group, and organizes workshops and tutorials. Three international verification workshops have been organized (Boulder, USA, August 2002; Montreal, Canada, September 2004; and Reading, UK, January 2007; see the JWGV web-page for web-links to the workshop agendas and presentations).

Through the lens of the Third International Verification Methods Workshop in Reading, this paper aims to review recent developments, discuss some of the unsolved issues, and suggest future research directions in verification. The paper is organized approximately according to the sessions of the workshop, as follows: Section 2 reviews verification techniques developed for forecasts defined over spatial domains and for high-resolution forecasts. Section 3 reviews methods for probabilistic forecasts and ensemble forecasts. Section 4 addresses the issues related to the verification of extreme events. Section 5 addresses some of the verification strategies relevant in an operational forecasting environment. Section 6 provides a review of available verification tools. Section 7 introduces some of the issues related to user-oriented verification, followed in Section 8 by a discussion of strategies for linking the forecast quality to its actual value. Finally, some concluding remarks on the state-of-the art of verification research and future recommendations are given in Section 9. While identified as an important verification issue in this paper, the methods used to estimate confidence in verification statistics are discussed in another paper in this issue (Mason, 2008).

2. Spatial verification methods

Weather variables are often predicted as fields defined over a spatial domain. Spatial fields are characterized by a coherent spatial structure and often by the presence of features, such as precipitation features. Standard verification methods based on a point by point comparison (e.g. Mean Squared Error, MSE) often do not account for the intrinsic spatial correlation existing within these fields. The results from such standard verification methods are often difficult to interpret in meaningful physical terms. Some new approaches that specifically address the verification of forecasts defined over spatial domains have been developed in the last decade. These approaches account for the spatial nature of forecast fields, and aim to provide feedback on the physical nature of the forecast error, adding new and complementary information to the traditional categorical and continuous verification methods.

The study by Hoffman *et al.* (1995) introduced the first verification technique based on an *optical flow* concept. The technique decomposes the forecast error into displacement, amplitude and residual errors. Displacement and amplitude errors correspond to two transformations of the forecast field, obtained by applying a velocity field and a scalar field, until the transformed forecast satisfies a best-fit criterion (e.g. maximization of spatial correlation) to the observation field. This feature calibration and alignment (FCA) approach sets a milestone in spatial forecast verification by measuring the error directly in physical units (e.g. displacement in km): verification results are therefore easily interpretable and address specific physical aspects of the forecast (e.g. the advection scheme). The technique inspired some further studies: for example, the technique has been applied to different weather variables (Du *et al.*, 2000). In addition, different formulations of the error decomposition have been developed (Douglas, 2000). Brill (2002) introduced a similar technique which evaluates the east–west phase (displacement) error and amplitude error in mean sea-level pressure forecasts by using cosine series trigonometric approximations. Nehrkorn *et al.* (2003) extend

DOI: 10.1002/met

the error-decomposition approach by using spherical harmonics and subsequently the method developed into a spectral-based variational analysis technique which was used for data assimilation (Brewster, 2003). Germann and Zawadzki (2002, 2004), and Turner *et al.* (2004) use a similar error-decomposition approach for the MAPLE nowcasting system and study the scale-dependence of precipitation predictability by using wavelets. Note that all the techniques based on the Hoffman *et al.* (1995) approach are performed over the whole field. Most of the recently developed FCA techniques are combined with a scale-decomposition approach, and are used mainly for data assimilation or nowcasting applications.

Feature-based approaches identify features in the forecast and observation fields and then assess different attributes associated with each individual pair of forecast-observed features (such as position, size, and intensity). The techniques differ according to the algorithms chosen to identify the features: Ebert and McBride (2000) used a simple thresholding, whereas Davis *et al.* (2006a,b) used a threshold after filtering by cylindrical convolution; Baldwin *et al.* (2002) used some data mining and image processing algorithms; Nachamkin (2004) and Nachamkin *et al.* (2005) used composites of several events; Marzban and Sandgathe (2006) used cluster analysis. The verification approach also varies with the different studies: Ebert and McBride (2000) were inspired by Hoffman *et al.* (1995) and assess displacement, volume and pattern errors for pairs of forecast and observed features. Grams *et al.* (2006) improved the technique of Ebert and McBride (2000) by re-formulating the error decomposition in terms of traditional continuous statistics (observation and forecast variance, bias, correlation, and MSE). They then applied the verification by stratifying the features within a convective system classification. Davis *et al.* (2006a) identified and verified attributes (e.g. intensity, area centroid location) associated with pairs of forecast and observed ‘objects’. Baldwin *et al.* (2002) measure the overall forecast performance by a single statistic obtained by a weighted combination of errors for different event attributes. In some cases, contingency tables based on feature-displacement criteria and associated categorical scores are also computed (e.g. Ebert and McBride, 2000; Davis *et al.*, 2006a,b; Marzban and Sandgathe, 2006; Nurmi *et al.*, 2007). Verification may be presented as a function of the feature size (e.g. Davis *et al.*, 2006a,b; Marzban and Sandgathe, 2006), so that feature-based verification can be interpreted within a scale-oriented context. Davis *et al.* (2006b) consider the time dimension, in addition to spatial coordinates, and use an object-based approach to assess rainfall systems of different time duration, evaluating timing error, as well as displacement errors. Wernli *et al.* (2006) introduced the SAL technique, which assesses the structure–area–location error for objects, without matching requirements.

Scale-decomposition approaches, in general, decompose forecast and observation fields into the sum of spatial components on different scales by using spatial filters, and then perform the verification on each scale component, separately. Verification on different scales can provide useful insight into Numerical Weather Prediction (NWP) model representation of the different physical processes associated with phenomena on different scales. Scale-verification approaches aim to assess quality and skill of the forecasts for different spatial scales, analyse the scale-dependency of the forecast predictability (e.g. evaluate the no skill – skill transition scale), and assess the forecast ability to reproduce scale spatial structure of observed precipitation fields. Briggs and Levine (1997) introduced a wavelet-based verification method on different spatial scales which uses continuous verification statistics (e.g. the MSE). Casati *et al.* (2004) developed an intensity-scale verification technique based again on 2D wavelet decomposition and on a categorical verification approach. A 2D wavelet filter was again used by Casati and Wilson (2007) to decompose the Brier score and Brier skill score, reliability and resolution on different scales, for the verification of probability forecasts defined over a spatial domain. Zepeda-Arce *et al.* (2000), Harris *et al.* (2001) and Tustison *et al.* (2003) assess the capability of forecasts to reproduce the observed spatio-temporal and multi-scale spatial structure of precipitation fields by the evaluation of scale-invariant parameters. De Elia *et al.* (2002) and Denis *et al.* (2003) evaluate the forecast time-scale predictability limits as a function of the scale for high-resolution regional climate models.

Neighbourhood-based (fuzzy) verification approaches consider values nearby in space and time in the forecast-observation matching process, and so relax the requirements for perfect time-space matching (see Ebert, 2008, for a review of neighbourhood-based approaches). These approaches enable one to account for the forecast and observation intrinsic time-space uncertainty by examining performance in a range of neighbourhood sizes, so that these approaches are particularly suitable for verifying high-resolution forecasts. Tremblay *et al.* (1996) evaluate categorical scores as a function of the allowance distance within which two grid-point values are considered a match. Atger (2001) verified and compared deterministic and Ensemble Prediction System (EPS) forecasts by evaluating spatial multi-event contingency tables and the corresponding relative operating characteristic (ROC) curves. Theis *et al.* (2005) transformed deterministic forecasts into probabilities by using neighbourhood grid-point values and then verify by using probabilistic verification approaches. Marsigli *et al.* (2006, 2008) verified the parameters describing the distributions of forecast and observations within neighbourhood square areas. Rezacova and Sokol (2005) verified high-resolution precipitation forecasts by using rank Root Mean Squared Error (RMSE) over grid-point neighbourhoods of different areas. Roberts and Lean (2007) consider within a grid-point neighbourhood the fraction of precipitation values exceeding some set thresholds, and define a ‘Fractions Skill Score’ with which the forecast is assessed for different intensities and scales. Note that the size of the neighbourhood naturally defines a ‘scale’, associated with

DOI: 10.1002/met

the verification: as the size of the neighbourhood (scale) is increased, forecast and observation fields are subjected to a filtering process and the time-space matching requirement becomes more and more relaxed. The scale of the neighbourhood-based approaches is therefore related to the resolution of the forecast and observation fields and to the looseness of the matching criteria, whereas the scale in the scale-oriented approaches relates to the scale of the features and error.

Metrics measuring the distance between binary images include 'detection performance (statistical) measures' (e.g. traditional categorical scores), 'localization performance (distance) measures' (e.g. the average distance), Pratt's figure of merit, Hausdorff metrics and the Baddeley metric (see Baddeley, 1992, and references therein). These metrics are sensitive to the distance and shape of forecast and observed features. A few verification methods are starting to make use of these metrics: Venugopal *et al.* (2005) define a new forecast quality index obtained by combining the Hausdorff distance with an amplitude-based error measure. Gilleland *et al.* (2008) use the Baddeley metric to match and merge features within their object-oriented technique.

Most of the spatial verification approaches need observations defined continuously over a spatial domain. Such methods rely, therefore, on a dense observation network and radar or satellite-based observations, which are often merged to produce an analysis. Exceptions are some neighbourhood verification approaches, which consider the neighbourhood of forecast grid points surrounding an observation at a specific location (e.g. Atger, 2001; Theis *et al.*, 2005; see also Ebert, 2008, and classification therein). Moreover, most of these methods do not allow missing values in the observations, except the Nachamkin (2004) and Nachamkin *et al.* (2005) composite approach. Future research could address the issues of missing values and sparse observation networks when applying spatial verification approaches.

3. Probabilistic forecasts and ensemble verification

Verification measures for probability forecasts were developed and studied many years ago for application to forecasts from statistical methods such as Model Output Statistics. The advent of ensemble forecasts in the early 1990s has given a strong impetus to the re-evaluation of existing verification methods for probability forecasts and to the development of new ones.

These methods are of three general types. Methods used to verify the distribution of an ensemble as the sample from a probability distribution function (pdf) include the rank histogram (Anderson, 1996; Hamill, 2001), the continuous rank probability score (CRPS) (e.g. Hersbach, 2000) and related skill score, the minimum spanning tree (MST) (Smith, 2001; Smith and Hansen, 2004; Wilks, 2004), and Bounding Boxes (Weisheimer *et al.*, 2004). Methods which evaluate the pdf of a generic probability forecast are the ignorance score (Good, 1952; Roulston and Smith, 2002) and the Wilson *et al.* (1999) probability score. Finally, forecasts of the probability of an event are evaluated using the Brier Score (Brier, 1950) and its decomposition (Murphy, 1973), the Brier skill score, reliability ('attributes') diagrams (Hsu and Murphy, 1986; Smith, 1997; Bröcker and Smith, 2007a), the ROC (Mason, 1982; Swets and Pickett, 1982), the rank probability score (Epstein, 1969) and related rank probability skill score.

The first category contains methods recently developed specifically for application to the verification of ensemble pdfs. The first four measures listed under the third category apply to forecasts of dichotomous variables ('yes/no'), while the last two apply to probability forecasts of multiple category variables. This is not an exhaustive list, but most of the commonly used methods are described here.

3.1. Verification of the ensemble pdf

Of the three pdf scoring methods which received some attention at the workshop, the CRPS has tended to become the method of first choice for the verification of operational ensemble forecasts, although the rank histogram is often used to evaluate the spread of the ensemble. Both the rank histogram and the MST are often presented graphically, but are also associated with summary measures of performance. The CRPS score summarizes the verification information into a single value.

The rank histogram (Talagrand diagram) is used to determine the extent to which the ensemble dispersion matches the dispersion of the distribution of verifying observations. It does not give meaningful results unless computed on a relatively large sample, and, as Hamill (2001) points out, it can obscure systematic biases in the forecast. The rank histogram and its interpretation is straightforward for continuous well-behaved (near normally distributed) variables, but its use and interpretation become more complicated for variables such as precipitation amount, which have highly skewed, bounded distributions. Hamill and Colucci (1997) describe a method for construction of a rank histogram which accounts for the frequent occurrence of zero in precipitation distributions, and this method was followed in results presented by Denhard *et al.* (2007). The interpretation of the rank histogram for precipitation amounts is difficult because observed precipitation amounts typically follow a gamma distribution which may have a variety of shapes. The dispersion of a Gamma distribution is highly dependent on the mean; thus, the average spread taken over a given set of observations will depend on the mean precipitation over the sample. Bimodal forms can also occur, for example, when some ensemble members predict no precipitation while others predict measurable precipitation. Thus, rank histograms for precipitation and other variables which follow a non-Gaussian distribution have a higher potential to be misleading. Rank histograms for

DOI: 10.1002/met

precipitation require additional evaluation on subsamples conditioned on the mean precipitation value.

Candille and Talagrand (2005) proposed an useful formula which allows quantitative assessment of the 'departure from flatness' of a rank histogram (see also Smith, 2001). This formula is a function of the ensemble size (number of bins) and the verification sample size, and takes into account the expected variations from flatness due only to random variation. This formula is useful for comparing rank histograms from different ensembles.

The CRPS essentially measures the difference between the cumulative distribution function (cdf) of the ensemble and the observation, also expressed as a cdf. Since the observation is a point value, its corresponding cdf is a Heaviside function with the step at the value of the observation. The CRPS has the nice property that it reduces to the mean absolute error for a deterministic forecast, which means that it can be used directly to compare the accuracy of an ensemble forecast with respect to the accuracy of a deterministic forecast.

The MST is a form of multidimensional rank histogram. Ensemble predictions and the corresponding observations are represented as points in a M-dimensional space, where M is the number of variables of interest. The trees are formed by computing the distance spanning all the points over the $N + 1$ ensembles formed by leaving out in turn each of the N-ensemble members and the observation. The minimum of these distances (the MST) is then tallied with respect to the left-out member to which it corresponds. The interpretation is similar as for the rank histogram: If the tree spanning the ensemble alone is smaller than all the trees spanning $N - 1$ ensemble members plus the observation more often than $1/(N + 1)$ proportion of the time, then the observation lies outside the cluster of ensemble points more often than expected by chance and the ensemble is underdispersive. The MST has not yet been widely used in practice. It has significant potential for the spatial verification of structures such as low centres, tropical storm positions, or precipitation maxima. In that case $M = 2$, and the metric would be the actual distance on the earth's surface.

3.2. Verification measures for pdfs of generic probability forecasts

Measures of this type are not specific for ensembles, as are the previous methods, and are not sensitive to the whole ensemble pdf. Instead, both the methods discussed below are local, evaluating the pdf at the observation value. Both methods reward sharp distributions which are also accurate. Smoothing the distribution (increasing the predicted uncertainty) results in limiting the 'best' attainable score, while predicting sharply (low spread) but with inaccurate placement of the pdf with respect to the observation leads to a heavy penalty.

The linear probability score proposed by Wilson *et al.* (1999) is easy to understand and can be computed meaningfully on a single case, but it was shown by Wilson and Gneiting (2007) that it is not strictly proper, confirming the theoretical presentation of Gneiting and Raftery (2007). This may not be an important limitation for this score since Wilson and Gneiting (2007) showed a potential improvement of only about 5% or so in the score value, by systematically predicting the ensemble mean. A proper linear score is discussed in the context of other p-scores in Bröcker and Smith (2007b).

The ignorance score (Good, 1952; Roulston and Smith, 2002) is a logarithmic score formulated in the same way as the Wilson *et al.* (1999) score. The ignorance score has been proven to be strictly proper (Good, 1952) and is the only proper local score for continuous variables (see Bröcker and Smith, 2007b for discussion). The ignorance score becomes infinite for forecast probabilities of zero, which has created problems in its application to ensemble forecasts (Gneiting and Raftery, 2007; Wilson and Gneiting, 2007). In the latter case, the problem was avoided by setting zero probability forecasts to a suitably small value (0.0001). One could argue that a probability forecast of 0 should never be assigned to an outcome which is known to be within the range of possible values. However, in practice, the forecast system, whether from an ensemble or other source, cannot distinguish between 0 and small probability values. Application of the ignorance score requires that this limiting value be selected. This choice can influence the score obtained, and it could be argued that this in effect renders the score improper, because the forecaster knows in advance the effect the choice of minimum value will have on his/her score. This drawback may be one reason why the ignorance score has not been widely used yet in ensemble verification.

3.3. Verification of forecasts of the probability of an event

The Brier score (Brier, 1950) probably the most commonly used score for verification of probability forecasts of dichotomous variables, has been subject to recent examination and some re-interpretation. An update to the Murphy (1973) three-way partition of the Brier score was proposed by Stephenson *et al.* (2008b), who pointed out that extra terms are needed in the partition to render the decomposition exact because of the effects of binning of the probabilities in the original decomposition. Ferro (2007a) and Ferro *et al.* (2008) evaluated the effect of the ensemble size on the Brier score and ranked probability score. They found that one needs to know the ensemble size to properly interpret Brier score results, and that it may be hazardous to compare Brier scores computed with different ensemble sizes. However, we note that this issue is really related to how the probability forecast is defined from the ensemble system: once the forecast is made, the Brier score will correctly reflect the accuracy of the forecast. In addition, it is possible to use bootstrap confidence intervals to generalize Brier score estimates over different sample sizes.

The ROC was brought from signal detection theory into meteorology by Mason (1982). Essentially the ROC

DOI: 10.1002/met

is a measure of the likelihood that probability forecasts for an event are higher for occurrences than for non-occurrences of the event. The ROC and its associated measure, the area under the curve, are indications of the ability of the forecast system to 'discriminate' occurrences and non-occurrences of the event (Murphy, 1993).

Some issues remain to be resolved concerning the method of computation of the ROC area and its interpretation. First, the trapezoidal rule often is not an appropriate method for estimating the ROC area. As shown by Wilson (2000), the trapezoidal rule may lead to underestimation of the ROC area by an amount which depends on the location of the points from which the ROC is estimated. For the sample sizes that are usually available in meteorological verification practice, the bi-normal method (Mason, 1982; Swets, 1986) is often the most appropriate approach, and its accuracy has been empirically validated in many different fields (Swets, 1986). When verification samples are small, specifically when the number of occurrences of the event is small, the trapezoidal rule is a correct method to compute the area. A general rule of thumb is that there should be more than 10 occurrences of the event in the sample if the bi-normal method is to be used.

Second, a clarification is needed regarding the roles of ensemble and sample sizes in the computation and interpretation of the ROC. For instance, Bowler *et al.* (2007) show that the fitted ROC represents the potential discrimination for an unlimited ensemble size, to support the use of the trapezoidal rule for real ensembles. Granularity in the forecast probabilities that arises with small ensembles may lead to fewer points from which to estimate the ROC, but this is a matter of sampling variation; the points still lie on a convex curve rather than on a set of trapezoids.

Third, the interpretation of the ROC becomes more complicated when rare events are considered. Wilson (2000) and others show a tendency for the points on the ROC to cluster toward the lower left corner of the graph for rare events. One solution to this problem is to subdivide the lowest-valued forecast probability bins. The verification sample can usually support subdividing the lower-valued probability bins when fitting the ROC for low base rates. Another issue concerning rare events and the ROC is that models can quite happily predict any probability value in the range (0, 1), but forecasters may experience more difficulty in discerning small differences in probabilities. It is clear that further research is needed into the application of the ROC to ensemble verification.

A shortcoming in verification practice as applied to probability forecasts has been given renewed attention recently: there is a tendency to overstate skill levels in summary verification results. This problem arises in verification scores which are referenced explicitly or implicitly to climatology, such as the Brier skill score and the ROC, and when the samples are drawn from inhomogeneous datasets (e.g. when summer and winter data are included in the same sample). This issue is discussed by Hamill and Juras (2006). This unrepresentative estimate of skill is best avoided by computing skill scores on stratified samples, by season and for single stations or homogeneous regions. If the sample size is not sufficient, then the score computation should use anomalies to remove the climatological signal before compositing over space and time; alternatively, climatological quantiles can be used to define the forecast events instead of actual values of the weather variable.

The reliability diagram (or 'attributes' diagram in its full version as described by Hsu and Murphy, 1986) has been and continues to be widely used in verification practice. Embodied in this one verification tool are measures of several attributes of probability forecasts: reliability, resolution, sharpness (when the forecast probability distribution is shown), and skill (Murphy, 1993). The consistency between observed frequencies and predicted probabilities can be included in the visual presentation of the graph (Smith, 1997; Bröcker and Smith, 2007a). Atger (2004) illustrates the tendency toward overestimation of the reliability component of the Brier score (underestimating the reliability) because of sampling variability when samples within the bins are small, or when the number of bins is large. Atger (2004) proposes the use of the ROC fitted by the bi-normal model, which involves unrestrictive assumptions, as an aid to offsetting the effects of sampling variability in the reliability table. This issue was also discussed by Weigel *et al.* (2007) in the context of the rank probability skill score and small ensemble sizes.

Discussion of the attributes of probability forecasts and their meaning has also increased in recent years. The most comprehensive discussion is that of Murphy (1993), who identifies nine attributes which apply not only to probability forecasts but also forecasts of continuous variables. These attributes can be related to the joint distribution of forecasts and observations and its factorizations (Murphy and Winkler, 1987), to define a complete general framework for forecast verification. A more recent general review of forecast verification methods (Toth *et al.*, 2006) would seem to be less comprehensive, presenting only the two attributes reliability and resolution as important to the evaluation of forecasts. Four of Murphy's (1993) attributes are important for a full diagnostic verification of probability forecasts. These are reliability, resolution, discrimination and sharpness. The first two are obtainable from the reliability table or the decomposition of the Brier score, the third is measured by the ROC area, and the fourth is determined by the forecast strategy alone.

There would seem to be some confusion between resolution and discrimination. While they are not independent attributes, they are not the same. The former is the conditional distribution of the observations given the forecasts while the latter is the conditional distribution of the forecasts given the observations (Murphy, 1993). Contrary to Toth *et al.* (2006), it is more important to measure the discrimination *via* the ROC area than the resolution, since the ROC and the reliability are independent diagnostic

DOI: 10.1002/met

measures of forecast performance, and express both factorizations of the joint distribution (Murphy and Winkler, 1987). The reliability table with its associated summary measures and the sharpness graph, and the ROC, with its associated skill measure, the ROC area, together form a reasonably complete basis for the diagnostic verification of probability forecasts. Moreover, the four attributes evaluated by these two tools can be translated to corresponding tools for other forecast types.

3.4. Other probabilistic verification issues considered at the workshop

As part of an evaluation of time-lagged ensembles of high-resolution precipitation forecasts, Mittermaier (2007) suggested that the distance between the two forecast probability distributions conditional on occurrences and non-occurrences might be a more useful measure of discrimination than the more often used ROC area.

Mason (2007) suggested that instead of using the contingency table to obtain one point on the ROC curve, it is fairer to rank the forecasts in the sample, assign equal probability to each interval, then compute hit rates and false alarm rates for each probability threshold in the sample. In such a way, one obtains a full ROC curve, which takes into account forecasts which are nearer to the threshold compared to those which are further away.

Hopson *et al.* (2007) argue that the 'standard' spread-skill correlation measure may be a misleading measure of the utility of ensemble dispersion forecasts. As alternatives, Hopson proposed normalizing the correlation using a constant climatological forecast, binning the spread-skill correlation, or computing a binned rank histogram.

4. Verification of extreme events

One of the important goals of weather prediction is to help forewarn society about *severe* or *high-impact* events that can incur large damages and losses. The loss caused by severe events depends in a complex manner on attributes of the event (e.g. the magnitude of the meteorological variables), the vulnerability of infrastructure, and the amount of exposure (Stephenson, 2008). In order that the mean loss is sustainable, severe events also have to be *rare events,* hence, the use of the term *rare severe event* (*RSE*) by Murphy (1991). Such events are also loosely referred to as *extreme events* in atmospheric science. The urge to develop some new verification approaches for extreme events has grown strongly in the last decade, partially due to the development of high-resolution models, which are capable of resolving the spatial and time-scale of extreme events, and because of the increased frequency of extreme events expected from anthropogenic climate change, as well as the increased vulnerability of societies to the occurrence of extremes (e.g. European heat wave in the summer 2003).

A simple and rather naïve way to consider an extreme event is to examine only its occurrence or non-occurrence. Examples include the occurrence of tornadoes in the widely debated Finley's (1884) tornado forecast verification (see Murphy, 1996 and references therein), or the IPCC (2001) definition of *simple extreme events* to be 'individual local variable exceeding critical levels on a continuous scale'. Performance of deterministic forecasts of such events is typically assessed using traditional categorical scores for binary events (see Jolliffe and Stephenson, 2003, chapter 3). In particular, the threat score (also known as the Critical Success Index) has been widely used for rare events because of its ability to be defined even if one does not know the number of correct no-event forecasts (Gilbert, 1884; Donaldson *et al.*, 1975). Several studies have analysed the categorical verification score limits for increasingly rare events. Schaefer (1990) showed that the Equitable Threat Score (also known as the Gilbert Skill core) converges to the threat score as the events become rarer. Doswell *et al.* (1990) showed that the Peirce Skill Score (also known as the Hanssen-Kuipers score or the True Skill Statistic) converges to the Hit Rate as the number of correct rejections becomes larger with respect to the other entries of the contingency table. Marzban (1998) also presented rare event limits of scores but noted that there was some ambiguity in how the limits of the cell counts could be taken. Stephenson *et al.* (2008a) avoided this problem by developing a simple asymptotic model for rare binary event forecasts, and then used it to show that these results depend on how the hit rate decreases as a function of increasing event rarity. Using this model, Stephenson *et al.* (2008a) demonstrated that all the traditional scores tend to non-informative limits such as 0 or 1 for increasingly rare events.

The trivial limit of scores can be avoided by using more appropriate association measures for extremes. For example, Göber *et al.* (2004) found that the skill as judged by association measures such as the odds ratio can actually increase for rarer events. Inspired by recent work in bivariate extreme value theory (Coles *et al.*, 1999), Stephenson *et al.* (2008a) have proposed a new score, known as the Extreme Dependency Score (EDS), for the assessment of skill in deterministic forecasts of rare binary events. The EDS has no explicit dependence on the bias of the forecasting system and so has the desirable property that it cannot be improved by hedging the forecasts (i.e. under- or over-forecasting the occurrence of the event). EDS is a measure of association for extreme events that is insensitive to the choice of threshold. Ferro (2007b) developed these ideas further by using a probability model from extreme value theory to model the bivariate probability distribution between the forecasts and observations. This distribution-oriented approach to verification, based on a parametric model, allows one to use the data at all thresholds to smoothly interpolate between thresholds and to make formal inference about the true skill of the system (e.g. confidence intervals on the scores).

The rarity of extreme events poses some specific challenges for verification. Firstly, rarity of sample events can

DOI: 10.1002/met

lead to large sampling uncertainty in verification statistics. This problem can be partially alleviated by calculating verification scores obtained by pooling observations and forecasts over larger space-time domains. However, larger domains can also bring with them potential problems of inhomogeneity and non-stationarity. When performing data pooling, one should take account of the possible variations within either the spatial domain or time-period. As mentioned earlier, pooling over inhomogeneous data can lead to unrepresentative estimates of skill (Hamill and Juras, 2006). To address this issue, statistical models that have spatio-temporal explanatory variables are required. Secondly, rarity can lead to small or zero counts when stratifying events into categories (e.g. contingency table cells or bins for reliability diagrams). Various methods exist in the statistical literature for dealing with such sparseness problems (see section 9.8 of Agresti, 2002). Thirdly, small sample sizes can be unduly influenced by outlier values that can corrupt the verification of extreme event forecasts. The effect of such outliers can be reduced by fitting appropriate extreme value models to *all* the available data (e.g. Ferro, 2007b).

Extreme value modelling approaches use a subset of large values from the data sample to infer the extreme properties of the underlying process that generated the data (e.g. Coles, 2001; Beirlant *et al*., 2004). Large values are selected in various ways, for example, peaks-over-threshold (e.g. values which exceed a pre-defined threshold), block maxima (e.g. annual maxima), or the r-largest values (e.g. the five largest events in the year). The large values are used to infer behaviour about the tail of the distribution rather than provide an absolute binary definition of what is an extreme event. Extremeness is a relative concept rather than an absolute dichotomy. However, regularity assumptions are not always valid, for example, for variables related to physical phenomena that do not exist in less extreme forms (e.g. tornadoes), or that are qualitatively different because of non-linear feedbacks (e.g. heat waves, fog). Nevertheless, extreme value theory concepts and models are highly relevant for the verification of extreme event forecasts.

Inference of skill for the higher thresholds, by issuing and assessing forecasts at lower thresholds, can provide forecasters with useful experience and feedback in forecasting and interpreting more extreme events (personal communication, Dr K. Kok at KNMI). For example, to have sufficient numbers of events for longer range forecasting systems such as current operational seasonal forecasting systems, forecasts are issued for (moderate) extreme events that are not very rare: tercile categories defined by the 0.15 and 0.85 empirical cumulative probabilities are currently used at the Met Office and European Centre for Medium-Range Weather Forecasts (ECMWF) for defining such events as extremes in summer mean temperatures (personal communication, Dr F.J. Doblas-Reyes at ECMWF).

Deterministic forecasts are often hedged to avoid missing warnings of severe events and this often leads to a large frequency bias. For extreme events it is therefore desirable to issue probability forecasts that cannot be as easily hedged (Murphy, 1991). Furthermore, probabilistic forecasts provide useful information on the forecast uncertainty, which is desirable to communicate for events that can incur large losses. Probabilistic forecasts can also be used to quantitatively assess risk using a cost–loss model (for example) and help in optimal decision-making for specific users. However, communication, understanding and perception of probability forecasts encounters more difficulties than for a deterministic single-value point forecast, and some users prefer forecasters to make a definitive statement, rather than being themselves involved in making an optimal decision. (For example, the aviation industry has resisted the use of probabilities in terminal area forecasts.) In fact, probabilistic forecasts push the decision for action away from the forecast community towards the user community and reacting agencies. Resistance to the use of probabilistic forecasts is sometimes due to the awareness that a better understanding of the forecast meaning and ownership of the problem would be required from such community and agencies. Extreme event verification is partially driven by the need for understanding and informing on the consequences and possibly elevated risks in responding to extreme event forecasts. One other important aspect is that the probability for an event should be expressed relative to climatological probabilities (Murphy, 1991), as has been recently implemented in operational warning systems for extreme events (Lalaurette, 2003). Very little published work has focused on methods specifically suited to the verification of probability forecasts of rare events.

It should be noted that real-world weather events are complex phenomena consisting of many complexly related attributes. For example, they can be characterized by variable temporal durations and spatial scales, and they often involve more than one non-independent weather variables (e.g. hurricanes that cause damage due to extreme wind speeds and precipitation). Forecasts of complex extreme events can involve the definition of probability indices obtained from the combination of probabilities of occurrence of extreme events for many different variables, with different weights. Verification of such forecasts should take care in matching the forecast with an 'observation' that is coherent with the definition of the extreme event being forecast. As an example of complex extreme event verification, Roy and Turcotte (2007) evaluated the average distance between observed and forecast extreme events for hail, gusts, torrential rain and tornados, and then evaluated categorical scores by defining the contingency table entries with respect to increasing radius distances, for each weather element. They also computed a severe weather probability index, as a weighted average of the different weather element extreme events, and verified it with a Brier score, again for increasing radius distances, where the observation probability was computed with the same weight used to compute the forecast probability index. Forecast–observation pairing for extreme event verification purposes can allow a certain tolerance in time and space, as for the neighbourhood verification approaches (Section 3).

Despite increasing concerns about extreme weather events, few studies have focussed on extreme events verification. Much work remains to be done to address the challenges facing verification of extreme event forecasts.

5. Operational verification

Forecast verification is an integral element in the operational forecasting environment. Developments in all forecasting, such as advancements in NWP models or improvements of end-forecasts produced by human forecasters, can be monitored and evaluated through verification. This evaluation process can provide valuable feedback to all stakeholders in the long research–development–production chain, when associated with a functional, properly designed operational verification system. The forecasters themselves, likewise model developers, research scientists, forecast office administration, end customers of the weather service, and the general public can be served by output from such an operational verification system. It can provide means for an active feedback and dialogue process for developers, forecasters, end users and decision makers. On the other hand, owing to the quite many different types of potential users of the system, it has to be constructed in a modular fashion to serve the various user needs and requirements.

Regardless of its acknowledged importance, state-of-the-art operational forecast verification systems are still rather rare within operational weather services. Maybe one reason is the aforementioned difficulty to serve (too) many prerequisites and the complexity of developing a functional but yet user-oriented package. Moreover, most of the recently developed verification methods presented in the previous sections of this paper are missing from the existing operational verification systems. This is partially due to the complexity of some of these methods. In addition, interpretation of the results from some of the methods is not always intuitive for some users, such as operational forecasters or some model developers, who are not dealing with verification issues within their normal duties, and are not involved in the development of new verification techniques. The verification research community needs to account for these issues while developing new methods; otherwise these new techniques risk to remain just theoretical studies confined to the (small) verification research community itself. It is quite rare for a new verification method to become commonly accepted: the ROC may be the only 'new' score to become a standard during recent years.

It is, however, highly possible to construct a functional operational verification framework even based on the more traditional verification methods. Standard categorical and continuous scores, as recommended by the WMO (2000), are evaluated daily in most National Meteorological Services (NMS). The quality of the forecasts is usually monitored for specified time steps, regions and time periods. Confidence intervals and statistical significance of the evaluated scores, which have been far too often neglected, have recently started appearing aside verification statistics.

Despite addressing in a basic fashion the many different requirements dictated by the wide user community of operational verification, traditional verification methods are familiar to many and can provide much valuable information, when their pros and cons are acknowledged and they are interpreted properly. However, even the more common measures and their features are not necessarily well understood within the meteorological community, especially so among the operational forecasters. A knowledge gap clearly exists between the verification research community and the operational meteorologists who are, after all, some of the most important users of feedback from the verification undertaking. Training in both traditional and recently developed verification methods is a major need in operational centres. A high-quality verification system should have a comprehensive training module attached to it in a form that facilitates a self-learning approach. As an example, P. Nurmi and L. Wilson have developed a computer-aided self-learning package, available from EUMETCAL (http://www.eumetcal.org). This comprehensive package includes tutorials covering the verification of continuous, categorical and probabilistic variables, and has particular emphasis on the interpretation of the verification results for meteorologists.

Surveys concerning operational verification activities have occasionally been conducted within the meteorological community. A somewhat outdated 1997 WMO (2000) global survey of member states' public weather services programmes suggested that 57% of NMSs had a formal verification programme. All NMSs who responded indicated that they passed the results to staff and about a quarter submitted them to government authorities and other users. Bougeault (2002) conducted a survey about activities on the verification of weather elements performed at a selection of operational NWP centres. More recently, the Royal Meteorological Society commissioned a survey (Mailier *et al*., 2006) on quality assessment of commercial weather forecasts in the UK, which highlighted the deficiencies and lack of standardized procedures for forecast quality assessment. The difficulty of engaging the whole market place in an open debate around issues of forecast quality was one of the interesting results of the survey. Moreover, the survey suggested promoting awareness of the importance of weather forecast quality and an open culture that would favour the raising of quality standards and public awareness of quality issues.

The ECMWF has the practice of conducting annual inquiries about their member states' verification activities which are then reported in the Centre's internal documents. All member states are requested to provide details on the major highlights of their verification activities and most relevant verification results. The feedback is expected to be based on the guidance provided as a set of common verification recommendations (Nurmi, 2003). While the technical report by Nurmi (2003) was prepared, five out of 24 ECMWF member states indicated they were running an on-line operational verification system. An additional three member states stated that they produce periodical verification summaries. To summarize, there is presently no clear widespread understanding about the status and features of operational verification systems that are operating within the weather services. It might be fruitful to launch a comprehensive survey to investigate the present situation.

In order to verify model output, it is necessary to have a description of what we believe is the true state of the atmosphere. Only a small fraction of surface observations, which describe the state of the atmosphere affecting activities of the general public, are currently used in the verification methods at NWP centres. Numerous difficulties are involved in the routine acquisition, quality control and processing of surface observation data, which often lead centres to use analyses based on the model's data assimilation system as 'truth data' for operational verification, instead of 'raw' observation data. Analyses have the advantage of retaining the model physical coherence, and are characterized therefore by a temporal and spatial structure. Moreover, analyses are already defined over the model domain, usually at the same locations as the model forecast being verified. However, analyses tend to filter and smooth observation data, both through the quality control and through the data assimilation itself. In fact, the observations are checked against and merged to a short-range background forecast from the model. An accurate observation may then be rejected (or rescaled) by the quality control (and merging procedure) because it contains information on small scales which the model cannot resolve; or it can be rejected because of, say, a position error in the short-range model forecast against which it is compared. Moreover, model-based analyses produced in areas with none or sparse observations will tend to resemble the model background field. This leads to an incestuous verification where the verification data have been processed to resemble the characteristics of the model being verified. As an example, verification could become unnaturally favourable, in the context of a comparison of different NWP forecasts, to the model used to produce the analysis. Operational meteorological services can often be tempted to use the model analysis for verification, since it is practical, handy, and leads to more positive scores. Despite being an easy short-term solution, this approach can affect decisions on the model development and could eventually lead to a long-term deterioration of forecast quality, since the model could slowly depart from reality. Within the verification strategy, the choice of the reference data representing the true state of the atmosphere can dramatically affect the verification results. Therefore, such choices should be made with awareness and caution. Moreover, the source and characteristics of reference verification data should always be clearly described, not only for operational verification systems but also in research-oriented verification activities.

When constructing an operational verification system, most of the workload is spent on data management issues (extracting the data from the archives, performing quality control, pairing forecasts with observations, stratifying or pooling the data), and relatively small effort is dedicated to the statistics computation. Operational verification systems are designed around the need for quick evaluation and display of statistics. Since the most time consuming process is data management, this issue is sometimes addressed by building a database of observation-forecast pairs. However, this solution is impractical because of space issues (it is a second archive of data), and because the algorithms to perform Quality Control (QC) and pairing are likely affected by changes with time, so that the archive needs to be rebuilt every time there is an algorithm update. A more efficient way of addressing this issue is by evaluating and archiving some basic summary statistics (e.g. contingency table counts), from which most of the verification statistics can be evaluated and aggregated, rapidly, over the desired time-period and region of interest.

During the Third International Workshop on Verification Methods examples were provided of new verification methods applied in an operational setting to assess various aspects of model performance from a new perspective. Schubiger *et al.* (2007) showed how new verification techniques can be integrated in existing verification suites to provide additional information on model performance on different spatial scales. Zingerle and Nurmi (2008) and Marsigli *et al.* (2008) have applied new verification methodologies in the operational environment to provide instant information on model behaviour to the forecasters. They have used high-resolution surface observations for precipitation and re-sampled satellite images to verify pseudo-satellite images produced by NWP models. A real time forecast verification (RTFV) system is being developed by E. Ebert for the Forecast Demonstration Project associated with the Beijing 2008 Olympic Games. The RTFV system is designed to produce and display real-time verification products to guide forecasters in formulating their final forecast. Both traditional scores and new feature-oriented and scale-verification techniques are included in the RTFV console.

Operational centres are encouraged to adopt some of the new verification approaches (in a parallel running framework), alongside the traditional techniques, both to gain from the added information about forecast quality that can be provided by these techniques, and to provide feedback to the verification developers for further improvement/tuning of the techniques. On the other side, the research community should try to develop their techniques with fast and efficient algorithms, to enable their use within operational verification systems. A dialogue and exchange between the two communities is necessary to achieve an optimal verification product.

6. Verification packages and open sources

The desire to share and promote the use of new more advanced verification approaches and the need for verification capabilities by many members of the meteorological community has led the verification community to develop several new verification packages (e.g. Holland *et al.*, 2007). Desirable properties of a package are that it is easily sharable (both in terms of code availability and data handling) and well documented. Demargne *et al.* (2007) underlined that, prior to building a verification package, it is important to define goals, customers, components, and desired capabilities of the system. Holland *et al.* (2007) extended the requirements further to include modularity, configurability, and flexibility. These characteristics are needed to appeal to a vast audience of diverse customers. Both authors insisted on concentrating efforts toward easily understandable informative metrics supported by customized graphical displays that help viewing and understanding the results.

The R statistical computing software (http://www.r-project.org/) is a well-documented, free and open-source programming language which includes the most robust and advanced (but well established) tools for statistical analysis. It operates on most operating systems (Windows, Linux and Apple OS) and it can incorporate codes from other languages (e.g. FORTRAN and C++). Since it is based on a user-interactive concept, R is suitable as research tool. On the other hand R codes can be run in batch-mode from shell scripts. However, R might encounter issues related to the size of large data sets and might not meet the efficiency requirements demanded in operational verification. Few packages related to verification have been developed in the R language.

The R Verification Package (Pocernich, 2007) includes many of the basic verification statistics outlined in Wilks (2006) and in Jolliffe and Stephenson (2003), such as traditional continuous and categorical verification scores and skill scores, conditional quantile plots, verification statistics for probability forecasts, ROC plots and reliability diagrams. In addition, some of the new verification techniques have been included in the package, such as the evaluation of skill scores which include measurement errors (Briggs and Ruppert, 2004), the intensity-scale technique (Casati *et al.*, 2004), and the circular CRPS (Grimit *et al.*, 2006). Recent updates have focused on calculating confidence intervals for skill scores, displaying intervals on attribute and ROC plots, and additional tools for statistical inference. Some other R packages (e.g. Carstensen *et al.*, 2007; Sing *et al.*, 2007), developed by the medical research community, include some advanced verification routines for the calculation of statistics related to the ROC curve.

In addition to the R package, software to run some of the new verification techniques is available for use by researchers, forecasters, or users who are interested in testing these methods with their data. The IDL code accompanying Ebert's (2008) article on the comparison of neighbourhood (fuzzy) verification techniques is available at http://www.bom.gov.au/bmrc/wefor/staff/eee/beth_ebert.htm, as is the code for the Ebert and McBride (2000) features-based technique. The C++ code for the MODE (Davis *et al.*, 2006a,b) object-based technique is available as a component of the model evaluation tools (Holland *et al.*, 2007) at http://www.dtcenter.org/met/users/. The R and FORTRAN code for the intensity-scale verification approach (Casati *et al.*, 2004) is available at http://www.met.rdg.ac.uk/~swr00bc/IS.verif.html. The R code for the Extreme Dependency Score of Ferro (2007b) is available at http://www.met.rdg.ac.uk/~sws02caf. A toolbox for the translation of ensembles of model runs into probabilistic weather forecasts (Roulston and Smith, 2003; Bröcker and Smith, 2008) and their evaluation (Roulston and Smith, 2002; Bröcker and Smith, 2007a,b) and value (Smith, 2003) can be found at http://www.lsecats.org. The verification community is strongly promoting algorithm sharing, to allow a more dynamic exchange of research tools and experiences, and a faster improvement and dissemination of new verification techniques.

7. User-oriented verification strategies

Different forecast users have many different needs with respect to forecast verification information. As an example, information on forecast quality can be used to learn about specific forecast deficiencies and refine particular aspects of NWP models, or to help define the post-processing needed to correct NWP forecast errors. Information on forecast performance can affect some meteorological services' administrative decisions. Finally, information on forecast quality can also be used by specific forecast users to interpret the forecast itself, to assess its level of trustworthiness (or uncertainty), to help make decisions regarding whether to take particular actions, and to estimate the value gained from the use of a forecast product for a specific purpose. Different verification strategies need to be tailored to the interests of specific users. The following paragraphs identify some user needs and illustrate how these needs affect the design and choice of verification strategies.

Model developers might need information about the behaviour of operational models both in a monitoring sense and in a scientific/development sense. For monitoring purposes, time series of accuracy measures and skill scores can be evaluated, and can be used as a baseline to evaluate the impact of proposed changes to the model. Spatial and temporal aggregation should be performed on regions characterized by similar weather regimes, seasonally or monthly, and for individual forecast lead times. Some of the new spatial verification techniques can be used to assess systematic displacement error or the scale-dependency of the forecast performance. Sometimes a conditional verification study can provide specific feedback on forecast quality for particular forecast or observed weather situations. Despite the fact that it is often ignored, information about the statistical significance of differences in verification results between older and new model versions is usually desirable.

Forecasters need to be aware of systematic model deficiencies in order to formulate their final forecasts and/or to issue warnings. Moreover, they might want to know about the capabilities of their own forecasts: optimally, each forecaster should be provided with feedback on his/her own performance. Forecasters can use the same type of verification information used by the modellers, in order to correct, as an example, systematic biases or errors in NWP model output related to specific weather conditions. Forecasters, however, are likely to require their verification information earlier, perhaps while their most recent forecast situation is still fresh in their minds. Thus, real-time verification systems which bring information to the forecasters based on the latest verification results could be appealing, for example, to use as a tool to choose the best-performing model among several candidate models.

The management of meteorological services may wish to know that their investment in research and development is indeed leading to improvements in the quality of weather forecasts. A few representative verification statistics providing information about overall forecast performance and long-term trends would normally meet this need. Aggregation of verification data is typically performed over large (administrative) regions (e.g. country boundaries, rather than regions characterized by similar weather regimes) and long periods.

Increasingly, owners or managers of businesses affected by the weather and who need guidance for their decisions have become users of weather and climate forecasts. These users all have an economic stake in the quality of the forecasts, to the extent that the weather sensitivity of such users can be defined in monetary terms. Verification information becomes an essential ingredient in the assessment of economic benefit (if any) of the forecast to them. Weather-sensitive business groups encompass a wide variety of social and economical sectors to individual decision makers, including, for example, the agricultural sector, transport and aviation, managers of electricity companies, operators of wind farms, managers of hydrological agencies, individual farmers, and retailers who might want to be sure to stock umbrellas when significant rain is forecast (see Smith *et al.*, 2001). Those users need information about the quality of forecasts in terms of the specific weather elements and categories of relevance for their application.

All this diversity in the user community for forecasts and verification information suggests a need for some sort of hierarchy of forecast verification methods, which reflects varying levels of needs to ensure user relevance of results. In general, it could be argued that the greater the tuning of the design of verification methods to specific users, the smaller the user group for which those results will be valid. The tendency has been to try to satisfy the largest number of potential forecast users with specific verification systems. Unfortunately, this can lead to verification efforts which really are of little use to anyone. The time has come to diversify verification methodology to match the diverse needs of users more closely, even if those user groups are relatively small. This goal also means the scientific community will need to work more closely with the user community in the design of new verification strategies.

8. Forecast value

The value of a forecast depends not only on how well it foreshadows future meteorological events but also upon its communication to decision makers and their ability to use information to mitigate impacts of the weather. From a user's perspective, it is usually those verification measures that tend to evaluate the accuracy and skill of a forecast in predicting the future paths of the weather, which are of most use. However, it is useful to distinguish the value of a forecast to a particular user from the verification of the quality of the weather forecast itself (Murphy, 1993). It is quite reasonable for a user to evaluate a forecast in terms of expected economic benefit. This cannot be done by using a skill score, but must be determined by evaluating whether or not the investment required to apply the information provided by the forecast is justified for that user. While forecast system improvement should be based on proper verification measures, a user's evaluation of forecast value should reflect the expected benefits gained from the use of the forecast, for that specific user.

The economic value of weather and climate forecasts has been of concern for more than a century (e.g. Liljas and Murphy, 1994; Katz and Murphy, 1997; Murphy, 1998). A common approach for estimating forecast value is based on the cost/loss decision model. This model assesses the forecast value by combining financial information from the user with verification results, generally framed in a contingency table format. The user's financial sensitivity to weather is generally quantified by the costs and losses to be faced if action or no-action is taken with respect a certain forecast. As an example, Richardson (2000) describes a method to estimate the value of the ECMWF ensemble prediction system with respect to a deterministic forecast, which depends on the user's cost and loss ratio. As noted by Zhu *et al.* (2002), such a model provides a relationship between value and the ROC curve. Rollins and Shaykewich (2003) proposed an alternative to the cost-avoidance approach, and assess the weather forecast value via a demand-based (or willingness-to-pay) approach. While the cost–loss model provides a straightforward approach for estimating value, and has been widely applied to meteorological forecasts (e.g., Katz and Murphy, 1997), this approach typically is overly simplistic and is difficult to apply to problems involving complex-decision situations or multiple-decision makers. Other methods that can be applied to this problem include survey techniques and econometric models (Katz and Murphy, 1997). Much more effort is needed to develop this area of forecast evaluation.

A parallel to the distinction between quality and value also exists in the context of public safety and severe weather: in particular, issuing statements to motivate action is quite distinct from issuing information to inform. The effectiveness of warning and evacuation orders depend on much more than the probability of the event, even when that can be forecast precisely (Roulston and Smith, 2004), and thus the value of the forecast in terms of public safety will also depend on more than the meteorological accuracy of the forecast system.

While the meteorological community tends to focus on severe weather, there is significant value in forecasting relatively boring weather with significant economic impact which can, in fact, be mitigated (Roulston *et al.*, 2003; Altalo and Hale, 2004; Altalo and Smith, 2004; Roulston *et al.*, 2005). Information from ensemble forecasts is used every day by companies around the world. In sectors such as energy generation and financial markets, the utility of the forecasts for non-severe weather can have very large impacts.

9. Conclusions

This paper has presented a survey of the 'state of the art' in verification practice, research and development, through the eyes of the Third International Verification Methods Workshop, Reading, UK, January 2007. The workshop and a survey of the recent literature reveals that forecast verification has been a very active field of research in the last decade or so. The field has been broadened with the development of new techniques; furthermore, existing techniques, particularly those relating to verification of extreme events and probability forecasts, have been re-evaluated and extended to new applications. Efforts to promote verification and the proper interpretation of verification output have been made through recent books (Wilks, 2006 and Jolliffe and Stephenson, 2003), through the series of workshops organized by the Joint Working Group on Verification, and through the verification efforts associated with international Forecast Demonstration Projects, such as those organized for the Sydney and Beijing Olympics.

Some general trends can be identified, for example, toward the development of techniques which account for the spatial structure and presence of features in meteorological fields, and toward the inclusion of confidence limits and consistency bounds in verification results. The last of these is long overdue, and unfortunately still incipient. Such a trend must continue. A renewed emphasis on user-oriented verification is evident, but it is also clear that much remains to be done before it can be said that most verification efforts really meet the needs of specific, identified user communities. Finally, the standardization of verification practice is being encouraged through the generation and dissemination of open-source code such as the 'R' verification package.

Specific issues have been identified for further research and development. These include the need to extend spatial verification methods to sparse observation networks; the importance of properness of scores in verification systems; issues surrounding the computation of the ROC area; the development of methods which are more appropriate than standard scores for extreme event verification; the need for operational verification systems which use truth datasets that are quality controlled independently of the model or models being verified; and last, but certainly not least, the need to tune verification efforts to a user or user community which is defined and consulted *a priori*.

With the increasing availability of open-source, well-documented, user-friendly codes for verification, it is becoming easier to carry out meaningful verification studies. Furthermore, the results of verification research and development are being made available to the wider research and operations community more quickly than ever before. These are excellent trends which will enhance the accessibility of verification methods and results for all users in the future.

Acknowledgements

The authors wish to thank L.A. Smith for his thoughtful comments on the section on probability forecasts and ensemble verification, and for his contribution to the forecast value section. We also would like to thank the World Meteorological Organization, the European Cooperation in the Field of Scientific and Technical Research (COST), and the European Centre for Medium-Range Weather Forecasts (ECMWF) for their support of the Third International Verification Methods Workshop, and the workshop participants for their excellent contributions.

References

Agresti A. 2002. *Categorical Data Analysis*, 2nd edn. Wiley: Chichester, UK; 710.

Altalo MG, Hale M. 2004. Turning weather forecasts into business forecasts. *Environmental Finance*, May: 20–21.

Altalo MG, Smith LA. 2004. Using ensemble weather forecasts to manage utilities risk. *Environmental Finance* **20**: 8–9.

Anderson JL. 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate* **9**: 1518–1530.

Atger F. 2001. Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Processes in Geophysics* **8**: 401–417.

Atger F. 2004. Estimation of the reliability of ensemble-based probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society* **130**: 627–646.

Baddeley AJ. 1992. An error metric for binary images. In *Robust Computer Vision*, Forstner W, Ruwiedel S (eds). Wichmann: Bonn, Germany; 59–78.

Baldwin ME, Lakshmivarahan S, Kain JS. 2002. Development of an "events-oriented" approach to forecast verification. *Preprints, 19th Conference on Weather Analysis and Forecasting*. American Meteorological Society: San Antonio, TX; 255–258.

Beirlant J, Goegebeur Y, Segers J, Teugels J. 2004. *Statistics of Extremes: Theory and Applications*. John Wiley and Sons: Chichester, UK; 490.

Bougeault P. 2002. WGNE survey of verification methods for numerical prediction of weather elements and severe weather events. CAS/JSC WGNE Report No. 18, Appendix C. WMO/TD.No.1173, Toulouse, France. Available on the internet at http://www.wmo.ch/pages/prog/wcrp/pdf/wgne18rpt.pdf.

Bowler NE, Pierce CE, Seed AW. 2007. STEPS: A probabilistic forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Quarterly Journal of the Royal Meteorological Society* **132**: 2127–2155.

Brewster KA. 2003. Phase correcting data assimilation and application to storm-scale numerical prediction. *Monthly Weather Review* **131**: 480–507.

Brier GW. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**: 1–3.

Briggs WM, Levine RA. 1997. Wavelets and field forecast verification. *Monthly Weather Review* **125**: 1329–1341.

Briggs WM, Ruppert D. 2004. Assessing the skill of Yes/No predictions. *Biometrics* **61**(3): 799–807.

Brill KF. 2002. Automated east-west phase error calculation. Technical report, US Department of Commerce/NOAA/NSW/NCEP/HPC. Available at http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/Brill/Brill_EW_PhaseError.html.

Bröcker J, Smith LA. 2007a. Increasing the reliability of reliability diagrams. *Weather and Forecasting* **22**(3): 651–661.

Bröcker J, Smith LA. 2007b. Scoring probabilistic forecasts: on the importance of being proper. *Weather and Forecasting* **22**(2): 382–388.

Bröcker J, Smith LA. 2008. From ensemble forecasts to predictive distribution functions. *Tellus* in press.

Candille G, Talagrand O. 2005. Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society* **131**: 2131–2150.

Carstensen B, Plummer M, Läärä E, Myatt M, Clayton D. 2007. Epi: A package for statistical analysis in epidemiology. R package version 0.7.0, http://www.pubhealth.ku.dk/~bxc/Epi.

Casati B, Wilson LJ. 2007. A new spatial scale decomposition of the Brier score: application to the verification of lightning probability forecasts. *Monthly Weather Review* **135**: 3052–3069.

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

Coles S. 2001. *An Introduction to Statistical Modeling of Extreme Values*. Springer-Verlag: London, UK; 208.

Coles S, Heffernan J, Tawn J. 1999. Dependence measures for extreme value analyses. *Extremes* **2**(4): 339–365.

Davis C, Brown B, Bullock R. 2006a. Object-based verification of precipitation forecasts. Part I: methodology and application to Mesoscale Rain Areas. *Monthly Weather Review* **134**: 1772–1784.

Davis C, Brown B, Bullock R. 2006b. Object-based verification of precipitation forecasts. Part II: application to convective rain systems. *Monthly Weather Review* **134**: 1785–1795.

De Elia R, Laprise R, Denis B. 2002. Forecasting skill limits of nested, limited-area models: a perfect model approach. *Monthly Weather Review* **130**: 2006–2023.

Demargne J, Seo DJ, Wu L, Schaake J, Brown J. 2007. Verifying hydrologic forecasts in the US National Weather Service. In *Oral Presentation at the Third International Workshop on Verification Methods*, Reading, UK. PDF available from: http://www.ecmwf.int/newsevents/meetings/workshops/2007/jwgv/index.html.

Denhard M, Trepte S, Göber M, Anger B. 2007. Verification of the SRNWP-PEPS. In *Oral Presentation at the Third International Workshop on Verification Methods*, Reading, UK. PDF available from: http://www.ecmwf.int/newsevents/meetings/workshops/2007/jwgv/index.html.

Denis B, Laprise R, Caya D. 2003. Sensitivity of a regional climate model to the resolution of the lateral boundary conditions. *Climate Dynamics* **20**: 107–126.

Donaldson RJ, Dyer RM, Kraus MJ. 1975. An objective evaluator of techniques for predicting severe weather events. In *Preprints, Ninth Conference on Severe Local Storms*. American Meteorological Society: Norman, OK; 321–326.

Doswell CA, David-Jones R, Keller DL. 1990. On summary measures of skill in rare event forecasting based on contingency tables. *Weather and Forecasting* **5**: 576–585.

Douglas RJ. 2000. Rearrangements of functions with applications to meteorology and ideal Fluid Flow. Internal Report 118, JCMM. Department of Meteorology, University of Reading, RG6 6BB, Reading, UK.

Du J, Mullen SL, Sanders F. 2000. Removal of distortion error from an ensemble forecast. *Journal of Applied Meteorology* **35**: 1177–1188.

Ebert EE. 2008. Fuzzy verification of high resolution gridded forecasts: a review and proposed framework. *Meteorological Applications* **15**: 53–66.

Ebert EE, McBride JL. 2000. Verification of precipitation in weather systems: Determination of systematic errors. *Journal of Hydrology* **239**: 179–202.

Ebert EE, Wilson LJ, Brown BG, Nurmi P, Brooks HE, Bally J, Jaeneke M. 2004. Verification of nowcasts from the WWRP Sydney 2000 forecast demonstration project. *Weather and Forecasting* **19**: 73–96.

Epstein ES. 1969. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology* **8**: 985–987.

Ferro CAT. 2007a. Comparing probabilistic forecasting systems with the Brier score. *Weather and Forecasting* **22**: 1076–1089.

Ferro CAT. 2007b. A probability model for verifying deterministic forecasts of extreme events. *Weather and Forecasting* **22**: 1089–1100.

Ferro CAT, Richardson DS, Weigel AP. 2008. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications* **15**: 19–24.

Finley JP. 1884. Tornado prediction. *American Meteorological Journal* **1**: 85–88.

Germann U, Zawadzki I. 2002. Scale-dependence of the predictability of precipitation from continental radar images. Part I: description of the methodology. *Monthly Weather Review* **130**: 2859–2873.

Germann U, Zawadzki I. 2004. Scale-dependence of the predictability of precipitation from continental radar images. Part II. *Journal of Applied Meteorology* **43**: 74–89.

Gilbert GK. 1884. Finley's tornado predictions. *American Meteorological Journal* **1**: 166–172.

Gilleland E, Lee T, Halley-Gotway J, Bullock R, Brown B. 2008. Computationally efficient spatial forecast verification using Baddeley's Δ image metric. *Monthly Weather Review* in press.

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

Göber M, Wilson CA, Milton SF, Stephenson DB. 2004. Fairplay in the verification of operational quantitative precipitation forecasts. *Journal of Hydrology* **288**: 225–236.

Good IJ. 1952. Rational decisions. *Journal of the Royal Statistical Society Series B-Methodological* **14**: 107–114.

Grams JS, Gallus WA, Koch SE, Wharton LS, Loughe A, Ebert EE. 2006. The use of a modified Ebert-McBride technique to evaluate mesoscale model QPF as a function of convective system morphology during IHOP 2002. *Weather and Forecasting* **21**: 288–306.

Grimit EP, Gneiting T, Berrocal VJ, Johnson NA. 2006. The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society* **132**: 1–17.

Hamill TM. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review* **129**: 550–560.

Hamill TM, Colucci SJ. 1997. Verification of Eta-RSM short range ensemble forecasts. *Monthly Weather Review* **125**: 711–724.

Hamill TM, Juras J. 2006. Measuring forecast skill: is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society* **132**: 2905–2923.

Harris D, Foufoula-Georgiou E, Droegemeier KK, Levit JJ. 2001. Multiscale statistical properties of a high-resolution precipitation forecast. *Journal of Hydrometeorology* **2**: 406–418.

Hersbach H. 2000. Decomposition of the continuous rank probability score for ensemble prediction systems. *Weather and Forecasting* **15**: 559–570.

Hoffman RN, Liu Z, Louis J-F, Grassotti C. 1995. Distortion representation of forecast errors. *Monthly Weather Review* **123**: 2758–2770.

Holland L, Fowler T, Brown B, Nance L. 2007. Designing a state-of-the-art verification system. In *Oral Presentation at the Third International Workshop on Verification Methods*, Reading, UK. PDF available from: http://www.ecmwf.int/newsevents/meetings/workshops/2007/ jwgv/index.html.

Hopson T, Weiss J, Webster P. 2007. Verifying the relationship between ensemble forecast spread and skill. In *Oral Presentation at the Third International Workshop on Verification Methods*, Reading, UK. PDF available from: http://www.ecmwf.int/newsevents/meetings/workshops/2007/jwgv/index.html.

Hsu WR, Murphy AH. 1986. The attributes diagram. A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting* **2**: 285–293.

Intergovernmental Panel on Climate Change. 2001. 3rd Assessment Report, Available from http://www.ipcc.ch.

Jolliffe IT, Stephenson DB. 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons: Chichester, UK; 240.

Katz RW, Murphy AH. 1997. *Economic Value of Weather and Climate Forecasts*. Cambridge University Press: New York; 222.

Keenan T, Joe P, Wilson J, Collier C, Golding B, Burgess D, May P, Pierce C, Bally J, Crook A, Sills D, Berry L, Bell I, Fox N, Pielke R Jr, Ebert EE, Eilts M, O'Loughlin K, Webb R, Carbone R, Browning K, Roberts R, Mueller C. 2002. The sydney 2000 world weather research programme forecast demonstration project: overview and current status. *Bulletin of the American Meteorological Society* **84**: 1041–1054.

Lalaurette F. 2003. Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quarterly Journal of the Royal Meteorological Society* **129**(594): 3037–3057.

Liljas E, Murphy A. 1994. Anders Ångström and his early papers on probability forecasting and the use/value of weather forecasts. *Bulletin of the American Meteorological Society* **75**: 1227–1236.

Mailier PJ, Jolliffe IT, Stephenson DB. 2006. Quality of weather forecasts: Review and Recommendations. Royal Meteorological Society Project Report, 89.

Marsigli C, Montani A, Paccagnella T. 2006. Verification of the COSMO-LEPS new suite in terms of precipitation distribution. COSMO Newsletter No. 6, Available at http://www.cosmo-model.org/public/newsLetters.htm.

Marsigli C, Montani A, Paccagnella T. 2008. A spatial verification method applied to the evaluation of high-resolution ensemble forecasts. *Meteorological Applications* **15**: 127–145.

Marzban C. 1998. Scalar measures of performance in rare-event situations. *Weather and Forecasting* **13**: 753–763.

Marzban C, Sandgathe S. 2006. Cluster analysis for verification of precipitation fields. *Weather and Forecasting* **21**: 824–838.

Mason I. 1982. A model for assessment of weather forecasts. *Australian Meteorological Magazine* **30**: 291–303.

Mason S. 2007. Do high skill scores mean good forecasts? In *Oral Presentation at the Third International Workshop on Verification Methods*, Reading, UK, PDF available from: http://www.ecmwf.int/newsevents/meetings/workshops/2007/jwgv/index.html.

Mason S. 2008. Understanding forecast verification statistics. *Meteorological Applications* **15**: 31–40.

Mittermaier M. 2007. Using time-lag ensemble techniques to assess behaviour of high-resolution precipitation forecasts. In *Oral Presentation at the Third International Workshop on Verification Methods*, Reading, UK, PDF available from: http://www.ecmwf.int/newsevents/meetings/workshops/2007/jwgv/index.html.

Murphy AH. 1973. A new vector partition of the probability score. *Journal of Applied Meteorology* **12**: 595–600.

Murphy AH. 1991. Probabilities, odds, and forecasts of rare events. *Weather and Forecasting* **6**: 302–307.

Murphy AH. 1993. What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting* **8**: 281–293.

Murphy AH. 1996. The Finley affair: a signal event in the history of forecast verification. *Weather and Forecasting* **11**: 3–20.

Murphy AH. 1998. The early history of probability forecasts: some extensions and clarifications. *Weather and Forecasting* **13**: 5–15.

Murphy AH, Winkler RL. 1987. A general framework for forecast verification. *Monthly Weather Review* **115**: 1330–1338.

Nachamkin JE. 2004. Mesoscale verification using meteorological composites. *Monthly Weather Review* **132**: 941–955.

Nachamkin JE, Chen S, Schmidt J. 2005. Evaluation of heavy precipitation forecasts using composite-based methods: a distributions-oriented approach. *Monthly Weather Review* **133**: 2163–2177.

Nehrkorn T, Hoffman RN, Grassotti C, Louis J-F. 2003. Feature calibration and alignment to represent model forecast errors: empirical regularization. *Quarterly Journal of the Royal Meteorological Society* **129**: 195–218.

Nurmi P. 2003. Recommendations on the verification of local weather forecasts. *ECMWF Technical Memorandum* **430**: 19.

Nurmi P, Näsman S, Zingerle C. 2007. Entity-based verification in the intercomparison of three NWP models during a heavy snowfall event. *Geophysical Research Abstracts* **9**: 09247, Oral presentation at the EGU General Assembly, Vienna, Austria, PDF available from: http://www.cosis.net/abstracts/EGU2007/09247/EGU2007-J-09247.pdf.

Pocernich M. 2007. Verification: forecast verification utilities. R package version 1.20. R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, ISBN 3-900051-07-0, URL http://www.r-project.org.

Rezacova D, Sokol Z. 2005. The use of radar data in the verification of a high resolution quantitative forecast of convective precipitation. In *Extended Abstract, WWRP Symposium on Nowcasting and Very Short Range Weather Forecasting*, Toulouse, 5–9 September 2005.

Richardson DS. 2000. Skill and economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society* **126**: 649–668.

Roberts NM, Lean HW. 2007. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review* **136**: 78–97.

Rollins KS, Shaykewich J. 2003. Using willingness-to-pay to assess the economic value of weather forecasts for multiple commercial sectors. *Meteorological Applications* **10**: 31–38.

Roulston MS, Smith LA. 2002. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review* **130**: 1653–1660.

Roulston MS, Smith LA. 2003. Combining dynamical and statistical ensembles. *Tellus Series A-Dynamic Meteorology and Oceanography* **55**: 16–30.

Roulston MS, Smith LA. 2004. The boy who cried wolf revisited: the impact of false alarm intolerance on cost-loss scenarios. *Weather and Forecasting* **19**(2): 391–397.

Roulston MS, Ellepola J, Smith LA. 2005. Forecasting wave height probabilities with numerical weather prediction models. *Ocean Engineering* **32**(14–15): 1841–1863.

Roulston MS, Kaplan DT, Hardenberg J, Smith LA. 2003. Using medium-range weather forecasts to improve the value of wind energy production. *Renewable Energy* **28**(4): 585–602.

Roy G, Turcotte V. 2007. Verification des algorithms radars du GemLam 2.5. Internal Report, Severe Weather National Lab, Environment Canada.

Schaefer JT. 1990. The critical success index as an indicator of warning skill. *Weather and Forecasting* **5**: 570–575.

Schubiger F, Ament F, Baehler T. 2007. Verification of the COSMO 7km model and new developments for the 2.2 Km model with a special emphasis on precipitation. In *Oral Presentation at the Third International Workshop on Verification Methods*, Reading, UK, PDF available from: http://www.ecmwf.int/newsevents/meetings/workshops/2007/jwgv/index.html.

Sing T, Sander O, Beerenwinkel N, Lengauer T. 2007. ROCR: Visualizing the performance of scoring classifiers. R package version 1.0-1, http://rocr.bioinf.mpi-sb.mpg.de.

Smith LA. 1997. The maintenance of uncertainty. *Proceedings International School of Physics "Enrico Fermi"*, Course CXXXIII. Societa Italiana di Fisica: Bologna; 177–246.

Smith LA. 2001. Disentangling uncertainty and error: on the predictability of nonlinear systems. In *Nonlinear Dynamics and Statistics*, Mees AI (ed.). Birkhauser: Boston, MA; 31–64.

Smith LA. 2003. Predictability past predictability present. In *Predictability of Weather and Climate*, Palmer T, Hagedorn R (eds). Cambridge University Press 2006: Cambridge, UK.

Smith LA, Hansen JA. 2004. Extending the limits of forecast verification with the minimum spanning tree. *Monthly Weather Review* **132**: 1522–1528.

Smith LA, Roulston M, von Hardenberg J. 2001. End-to-end ensemble forecasting: towards evaluating the economic value of an ensemble prediction system. *ECMWF Technical Memorandum* **336**: 29.

Stanski HR, Wilson LJ, Burrows WR. 1989. Survey of common verification methods in meteorology. World Weather Watch Technical Report No. 8, WMO/TD No.358. WMO: Geneva, 114.

Stephenson DB. 2008. Definition, diagnosis, and origin of extreme weather and climate events. In *Climate Extremes and Society*, Diaz HF, Murnane RJ (eds). Cambridge University Press: New York; 348.

Stephenson DB, Casati B, Ferro CAT, Wilson C. 2008a. The extreme dependency score: a non-vanishing score for forecasts of rare events. *Meteorological Applications* **15**: 41–51.

Stephenson DB, Coelho CAS, Jolliffe IT. 2008b. Two extra components in the Brier Score Decomposition. *Weather and Forecasting* in press.

Swets JA. 1986. Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psychological Bulletin* **99**: 181–198.

Swets JA, Pickett RM. 1982. *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory*. Academic Press: New York.

Theis SE, Hense A, Damrath U. 2005. Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorological Applications* **12**: 257–268.

Toth Z, Talagrand O, Zhu Y. 2006. The attributes of forecast systems: a general framework for the evaluation and calibration of weather forecasts. In *Predictability of Weather and Climate*, Palmer T, Hagedorn R (eds). Cambridge University Press: Cambridge, UK; 584–595.

Tremblay A, Cober SG, Glazer A, Isaac G, Mailhot J. 1996. An intercomparison of mesoscale forecasts of aircraft icing using SSM/I retrievals. *Weather and Forecasting* **11**: 66–77.

Turner BJ, Zawadzki I, Germann U. 2004. Predictability of precipitation from continental radar images. Part III: operational nowcasting implementation (MAPLE). *Journal of Applied Meteorology* **43**: 231–248.

Tustison B, Foufoula-Georgiou E, Harris D. 2003. Scale-recursive estimation for multisensor Quantitative Precipitation Forecast verification: a preliminary assessment. *Journal of Geophysical Research* **108**: 11775–11784.

Venugopal V, Basu S, Foufoula-Georgiu E. 2005. A new metric for comparing precipitation patterns with an application to ensemble forecasts. *Journal of Geophysical Research* **110**: D08111.

Volkert H. 2005. The Mesoscale Alpine Programme (MAP) – a multi-facetted success story. Preprints ICAM/MAP 2005, Zadar, Croatia, 23–27 May 2005, 226–230.

Weigel AP, Liniger MA, Appenzeller C. 2007. The discrete Brier and ranked probability skill scores. *Monthly Weather Review* **135**: 118–124.

Weisheimer A, Smith LA, Judd K. 2004. A new view of forecast skill: bounding boxes from the DEMETER ensemble seasonal forecasts. *Tellus* **57**(3): 265–279.

Wernli H, Paulat M, Frei C. 2006. The concept of a new error score SAL for the verification of high-resolution QPF. *Presented at the 2nd International Symposium on Quantitative Precipitation Forecasting and Hydrology*, Boulder, US, 4–8 June 2006.

Wilks DS. 2004. The Minimum Spanning Tree (MST) histogram as a verification tool for multidimensional ensemble forecasts. *Monthly Weather Review* **132**: 1329–1340.

Wilks DS. 2006. *Statistical Methods in Atmospheric Science*, 2nd edn. Academic Press: Burlington, MA; 627.

Wilson LJ. 2000. Comments on "Probabilistic predictions of precipitation using the ECMWF ensemble prediction system". *Weather and Forecasting* **15**: 361–364.

Wilson LJ, Gneiting T. 2007. Another look at proper scoring rules. In *Oral Presentation at the Third International Workshop on Verification Methods*, Reading, UK, PDF available from: http://www.ecmwf.int/new-sevents/meetings/workshops/2007/jwgv/index.html.

Wilson LJ, Burrows WR, Lanzinger A. 1999. A strategy for verification of weather element forecasts from an ensemble prediction system. *Monthly Weather Review* **127**: 956–970.

WMO. 2000. Guidelines of performance assessment of public weather services. Document available at http://www.wmo.ch/web/aom/pwsp/downloads/guidelines/TD-1023.pdf.

Yu X. 2005. Overview of Beijing 2008 Olympics WWRP Nowcast FDP and implementation plan. In *Presented at the WWRP Symposium on Nowcasting and Very Short Range Forecasting*, Toulouse, 5–9 September 2005.

Zepeda-Arce J, Foufoula-Georgiou E, Droegemeier KK. 2000. Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *Journal of Geophysical Research* **105**(D8): 10129–10146.

Zhu Y, Toth Z, Wobus R, Richardson D, Mylne K. 2002. The economic value of ensemble-based weather forecasts. *Bulletin of the American Meteorological Society* **83**: 73–83.

Zingerle C, Nurmi P. 2008. Monitoring and verifying cloud forecasts originating from operational numerical models. *Meteorological Applications* submitted.