The WGNE survey of verification methods for numerical prediction of weather elements and severe weather events

Philippe Bougeault
Météo-France, Toulouse

January 2003

1. Introduction

In its 13th session, the WMO/Commission for Atmospheric Science tasked WGNE to prepare a position paper on high-resolution model verifications, oriented towards weather elements and severe weather events (item 5.3.10 of the abridged final report, document WMO-N°941). This recognizes the specific difficulty of traditional verification methods in providing a useful measure of model performance at high resolution and for intense events. First, the verification of mesoscale events is limited by the insufficient density and quality of the observing networks. Second, the related weather elements may be on the edge of predictability, or entirely stochastic from the perspective of current NWP models. As such, the traditional verification methods based on instantaneous comparison of analyzed and predicted fields may not yield useful information, and new methods are needed. Third, there is a great expectation that mesoscale models will deliver products of direct relevance to end- users, and consequently much work is done on the development of user-oriented verifications, but the needs are not the same for user-oriented and developers-oriented verifications.

The verification of numerical models against observations has several purposes. For instance: (i) provide a measure of the progress of the forecast skill over the years; (ii) compare the merits of two versions of a forecasting system in order to decide which is the best for operations; (iii) understand where the problems are and what aspects of the system need refinements; (iv) compare the relative value of two different systems for a specific category of users. No single verification system can be optimal for all of these tasks and there is a need to issue guidance on what methods are good for what purpose. The purpose of the present paper is to report on a survey of methods currently in use or under development in many operational NWP centers, and to provide guidance on desirable features for verification methods, based on shared experience.

The organization of the paper is as follows: Section 2 is a list of the available sources and recent discussions. Section 3 summarizes the logical process of verification and discusses some "recommended" methods, depending on a range of issues. Section 4 focuses on the topic of severe weather. Finally Section 5 summarizes the replies of various centers to the survey.

2. A short review of available sources

The subject of verification is a very active area. The most common methods are presented by Stanski et al. (1989). A quick overview of recent developments can be obtained from the Internet site on Verification Methods maintained by E. Ebert at BMRC, see http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html. A detailed glossary is available at http://www.sel.noaa.gov/forecast_verification/verif_glossary.html#catfcst. An early discussion of verification techniques for high resolution models and related problems can be found in Anthes (1983). Some general concepts are discussed by Murphy (1991, 1993) and Murphy and Winkler (1987). A classic book on statistical methods is Wilks (1995). The subject was discussed in 1998 at a NCAR Workshop on Mesoscale Model Verification (Davis and Carr, 2000). A very recent discussion is given by Mass et al. (2002) in the context of the evaluation of a mesoscale model over the Pacific Northwest. Under the auspices of WGNE, a systematic inter-comparison of model precipitation forecasts against high resolution rain gauges (and sometimes radars) is now conducted in several centers (Ebert et al., 2002). These papers also contain some discussions of the best approach to verification at the mesoscale.

Verification methods at high-resolution are currently a subject of debate, with many on-going meetings. Here are a few recent examples: The European COST717 action on the use of radars in NWP has published a review of current methods and tools for verification of numerical forecasts of precipitation (written by C. Wilson, MO). This is available on http://www.smhi.se/cost717 . The European Short Range Network on Numerical Weather Prediction held a workshop in De Bilt in April 2001. Their discussion on verification methods can be found on http://srnwp.cscs.ch/leadcenters/ReportVerification.html The World Weather Research Program (WWRP) organized a workshop on QPF verification methods (Prag, May 2001). The report may be found on http://www.chmi.cz/meteo/ov/wmo . Another workshop devoted to the definition of more meaningful methods took place at NCAR in August 2002, see http://www.rap.ucar.edu/research/verification/ver_wkshp1.html . The WWRP Conference on Quantitative Precipitation Forecasting (Reading, September 2002) also had a session on verification methods. A summary of the session can be found on http://www.royal-met-soc.org.uk/qpfproc.html . The WWRP and WGNE have recently initiated a joint Working Group on Verification Methods.

A general consensus of these discussions seems to be: (i) new methods are needed to deal with the verification of mesoscale models; (ii) the international exchange of observations need to be enhanced; (iii) the intercomparison of model scores can be useful, but only if it is done with great care.

3. Methodology of Verification

The logical process of verification against observations can be divided in five steps: (i) the choice of a set of observations for verification; (ii) the technique to compare a single model forecast to a single observation; (iii) the aggregation of model/observation pairs in ensembles of a convenient size; (iv) the use of statistics to condense the information contained in the joint distribution of model/observation pairs (v) the use of additional information to help interpret the scores, in particular their statistical significance.

3.1 Observations available for verification of weather elements

The most commonly used observations for verification of weather elements are surface precipitation from rain gauges. The accumulation period is quite variable, from a few minutes to 24 hours. The use of the shorter accumulation periods should be encouraged for high-resolution models, with a view of matching the accumulation period to the model resolved time scales. Surface air temperatures, humidity and winds are also widely used. Cloud cover reports from surface stations are sometimes cited. The use of more advanced observation systems, such as meteorological radars and satellite cloud cover, is incipient and should be encouraged, although they are posing an obvious problem of accuracy, especially in mountainous areas. The use of a standard Z-R relationships for radar data is insufficient for heavy rainfall because of attenuation. The observation uncertainty should always be kept in mind when building a verification system. A few centers are developing verifications of other weather elements: Hail is reported in Synop observations, and specific detection networks exist in some parts of the world. Visibility is a subject of much interest, and reliable measurements are now available. Wind gusts are also commonly measured and predicted, and so deserve a specific verification. Ground skin temperature can be measured by satellite and is predicted by models, it should therefore also be verified.

3.2 Controlling the quality of the observations

This is a key step in the whole process. Most modern NWP systems have adopted a double quality-check procedure. In a first step, observations are checked for gross errors (unit problems, unphysical values, internal lack of consistency). Then they are compared to the model (see next subsection) and in case of a large differences between a model-derived value and the actual observation, other observations close-by are checked to ascertain whether the suspicious value is isolated, in which case it is discarded. This involves a considerable degree of empiricism, and could be at the origin of large differences in the results of various verification packages. There is a need for international exchange and comparison of the procedures involved in the quality control of observations, with due regards to differences inherent to the diversity of observing networks. The quality control methods might also be different for various verification purposes. For instance, an observation unrepresentative of the scale resolved by a model could be discarded as part of the quality control procedure when the verification is oriented towards model assessment, while it should be retained when the verification is user-oriented. This problem is even more important for high resolution models.

3.3 Comparing the model with the observations

The way in which forecasts and observations are matched becomes more important for mesoscale verification because of the sampling limitations of both observations and forecasts for small scale structures and processes. The best strategy obviously depends on the density and quality of the observing network, the resolution of the model, the type of observation considered, etc… This is highly variable around the world, so it is no surprise that meteorologists facing different situations in different countries have developed a large variety of methods, and sometime even vocabulary. Point observations contain information on all space and time scales, but usually drastically under-sample finer space and time scales. It is often considered preferable to treat the observations as estimates of area or time averages rather than to carry out an analysis of under-sampled fields. Such analyses artificially treat the point observations as if they contain information on only those scales which can be represented by the grid on which the analysis is done. Analysis of observations has the effect of eliminating from the verification the component of the error due to the inability of the model to represent scales smaller than its grid allows.

When the resolution of the observing network is larger than the model, the observations should clearly be up-scaled to the model resolution. A simple and efficient technique has recently been described by Cherubini et al. (2001) in the context of ECMWF model 24h accumulated rainfall verification: the climate observing network for 24h rainfall is significantly denser over most of Europe than the ECMWF model grid. It is therefore adequate to compute the arithmetic average of all the climate stations falling inside each model grid box. This more representative “super-observation” is then compared to the model grid value. This dramatically improves the model performance (especially the FBI and ETS scores at threshold 0.1 mm), and shows that the previous comparisons to the closest SYNOP rain observation were misleading.

A more common case is, however, when the model resolution is higher than the observing network. This will be true for most meso-scale models and weather parameters. One simple technique is then to interpolate the model prediction to the location of the observation, but this has the effect of smoothing the model result, and could result in a biased interpretation of its capacity to deal with extreme events. A common technique is to use the value at the nearest grid point to the observation location, ignoring the corresponding error on location.

Observations may not always be representative for the average model grid box (in fact they rarely are). Various representativity problems are due to the ground altitude (for temperature), to exposure effects (for wind and rain), to land cover heterogeneity (for temperature and humidity). Most centers use a standard vertical gradient of temperature to correct for altitude differences. Some schemes have been developed to correct wind forecasts for exposure effects and rain observations for altitude effects. With the rapid development of surface schemes using 'tiles', it may become possible to compare an observed temperature with one of several temperatures within the grid box (the one corresponding to the model land cover type matching the observation best). This may generate a need to have additional meta-data attached to the surface observations, indicating what is the immediate environment of the observation station (e.g. crops, lake, forest, urban, etc…).

The computation of the 'model equivalent' to the observations for verification purposes shares many aspects with the computation of the 'observation operators' in the variational data assimilation techniques. The development of common software for these two aspects of the NWP suite is encouraged, for instance for the radar and satellite observations. Furthermore, the differences between observations and model, in observation space, is already computed to evaluate the cost function which is minimized in variational procedures. These computations may not need to be done again for model verification (Davis and Carr, 2000). However, differences in the set of considered observations or in the detail of these computations may become necessary for user-oriented and model-oriented verifications.

3.4 How to aggregate/stratify the results?

There is a need to find a trade-off between various constraints: ensembles of forecast/observation pairs should be large enough to carry a good statistical significance, but small enough to distinguish between various areas or time periods prone to different types of errors (eg various climate, or altitudes). Stratifying results by time of the day will allow one to spot errors on the diurnal cycle of temperature and other variables, presumably linked to deficiencies in the surface energy budget parameterization, or the soil humidity. Stratifying by lead time tells about how fast the model is deriving from the truth. Stratifying by the values of the observed parameter shows how the model performance degrades towards extreme values. Stratifying by geographical area, or altitude above sea level, helps to point out the relations between model errors and the terrain. Finally, the available manpower to inspect the results will usually set a practical upper limitation to the number of scores. It is impossible to know in advance what combination of parameters will be needed to solve rapidly any new problem, so it is advisable to store individual values in a relational database for the purpose of quickly forming new combinations. This approach is now used in several centers.

The full examination of the joint distribution of the forecasts/observations pairs is a powerful way to acquire a detailed understanding of the characteristics of a forecast system (Murphy and Winkler, 1987). The bi-variate distribution p(f,o) can be factorized in marginal distributions for observations p(o), forecasts p(f), and the conditional distribution of observations given the forecasts p(o|f) or forecasts given the observation p(f|o). An approach used at the Met Office is to look at the distribution of observations for given forecast events. This can be interpreted as the probability distribution of observations given a specific forecast. For a perfectly accurate NWP model, we would expect to observe a parameter in a given interval on every occasion when the forecast is in that interval. A recent example of a distribution-oriented analysis of forecasts/observations pairs is provided by de Elia and Laprise (2002) (though they used only virtual observations supplied by a reference model run). They point to the fact that even for a globally unbiased forecast, the conditional bias (the bias of the forecast for a given value of the observed parameter) is in all cases towards the mean of the marginal forecast distribution. This should not be interpreted as an indication that the model is under-predicting. In fact, the conditional bias of the observations for a given forecast value is also towards the mean of the marginal observation distribution. This behavior is known as Galton's law in statistics.

3.5 Scoring deterministic forecasts

In practice, the bi-variate distribution often carries too much information and must be condensed by use of statistics. A large variety of statistical scores has been described in the literature, each of them having advantages and shortcomings. No single score can convey the full information, but it is often believed that a combination of a small number of well chosen scores can provide a reasonable assessment of most model error distributions.

The definitions and main properties of the most common scores are explained e.g. on http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html . Here is a short summary. Continuous statistics allow to measure how the values of forecasts variables differ in average from the values of observed variables. The mean error, or bias, is a useful basic information, but it does not measure the magnitude of the errors. The mean absolute error, the root mean square error, or the mean squared error all measure the average magnitude of the errors, with different weights of the largest errors. The anomaly correlation measures the correspondence or phase difference between the forecasts and observations without being sensitive to their absolute value. Categorical statistics are more appropriate to evaluate yes/no forecasts. They are often used to evaluate the capacity of models to predict that weather parameters will exceed a given threshold. A contingency table is constructed to count the correct predictions of observed events (hits), their non prediction (misses), the prediction of a non-observed event (false alarms) and the correct prediction of non-observed events (correct negatives). These quantities are combined in various categorical statistics. The Accuracy (ACC= hits + correct negative divided by total) measure the fraction of all forecasts that were correct. It can be misleading because it is heavily influenced by the most common category, usually the "no event" in the case of weather. The Frequency Bias Index (FBI) measures the ratio of the frequency of forecast events (hits + false alarms) to the frequency of observed events (hits + misses). It indicates whether the forecast system has a tendency to underforecast (FBI<1) or overforecast (FBI>1) events. It does not measure how well the forecast corresponds to the observations, only relative frequencies. The Probability of Detection (POD= hits/ hits + misses) measures the fraction of observed events that were correctly forecast. It is sensitive to hits, good for rare events, but ignore false alarms. It can be artificially increased by issuing more "yes" forecasts to improve the number of hits. The False Alarm Ratio (FAR= false alarms/hits + false alarms) measures the fraction of "yes" forecasts in which the event did not occur. It ignores misses and can be artificially improved by issuing more "no" forecasts to reduce the number of false alarms. The Threat Score (or Critical Success Index) (TS= hits/ hits + misses + false alarms) measures the fraction of observed and/or forecast events that were correctly forecast. It is sensitive to hits, but penalizes both misses and false alarms. However, it does not distinguish the source of forecast error, and is sensitive to the frequency of events, since some hits can occur due to random chance. Thus in general, the Threat Score will be higher for a sequence of unusually numerous events, and this should not be interpreted as an indication that the forecasting system is becoming better. In order to correct for this effect, the Equitable Threat Score (ETS) measures the fraction of observed and/or forecast events that were correctly predicted, adjusted for hits associated with random chance in the forecast (ETS=(hits – hits(random))/ hits + misses + false alarms – hits(random), where hits(random) = (hits + misses)x(hits + false alarms)/total). This score is often used in the verification of rainfall forecasts because its "equitability" allows scores to be compared more fairly across different precipitation regimes. Along the same ideas, the Heidke Skill Score (HSS) measures the fraction of correct forecasts after eliminating those forecasts due purely to random chance. It measures the improvement over random chance. However, random chance is usually not the best forecast to compare to, and the HSS is sometimes computed with respect to climate or to persistence. More recently, the merits of the Odds Ratio (OR=hits * correct negative / misses * false alarms) have been argued (Stephenson, 2000; Goeber and Milton, 2002). The OR measures the ratio of the probability of making a hit to the probability of making a false alarm. It is appropriate for rare events, does not depend on marginal totals, and is therefore "equitable". It can easily be used to test whether the forecast skill is significant.

Multi-category forecasts can also be verified by building multi-category contingency tables. Scores can then be defined to quantify the degree of fit between the distributions of forecasts and verifying observations. The Accuracy and Heidke Skill Score are two examples of scores that can be easily generalized to account for multi-category forecasts.

Specific user-oriented scores are easily developed based on the above principles. Very often the main interests of users can be summarized in the two following questions: what is the probability that an event will occur when it is forecast? What is the probability that an event has been forecast when it occurs?

3.6 The double penalty problem

It is common observation that the objective scores for weather parameters can be worse for high resolution models than for low resolution models. Indeed, increased resolution generally produces better defined mesoscale structures, greater amplitude features and larger gradients. Thus, inevitable space and timing errors for weather-related parameters will lead to a larger RMSE than the smoother forecasts of a low resolution model. This is generally known as the 'double penalty' problem (see e.g. Anthès, 1983, or Mass et al., 2002). At the same time, there is a consensus that high-resolution numerical predictions are very useful to forecasters, even with small space and timing errors, because they point to the possibility of some important weather patterns happening in a given area, and because they convey some explanation of why and how this may happen (a conceptual model). A classical example is the forecast of isolated thunderstorms, where models are not expected to provide a very accurate location, but can be very informative regarding timing and severity. The need for verification techniques that allow for some tolerance to reasonably small space and time errors is universally recognized and central to much of the recent literature on the subject. One approach is to average the output of the high-resolution model to a lower resolution before applying the deterministic scores (this is sometimes called "hedging"). This may reveal the superiority of high-resolution models over low-resolution models, while direct comparison of model outputs interpolated to the station point would in general give a more favorable result for the low resolution model (Damrath, personal communication). However, smoothing model outputs will in general deteriorate their intrinsic behavior, such as forecast variance, spectrum of energy, and the frequency of intense events. This detrimental effect can be measured by other indicators, such as the Frequency Bias Score (see definition below). In general, it is recommended to consider several indicators to assess the quality of a model (e.g. the forecast variance should be close to the observed variance, the forecast bias should be very small, and the root mean square error should be reasonably small). An early paper on the usefulness of Control Statistics to avoid "playing the scores" is Glahn (1976).

Other approaches to circumvent the double penalty are reviewed by Davis and Carr (2000). Brooks et al. (1998) compute the probability density distribution associated with local severe weather reports on a single day, and evaluate the maximum skill of a forecast based on simple spatial averaging. This turns out to be fairly low (a CSI of 0.24 in his example). Thus, a hypothetical numerical forecast having a CSI of 0.09, despite being rather low in absolute value, represents 38% of the upper bound, and must be considered as relatively successful forecast. A most simple method used by de Elia and Laprise (2002) consists in allowing for a tolerance of one grid point to find the best match between the forecast and the observation. A more elaborated version of the procedure is to consider that all grid points within a given distance of a point of interest are equally likely forecasts of an event at this point. Thus, a probability of some threshold being exceeded at this point can be computed as the ratio of the number of neighboring grid points where it happens over the total number of grid points considered. The size of the area for these counts is subject to optimization. This probabilistic forecast must then be evaluated through appropriate scoring. An example of this approach is discussed by Atger (2001).

3.7 Scoring probabilistic forecasts

A good probability forecast system has three attributes: (i) Reliability is the agreement between the forecast probability and the mean observed frequency; (ii) Sharpness is the capacity of the system to forecast probabilities close to 0 or 1; (iii) Resolution is the ability of the system to resolve the set of sample events into subsets with characteristically different frequencies. Sharpness and resolution are somewhat redundant, and become identical when reliability is perfect. The most common measure of the quality of probabilistic forecasts is the Brier Score (Brier, 1950). It measures the mean squared probability error, and ranges from 0 to 1, with perfect score 0. Murphy (1973) showed that the Brier Score can be partitioned into three terms accounting respectively for reliability, resolution, and uncertainty. The Brier Score is sensitive to the frequency of the event: the more rare the event, the easier it is to get a good BS without having any real skill. The Ranked Probability Score (RPS) measures the sum of squared differences in cumulative probability space for a multi-category probabilistic forecast. It penalizes forecasts more severely when their probabilities are further from actual outcome. As the BS, it ranges from 0 to 1, with perfect score 0.

Reliability is specifically measured by reliability diagrams, where the observed frequency is plotted against forecast probability, divided into a certain number of groups (Wilks, 1995). Perfect reliability is achieved when the results are aligned along the diagonal of the diagram. A shortcoming of reliability diagrams is that one needs a large number of forecasts to generate a meaningful diagram. An alternative approach known as the multi-category reliability diagram (Hamill, 1997) allows to accumulate statistics from a reduced number of forecasts. In the case of ensemble forecasts, an additional useful evaluation is given by the Rank Histograms (also called Talagrand diagrams). The rank is the position of the verifying observation relative to the ranked ensemble forecast values. For a reliable ensemble, the Rank Histogram should be approximately uniform, meaning that an observation is equally likely to occur near any ensemble member.

Probabilistic forecasts can be tailored to the use of any specific user category, by adjusting the probability threshold required to make a yes/no decision. Of course, this will induce a simultaneous change of the POD and of the FAR. An increase of POD will be achieved at the cost of an increase in FAR. Diagrams showing how the POD and FAR rate change with the decision criteria are called Relative Operating Characteristics curves (ROC). They describe how a forecast system can meet simultaneously the needs of various users categories, and therefore contain a lot of information. In contrast, a deterministic forecast system will be represented by a single point on such a diagram. It is expected that the curve describing the probabilistic system results pass above the point describing the deterministic system, showing the superiority of the probabilistic approach. The area under the ROC curve is frequently used as a global indicator of the quality of a probabilistic forecast system. However it tells nothing about reliability. Another increasingly used measure of the quality of probabilistic forecasts is the Potential Economic Value, which conveys about the same information as the ROC curve, translated in potential gain for any category of users, stratified by their Cost/Loss parameter (e.g. Richardson, 2000).

3.8 Additional information necessary to interpret the scores

An essential information is the uncertainty associated with the above statistics. A related question is the statistical significance of the comparison between two forecasting systems on a given series of weather events. This is especially important for severe weather, since the number of events is often small. Hamill (1999) discusses a number of limitations of common hypothesis tests in weather forecast verification, such as spatial correlations and non-normality of errors. He proposes new methods such as re-sampling techniques, that allow to evaluate the uncertainty associated to statistical scores such as the widely used ETS. Similar techniques are also applicable to the probabilistic scores. Atger (2001) has applied this method in the context of QPF. The sample of events was randomly halved into two sub-samples, and the score differences between the two sub-samples were evaluated. The process was repeated a large number of times and resulted in an evaluation of the uncertainty in the scores.

Another recommended point of comparison is with straightforward forecasting techniques, such as climate, persistence, or chance. This is embodied in a number of the above-mentioned scores. Finally, it is considered that computing scores on the verifying analysis (or on the model initial state) is a good point of comparison.

3.9 Research in verification methods

The development of new verification methods is an active area of research. Most of the methods discussed above will tell little about exactly what the error is, or why there is an error. Therefore recent efforts are directed towards the development of methods that could help the modelers to improve models. The need to identify the spatial scales involved in a given error was already mentioned by Anthès (1983). Scale separation techniques are being developed, base e.g. on wavelets (Briggs and Levine, 1997). The objective is to identify at what scales the greatest error is occurring, and whether the model resolves all of the scales that can be measured in the observations. Zepeda-Arce et al. (2000) propose a method consisting in up-scaling from fine to coarse resolution by simple averaging, and computing verification scores as a function of both threshold and resolution. If the scores improve very quickly towards coarser resolution, there is an indication that the forecast is good. Fuzzy verification techniques under developments at BMRC and DWD try to deal with uncertainty in both forecasts and observations. Finally, the examination of model energy spectra and their evolution over time has often been recommended to verify the realism of simulations.

The use of object-oriented techniques is also developing rapidly. This is making sense when it is possible to associate unambiguously an observed weather object with its forecasted counterpart. A most classical application of this is the verification of the skill in forecasting the track of tropical cyclones. The score is based on the distance between the observed and forecasted tracks of the cyclone center, assuming that the association between the observed and forecasted cyclone is a simple issue. Hoffman et al. (1995) have proposed a generalization of this approach to other types of events, and Ebert and McBride (2000) have implemented a similar system, the Contiguous Rain Area method (CRA), now routinely used at BMRC, Australia. They show that the total RMSE of a precipitation forecast can be decomposed into three components, describing respectively a displacement error, a volume error and a pattern error. A systematic evaluation of these three components of error on a long period helps in understanding what the problems of the model are. It also allows to define the 'hits' and 'false alarms' cases with a certain tolerance, consistent with the forecasters opinion of a useful forecast. It should be noted that full application of this technique is only possible when the forecast and observed rain systems are completely enclosed within the verification domain. Moreover, application to local storms would probably be hampered by the difficulty of associating unambiguously an observed and a forecast event without a human intervention, except for the strongest cases.

4. The severe weather problem

Severe weather poses a special problem because it is unfrequent, poorly documented by observations, and at the limit of predictability. Quantitative verifications are therefore more difficult and their statistical significance is always poor. At the same time, it is recognized that a poor numerical forecast in absolute terms can be of great value if it is well interpreted by an experienced forecaster. This may be seen as an extreme example of the "double penalty" problem. In addition of a tolerance on space and time, a tolerance on the value of weather-related parameters must often be accepted in the case of extreme values. For instance, in a region where a daily accumulated precipitation larger than 200 mm is a rare event, a 200 mm forecast represents a bad forecast if the observed value is more frequent (say, 50 mm), but a useful forecast if the observed value is 350 mm. So, the same absolute error can have varying significance depending on how the forecast is placed with respect to climate. The issue is made more complex by the scale difference between model and observations. In many cases indeed, we should not expect the current models to reproduce the maximum values of weather parameters observed in extreme events because their resolution is too low. We should however design methods to diagnose severe weather based on the existing models, and thoroughly verify the validity of these diagnostics.

The linear error in probability space method (LEPS, Ward and Folland, 1991) is an early attempt to deal with this problem. If f is the forecast, o the observation, and F(o) the cumulative probability density function of o, (i.e. the probability that the observation is smaller then o), the LEPS measure of the error is the difference F(f)-F(o). Therefore large differences between f and o are less penalized if they occur near extreme values of the distribution of o. The minimum error is 0 and the maximum error is 1.

The Extreme Forecast Index, developed recently at ECMWF (Lalaurette, 2002) provides a generalization to probabilistic forecasts. The extreme forecast index (EFI) is a measure of the difference between a probabilistic forecast and a model climate distribution. In order to avoid a dependence on the climate of the region under study, it is desirable that such an index do not scale like the forecast parameter, but varies from –1 (an extreme negative value) to +1 (an extreme positive value). To achieve this goal, the EFI is formulated in the probability space: for a given location on Earth and a given meteorological parameter, one associates to each proportion p of the ranked model climate records a parameter threshold $q_c(p)$, known as the percentile of the distribution: $q_c(0)$ is the absolute minimum, $q_c(0.5)$ the median, $q_c(1)$ the absolute maximum. We then define $F_f(p)$ as the probability with which a probabilistic forecast predicts that the observation will be below $q_c(p)$, and write $EFI= \int (p-F_f(p)\, dp$. The index cumulates the differences between the climate and forecast distributions. $F_f(p)=p$ only in the case where forecast probability distribution is exactly the same as the climate, and in this case EFI=0. This will be also true for a deterministic forecast calling for the median value of the climate record. Furthermore, EFI=+1(-1) only if all possible values in the forecast are above (below) the highest (lowest) value of the climate record. In practice, an exponent (3) is used in order to have the EFI varying more rapidly near the extreme values. One limitation of the EFI is the need to have a good representation of the model climate. In practice, this can only be obtained by running a constant version of the model (or nearly constant) during several years. There is some hope that the time period needed to accumulate enough statistics can be considerably reduced by using ensemble predictions, providing many realizations of the forecast every day. In order to verify the EFI forecast, the model analysis or short-range forecast can be used. Contingency tables can be constructed to count the number of occasions when the EFI prediction performed well or bad in exceeding a given value. Thus, categorical scores can be produced for the EFI prediction. Also, to account for under- or over-prediction of extreme events by a model, one may decide to issue a warning when the EFI forecast exceeds a value lower or higher than the target, and construct ROC curves. This type of verification is believed to be extremely useful to increase and assess the capacity of a NWP model to predict extreme events, with due regards to its systematic biases. The ECMWF EFI system is being developed in the frame of a medium resolution ensemble prediction system, but it is believed that a similar approach could be adopted for deterministic or probabilistic forecasts from a high resolution model, provided a convenient knowledge of the model climate is at hand.

5. Main verifications of weather elements currently performed at operational centers

A survey of methods currently in use or in development has been performed, focusing on the verification of weather elements. The following is a summary of replies by operational NWP centers, indicating only the major efforts. There is no intention to provide an exhaustive list of verifications performed by these centers.

Australia: the rainfall forecast is verified against an analysis of 24-hour rain gauge data over continental Australia. The resolution of the analysis is 0.25 degrees, and the analysis is remapped to the model resolution. The basic verification relies on bias, RMSE, and contingency tables from which various categorical scores are computed. The statistics are written to files and saved for various aggregation and display schemes. An object-oriented verification (Contiguous Rain Area method, Ebert and McBride, 2000) is also performed on up to four individual rain systems per day. The location, volume, and pattern errors are computed, as well as errors in rain area and intensity. This is considered very useful for extreme events. Some work is in progress with radar data.

Canada: Bias and RMSE of wind, temperature, dew point, and surface pressure to surface and upper air stations are routinely monitored. For precipitation, bias and threat scores for various thresholds are computed to the synop stations, and more recently to a higher resolution SHEF (standard hydrometeorological exchange format) network (Belair et al., 2000). Work on the North American radar data has started and will be used to assess the relative importance of the various physical processes in the model.

China: 400 stations were carefully chosen over China’s territory for precipitation forecasts verification. Both NWP models and subjective forecasts are interpolated to the location of these stations, and the verification is done routinely. It is based on threat scores and bias for various thresholds (0.1, 10, 25, 50 and 100 mm/24 hours).

France: About 1200 synoptic and automated surface weather stations are used. The parameters subject to systematic verification are the precipitation, the cloud cover, the temperature and humidity at 2m, the wind speed and direction and the intensity of the wind gusts. The nearest model grid point is used to compare with the point observations. The biases and RMSE are computed. In addition contingency tables are computed for precipitation (4 classes) and cloud cover (3 classes). All observations and forecasts at each point station are retained in a single database in order to conduct analyses of the model performance by sorting stations according to various criteria. Further contingency tables are being developed for wind speed and wind gusts. Work is in progress concerning the use of radar data to verify the precipitation forecast, and object-oriented methods.

Germany: At DWD, verification of precipitation is done using a high density network of observations (around 3600 sites with daily totals). The following verification strategies are used: (i) user-oriented verification: comparison of observations with forecasted values at the nearest grid point of the model or with an interpolated value from the surrounding grid points; (ii) modeler-oriented verification: computation of super-observations in different grids (1°x1° grid for WGNE; in the grid of the global model; in the grid of the regional model). Verification using Synop data is also done for the operational models.

Japan: JMA operates a high resolution surface observation network named the Automated Meteorological Data Acquisition System (AMeDAS), which consists of 1300 raingauges, 200 snowgauges and 800 thermometers, aerovanes and heliographs all over Japan. Its estimated grid spacing is about 17km for raingauges and 20 km for other facilities. The AMeDAS data are used to verify forecast performance on both precipitation and surface temperature. The observational data are converted into a set of uniform grid data in 80 km mesh and the forecasts are compared with the gridded observations. This method is adopted to avoid discontinuity caused by changes in model resolution and to reduce sampling error of observation. JMA also operates 20 radar sites and produces a precipitation analysis over Japan by compositing radar reflectivities and AMeDAS raingauge data. This analysis is used to evaluate the forecast skill of the mesoscale model at three different resolutions: 10, 40 and 80km, and for time periods of 1, 3, and 6 hours. The regional spectral model is verified with the same data at 20, 40, and 80km resolution. Standard categorical scores are computed, such as threat score, bias score and equitable threat score.

UK: Operationally, the UK mesoscale forecasts are assessed by a summary index based on five parameters: 1.5m temperature, 10m wind, 6h accumulated precipitation, total cloud cover and visibility. Skill scores from T+6 to T+24 are used, with 42 stations used as truth. For temperature and wind the skill scores are based on mean square errors compared to persistence. Equitable threat scores are used for precipitation, cloud cover and visibility with thresholds of 0.2, 1.0, 4.0 mm/6h, 2.5, 4.5, 6.5 oktas for clouds, and 5km, 1km, and 200m for visibility. UK is also making a considerable effort to use radar composites to verify precipitation. Within the NIMROD system, ground clutter, corrupt images, and anomalous propagation effects are removed, the vertical profile of reflectivity is taken into account, and calibration against gauges is adjusted once per week.

United States: at the U.S. National Centers for Environmental Prediction, model forecasts of surface and upper-air fields are verified against a myriad of observational data, including height, temperature, wind and moisture observations from radiosondes, dropsondes, land and marine observation stations; temperature and wind from aircraft at flight level and during ascent/descent; and upper air winds from pibals, profilers, satellite derivations and doppler radar VAD product. Model fields are interpolated to the location of the observation for the comparison. The extensive verification database allows evaluation of model performance from a variety of angles. Daily (12Z-12Z) precipitation verification is performed using a 0.125 degree precipitation analyses over the contiguous United States based on 7,000-8,000 daily gauge reports which are quality-controlled with radar and climatological data. The verification is done on 80-km and 40-km grids for NCEP operational models and various international models, and on a 12-km grid for NCEP's mesoscale non-hydrostatic nested model runs. Precipitation fields (forecast and observed) are mapped to the verification grids. From the precipitation forecast/observed/hit statistics collected for the verification domains (Continental US and 13 subregions), 26 different scores can be calculated, among which are equitable threat, bias, probability of detection, false alarm rate and odds ratio. Limited 3-hourly verification is also performed using NCEP's hourly 4-km multi-sensor precipitation analysis based on radar and automated hourly gauge reports. Monthly/month-to-date precipitation verification graphics are available at:
http://wwwt.emc.ncep.noaa.gov/mmb/ylin/pcpverif/scores/ .

Except for precipitation, where raingage observations are used in a procedure similar to that described above, global (medium range, 15 days) and regional (short-range, 48 h) ensemble forecasts generated operationally at NCEP are currently evaluated against gridded NWP analysis fields. Beyond the traditional scores (root mean square error and anomaly correlation for the individual members and the ensemble mean field) analysis rank histograms and histograms assessing the time consistency of consecutive ensemble forecasts are also computed (Toth et al, 2003). Probabilistic forecasts derived from the ensemble, including spread and reliability measures, are evaluated using a variety of standard probabilistic verification scores including the Brier Skill Score, Ranked Probability Skill Score, Relative Operating Characteristics, and Economic Value of forecasts. The latter two measures are also computed for single higher and equivalent resolution "control" forecasts originating from unperturbed initial fields, allowing for a comparative analysis of the value of a single higher resolution forecast and a lower resolution ensemble of forecasts.

Russia: The grid-point values of the non-hydrostatic meso-scale model are compared with nearby stations directly. The verified parameters are: surface pressure (bias, mean absolute and rms errors); surface temperature (bias, mean absolute and rms errors, relative error); wind (mean absolute vector error, mean absolute speed and direction errors, speed bias, scalar and vector RMSE); precipitation (an ensemble of scores based on contingency tables).

6. Conclusions

While it is impossible to cover the whole field of verification techniques in this survey, several conclusions emerge from the review of current works and debates:

1. As high resolution models are expected to provide results in direct relevance to user needs, there is a growing pressure to develop ‘user-oriented’ verifications. These can depart significantly from the more traditional model-oriented verifications, for instance in the choice of actually used observations or scrutinized model

scales. Since it is difficult to accommodate several different needs in the same software, it is recommended to separate clearly the user-oriented and model-oriented parts of the verification packages.

2. The resolution of the observing networks is now often inferior to the resolution of NWP models. This calls for an improvement of observing networks, and design of more adequate verification techniques, especially for weather elements. Enhanced international exchange of high resolution data should also be encouraged.
3. The difference in horizontal scale between the forecast and the observations is too often neglected. However, no really adequate technique appears to exist to deal with this problem, except in some very special cases where up-scaling of observations is possible.
4. The detailed prediction of some weather elements often appears at the limit of current NWP models capacity. This is due to specific predictability problems, and to remaining weaknesses in the model formulation, observations, and data assimilation techniques.
5. The double penalty problem remains a central issue with which many verification scientists are struggling. Several approaches to this problem are pursued. (i) Use of convenient battery of scores (e.g. ETS and FBI); (ii) up-scaling the verification; (iii) formulate the forecast in a probabilistic way, either by use of ensembles, or by use of a collection of neighbor grid points and time steps.
6. Because of the intrinsic predictability limitation, the verification problem for weather elements at high resolution is better posed in a probabilistic way. There is a need to develop probabilistic formulations of the forecast and adequate verifications.
7. Severe weather verification poses a specific problem, and currently requires a verification in the probability space (such as LEPS or EFI). The relative frequency of severe events should be matched between model and observation rather than their quantitative representation. This requires a good knowledge of the model climate.
8. The verification problem shares many aspects with the data assimilation problem, for instance the computation of observation operators and of differences between forecasts and observations. This should be recognized and exploited in the development and maintenance of software.
9. A set of standard verifications should be defined for weather elements from high resolution NWP. This may be the subject of future work of the WGNE.
10. Verification scores should always be accompanied by information on the uncertainty and/or statistical significance. Extreme cases are very limited in number, and verification without proper account of uncertainty may easily result in wrong conclusions. Comparison with more simple forecasting methods, such as climate, persistence, and chance should also be provided as a reference. Also, the model analysis (initial field) should be scored with the same technique as the forecast in order to provide a reference. Some scores are more easily amenable to uncertainty computation and more "equitable" (in the sense that they are less sensitive to the sample composition). Recently, the Odds Ratio has been claimed to possess those qualities.

Advanced verification techniques are under development in various centers to provide model developers with more appropriate information on the origin of errors and the realism of models. Verification can have a number of different objectives and no single technique can address all objectives at once. Verification will remain a complex and important subject. It is believed that the search for better verification methods is one powerful way to reach better forecasts.

Acknowledgements: This paper has benefited of much input from various people. I am particularly indebted to Elizabeth Ebert (Australia), René Laprise, Jocelyn Mailhot and Laurie Wilson (Canada), Chen De Hui (China), François Lalaurette (ECMWF), Samuel Westrelin and Frederic Atger (France), Ulrich Damrath (Germany), Goda Harutoshi (Japan), Gerard Kats (Netherlands), Dmitry Kiktev (Russia), Martin Goeber, Andrew Lorenc and Clive Wilson (UK), Stephen Lord and David Williamson (USA).

References

Anthes, R.A., 1983: Regional models of the atmosphere in middle latitudes. Mon. Wea. Rev., 111, 1306-1330.

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

Belair, S., A. Methot, J. Mailhot, B. Bilodeau, A.Patoine, P. Pellerin and J. Coté, 2000 : Operational implementation of the Fritsch-Chappel scheme in the 24-km Canadian regional model. Wea. Forecasting, 15, 257-274.

Brier, G.W., 1950: Verification of forecasts expressed in terms of probability. Mon. Wea. Rev., 78, 1-3.

Briggs, W.M., and R.A. Levine, 1997: Wavelets and field forecast verification. Mon. Wea. Rev., 125, 1329-1341.

Brooks, H.E., M. Kay and J.A. Hart, 1998: Objective limits to forecasting skill of rare events. Preprints, 19th Conf. On Severe Local Storms, Minneapolis, MN, Amer. Meteor. Soc, 552-555.

Cherubini, T., A. Ghelli and F. Lalaurette, 2001: Verification of precipitation forecasts over the Alpine region using a high density observing network. ECMWF Technical Memorandum, n°340, 18pp.

Davis, C., and F. Carr, 2000: Summary of the 1998 Workshop on mesoscale model verification. Bull. Amer. Meteor. Soc., 81, 809-819.

De Elia, R., and R. Laprise, 2002: Distribution-oriented verification of limited-area model forecasts in a perfect-model framework. Mon. Wea. Rev., submitted.

Ebert, E.E., and J.L. McBride, 2000: Verification of precipitation in weather systems: determination of systematic errors. J. Hydrology,239, 179-202.

Ebert, E.E., U. Damrath, W. Wergen, and M. Baldwin, 2002: The WGNE assessment of short-term quantitative precipitation forecasts from operational NWP models. Bull. Amer. Meteor. Soc., to appear.

Glahn, H.R., 1976: Forecast evaluation at techniques development laboratory. In "Weather Forecasting and Weather Forecasts: Models, Systems and Users". NCAR Colloquium, Summer 1976, pages 831-838.

Goeber, M., and S. Milton, 2002: Verifying precipitation events. NWP Gazette, March 2002, Met Office, 9-11.

Hamill, T.M., 1997: Reliability diagrams for multi-category probabilistic forecasts. Wea. Forecasting, 12, 736-741.

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

Hoffman, R.N. Z. Liu, J.F. Louis, and C. Grassoti, 1995: Distortion representation of forecast errors. Mon. Wea. Rev., 123, 2758-2770.

Lalaurette, F., 2002: Early detection of abnormal weather using a probabilistic Extreme Forecast Index. ECMWF Technical Memorandum, n°373, 27pp.

Mass, C.F., D.Ovens, K. Westrick and B.A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? Bull. Amer. Meteor. Soc., March 2002, 407-430.

Murphy, A.H., 1991: Forecast verification: its complexity and dimensionality. Mon. Wea. Rev., 119, 1590-1601.

Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. Weather and Forecasting, 8, 281-293.

Murphy, A.H., and R.L. Winkler, 1987: A general framework for forecast verification. Mon. Wea. Rev., 115, 1330-1338.

Richardson, D.S., 2000: Skill and economic value of the ECMWF Ensemble Prediction System. Quart. J. Royal Meteor. Soc., 126, 649-668.

Stanski, H.R., L.J. Wilson, and and W.R. Burrows, 1989: Survey of common verification methods in meteorology, WMO/WWW Tech. Rep. n° 8.

Stephenson, D.B., 2000: Use of the "Odds Ratio" for diagnosing forecast skill. Weather and Forecasting, 15, 221-232.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. In: Environmental Forecast Verification: A practitioner's guide in atmospheric science. Ed.: I. T. Jolliffe and D. B. Stephenson. Wiley, in print.

Ward, M.N., and C.K. Folland, 1991: Prediction of seasonal rainfall in the north nordeste of brazil using eigenvectors of sea surface temperature. Int. J. Climatology, 11, 711-743.

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

Zepeda-Arce, J., E. Foufoula-Georgiou and K.K. Droegemeier, 2000: Space-time rainfall organization and its role in validating quantitative precipitation forecasts. J. Geophys. Res., 105, 10129-10146.