

User-Relevant Verification

Barbara Brown¹

National Center for Atmospheric Research
Research Applications Program
P.O. Box 3000
Boulder CO 80307-3000
bgb@ucar.edu

Prepared for the North American THORPEX Societal and Economic
Research and Applications (NAT SERA) Workshop
August 14–16, 2006
Boulder, Colorado

December 24, 2006

¹ Corresponding author

Theme 2: User-Relevant Verification

1. Introduction

Late in the 19th century, a sergeant in the U.S. Army Signal Corps was responsible for an event that gave rise to numerous developments related to the methods used for evaluating (or verifying) the performance of weather forecasts. In 1884 Sergeant John Finley published the results of an experiment in which he forecasted the occurrence/non-occurrence of tornadoes in a large number of U.S. districts (Finley 1884). Finley evaluated the quality² or performance of the tornado forecasts using a reasonable statistical measure, the percentage of correct forecasts. With this statistic, Finley found that his tornado forecasts were 96.6% accurate. Unfortunately—as was pointed out in numerous articles and discussions that followed—Finley’s forecasts would have fared better, with 98.2% correct, had he never forecasted a tornado (Murphy 1996).³

The Finley example is relevant and compelling because (a) it clearly demonstrates that some measures of forecast quality may not indicate much about the usefulness of a forecast; (b) it led to a set of standard verification measures that are still heavily relied on today; and (c) a number of issues that were raised in the resulting discussions have still not been resolved or adequately considered by the forecast verification and forecasting communities, even 120 years later.⁴

In particular, the “Finley affair” led to the first scientific discussion of forecast verification (described in Murphy 1996), the development of numerous verification measures (e.g., Equitable Threat Score, also known as the Gilbert skill score; Gilbert 1884), and the awareness and discussion of many issues related to how forecasts should be evaluated. For example, Peirce (1884) and others considered the dimensionality of the forecast verification problem—the fact that the quality of a forecast cannot be summarized using a single measure; in fact, representing the full information in the joint distribution of forecasts and observations requires a large number of measures (Murphy 1991). Köppen (1893) and others raised questions regarding the relationship between forecast performance and forecast use and value, and Nichols (1890) considered the fact that the costs and losses associated with misses and false alarms may be different, and they may differ for particular sets of users. Finally, Clayton (1889) discussed the need to specify the purpose of verification before designing verification metrics.

Much later, Brier and Allen (1951) specified three purposes for forecast verification — administrative, scientific, and economic—which are still relevant and noted today (e.g., Wilks 2006; Jolliffe and Stephenson 2003). The *administrative* purpose refers to the need to for managers and others (e.g., funding agencies) to monitor forecast performance, to ensure that forecasts are not degrading, and to select among alternative forecasting systems. The *scientific* purpose refers to the need by forecast developers (or forecasters) for information that will help improve the forecasts. The *economic* purpose refers to the need for users to have information that will improve decisions, or allow them to evaluate forecast benefits.

² Here the word “quality” refers to the degree of correspondence between the forecasts and observations (Murphy 1993).

³ The Finley example is beautifully described, and the aftermath discussed, in Murphy (1996). Because the Finley saga represents an important stage in the development and understanding of several aspects of forecast verification, many texts, including Jolliffe and Stephenson (2003) and Wilks (2006), also discuss this event and its consequences.

⁴ Murphy (1996) called this situation “the benign neglect afforded the subdiscipline of forecast verification by the meteorological community as a whole during this period.”

Unfortunately, most current verification activities, especially at operational prediction centers, serve only the first of these purposes (i.e., administrative) and many of the lessons learned and issues raised by the Finley affair are still not given adequate consideration today.

The thesis of this paper is that the atmospheric sciences community, with the guidance of social scientists and application area experts, can—and *should*—do a much better job of evaluating forecasts in a meaningful way (even taking guidance from the late 19th century). By making forecast verification more user-relevant in both operational settings and in forecast development, information can be provided that is beneficial for a variety of end users (including forecast developers).

The paper first considers the current situation and widely applied “traditional” verification approaches, and then discusses the problems and limitations of these approaches. A diagnostic approach to verification is then considered and demonstrated using examples. The next section proposes a hierarchy of “user relevance” to help define how forecast verification approaches can mature into approaches that truly produce meaningful information for forecast users. Finally, we propose set of research objectives related to this hierarchy, which could be accomplished as part of a North American THORPEX Societal and Economic Research and Applications (SERA) research agenda.

In the context of this paper it is important to distinguish between forecast *quality* and forecast *value* or *benefits*. As eloquently discussed by Murphy (1993), improvements in forecast quality (which measures the correspondence between forecasts and observations) do not necessarily result in increases in value or benefits for any or all users. The relationship between forecast quality and value typically is very complex, and depends on the parameters defining the use of the forecasts or the decision-making situation, and on attributes of the forecast user. Measuring forecast value thus is not as straightforward as measuring quality, and requires application of economic or decision models, or sophisticated survey research methods⁵. Estimates of forecast quality and forecast value both contribute to forecast *evaluation*.

2. Traditional (measures-oriented) verification approaches

The traditional forecast verification approaches that are commonly applied in practice in both research and operational settings are *measures-oriented*. That is, they rely on simple scalar estimates of forecast quality. Typically, one or two measures or skill scores are used to summarize the performance characteristics of a set of forecasts. Each measure estimates quality in terms of a single specific attribute⁶ (e.g., accuracy, skill). The selection of specific measures typically depends on characteristics of the forecasts (e.g., continuous, categorical, probabilistic) and observations, but may include such statistics as the Heidke skill score, Root-Mean Squared Error (RMSE), correlation, Brier score, Critical Success Index (CSI), Probability of Detection (POD), False Alarm Ratio (FAR), and Brier score⁷. Such measures are typically applied even if the forecast has a spatial or gridded form (i.e., as opposed to representing conditions at a

⁵ The Theme 3 discussion paper for this workshop discusses these methods.

⁶ A forecast attribute is a characteristic of a forecast that represents one aspect of forecast quality. For example, skill measures the relative accuracy of a forecast compared to the accuracy associated with a naïve standard such as climatology.

⁷ See Jolliffe and Stephenson (2003), Stanski et al. (1989), and Wilks (2006) for details about these and other verification measures.

particular point or station). For example, gridded precipitation forecasts are generally evaluated as a set of point forecasts.

Model developers and managers commonly use scalar verification measures such as these to make decisions regarding model parameterizations and other attributes of operational forecasting systems. By applying such criteria to make these choices, they are implicitly deciding which aspects of the model are important and should be optimized. For example, by keying on the RMSE associated with predictions of the height of the 500 mb pressure surface as an important measure of model performance, model developers (and managers of model development) implicitly assume that this attribute of the forecasts (500 mb height) is (or represents) the most important attribute of the model forecasts; this approach results in a model that is good at making these types of forecasts—but may not be as good at making other types of forecasts (e.g., forecasts of surface variables, such as temperature or precipitation extremes that are relevant for end users). Moreover, using the RMSE to measure quality implies that small errors are much less important than large errors, and rewards conservatism in the model predictions. Choices like these generally are made without considering the needs or values of end users. Thus, forecast model development generally is controlled by simplistic evaluation of one or two parameters that are not likely to be relevant for any particular set of users.⁸

Typically, operational forecast verification and verification analyses applied in forecast development studies provide *aggregated statistics* for forecasts collected across large spatial and temporal domains rather than for homogeneous subsets, even though such aggregation tends to hide important information about forecast performance. Moreover, stratification of results according to regional domains and sub-periods (and other relevant categories) might provide information that is more meaningful for particular users.

Forecast verification studies also rarely include information about the uncertainty associated with the verification measures themselves. This uncertainty arises from measurement errors in the observations used in the verification analyses, as well as sampling variability associated with the selection of forecasts to be included in the analysis. Ignoring this uncertainty may lead, for example, to arbitrary choices for operational models, based on differences in performance that are not statistically meaningful.

3. Motivations for alternative verification approaches

Application of traditional forecast verification measures allows overall monitoring of forecasting performance according to specific criteria (e.g., these measures may meet certain administrative needs for forecast verification information). However, the traditional measures have several limitations and flaws. For example, they measure only a very limited set of attributes of forecast quality. In addition, they tend to reward “smooth” forecasts over forecasts with more detail (including more realistic ranges of values) and they do not provide information about *what* went wrong with a forecast (i.e., they only have the capability to indicate that it was wrong). In addition, using the measures-based approach does not allow diagnosis of how the

⁸ From a historical perspective, the convention of evaluating forecasting models using variables that do not have direct human impacts (e.g., 500 mb height) and the reliance on simple scalar verification measures (e.g., RMSE, anomaly correlation) may have been reasonable in the early years of numerical weather prediction, when weather prediction models were much simpler and had limited capabilities. However, many model forecast evaluations still rely on these measures and variables even though the capabilities of models are now much richer, and methods are available to evaluate attributes of forecasting performance that are much more relevant for human activities.

forecast can be “fixed,” or provide information to feed into the forecast development process. Finally, such approaches frequently are not “informative” to users—the relevancy of many of the measures to decision making is often unclear.

What makes a good forecast? Traditional verification measures simply evaluate goodness in terms of the overall correspondence between forecasts and observations. However, as a couple of simple examples demonstrate, forecast goodness depends critically on the forecast application (i.e., how the forecast is used) as well as the information about performance that is needed by forecast users or decision making systems.

The first example, in Figure 1, shows a simple spatial forecast, say, of precipitation. When considering only the forecast (Fig. 1a), all standard verification measures would indicate this is a bad forecast; the RMSE would be large, POD would be 0, FAR would be 1, and CSI would be 0. None of these measures indicates that the forecast is the correct size and shape, and that it is displaced only slightly from the observed precipitation region. In the context of predicting precipitation over a watershed (Fig. 1b) the forecast would, indeed, be considered a poor forecast because no precipitation was predicted to fall over the watershed. In contrast, an air traffic manager might consider the forecast to be pretty good (Fig. 1c), because it indicates that the route from A to B will be blocked, which turns out to be the case.

In the second example (Fig. 2), the first forecast and observation pair (Fig. 2a) are equivalent to the example in Fig. 1a—the forecast region is the correct size and shape and is just displaced slightly from the observed region. In the examples in Figs. 2b-d, the forecasts have various additional flaws—having the wrong shape and orientation, and/or greater displacement than in Fig. 2a. In the fifth example (Fig. 2e), the forecast area is much larger than the observed area, but some of the forecast region overlaps the observed region. The traditional verification statistics for the examples in Figs. 2a-d indicate that none of the forecasts has any skill; moreover, all of the forecasts are assigned the same numerical scores (i.e., $POD=0$, $FAR=1$, $CSI=1$; large RMSE), regardless of the form of the forecast error. Moreover, the fifth forecast (Fig. 2e)—which may be considered by many/most users to be the *worst* forecast—has *positive* skill (i.e., $POD>0$, $FAR<1$, $CSI>0$, smaller RMSE). Thus, the traditional scores do not offer information about the relative goodness of these forecasts, or information that could be used to distinguish one forecast from another.

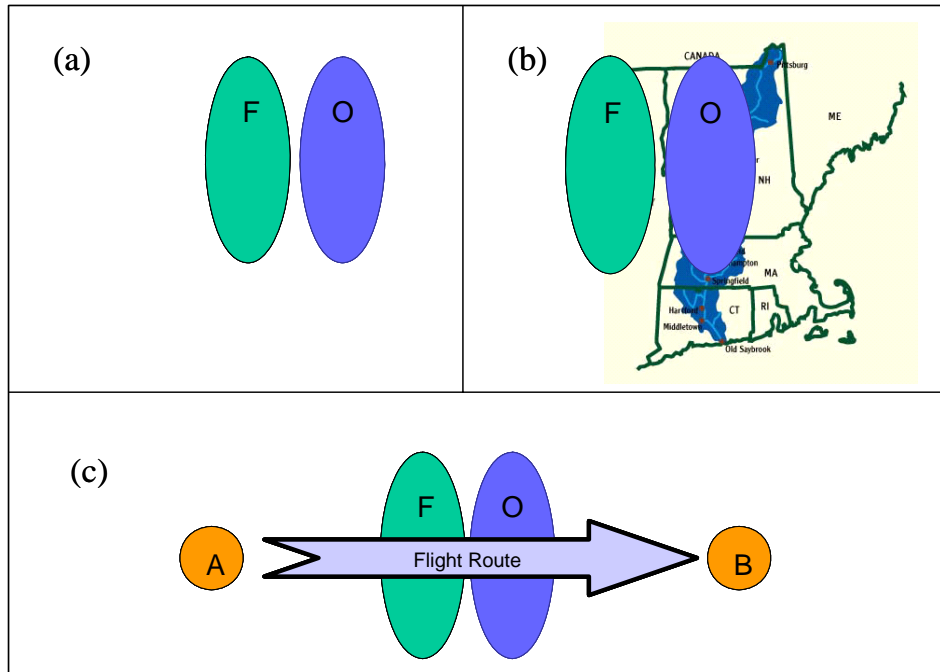


Figure 1. Example showing the importance of how a forecast is used as a determinant of whether it is “good”.

High-resolution forecasts typically show more spatial variation and often include more extreme values than smoother lower-resolution forecasts. Often a subjective evaluation indicates that these forecasts provide a more realistic depiction of the forecasted conditions than is provided by a smoother, lower resolution forecast. However, the high-resolution forecasts typically do not score as well as the low-resolution forecasts when traditional metrics are used (e.g., Mass et al. 2002); in some respects, these statistics penalize the high-resolution forecasts for their increased variation and more extreme values. For example, because the RMSE is based on squared

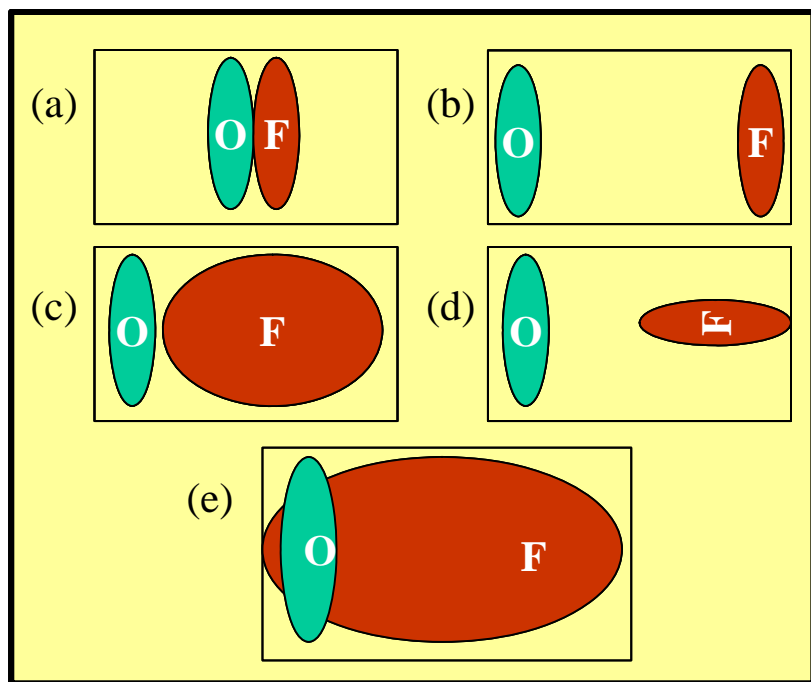


Figure 2. Simple example of the impacts of different types of errors on verification results. Each of the forecasts in (a) through (e) has 0 skill according to traditional verification approaches, whereas the forecast in (e) has positive skill. From Davis et al. (2006a).

errors, large errors have a much greater effect on the scores than smaller errors. Thus, if a correct large value is forecasted but is slightly misplaced, the squared error is much larger than if a smaller (incorrect) extreme value were forecasted in the wrong place (as could happen with a low-resolution forecast). Moreover, as shown by the examples in Figs. 1 and 2, traditional metrics are insensitive to the character or cause of an error, so that even if a detailed forecast correctly *characterizes* future weather conditions (which could be useful information for many applications), standard verification scores will indicate the forecast is poor—or even that it has no skill—if there is any error in the placement of the forecast.

4. Diagnostic verification approaches

The goal of a diagnostic forecast verification approach⁹ is to provide a more “holistic” evaluation of forecasts, by viewing multiple attributes of forecast performance, rather than summarizing performance in a single or few measures. These attributes can represent statistical quantities such as reliability, resolution, accuracy, or discrimination, or they can represent features that are of interest to answer specific questions about the performance of the forecasts. The attributes can be designed to answer specific questions (e.g., What is the average timing error in the forecasts? Is the maximum intensity correct?). In general, diagnostic approaches provide detailed information about forecast quality, and allow the capability to understand what went wrong with the forecasts, as well as what went right. They also provide information to help determine how a forecasting system can be improved, and allow meaningful comparisons to understand how forecasting systems differ from one another.

Figure 3 shows an example of a basic diagnostic technique. This conditional quantile plot shows the distributions of observations given particular forecasts, for a generic forecasting system. This diagnostic plot gives information about the ranges of observed values associated with particular forecast values. For example, in Figure 3, for a forecast of 15, most of the observations (between the 10th and 90th

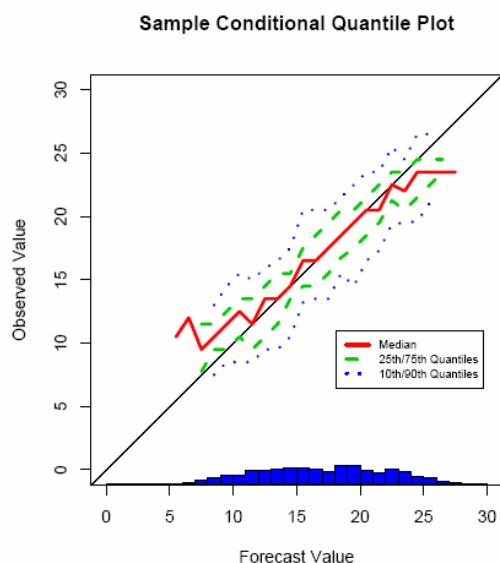


Figure 3. Example conditional quantile plot showing quantiles of the conditional distributions of observations associated with particular forecast values. The median line represents the middle of the distributions, with 50% of the observed values less than this value and 50% larger than this value. The 10th (90th) quantiles represent the values for which only 10% of the observations are smaller (larger); and the 25th (75th) quantiles represent the values for which only 25% are smaller (larger). (Figure courtesy of Matt Pocerlich)

⁹ Murphy and Winkler (1987) set the stage for diagnostic approaches for verification by considering verification in the context of the joint distribution of forecasts and observations. This work was extended to discussions of diagnostic approaches for probabilistic and continuous forecasts (Murphy and Winkler 1992; Murphy et al. 1989). The distributions-oriented diagnostic approach is described by Wilks (2006) and has been applied by Brooks and Doswell (1996) and others.

percentile) are between about 12 and 18. The median values indicate that small forecasts typically underestimate the observed value, whereas large forecasts are overestimates—that is, the forecasts are conditionally biased for extreme values (note that the bias would appear to be 0 if all the forecasts were combined). This kind of information could be quite useful for decision makers in determining how to react, for example, to forecasts of low temperatures. Moreover, it could be used to calibrate and improve a forecasting system.

In addition to the basic diagnostic methods that have been developed for forecasts at a particular location (e.g., “temperature in Denver”), new methods are under development that are specifically focused on diagnostic evaluation of high-resolution spatial forecasts. These methods consider a variety of characteristics of the performance of spatial forecasts (e.g., for precipitation and similar types of weather events), including characteristics that could help improve the forecasts and/or aid in users’ applications of the forecasts.¹⁰

The object-based approach, in particular, focuses on identifying and evaluating forecast attributes that indicate which aspects of a forecast are good and bad; the approach can incorporate particular attributes that are of interest to individual types of users. Figure 4 shows an example of this approach; individual precipitation objects are identified in the forecast and observed grids, and are grouped into “composite” objects in each field. The composite forecast and observed objects are then matched to one another, and a variety of attributes are compared. The results of the object-based comparisons provide many details about particular characteristics of the forecasts. For example, all of the forecast objects in this example were located too far north and too far west, the median forecast precipitation intensities were too large, and the forecast extreme precipitation intensities were too small. In contrast, a traditional verification analysis for this example—POD=0.27, FAR=0.75, CSI= 0.34—clearly provides information that is much less rich and informative.

5. Moving toward increased user focus

Diagnostic approaches are one way to make verification information more meaningful to users. Certain forecast quality attributes are clearly more meaningful to certain users than others, and the diagnostic approaches allow a wide variety of attributes to be computed. Nevertheless, by themselves, diagnostic approaches only go part of the way toward providing truly beneficial user-relevant information.

An obvious additional step is to work directly with users to determine what forecast performance information would be most useful, and to produce verification information that is tailored to specific applications. This approach would be especially beneficial, for example, for identifying the forecast performance information required by an objective decision support system.

Stratifying verification results according to meaningful subgroups is a second—and easily implemented—approach for making verification results more user-focused. Aggregation of results across large regions and time periods can lead to the loss of a great deal of useful information. For example, verification results for long-range and seasonal forecasts issued by the

¹⁰ Examples of these methodologies include scale separation approaches (e.g., Casati et al. 2004; Tustison et al. 2003; Zepeda-Arce et al.2000); an entity approach (e.g., Ebert and McBride 2000); a composite approach (e.g., Nachamkin 2004; Nachamkin et al. 2005); fuzzy approaches, which take into account uncertainties in the forecasts and/or observations (e.g., Atger 2001, Brooks et al. 1998); and several object-based approaches (e.g., Baldwin 2004; Brown et al. 2004; Davis et al.2006a,b; Marzban and Sandgathe 2006).

NWS’s Climate Prediction Center are currently presented on the NWS web page as aggregate statistics over the entire continental United States; breaking these results into regional statistics would allow users to know in which parts of the country the forecast performance is good enough for the forecasts to possibly be useful.

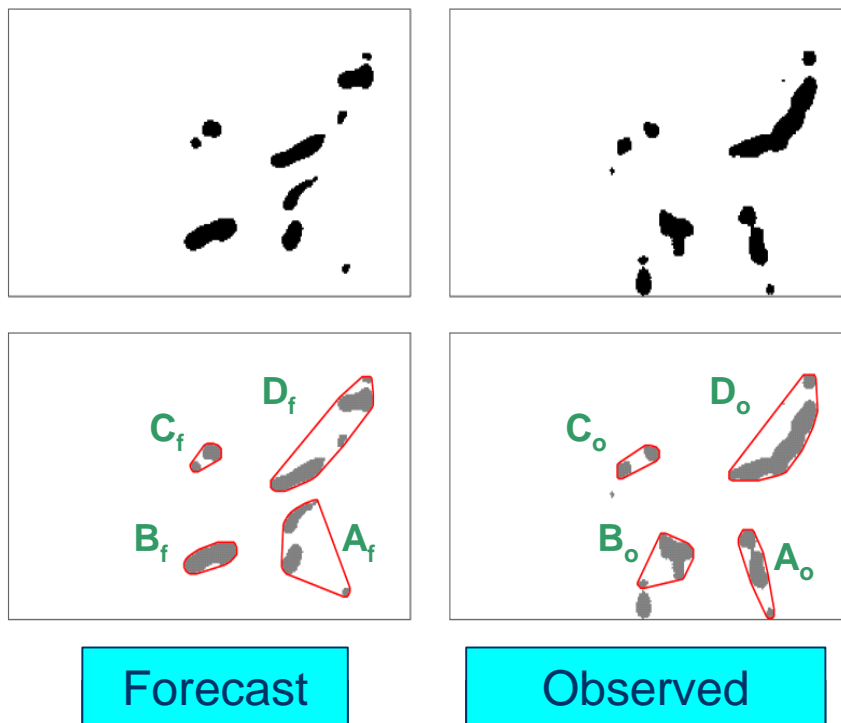


Figure 4. Example of an application of an object-based verification approach for precipitation forecasts.

Another important aspect of verification that has usually been ignored, as noted in Section 2, is the uncertainty in the verification measures themselves. For users to make rational decisions, and for managers and forecast developers to make appropriate choices of model enhancements, this uncertainty must be computed and communicated.

Finally, it is critical to consider how users can make the best use of forecast verification information. Objective decision support systems can directly apply appropriate types of verification information—in fact, verification information provides the link between forecasts and their value for problems where the decision-making situation is well understood. A simple example is the reliability of probability forecasts, which is an obvious direct input to a decision support system that employs probabilistic weather forecasts¹¹. In addition, verification information represents base-level uncertainty information about forecasts, and should at the very least be provided in that context. For example, the fact that the performance of temperature forecasts degrades with lead time is relevant and meaningful information for most users. Much more investigation is needed regarding users’ needs for and applications of this type of uncertainty information, to help understand the forms and types of information that should be provided for specific users.

¹¹ Reliability indicates the probability that the forecasted event will occur, given a particular forecast.

6. Levels of user focus in verification

Although traditional verification approaches have limited benefits for most users (other than for the administrative monitoring activity) it clearly is possible to measure the quality of various forecast attributes in ways that would provide meaningful information for a variety of users. Moreover, although the relationship between forecast quality and value is complex, verification approaches can be designed that have relevance for the estimation (or at least understanding) of forecast value. To foster development of approaches that will meet the needs of specific users and applications of forecast quality information, we propose a hierarchy of five levels of user-relevant verification information (see Table 1), where Level 0 represents the conventional application of scalar measures. Moving through this hierarchy leads to verification approaches that produce information that is meaningful for specific forecast users, and finally to Level 4 where forecast value and usefulness are estimated directly. The five levels are characterized as follows:

Level 0

Only conventional measures-oriented verification is applied, with only one or two measures reported. The verification results are aggregated across broad regions and time periods, with no information about the uncertainty in the verification statistics.

Level 1

Broad diagnostic approaches are applied, which provide a great deal of information about a variety of attributes of forecast performance. Results are stratified into meaningful categories (e.g., for climatologically homogeneous regions), and uncertainty information (associated with the verification statistics themselves) is provided either directly or through distributions of errors. Much of the information is presented graphically and with user-selectable thresholds and stratifications. Even this level of verification is a very large advance approaches commonly applied today.

Level 2

Features-based and advanced spatial approaches are incorporated for the evaluation of spatial forecasts or forecasts across time; enhanced diagnostic approaches are also applied. The features evaluated have implications for the utilization of the forecasts by a variety of users, and can be defined to meet the needs of particular types of users. These users can infer information about the value/usefulness of the forecasts.

Level 3

Users are included in the development process for verification approaches. Specific verification information is designed to meet the needs of these users, and verification results are stratified into categories that users identify as meaningful. Users also specify particular thresholds of interest for their application. Specialized datasets that are relevant for a particular decision-making situation may also be incorporated into the verification analyses. In addition to enabling the development of user-focused verification approaches, the interactions with users will facilitate the development of bridges between the weather community and user groups, which will benefit many aspects of the forecasting process.

Table 1. Hierarchy of levels of forecast evaluation approaches.

Level	Description	Uses/users	Aggregation	Uncertainty in verification statistics	Value
0	<ul style="list-style-type: none"> ▪ Measures-oriented verification ▪ One or two statistics provided 	Administrative only	Broadly aggregated both spatially and temporally	Not considered	No information
1	Broad diagnostic approaches	<ul style="list-style-type: none"> ▪ Administrative ▪ Forecast developers ▪ Some users 	Stratified into relevant categories	Provided directly, or indirectly through distributions of errors	No direct information
2	Features-based and enhanced diagnostic approaches added to Level 1	<ul style="list-style-type: none"> ▪ Administrative ▪ Forecast developers ▪ Broad range of users 	Stratified into relevant categories	Provided	May be inferred by individual users
3	Level 2, plus specific information provided for particular users	<ul style="list-style-type: none"> ▪ Administrative ▪ Forecast developers ▪ Broad range of users ▪ Specific users 	Stratified into user-specified categories	Provided	Can be inferred by individual users
4	In addition to above, economic or cost-loss models, or survey methods, are used to assess the value or benefits of forecasts for specific users	<ul style="list-style-type: none"> ▪ Administrative ▪ Forecast developers ▪ Broad range of users ▪ Specific users 	<ul style="list-style-type: none"> ▪ Stratified into user-specified categories ▪ May aggregate across and among sectors 	Provided	Directly estimated

Level 4

The ultimate level of forecast evaluation represents the direct evaluation of forecast value for specific users, in addition to the verification information from Level 3. Verification information from Level 3 is utilized by decision-making and economic models. Information from this level can directly guide users on how to optimally use forecast information. Because it will be difficult to produce forecast value information for all user groups and all types of forecasts, users will continue to need information from Levels 1-3 as guidance.

This hierarchy offers a model for how the process of forecast verification/evaluation can develop and mature to provide more meaningful and relevant information for users. As discussed in previous sections, most verification currently addresses Level 0, with some examples at Level 1 and new research at Level 2. A few efforts have focused on Level 4, but only very limited efforts have included Level 3 activities.

7. Research challenges

Through its focus on societal and economic impacts of high impact weather forecasts, THORPEX offers an exciting opportunity to develop improved forecast evaluation approaches at Levels 1 through 4, as outlined in Table 1.

First, new methods should be developed and applied to produce a wide variety of verification information at Level 1 (broad diagnostic approaches). Current methods were developed many years ago and could easily be supplemented with a variety of new approaches (e.g., ways to view and summarize forecast, observation, and error distributions). Second, features-based approaches (Level 2) need further development so that they can be easily applied in practice. Both the diagnostic and features-based approaches need to be expanded to consider ensemble and probabilistic forecasts.

Third, initial efforts should begin soon to connect and work with specific sets of users to identify meaningful forecast attributes and verification approaches that will provide useful information for these users. The focus should initially be on individuals and groups in one or two sectors who are relatively sophisticated users of forecast information (e.g., water managers, electric utility forecasters, public weather forecasters). The team working on development at this level, as well as Level 4, must include both social scientists and application area experts.

Finally, many more studies and much more method development must be undertaken to achieve Level 4 (estimation of forecast value). Although several studies have focused on evaluations of forecast value for certain applications or sectors, these efforts have typically been very limited in scope and have not been done frequently enough to have an impact on the forecast development process. Moreover, these studies have not had the benefit of advanced verification methods as input to the forecast value estimation process, and so may not appropriately represent the benefits associated with meaningful improvements to the forecasts. Thus, in order for Level 4 efforts to be successful, Levels 1-3 must be appropriately addressed. Moreover, because the forecast development process cannot be directed by the needs of all users and the forecast value associated with all applications, a broad range attributes of forecast performance must be considered, in order to appropriately assess directions for future forecast development.

THORPEX field programs such as the Pacific-Asian Regional Campaign (PARC) and the THORPEX Interactive Global Grand Ensemble (TIGGE) project both provide opportunities for advancing verification methods and developing user-focused approaches. For example, field programs provide an opportunity to test new methods in practice, and TIGGE will produce a wealth of forecasts that can be used for extensive development and testing of new verification approaches. In both cases, the application of social science methods will be required to optimally develop the new verification methods.

Suggestions for specific research activities include the following:

- Select one or more forecasting systems (e.g., TIGGE, PARC) to utilize in the testing and development of diagnostic verification approaches (Level 1), to extend currently available approaches and answer new questions about forecasting performance.
- Enhance features-based approaches to be able to answer a wide variety of questions about forecasting performance for a number of different weather variables. Extend features-based approaches to the evaluation of ensemble forecasts.

- With one or more user groups (e.g., aviation, water resources) develop a process for identifying user-focused verification attributes. Develop a common language for discussing forecasting performance, investigate decision-making strategies, and translate the information needs into verification attributes.
- Coordinate with specific studies of forecast use and value for which appropriate verification information is a vital component. Develop meaningful verification attributes for use in these studies and investigate the relationship between forecast quality (as measured by these attributes) and forecast value.

8. Acknowledgments

This paper was greatly improved through comments from and discussions with many colleagues, including Mike Baldwin, Harold Brooks, Barbara Casati, Chris Davis, Beth Ebert, Jeff Lazo, Jennifer Mahoney, Rebecca Morss, Paul Roebber, and Laurie Wilson. I would like to thank them all for their important contributions to the ideas presented.

9. References

- Atger, F., 2001: Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Proceedings in Geophysics*, **8**, 401-417.
- Baldwin, M., 2004: Object identification techniques for object-oriented verification. Int'l Verification Methods Workshop, 15-17 Sept. 2004, Montreal. Presentation available at <http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/Workshop2004/MeetingProgram.htm>.
- Brier, G.W., and R.A. Allen, 1951: Verification of weather forecasts. In *Compendium of Meteorology*, American Meteorological Society, 841-848.
- Brooks, H.E. and C. A. Doswell III, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Weather and Forecasting*, **11**, 288-302.
- Brooks, H.E., M. Kay and J.A. Hart, 1998: Objective limits on forecasting skill of rare events. 19th Conference on Severe Local Storms, American Meteorological Society, 552-555.
- Brown, B.G., R.R. Bullock, C.A. David, J.H. Gotway, M.B. Chapman, A. Takacs, E. Gilleland, K. Manning, J. Mahoney, 2004: New verification approaches for convective weather forecasts. 11th Conference on Aviation, Range, and Aerospace Meteorology, 4-8 Oct 2004, Hyannis, MA. Available at <http://ams.confex.com/ams/pdfpapers/82068.pdf>.
- Casati, B., G. Ross, D.B. Stephenson, 2004: A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorological Applications*, **11**, 141-154.
- Clayton, H. H., 1891: Verification of weather forecasts. *American Meteorological Journal*, **8**, 369-375.

- Davis, C., B. Brown, and R. Bullock, 2006a: Object-based verification of precipitation forecasts, Part I: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, **134**, 1772-1784.
- Davis, C., B. Brown, and R. Bullock, 2006b: Object-based verification of precipitation forecasts, Part II: Application to convective rain systems. *Monthly Weather Review*, **134**, 1785-1795.
- Ebert, E.E. and J.L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *Journal of Hydrology*, **239**, 179-202.
- Finley, J. P., 1884: Tornado predictions. *American Meteorological Journal*, **1**, 85–88.
- Gilbert, G. K., 1884: Finley’s tornado predictions. *American Meteorological Journal*, **1**, 166–172.
- Jolliffe, I.T., and D.B. Stephenson, 2003: *Forecast Verification: A Practitioner’s Guide in Atmospheric Science*. Wiley, 240 pp.
- Köppen, W., 1893: The best method of testing weather predictions. *U.S. Weather Bureau Bulletin*, **11**, pp. 29–34.
- Marzban, C. and S. Sandgathe, 2006: Cluster analysis for verification of precipitation fields. *Weather and Forecasting*, submitted.
- Mass, C.F., D. Ovens, K. Westrick and B.A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bulletin of the American Meteorological Society*, **83**, 407-430.
- Murphy, A.H., 1991: Forecast verification: its complexity and dimensionality. *Monthly Weather Review*, **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., 1996: The Finley Affair: a signal event in the history of forecast verification. *Weather and Forecasting*, **11**, 3-20.
- Murphy, A.H., and R.L. Winkler, 1987: A general framework for forecast verification. *Monthly Weather Review*, **115**, 1330-1338.
- Murphy, A.H., and R.L. Winkler, 1992: Diagnostic verification of probability forecasts. *International Journal of Forecasting*, **7**, 435-455.
- Murphy, A.H., B.G. Brown and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Weather and Forecasting*, **4**, 485-501.

- Nachamkin, J.E., 2004: Mesoscale verification using meteorological composites. *Monthly Weather Review*, **132**, 941-955.
- Nachamkin, J.E., S. Chen and J. Schmidt, 2005: Evaluation of Heavy Precipitation Forecasts Using Composite-Based Methods: A Distributions-Oriented Approach. *Monthly Weather Review*, **133**, 2163–2177.
- Nehrkorn, T., R.N. Hoffman, C.Grassotti and J.-F. Louis, 2003: Feature calibration and alignment to represent model forecast errors: Empirical regularization. *Quarterly Journal of the Royal Meteorological Society*, **129**, 195-218.
- Nichols, W. S., 1890: The mathematical elements in the estimation of the Signal Service reports. *American Meteorological Journal*, **6**, 386–392.
- Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454.
- Stanski, H.R., L. J. Wilson and W. R. Burrows, 1989: Survey of common verification methods in meteorology. Atmospheric Environment Service, Forecast Research Division, WMO World Weather Watch Technical Report No.8, WMO/TD No. 358 (Available at http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/Stanski_et_al/Stanski_et_al.html).
- Tustison, B., E. Foufoula-Georgiou, and D. Harris, 2003: Scale-recursive estimation for multisensor Quantitative Precipitation Forecast verification: A preliminary assessment. *Journal of Geophysical Research*, **108**, D8, 8377.
- Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences*. Elsevier, San Diego.
- Zepeda-Arce, J., E. Foufoula-Georgiou, and K.K. Droegemeier, 2000: Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *Journal of Geophysical Research*, **105** (D8), 10,129-10,146.