METEOROLOGICAL APPLICATIONS
Meteorol. Appl. **20**: 130–139 (2013)
Published online 22 May 2013 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/met.1392

Royal Meteorological Society

Progress and challenges in forecast verification

E. Ebert,[a]* L. Wilson,[b] A. Weigel,[c] M. Mittermaier,[d] P. Nurmi,[e] P. Gill,[d] M. Göber,[f] S. Joslyn,[g] B. Brown,[h] T. Fowler[h] and A. Watkins[i]

[a] *Centre for Australian Weather and Climate Research (CAWCR), Melbourne, Australia*
[b] *Environment Canada, Dorval, Canada*
[c] *MeteoSwiss, Geneva, Switzerland*
[d] *Met Office, Exeter, UK*
[e] *Finnish Meteorological Institute, Helsinki, Finland*
[f] *Deutscher Wetterdienst, Offenbach, Germany*
[g] *Department of Psychology, University of Washington, Seattle, Washington, USA*
[h] *National Center for Atmospheric Research, Boulder, CO, USA*
[i] *Bureau of Meteorology, Melbourne, Australia*

ABSTRACT: Verification scientists and practitioners came together at the 5th International Verification Methods Workshop in Melbourne, Australia, in December 2011 to discuss methods for evaluating forecasts within a wide variety of applications. Progress has been made in many areas including improved verification reporting, wider use of diagnostic verification, development of new scores and techniques for difficult problems, and evaluation of forecasts for applications using meteorological information. There are many interesting challenges, particularly the improvement of methods to verify high resolution ensemble forecasts, seamless predictions spanning multiple spatial and temporal scales, and multivariate forecasts. Greater efforts are needed to make best use of new observations, forge greater links between data assimilation and verification, and develop better and more intuitive forecast verification products for end-users. Copyright © 2013 Royal Meteorological Society

KEY WORDS forecast verification; evaluation; challenges

Received 15 October 2012; Revised 4 February 2013; Accepted 27 February 2013

1. Introduction

Verification of meteorological and oceanographic forecasts is essential for monitoring their accuracy, understanding their errors, and making improvements in forecasting systems. Recent years have seen a growing interest in new approaches for forecast verification, as evidenced by the rising number of publications on newly developed verification techniques. The increased resolution of numerical models has created a need for new and different diagnostic methods to understand their limitations. Ensemble prediction has become widespread, calling for ways to assess not only their spread but also the probabilistic and scenario products that can be generated from ensembles. Prediction of extreme weather, while always of interest, has taken on new importance in light of our improved understanding of weather and climate variability and change, requiring verification methods that are targeted to extreme and rare events. Forecasting applications based on meteorological predictions, for example, continuous streamflow, wildfire behaviour, crop yields, and renewable energy prediction, call for new ways of evaluating these forecasts that are more relevant to the downstream application.

New types of observations offer the opportunity to evaluate model predictions in new ways. For example, active remote sensing of clouds from space-based and ground-based radar and lidar measurements provides detailed information on cloud vertical structures that can be used to verify model clouds both structurally and statistically. Existing observations, when carefully quality controlled and harmonized, can now be used for more extensive and detailed verification than was possible in the past (e.g., EUMETNET's OPERA radar programme; Huuskonen *et al*., 2012).

Finally, the demands of governments for greater accountability, combined with the need to justify funding expenditures on observing and forecasting system improvements, have stepped up the forecast verification effort of meteorological centres, in some cases just to demonstrate the value of their existing levels of activity in the face of competition for resources.

The World Meteorological Organization (WMO) Joint Working Group on Forecast Verification Research (JWGFVR) was formed in 2002 to (among other things), 'serve as a focal point for the development and dissemination of new verification methods ... and facilitate and encourage training and dissemination of information on verification methodologies'. The JWGFVR has run a series of international verification methods workshops with linked tutorial courses and scientific symposia. The fifth of these workshops was held at the Bureau of Meteorology in Melbourne, Australia, in December 2011. It brought together over 150 experts and interested scientists to discuss recent advances in the theory and practice of verification of weather and climate forecasts.

This paper summarizes recent progress in forecast verification that was highlighted at the December 2011 workshop.

* Correspondence: E. Ebert, Centre for Australian Weather and Climate Research (CAWCR), Melbourne 3001, Australia. E-mail: e.ebert@bom.gov.au

Challenges are identified and discussed in light of changing observations, evolving forecast types, and the increasing need for quantitative information on forecast uncertainty. We conclude with an outlook for future activities in this science.

2. Recent progress

2.1. Improved verification practice

The practice of forecast verification is clearly improving. For example, in the past verification results for meteorological forecasts have tended to be computed (or at least reported) without any information on the significance or confidence in the results. This has made it difficult to judge the robustness of results and especially to compare competing forecasts. Perhaps due to the influence of relevant books and online resources on verification (e.g., Wilks, 2005; Jolliffe and Stephenson, 2012[1] ; JWGFVR, 2013), articles (e.g., Jolliffe, 2007; Casati *et al*., 2008), recent training activities, and interactions between weather and climate communities, we are now seeing more widespread use of confidence intervals and significance tests to quantify the uncertainty of the verification results. Studies showing the distribution of errors, for example using box-whiskers plots or histograms, are also becoming more common and provide additional useful information beyond simply the average error.

Another example is the increased availability of verification information. Not only are performance metrics being calculated on a routine basis, they are being displayed in user-friendly fashion on internal and public websites. Some notable examples are the Finnish Meteorological Institute's internal pages (Nurmi, 2011) and the revised WMO Commission for Basic Systems (CBS) web portal hosted by the lead centre European Centre for Medium-Range Weather Forecasts (ECMWF) (http://apps.ecmwf.int/wmolcdnv/). The revised CBS metrics aim to include bootstrapped confidence intervals and greater consistency in the calculation of scores and the observations (radiosondes) used. Even greater exposure of verification results to the public would be desirable.

2.2. New diagnostic methods becoming mainstream

The use of spatial verification methods has become far more frequent and more routine, especially for evaluating output from high resolution numerical weather prediction (NWP) models where traditional verification metrics are strongly affected by the 'double penalty' associated with errors in the predicted location of weather features. In response to a proliferation of newly developed spatial verification methods, the Spatial Verification Methods Intercomparison Project (ICP; Gilleland *et al*., 2009) tested their ability to provide intuitive and useful forecast quality information using high resolution precipitation forecasts as an example. The methods were grouped into (1) neighbourhood methods that give credit for close forecasts, (2) scale separation methods that isolate scale-dependent errors, (3) object-based methods that evaluate attributes of coherent features, and (4) field deformation methods that measure phase and amplitude errors. There was no 'winner'; rather the best approach depends on the nature of the forecast and the questions being asked (Gilleland *et al*., 2009).

Among the neighbourhood methods, the Fractions Skill Score (FSS; Roberts and Lean, 2008), is now widely used in modelling centres to evaluate high resolution NWP precipitation forecasts against radar observations (e.g., Weusthoff *et al*., 2010; Duc *et al*., 2011; Mittermaier *et al*., 2011). This approach compares fractions of modelled and observed rain exceeding a given threshold within spatial neighbourhoods. A nice by-product of this technique is the determination of a 'skillful scale' at which an acceptable level of skill is reached. When plotted as a time series this intuitive measure can be used to monitor forecast performance (Mittermaier and Roberts, 2010).

The features-based methods have also continued to increase in popularity. The two most mature methods, namely the Contiguous Rain Area (CRA; Ebert and McBride, 2000) and the Method for Object-based Diagnostic Evaluation (MODE; Davis *et al*., 2006) are most often used for precipitation evaluation (e.g., Gallus, 2010), although they are now being applied in other contexts for evaluating climate features (e.g., Moise and Delage, 2011; Rikus *et al*., 2011), atmospheric rivers (Clark *et al*., 2011) and cloud structures (Kucera *et al*., 2011; Mittermaier and Bullock, 2013).

These newer diagnostic methods are generally more complex to calculate than the traditional verification scores. Their uptake has been greatly facilitated by the sharing of code by the intercomparison project, and especially through such freely available packages as the Model Evaluation Tools (MET; Fowler *et al*., 2010) and the R statistical package (R Development Core Team, 2011).

The categorical performance diagram (Roebber, 2009) is also becoming popular for visualizing forecast quality at a glance. It entails the plotting of the probability of detection (POD) *versus* the success ratio (SR = 1 − false alarm ratio). Since these are related to the frequency bias (FBI) and critical success index (CSI, also called the threat score) through the elements of the 2 × 2 contingency table, one can see all four scores at once. For a given level of accuracy or bias, trade-offs in performance, such as increased detections at the expense of more false alarms, can be easily seen. For good forecasts, POD, SR, FBI and CSI all approach unity, located in the upper right corner of the diagram. The performance diagram is similar in concept to the ROC plot, which is also a plot of two contingency table metrics, the POD and the false alarm rate. To illustrate a practical example of its use, Roberts *et al*. (2012) highlight differences in nowcasting skill before and after human intervention by 'connecting the dots' (i.e., the scores for both forecasts) in performance diagrams.

2.3. New scores for difficult issues

New scores and metrics have been developed in the last few years to deal with three particular issues: (1) how to make aggregated scores for precipitation climatologically relevant while enabling consistent forecast system monitoring across regions and seasons, (2) how to assess forecasts for rare events where the base rate tends to zero, and (3) how to give consistent performance information across forecasts of different types (e.g., categorical *vs* real-valued, or deterministic *vs* probabilistic).

The Stable Equitable Error in Probability Space (SEEPS) was developed by Rodwell *et al*. (2010, 2011) to enable long-term monitoring of trends in forecast quality. Based on the older Linear Error in Probability Space (LEPS) score (Potts *et al*., 1996), SEEPS was designed specifically to verify quantitative precipitation forecasts to give accuracy information on the occurrence

[1] 1st edition published in 2003.

of both measureable and heavy precipitation, where the threshold for 'heavy' precipitation depends on the observed long-term precipitation climatology. This makes it possible to meaningfully compare model performance across different regions and seasons. Since its introduction, SEEPS has been adopted as a headline measure at ECMWF and the Finnish Meteorological Institute (FMI), where target setting based on SEEPS is one key performance measures for the Institute's 5-year planning. This kind of concrete application of new verification metrics is a demonstration of the linkage between verification methodology research, verification for administrative purposes, and user-oriented verification, where the user community in this case includes both model developers and the national weather services of the member and cooperating states of ECMWF (see Section 3.6). SEEPS has recently been demonstrated for global precipitation forecasts (Haiden *et al.*, 2012) and regional 6 h precipitation forecasts (North *et al.*, 2013).

Stephenson *et al.* (2008) proposed the Extreme Dependency Score (EDS) to address the verification of rare (low base rate) categorical events, those events which are often the subject of weather warnings. Since most existing contingency table-based scores tend towards zero when the base rate becomes small, they are hard to use and interpret for rare events, or in situations where the base rate varies among forecasts to be compared, e.g., precipitation forecasts for arid areas and wet areas. While the EDS does not vanish for low base rate events, it depends on the base rate and can be increased by over-forecasting (Ghelli and Primo, 2009; Primo and Ghelli, 2009). These two deficiencies are accounted for by the Stable Extreme Dependency Score (SEDS), proposed by Hogan *et al.* (2009). The properties of both the EDS and SEDS were further explored by Ferro and Stephenson (2011), who introduced two more variants, the Extremal Dependency Index (EDI) and the Symmetric Extremal Dependency Index (SEDI), both of which are absolutely independent of base rate. The EDI and SEDI are functions of the hit rate and false alarm rate only, which makes them of interest in user-focused verification where attributes such as discrimination are important. The SEDS, on the other hand, uses the forecast frequency, which makes it a useful diagnostic tool for forecasters. The behaviour of this 'EDS family' of scores has been examined using extended datasets for precipitation (Nurmi, 2010; North *et al.*, 2013) and weather warnings (Wilson and Giles, 2013). They show promise for differentiating the performance of competing forecast systems for extreme events.

The generalized discrimination score (GDS) was introduced by Mason and Weigel (2009) and further discussed by Weigel and Mason (2011) as a verification framework to measure the discriminative power of a set of forecasts. The GDS has the appealing property of being applicable to most types of forecast and observation data, and is suitable for administrative purposes as well as scientific verification. Formulations of the GDS have been derived for observation data that are binary (e.g., 'precipitation' *vs* 'no precipitation'), categorical (e.g., temperature in lower, middle, or upper tercile), or continuous (e.g., temperature measured in °C); and for forecast data that are binary, categorical, continuous, ensemble distributions, discrete probabilistic (e.g., probability for temperature being in upper tercile) or continuous probabilistic (e.g., continuous probability distribution for temperature in °C). One of the most appealing properties of the GDS is its simple and intuitive interpretation: the score measures the probability that any two (distinguishable) observations can be correctly discriminated by the corresponding forecasts. Thus, the GDS can be interpreted as an indication of how often the forecasts are 'correct', regardless of whether forecasts are binary, categorical, continuous, or probabilistic.

For some data types, the GDS is equivalent or similar to tests and scores that are already widely used in forecast verification and known under different names. For instance, if binary forecasts and observations are considered, the GDS is a transformed version of the true skill statistic, also known as Pierce's skill score (Pierce, 1884). If forecasts and observations are measured on a continuous scale, the GDS is a transformed version of Kendall's ranked correlation co-efficient t (Sheskin, 2007). And if the forecasts are issued as discrete probabilities of binary outcomes, the GDS is equivalent to the trapezoidal area under the relative operating characteristic (ROC) curve and to a transformation of the Mann–Whitney U statistic (Mason and Graham, 2002).

2.4. Verification of 'downstream products'

Meteorological forecasts are increasingly being used as input for other types of forecasts. For example, energy production based on weather dependent renewable energies like wind and solar radiation has grown immensely in the last decade (Foley *et al.*, 2012). Thus the influence of the quality of weather forecasts on energy system operations has extended from the traditional side of the estimate of the consumption of energy to the production of energy. Verification of forecasts has been an integral part from the very beginning. The verification is done on both the meteorological input as well as the final product, i.e., the predicted power generation (Madsen *et al.*, 2005). Foley *et al.* (2012) consider forecasts for single wind turbines *versus* whole farms, using physical *versus* various statistical models, for very short ranges up to seasonal forecasts. Mainly traditional measures are used, e.g., bias and normalized mean square error, sometimes with weighting functions which act to penalize underestimates more than overestimates or to increase penalties for errors when observed values are small (Zhu and Genton, 2012). An important diagnostic approach is the verification of ramping events, which evaluates the temporal changes in wind speed (Pocernich, 2010).

Hydrological prediction is another important application of meteorological forecasts. The value of diagnostic verification has been shown by Welles and Sorooshian (2009) for river stage forecasts, e.g., by isolating the influence of errors in catchment precipitation forecasts. Zimmer and Wernli (2011) describe the extension of an object-based quality measure for precipitation fields in river catchments to also include the important timing errors. Demargne *et al.* (2009) propose a comprehensive river forecast verification service comprising both the meteorological input as well as the hydrological forecasts on multiple space-time scales.

Hazard forecasting for the aviation industry is a third area in which downstream products are generated and require verification. Large organizations such as NOAA and the Met Office produce a suite of aviation hazard forecasts for ceiling and visibility, icing, turbulence, and convection. The forecasts themselves may apply to points (airports), grids, or sectors. Because they are highly relevant to operational decisions made by the airlines, it is important that the verification considers the nature and context of these decisions. An interesting verification approach being developed at NOAA/ESRL/GSD for spatial hazard forecasts involves computing a flow constraint index (FCI), which is based on graph theory and measures the reduction in the potential flow through the air traffic corridor in a hexagonal grid box (Layne and Lack, 2010). Many kinds of hazards can lead to a flow reduction, making this approach attractive for comparing the effectiveness of different forecasting strategies.

The recent development of an objective verification system for aviation turbulence forecasts (Gill, 2013) at the Met Office has facilitated rapid improvements in forecast accuracy. The verification system uses high resolution automated aircraft observations from the British Airways fleet of Boeing 747–400 aircraft to produce an indicator of observed turbulence over segments of each aircraft track. Convectively induced turbulence has been studied and improvements have been suggested using a combination of convective and general turbulence indicators (Gill and Stirling, 2013). Further research has been carried out using the Met Office Global and Regional Ensemble System (MOGREPS) to create probabilistic aviation hazard forecasts for turbulence (Gill and Buchanan, 2013) and cumulonimbus cloud. The verification system has again proved useful in evaluating and optimizing the trial forecasts and in demonstrating the greater skill and value of the probabilistic forecasts.

3. Challenges

3.1. Observations

Since high quality observation datasets are crucial for all successful forecast verification, some of the biggest verification challenges are associated with the quantity and quality of observations. Observation networks are constantly changing, as manual observations are replaced by automatic ones or discontinued, new ground-based and space-based remotely sensed observations become available, and new measurement technologies potentially enable more accurate observations. In many respects these changes in observation capability are positive and improve the quality of our verification, though network changes would be expected to have an impact on the ability to identify trends in forecast accuracy over a long period of time.

One challenge is to make greater use in verification of several exciting new observations which are or soon will be available. The A-train satellite constellation features a variety of imagers, scanners, and active sensors providing near-simultaneous coordinated measurements of cloud and atmospheric structure and composition. These data are available for physical and statistical evaluation of atmospheric models. New geostationary satellites operated by the US, Japan, China, and Korea will soon join Meteosat in providing high spatial and temporal resolution multi-spectral imagery to improve the detection and analysis of clouds, aerosols and surface properties. Closer to the ground, many nations are polarizing their radar network, which should provide more accurate precipitation characterization and quantitative amounts. Phased array and CASA (Collaborative Adaptive Sensing of the Atmosphere) radar networks will provide denser spatial and temporal coverage. Global Positioning System (GPS) tomography is emerging as a new technology for measuring the three-dimensional structure of atmospheric water vapour.

All of these remotely sensed observations measure a signal that is related to, but fundamentally different from, the meteorological variable(s) of interest. Retrieval of those variables from radiance and backscatter measurements involves many assumptions and can result in non-negligible biases and errors. When using remotely sensed data to verify forecasts from models, it may be preferable to convert the model variable into measurement space though the use of an appropriate forward model. This process is closely related to data assimilation, which will be discussed in the next section.

Advances in internet and multi-media communication are enabling many other new types of observations. Although weather observations from ships and planes have been available for some time, GPS and smart phone technology will soon allow road vehicles, pedestrians, and virtually anyone with a smart device to measure and transmit atmospheric observations (Mass, 2012). Third party data will entail a far greater level of quality control than is normally required for data collected by meteorological agencies, and it remains to be seen how best to use these new data types for verification. Nevertheless the rapid uptake of smart phone technology, especially in traditionally data sparse regions such as Africa, offers great promise for enhancement of the surface-based observation network.

Analyses based on single instruments usually contain less information than multi-sensor analyses where the strengths of some instruments can compensate for weaknesses in others. For example, many countries now produce blended radar-rain gauge precipitation analyses. The ease with which data is now exchanged between organizations and across borders will encourage the development of new and better multi-sensor analyses. A good example is the OPERA programme, which supports radar data providers from 30 European nations in their use of agreed-upon best practices for the production and exchange of harmonized, quality-controlled 3D radar data for operational applications and research (Huuskonen *et al*., 2012). A challenge will be to adequately characterize the errors in these analyses, and incorporate that knowledge into the verification process.

It is convenient to verify NWP against gridded analyses. Point measurements in general under-sample the model forecast space, yet surface observations still represent the most important data source for verifying model forecasts from the perspective of the user on the ground. Generating the gridded analyses, however, involves processing the raw observation data, with the result that the verification with respect to the analysis may be less relevant to some users while more relevant to others. For example, forecast users may be more interested in point verification for their location, while modellers can take advantage of the analysis process to identify those components of the difference between observations and model forecasts which are due to unresolvable subgrid scale information represented in the point observations (sometimes called representativeness error). Haiden *et al*. (2012) demonstrated this rather dramatically for precipitation forecasts, showing that nearly 50% of the forecast error at day 1 is due to the comparison of grid box averages to point observations.

A challenge that many verification practitioners in modelling centres are beginning to feel is the shrinking difference between the model error and the observation/analysis error. In many cases it is no longer safe to assume that the forecast error overpowers the observation error. This is especially true when verifying against analyses in data sparse regions (with or without model first guess; see next section), or observations and analyses containing a large component of instrument or retrieval error. Methods are needed to account for, and ideally remove, the impact of observation errors on the verification results. Some small progress has been made in the areas of categorical (Bowler, 2006) and ensemble verification (Saetra *et al*., 2004; Bowler, 2008; Candille and Talagrand, 2008; Santos and Ghelli, 2012), but this is proving a difficult problem to solve more generally. A promising approach may be to treat observations probabilistically, assuming the observation uncertainty is known (e.g., Friederichs *et al*., 2009). For example, one can consider an ensemble of analyses to compensate for the uncertainties and errors introduced through the choice of interpolation method and grid resolution, as well as observation density and errors (Gorgas and Dorninger, 2012a, 2012b).

3.2. Synergy of verification with data assimilation

Some say that data assimilation and verification are the two sides of a coin. For many institutions this is indeed true since the data assimilation system serves as the most efficient and convenient quality control mechanism for meteorological observations for verification. Moreover, the data assimilation process leads to an analysis onto a regular grid, considering all the available data, and this grid is by default spatially matched to the model forecasts to be verified. This analysis is considered to be the 'best' estimate of the atmospheric state for the purpose of initializing the model forecast run.

Despite these advantages, there are some important shortcomings to the use of data assimilation in verification, traceable to the use of a model background (first guess) field. Observations with large deviations from the model background, whether due to observed local effects or to model errors, may be inappropriately rejected from the analysis and consequently from the verification dataset. Moreover, the use of the background field in the data assimilation process nudges the observation field toward the model climatology, especially in data sparse areas. Both of these processes lead to underestimation of the forecast error; this may be particularly problematic for extreme events.

While the use of the model analysis in verification has been considered appropriate for modellers wishing to compare different versions of a single model (recent experiments at the Met Office are leading to this assumption being questioned; see Clayton *et al*., 2012), the problems become very important whenever comparisons are carried out among models from different centres. Park *et al*. (2008) showed that for 850 hPa temperature forecasts the advantage from verifying against one's own analysis can be quite large, overwhelming any real differences in accuracy among competing models. Differences were significant even for forecasts in the 15 day range, long after the models can be considered to have 'forgotten' the initial state. More recent work by Hamill (2011) using a larger dataset from TIGGE shows that the differences among analyses from different global centres are rather large even when averaged over as long a period as a year and even though all centres in principle have access to the same observations.

The WMO standard NWP verification (WMO Manual on the GDPFS, No.485) stipulates the use of 'one's own analysis', which has an element of fairness as well as convenience, but analyses differ greatly as discussed above, so this sacrifices the notion of standardization of the truth dataset. Furthermore, for verification against radiosondes a standard set of stations is specified, but as the radiosonde data is filtered through each centre's model-based quality control, even the standard verification with respect to radiosondes is not truly standard. Efforts are underway to determine the magnitude and impact of the differences among radiosonde datasets used in the WMO standard verification.

Turning to the practical advantages of the use of the data assimilation system and the analysis in verification, one wonders how these could be reconciled with the problems discussed above. One might for example consider verifying with respect to the analysis, but including only those grid points which are adequately supported by observation data. The variational component of data assimilation is useful for blending information from different observation sources to create gridded verification datasets, for example surface-based and satellite observations of clouds. A third possibility is to form some sort of ensemble or consensus analysis from independent analyses, or otherwise estimate the uncertainty in the observed value at specific locations (e.g., Gorgas and Dorninger, 2012b).

3.3. Verification of 'seamless' predictions

One of the key scientific drivers for model development is the desirability of running the same model for all time scales, from short-range NWP to decadal prediction. This is known as 'unified modelling', and the predictions from such a system are often called 'seamless'. Pragmatically, this introduces the problem of how to maintain and improve short-range skill while meeting the requirements for skill at monthly, seasonal ranges and beyond. The question then arises as to how we are to assess the relative skill at these disparate time scales in a consistent way when the questions that need to be answered are so different? Perhaps different metrics may be used but the effect of any optimization of model formulation developed with verification information from one application must be checked across the seamless system for other applications.

It is as yet unclear whether will it be possible to assess exactly the same parameters, or the same parameters in the same way, but one of the key attributes is to determine how far in advance a certain event can be predicted reliably. In the short range, the impact of less predictable scales can to some degree be reduced through spatial and temporal averaging and verification of deterministic forecasts is a valid approach.

However, since small errors in a model's initial state have a tendency to propagate to larger scales and to act as quickly growing noise terms (Lorenz, 1993), forecasts ranging more than a couple of days into the future are intrinsically probabilistic. Consequently, seamless predictions that are designed to cover a range of temporal scales need to consider the evolution of the entire uncertainty range of possible initial states, rather than integrating a single trajectory from a best guess of the initial state. As longer time scales of several weeks to seasons are considered, prediction uncertainty is increasingly dominated by uncertainties in relevant boundary conditions, such as sea surface temperatures, and model feedback processes such as cloud-radiation interactions. By including these uncertainties in the ensemble generation and modelling process, the concept of ensemble forecasting has been extended to climate time scales, allowing for example the computation of seasonal ensemble forecasts (e.g., Stockdale *et al*., 1998).

The design of verification approaches that apply across different time ranges (e.g., in the context of seamless predictions) poses a difficult problem. A verification procedure should always be chosen according to the specific question being asked. However, the nature of forecasts changes as we move from short to long time scales, and so do the questions being asked by forecast users. For example, while forecast users usually expect sharp and accurate statements on whether or not it will rain tomorrow, they may be happy with somewhat fuzzier statements as far as the tendency of the following 4 weeks is concerned. That is, short-range forecasts may be interpreted deterministically, but medium and extended-range forecasts are inherently probabilistic. While the magnitude of systematic forecast bias is an essential performance characteristic of short-range forecasts, it is often less relevant for calibrated monthly to seasonal forecasts since the latter are typically issued as anomalies with respect to model climatology. While short-range forecasts are often linked to high levels of predictability, the signal to noise ratio usually decreases as we move to longer time scales, requiring spatial and temporal aggregation of the forecasts as discussed earlier. Finally, while a sufficient amount of verification data is, at least in principle, available for the validation of short-term weather forecasts, sample sizes get smaller as we move to long lead-times of several months or even years. The situation gets even more complicated for the recently evolving field of decadal forecasting, where sample sizes are so small that a classical verification may not be feasible at the present time (Mason, 2011; Gangsto *et al*., 2013).

Thus, if one wants to design a verification approach that is meaningful and applicable across many scales, it would need to be applicable both to deterministic and probabilistic forecasts, it would need to have good signal detection properties even if predictability is low and sample sizes are small, and it would need to penalize model bias only on the short time scales where they are relevant (since long-range forecasts are typically issued as anomalies). Even if it is possible to find a validation metric that simultaneously satisfies all these requirements, one must be aware that this metric would only measure the specific forecast attributes for which it is designed, and that it would probably fail to capture all aspects of accuracy that are relevant for forecast users. For example, the recently developed generalized discrimination score (GDS) is both applicable to deterministic forecasts (Mason and Weigel, 2009) and probabilistic and ensemble forecasts (Weigel and Mason, 2011), thus satisfying a key requirement for applicability across several time scales. However, the GDS only measures the skill attribute of discrimination. The GDS is neither sensitive to reliability (a key attribute of probabilistic long-range forecasts), nor to systematic forecast bias (a key attribute of short-range forecasts). Since different forecast attributes are relevant at different time scales, each time scale requires its own 'blend' of additional skill scores if one wants to give due consideration of all relevant attributes.

A key aspect for long-lead predictions is the impact of feedbacks. To investigate whether the physical processes are being correctly simulated, coupled climate models can be run in NWP mode and verified using standard and diagnostic approaches. The WMO Working Group on Numerical Experimentation's Transpose Atmospheric Model Intercomparison Project (Transpose-AMIP) exposes parameterization errors through this type of seamless model evaluation (Williamson *et al*., 2009). Biased behaviour is usually evident early in the simulation, enabling changes to the model parameterizations to be tested without the need for long model runs. A challenge is to introduce a multivariate, possibly multi-dimensional framework whereby the complex interactions, e.g., between temperature and cloud, can be assessed together. This is discussed next.

3.4. Multi-dimensional verification

For many applications multi-dimensional forecasts are needed, that is joint forecasts of several variables at one or more locations. This then obviously also requires information on the collective reliability of such multi-dimensional forecasts. It is thereby not sufficient to evaluate the individual forecast dimensions separately from each other. For instance, a model may on average yield good temperature and precipitation forecasts, but the simultaneous forecasts of these two variables may be inconsistent and unrealistic.

One approach might be to construct a derived variable that incorporates the variables of interest (e.g., the wind chill takes both temperature and wind speed into account). However, sensible combinations are not always possible, nor do they generally take into account the complexities of the variables' joint distributions. To reduce the high dimensionality of geophysical data the climate community frequently uses principal component analysis to derive spatial patterns or modes of coherent variability, commonly known as empirical orthogonal functions (EOFs). The strength of each mode can then be verified. Beyond the leading mode, however, it can become difficult to assign a physical interpretation to the lesser modes, as they must be strictly orthogonal in space and time whereas real processes are often interrelated to some degree. The lesser modes with their more detailed physical structure are also increasingly sensitive to any errors in the underpinning analysis.

If enough verification samples are available, conditional verification can yield more detailed information on the performance for forecast-observation pairs stratified according to specific criteria of interest, such as specific temperature ranges, weather patterns, or flow regimes. For ensemble predictions the concept of conditional exceedance probabilities (CEPs) represents a generalization of the classical rank histogram and characterizes the dispersion characteristics of an ensemble prediction system dependent on the observed outcome values (e.g., as a function of observed temperatures) (Mason *et al*., 2007).

Three approaches of assessing multi-dimensional ensemble reliability have been suggested: (1) minimum spanning tree histograms, (2) multivariate rank histograms, and (3) bounding boxes (a more detailed discussion of these three methods is provided in Weigel, 2011). The minimum spanning tree (MST) histogram as a tool for assessing multi-dimensional reliability was proposed by Smith (2001) and then further discussed by Smith and Hansen (2004), Wilks (2004) and Gombos *et al*. (2007). In essence, MST histograms assess how the mutual 'closeness' (in the multi-dimensional prediction space) of the ensemble members relates to their 'distance' from the observations. Another approach to validate multi-dimensional reliability is given by multivariate rank (MVR) histograms (Gneiting *et al*., 2008), which assess how the observations *rank* with respect to the individual ensemble members. MVR histograms therefore represent a generalization of the widely used scalar rank histogram.

A less stringent criterion to validate the reliability of ensemble forecasts is applied in the 'bounding box' (BB) approach, which has been suggested and discussed by Weisheimer *et al*. (2005) and Judd *et al*. (2007). In this approach, an observed outcome is considered to be consistent with the corresponding ensemble forecast if it falls into the BB spanned by the ensemble forecast. BBs have the additional advantage of being easy to compute for any number of dimensions, making them particularly useful for assessing high-dimensional ensemble forecasts. Moreover, due to their intuitive interpretation, they are easy to communicate to non-experts. However, there is the obvious disadvantage that capture rates can always be artificially enhanced by simply inflating the ensemble forecasts to unrealistically large spread values. In other words, overdispersion is not penalized. Moreover, being defined by the minimum and maximum values of an ensemble, BBs are unduly affected by outliers and may fail in characterizing the bulk ensemble properties appropriately. Thus, it is not recommended to use BBs instead of other verification metrics, but rather in complement to them.

3.5. Diagnostic verification

Diagnostic verification should give information about the nature of forecast errors, along with clues as to the sources of the errors. Ideally it should also be conceptually intuitive to ensure that it is understood by the widest possible audience. Developing diagnostic verification approaches that meet all of these criteria is not easy. Activities that assess and intercompare methods, such as the ICP mentioned in Section 2.2, can help with this process.

Most of the newer diagnostic verification methods have been developed and applied to the verification of model precipitation against radar rain fields. Wind also causes large amounts of damage but is verified less frequently, due both to the availability of fewer observations and the challenges of verifying vector quantities. Koh *et al.* (2012) recently proposed a diagnostic suite for assessing NWP performance that is applicable to both scalar and vector variables. It decomposes the model errors into bias, phase and amplitude errors and introduces three diagrams to visualize the results: the error decomposition diagram, the correlation-similarity diagram (similar to the Taylor diagram), and the anisotropy diagram.

Many forecasts contain a large component of timing error, and this aspect of verification clearly needs greater attention. Spatial verification approaches are starting to be extended into the temporal domain. Neighbourhood verification easily accommodates closeness in time (e.g., Weusthoff, 2011), while object-based methods like MODE can be used to investigate spatiotemporal errors of coherent weather features (Bullock, 2011). Zimmer and Wernli (2011) added a sliding time shift to the structure-amplitude-location (SAL) method to estimate precipitation timing errors.

The spatial verification approaches also lend themselves well to the evaluation of ensembles, and we are starting to see some innovative work in this area. Gallus (2010) applied the MODE and CRA object-based methods to individual members of ensemble precipitation forecasts and verified the spread of the object attributes (size, mean rainfall, etc.). Zacharov and Rezacova (2009) investigated ensemble spread-skill using the fractions skill score (FSS). In an interesting combination of deterministic-probabilistic thinking, Duc *et al.* (2011) extended the spatial FSS to include two new dimensions, namely time and model uncertainty (as represented by the ensemble members). This enabled them to verify individual forecasts from high resolution ensembles, which is something that has generally been frowned upon. (Note that neighbourhood approaches are gaining popularity for post-processing high resolution ensembles (e.g., Schwartz *et al.*, 2010; Ben Bouallègue, 2011).) Ebert (2011) proposed quantifying location-based uncertainty errors using a 'radius of reliability' and applied this concept to ensemble predictions of heavy rainfall in tropical cyclones.

Diagnostic approaches are needed for verifying new and difficult types of forecasts. Examples include forecasts of tropical cyclone genesis (likelihood? where? when?), graphical weather warnings (likelihood? how extensive? how strong? onset and cessation?), and forecasts for weather in complex topography where observations present a challenge.

3.6. User-oriented verification

Verification should benefit the general public and other forecast end users by improving their confidence in the forecast, by allowing them to identify situations in which the forecast is more or less reliable, and by identifying the extent to which the forecast is useful as a basis for decision-making in weather-sensitive activities. Psychology plays a role in understanding how verification information is perceived and understood. For example, verification may actually increase confidence in general public end-users when the forecast also includes an uncertainty estimate. Verification stresses the fact, of which most users are already aware (Joslyn and Savelli, 2010), that deterministic forecasts rarely verify exactly. A forecast that also includes a reliable uncertainty estimate will imply that the forecaster expects a range of possible outcomes. Combined with verification, such a forecast may seem more plausible as well as more reliable than deterministic forecasts.

But will enough people be able to understand the verification? Despite the fact that this information is both abstract and complex, there is some preliminary evidence suggesting that non-experts can understand simple verification graphics for both deterministic and probabilistic forecasts (Joslyn *et al.*, 2013). With no explicit training, using only a simple key, college undergraduates were readily able to identify forecasts that performed better over various forecast periods. Furthermore they understood, although the principle was never explained to them, that an 80% predictive interval was reliable when observations that differed from the single value forecast were within the interval the majority of the time. In this example the forecast graphic, a bracket encompassing the 80% predictive interval for temperature, was repeated in the verification graphic, clarifying the connection between the two. This likely facilitated understanding of the verification graphics. Thus, every day users may be able to understand even complex forecast information and graphics, if care is taken in how they are presented. Standard procedures and presentation formats for verification would better allow users to compare between weather forecast providers.

For users of forecasts, forecast quality can mean more than just accuracy. Nurmi *et al.* (2013) noted that accuracy is the first phase in a weather service value chain that includes provision of appropriate data to the decision maker, timely access, understanding of the information, ability to adapt his/her behaviour, responses that actually mitigate damage, and transfer of benefits to other economic agents. If information is available for each phase, then this weather service value chain can provide managerial guidance in a quantitative analytical framework. Ambühl (2010) shows how warning thresholds can be optimized for individual users, and how the warning performance can be understood by users in terms of their efficiency in risk mitigation.

Verification scores used as performance indicators and targets can encourage action to improve forecasting. Such metrics may look quite different from the diagnostic and model-oriented verification discussed throughout most of this paper, and the challenge is to come up with metrics that truly reflect the desired improvements. ECMWF and FMI now use the SEEPS score as a headline measure (among others) to monitor forecast performance over time. The Met Office computes and tracks global and UK 'NWP indices' that are based on combined model performance for a set of key meteorological variables (Met Office, 2010). Louey (2011) demonstrated the use of a balanced scorecard, which is a strategic performance management tool, to monitor official forecast performance in the Australian Bureau of Meteorology.

For corporate clients in the aviation industry the Met Office produces a suite of verification metrics (Gill, 2011). For probabilistic Terminal Aerodrome Forecasts (TAFs), a Service Quality Index is computed that measures the reliability of the forecasts at important thresholds for cloudbase height and visibility. The Flight Time Error (FTE) is an alternative accuracy measure for upper-air wind that is currently being trialled. It is the difference between the observed flight time and the forecast flight time calculated using the track that the aircraft actually took and then recalculating the time to fly each segment, replacing the actual winds with the forecast winds (Rickard *et al*., 2001). FTE is directly relevant to aircraft operations in estimating arrival times and in determining the amount of contingency fuel that is needed. The participation of British Airways in providing meteorological and other data to the Met Office has helped in the design, development and testing of these user-oriented metrics.

Verification systems may need to evolve as users 'self-educate' with experience over time. As they gain confidence with using the basic metrics, it may be desirable to introduce them to more sophisticated diagnostic verification measures that provide more detailed information on forecast performance.

4. Outlook

Following the 3rd International Verification Methods Workshop in 2007 Casati *et al*. (2008) reviewed the (then) current status and future directions for forecast verification. Since that time there has been progress in many areas, including confidence information with verification results; the transition of diagnostic spatial verification into mainstream use; the development of new scores for monitoring forecast performance, verifying forecasts for extreme rare events, and measuring the discrimination power of forecasts of different types in a consistent manner; and the increasing activity to verify downstream products derived from meteorological forecasts.

Recognition of the importance of forecast verification is increasing worldwide. Not only has verification been receiving greater attention in the national meteorological services of developed countries, motivated for example by the drive towards ISO (International Organization for Standardization) certification, but also in developing countries and by focusing attention on remote regions on the globe such as the polar areas.

The WMO-led Severe Weather Forecast Demonstration Project (SWFDP; WMO, 2008) has been the main driver for the spread of verification activity to the weather services of countries of Southern and Eastern Africa, the South Pacific islands, and southeast Asia. The SWFDP involves the real time transfer of mostly model-based forecast products from global and regional modelling centres to the national meteorological services of the participating nations, where they are used as guidance to prepare warnings of severe weather. As part of the project, meteorologists from the participating countries are trained in verification methods appropriate to weather warnings. As a result many have been using these methods to verify their own forecasts. Verification of the model output products from the global centres has been slower to develop, but efforts are underway to accomplish this to help the forecasters decide which of the many available products are most reliable and useful. Assessment of model forecasts for different meteorological conditions, especially severe weather conditions, would be extremely useful in this context.

There is a growing interest towards improved weather services in polar regions of the globe because of increased economic and transport activities there, and also due to climate change issues. The WMO World Climate and World Weather Research Programmes have recently initiated a 10 year Polar Prediction Project (PPP) with its main mission to promote cooperative international research to develop prediction services for the Polar Regions on temporal scales from hourly to seasonal (Jung, 2012). Major research efforts during PPP will be dedicated to evaluating polar predictability, diagnosing various forecast systems and taking into account user needs. One of the eight designated key research goals is to establish and apply appropriate verification methods. Establishment of optimal, high-quality observational networks, access to relevant reference data sets, and verification of high-impact weather and climate events in the Polar Regions are some of the expected major challenges of PPP.

Many important challenges lie ahead for conducting more effective forecast verification:

- The quantity and quality of observations must be maintained and improved, and efforts continued to harmonize them across national borders and across disciplines and make them easily accessible for use in verification.
- The synergy of verification and data assimilation must be treated carefully. Monitoring efforts at NWP centres must continue in order to minimize the rejection of good observations and reduce the influence of the model's climatology on the analysis and thus on the verification results.
- The trend toward seamless prediction across spatial and temporal scales calls for efforts to develop verification approaches that can also be used in a seamless fashion. This includes the emerging field of decadal predictions, a time scale that poses many new and yet unresolved questions with respect to their appropriate verification.
- The development of intuitive methods to verify forecasts of joint distributions of multiple variables should be encouraged.
- As ensemble prediction becomes increasingly widespread and diversifies into new applications, new diagnostic approaches for characterizing their errors will be required.
- Easy-to-understand user-focused verification, some aimed at evaluating user impacts, will also be needed to convey the quality of forecasts to non-specialists.

The meteorological community has led the science of forecast verification for many decades, with a strong and growing body of verification experts and practitioners. Valuable ideas have been brought into meteorological forecast verification from other disciplines such as medicine, psychology, economics and engineering (the relative operating characteristic, ROC, is an example). As the global scientific endeavour becomes increasingly more integrated and interdisciplinary, it will be interesting to see how the need for new approaches, and the perspectives brought together from different fields, will lead to further advances in the science of forecast verification.

Acknowledgements

The authors would like to thank Robert Fawcett, Peter Steinle, and two anonymous reviewers for helpful feedback that helped improve the quality of the manuscript.

References

Ambühl J. 2010. Customer oriented warning systems. *Veröffentlichung MeteoSchweiz* **84**: 93 pp.

Ben Bouallègue Z. 2011. Upscaled and Fuzzy Probabilistic Forecasts: Verification Results. *COSMO Newsletter*, No. 11. http://cosmo-model.cscs.ch/content/model/documentation/newsLetters/newsLetter11/4_bouallengue.pdf (accessed 1 April 2013).

Bowler NE. 2006. Explicitly accounting for observation error in categorical verification of forecasts. *Mon. Weather Rev.* **134**: 1600–1606.

Bowler NE. 2008. Accounting for the effect of observation errors on verification of MOGREPS. *Meteorol. Appl.* **15**: 199–205.

Bullock R. 2011. Development and implementation of MODE time domain object-based verification. In *24th Conference Weather and Forecasting, 24–27 January 2011, Seattle, Washington*. American Meteorological Society: Boston, MA.

Candille G, Talagrand O. 2008. Impact of observational error on the validation of ensemble prediction systems. *Q. J. R. Meteorol. Soc.* **134**: 959–971.

Casati B, Wilson LJ, Stephenson DB, Ghelli A, Pocernich M, Damrath U, Ebert EE, Brown BG, Mason S. 2008. Forecast verification: current status and future directions. *Meteorol. Appl.* **15**: 3–18.

Clark WL, Yuan H, Jensen TL, Wick G, Tollerud EI, Bullock RG, Sukovich E. 2011. Evaluation of GFS water vapor forecast errors during the 2009–2010 West Coast cool season using the MET/MODE object analyses package. In *25th Conference Hydrology, 24–27 January 2011, Seattle, Washington*. American Meteorological Society: Boston, MA.

Clayton AM, Lorenc AC, Barker DM. 2012. Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Q. J. R. Meteorol. Soc.* DOI: 10.1002/qj.2054.

Davis C, Brown B, Bullock R. 2006. Object-based verification of precipitation forecasts. Part I: methods and application to mesoscale rain areas. *Mon. Weather Rev.* **134**: 1772–1784.

Demargne J, Mullusky M, Werner K, Adams T, Lindsey S, Schwein N, Marosi W, Welles E. 2009. Application of forecast verification science to operational river forecasting in the U.S. National Weather Service. *Bull. Am. Meteorol. Soc.* **90**: 779–784.

Duc L, Saito K, Seko H. 2011. Application of spatial-temporal fractions skill score to high-resolution ensemble forecast verification. In*5th International Verification Methods Workshop*, 1–7 December 2011, Melbourne, Australia. http://cawcr.gov.au/events/verif2011/posters/48_Duc_L.pdf (accessed 1 April 2013).

Ebert E. 2011. Radius of reliability: a distance metric for interpreting and verifying spatial probabilistic warnings. *CAWCR Res. Lett.* **6**: 4–10.

Ebert EE, McBride JL. 2000. Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrol.* **239**: 179–202.

Ferro CAT, Stephenson DB. 2011. Extremal dependence indices: improved verification measures for deterministic forecasts of rare binary events. *Weather Forecast.* **26**: 699–713.

Foley AM, Leahy PG, Marvuglia A, McKeogh EJ. 2012. Current methods and advances in forecasting wind power generation. *Renew. Energy* **37**: 1–8.

Fowler TL, Jensen T, Tollerud EI, Halley Gotway J, Oldenburg P, Bullock R. 2010. New Model Evaluation Tools (MET) software capabilities for QPF verification. *Preprints, 3rd International Conference on QPE, QPF*, 18–22 October 2010, Nanjing, China.

Friederichs P, Göber M, Bentzien S, Lenz A, Krampitz R. 2009. A probabilistic analysis of wind gusts using extreme value statistics. *Meteorol. Z.* **18**: 615–629.

Gallus WA Jr. 2010. Application of object-based verification techniques to ensemble precipitation forecasts. *Weather Forecast.* **25**: 144–158.

Gangsto R, Weigel AP, Liniger MA, Appenzeller C. 2013. Meteorological aspects of the evaluation of decadal predictions. *Clim. Res.* **55**: 181–200.

Ghelli A, Primo C. 2009. On the use of the extreme dependency score to investigate the performance of an NWP model for rare events. *Meteorol. Appl.* **16**: 537–544.

Gill PG. 2011. Aviation verification – Recent developments at the UK Met Office. *5th International Verification Methods Workshop*, 1–7 December 2011, Melbourne, Australia. http://cawcr.gov.au/events/verif2011/ppt/P_Gill.pdf (accessed 1 April 2013).

Gill PG. 2013. Objective verification of World Area Forecast Centre clear air turbulence forecasts. *Meteorol. Appl.* DOI: 10.1002/met.1288.

Gill PG, Buchanan P. 2013. An ensemble based turbulence forecasting system. *Meteorol. Appl.* DOI: 10.1002/met.1373.

Gill PG, Stirling AJ. 2013. Including convection in global turbulence forecasts. *Meteorol. Appl.*, **20**: 107–114. DOI: 10.1002/met.1315.

Gilleland E, Ahijevych D, Brown BG, Casati B, Ebert EE. 2009. Intercomparison of spatial forecast verification methods. *Weather Forecast.* **24**: 1416–1430.

Gneiting T, Stanberry LI, Grimit EP, Held L, Johnson NA. 2008. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test* **17**: 211–235.

Gombos D, Hansen JA, Du J, McQueen J. 2007. Theory and applications of the minimum spanning tree rank histogram. *Mon. Weather Rev.* **135**: 1490–1505.

Gorgas T, Dorninger M. 2012a. Quantifying verification uncertainty by reference data variation. *Meteorol. Z.* **21**: 259–277.

Gorgas T, Dorninger M. 2012b. Concepts for a pattern-oriented analysis ensemble based on observational uncertainties. *Q. J. R. Meteorol. Soc.* **138**: 769–784.

Haiden T, Rodwell MJ, Richardson DS, Okagaki A, Robinson T, Hewson T. 2012. Intercomparison of global model precipitation forecast skill in 2010/11 using the SEEPS score. *Mon. Weather Rev.* **140**: 2720–2733.

Hamill T. 2011. Examining characteristics of analysis and short-range forecast errors using TIGGE. *Presentation to NCEP/EMC/GMB*, December, 2011, http://www.esrl.noaa.gov/psd/people/tom.hamill/analysis-errors-wgne-hamill.pdf (accessed 1 April 2013).

Hogan R, O'Connor EJ, Illingworth AJ. 2009. Verification of cloud-fraction forecasts. *Q. J. R. Meteorol. Soc.* **135**: 1494–1511.

Huuskonen A, Delobbe L, Urban B. 2012. EUMETNET OPERA: achievements of OPERA-3 and challenges ahead. *Proceedings of ERAD 2012*, 25–29 June 2012, Toulouse, France.

Jolliffe IT. 2007. Uncertainty and inference for verification measures. *Weather Forecast.* **22**: 637–650.

Jolliffe IT, Stephenson DB. 2012. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd edn. Wiley and Sons Ltd.: Chichester, UK; 274pp.

Joslyn S, Nemec L, Savelli S. 2013. The benefits and challenges of predictive interval forecasts and verification graphics for end-users. *Weather, Clim. Soc.* (in press).

Joslyn S, Savelli S. 2010. Communicating forecast uncertainty: public perception of weather forecast uncertainty. *Meteorol. Appl.* **17**: 180–195.

Judd K, Smith LA, Weisheimer A. 2007. How good is an ensemble at capturing truth? Using bounding boxes for forecast evaluation. *Q. J. R. Meteorol. Soc.* **133**: 1309–1325.

Jung T. 2012. *The WWRP Polar Prediction Project.* Planning Workshop of the WCRP Polar Prediction Initiative, 12–14 April 2012, Toronto, Canada.

JWGFVR. 2013. Forecast verification: issues, methods and FAQ. http://www.cawcr.gov.au/projects/verification (accessed 28 January 2013).

Koh T-Y, Wang S, Bhatt BC. 2012. A diagnostic suite to assess NWP performance. *J. Geophys. Res.* **117**: D13109, DOI: 10.1029/2011JD017103.

Kucera PA, Weeks C, Brown B, Bullock R. 2011. Development of new diagnostic tools to evaluate NWP cloud and precipitation products using A-Train satellite observations. *5th International Verification Methods Workshop*, 1–7 December 2011, Melbourne, Australia. http://cawcr.gov.au/events/verif2011/ppt/P_Kucera.pdf (accessed 1 April 2013).

Layne GJ, Lack SA. 2010. Methods for estimating air traffic capacity reductions due to convective weather for verification. *14th Conference on Aviation, Range, and Aerospace Meteorology (ARAM)*, Atlanta, Georgia.

Lorenz EN. 1993. *The Essence of Chaos*. University of Washington Press: Seattle, WA; 226pp.

Louey J. 2011. Verification presentation in management reporting. *5th International Verification Methods Workshop*, 1–7 December 2011, Melbourne, Australia. http://cawcr.gov.au/events/verif2011/ppt/J_Louey.pdf (accessed 1 April 2013).

Madsen H, Pinson P, Kariniotakis G, Nielsen HA, Nielsen TS. 2005. Standardizing the performance evaluation of short term wind power prediction models. *Wind Eng.* **29**: 475–489.

Mason SJ. 2011. Seasonal and longer-range forecasts. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd edn, Jolliffe IT, Stephenson DB (eds). John Wiley & Sons Ltd.: Chichester, UK; 203–220.

Mason SJ, Galpin JS, Goddard L, Graham NE, Rajartnam B. 2007. Conditional exceedance probabilities. *Mon. Weather Rev.* **135**: 363–372.

Mason SJ, Graham NE. 2002. Areas beneath the relative operating characteristics (ROC) and levels (ROL) curves: statistical significance and interpretation. *Q. J. R. Meteorol. Soc.* **128**: 2145–2166.

Mason SJ, Weigel AP. 2009. A generic forecast verification framework for administrative purposes. *Mon. Weather Rev.* **137**: 331–349.

Mass C. 2012. Nowcasting: the promise of new technologies of communication, modelling, and observation. *Bull. Am. Meteorol. Soc.* **93**: 797–809.

Met Office. 2010. Global NWP index documentation. http://www.wmo.int/pages/prog/www/DPFS/Meetings/CG-FV_Montreal2011/Doc4-6-Annex_Global_index_documentation_2010-2.pdf (accessed 1 April 2013).

Mittermaier MP, Bullock R. 2013. Using MODE to explore the spatial and temporal characteristics of cloud cover forecasts from high-resolution NWP models. *Meteorol. Appl.* **20**: 187–196, DOI: 10.1002/met.1393 (this issue).

Mittermaier M, Roberts N. 2010. Intercomparison of spatial forecast verification methods: identifying skillful spatial scales using the fractions skill score. *Weather Forecast.* **25**: 343–354.

Mittermaier MP, Roberts N, Thompson SA. 2011. A long-term assessment of precipitation forecast skill using the Fractions Skill Score. *Meteorol. Appl.* **20**: 176–186, DOI: 10.1002/met.296 (this issue).

Moise AF, Delage FP. 2011. New climate model metrics based on object-orientated pattern matching of rainfall. *J. Geophys. Res.* **116**: D12108, DOI: 10.1029/2010JD015318.

North R, Trueman M, Mittermaier M, Rodwell M. 2013. An assessment of the SEEPS and SEDI metrics for the verification of 6h forecast precipitation accumulations. *Meteorol. Appl.* **20**: 164–175, DOI: 10.1002/met.1405 (this issue).

Nurmi P. 2010. Experimentation with new verification measures for categorized QPFs in the verification of high impact precipitation events. *Proceedings 3rd WMO International Conference Quantitative Precipitation Estimation and Quantitative Precipitation Forecasting and Hydrology*, 18–22 October 2010, Nanjing, China; 222–226.

Nurmi P. 2011. Operational verification systems at FMI. *5th International Verification Methods Workshop*. http://cawcr.gov.au/events/verif2011/ppt1/Nurmi_P.pdf (accessed 1 April 2013).

Nurmi P, Perrels A, Nurmi V. 2013. Expected impacts and value of improvements in weather forecasting on the road transport sector. *Meteorol. Appl*. **20**: 217–223, DOI: 10.1002/met.1399 (this issue).

Park YY, Buizza R, Leutbecher M. 2008. TIGGE: preliminary results on comparing and combining ensembles. *Q. J. R. Meteorol. Soc.* **134**: 2029–2050.

Pierce CS. 1884. The numerical measure of success in predictions. *Science* **4**: 453–454.

Pocernich M. 2010. Verification of wind forecasts of ramping events. http://ral.ucar.edu/projects/wind_energy_workshop/presentations/Ramp_Verification_Methods_Pocernich_28.pdf (accessed 1 April 2013).

Potts JM, Folland CK, Jolliffe IT, Sexton D. 1996. Revised "LEPS" scores for assessing climate model simulations and long-range forecasts. *J. Climate* **9**: 34–53.

Primo C, Ghelli A. 2009. The effect of the base rate on the extreme dependency score. *Meteorol. Appl.* **16**: 533–535.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: ViennaISBN 3-900051-07-0, http://www.R-project.org (accessed 1 April 2013).

Rickard GJ, Lunnon RW, Tenenbaum J. 2001. The Met Office upper air winds: prediction and verification in the context of commercial aviation data. *Meteorol. Appl.* **8**: 351–360.

Rikus L, Elliott T, Dietachmayer G. 2011. Using image mapping techniques to generate model evaluation metrics. *5th International Verification Methods Workshop*, 1–7 December 2011, Melbourne, Australia. http://cawcr.gov.au/events/verif2011/posters/46_Rikus_L.pdf (accessed 1 April 2013).

Roberts RD, Anderson ARS, Nelson E, Brown BG, Wilson JW, Pocernich M, Saxen T. 2012. Impacts of forecaster involvement on convective storm initiation and evolution nowcasting. *Weather Forecast.* **27**: 1061–1089.

Roberts NM, Lean HW. 2008. Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Weather Rev.* **136**: 78–97.

Rodwell M, Haiden T, Richardson D. 2011. Developments in precipitation verification. *ECMWF Newsl.* **128**: 12–16.

Rodwell MJ, Richardson DS, Hewson TD, Haiden T. 2010. A new equitable score suitable for verifying precipitation in numerical weather prediction. *Q. J. R. Meteorol. Soc.* **136**: 1344–1363.

Roebber PJ. 2009. Visualizing multiple measures of forecast quality. *Weather Forecast.* **24**: 601–608.

Saetra O, Hersbach H, Bidlot J-R, Richardson D. 2004. Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Weather Rev.* **132**: 1487–1501.

Santos C, Ghelli A. 2012. Observational probability method to assess ensemble precipitation forecasts. *Q. J. R. Meteorol. Soc.* **138**: 209–221.

Schwartz CS, Kain JS, Weiss SJ, Xue M, Bright DR, Kong F, Thomas KW, Levit JJ, Coniglio MC, Wandishin MS. 2010. Toward improved convection-allowing ensembles: model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Weather Forecast.* **25**: 263–280.

Sheskin DJ. 2007. *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman and Hall/CRC: Boca Raton, FL; 1776pp.

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

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

Stephenson DB, Casati B, Ferro CAT, Wilson CA. 2008. The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteorol. Appl.* **15**: 41–50.

Stockdale TN, Anderson DLT, Alves JOS, Balmaseda MA. 1998. Global seasonal rainfall forecasts using a coupled ocean–atmosphere model. *Nature* **392**: 370–373.

Weigel AP. 2011. Verification of ensemble forecasts. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd edn, Jolliffe IT, Stephenson DB (eds). John Wiley & Sons Ltd.: Chichester, UK; 141–166.

Weigel AP, Mason SJ. 2011. The generalized discrimination score for ensemble forecasts. *Mon. Weather Rev.* **139**: 3069–3074.

Weisheimer A, Smith LA, Judd K. 2005. A new view of seasonal forecast skill: bounding boxes from the DEMETER ensemble forecasts. *Tellus* **57A**: 265–279.

Welles E, Sorooshian S. 2009. Scientific verification of deterministic river stage forecasts. *J. Hydrometeorol.* **10**: 507–520.

Weusthoff T. 2011. Neighbourhood verification in space and time. *5th International Verification Methods Workshop*, 1–7 December 2011, Melbourne, Australia. http://cawcr.gov.au/events/verif2011/posters/49_Weusthoff_T.pdf (accessed 1 April 2013).

Weusthoff T, Ament F, Arpagaus A, Rotach MW. 2010. Assessing the benefits of convection-permitting models by neighborhood verification: examples from MAP D-PHASE. *Mon. Weather Rev.* **138**: 3418–3433.

Wilks DS. 2004. The minimum spanning tree histogram as a verification tool for multidimensional ensemble forecasts. *Mon. Weather Rev.* **132**: 1329–1340.

Wilks DS. 2005. *Statistical Methods in the Atmospheric Sciences*, 2nd edn. Elsevier: Amsterdam; 627pp.

Williamson D, Nakagawa M, Klein S, Earnshaw P, Nunes A, Roads J. 2009. Transpose AMIP: a process oriented climate model evaluation and intercomparison using model weather forecasts and field campaign observations. *EGU General Assembly*, 19–24 April 2009, Vienna, Austria.

Wilson LJ, Giles A. 2013. A new index for the verification of accuracy and timeliness of weather warnings. *Meteorol. Appl.* **20**: 206–216, DOI: 10.1002/met.1404 (this issue).

WMO. 2008. *Severe Weather Forecasting Demonstration Project (SWFDP) Guidebook on Planning Regional Subprojects*. World Meteorological Organization Commission for Basic Systems. World Meteorological Organization: Geneva, Switzerland; 60pp.

Zacharov P, Rezacova D. 2009. Using the fractions skill score to assess the relationship between an ensemble QPF spread and skill. *Atmos. Res.* **94**: 684–693.

Zhu X, Genton MG. 2012. Short-term wind speed forecasting for power system operations. *Int. Stat. Rev.* **80**: 2–23.

Zimmer M, Wernli H. 2011. Verification of quantitative precipitation forecasts on short time-scales: a fuzzy approach to handle timing errors with SAL. *Meteorol. Z.* **20**: 95–105.