Summary of Workshop on Forecast Verification:

Making Verification More Meaningful

Barbara Brown[1], Agnes Takacs[1], Jennifer Mahoney[2],

and Tressa Fowler[1]

December 2002

[1] Research Applications Program, National Center for Atmospheric Research, Boulder, CO 80307-3000
[2] Forecast Systems Laboratory, National Oceanic and Atmospheric Administration, Boulder, CO 80303

ABSTRACT

A workshop titled “Making Verification More Meaningful,” which took place from 30 July to 1 August 2002, brought together an international group of researchers and operational meteorologists and hydrologists at the National Center for Atmospheric Research in Boulder, Colorado. The workshop focused on development of advanced diagnostic verification approaches, operational and user issues, observational concerns, and verification of ensemble forecasts. A number of issues in these areas were discussed, and a variety of options for solutions were presented. The workshop was sponsored by the Federal Aviation Administration’s Aviation Weather Research Program in response to the many difficulties are associated with evaluation of aviation weather forecasts. A number of important issues that are common across fields (e.g., in model development, severe storm forecasting, hydrology, aviation weather) were raised and discussed. Some of these issues, which require further investigation and development, include observational and forecast scaling questions; development of operationally-relevant verification approaches; and the need for improved approaches for verification of spatial forecasts, particularly as model/forecast resolution increases. Workshop participants also identified the need for additional educational opportunities in forecast verification (e.g., through AMS short courses, university curricula, COMET classes) and the desirability of future workshops with a similar theme.

1. Introduction

A workshop on forecast verification, titled "Making Verification More Meaningful," was held on 30 July – 1 August 2002 at the National Center for Atmospheric Research (NCAR), in Boulder, CO. The workshop was sponsored by the Federal Aviation Administration's Aviation Weather Research Program (AWRP) and was organized by the verification teams at NCAR and the National Oceanic and Atmospheric Administration's Forecast System Laboratory (NOAA/FSL). NCAR also provided travel support for some of the invited speakers and students. (The workshop web page is: http://www.rap.ucar.edu/research/verification/ver_wkshp1.html.)

The primary goals of the workshop were to bring together individuals who are working on difficult issues in verification, to establish collaborations and to share ideas, and to discuss particular problem areas. For the AWRP, the workshop was an opportunity to obtain new ideas from the general verification community, as well as to discuss our intrinsic verification concerns.

For a number of reasons, the workshop generated a great deal of interest from the community and was very well attended. Specifically, these reasons include the following: (a) the verification community is fairly small, with relatively few opportunities for collaboration; (b) forecast verification is an essential component of any weather forecasting system; (c) establishing the credibility of forecasting systems has become more critical, in an era when program benefits must be demonstrated; (d) the need for "operationally-relevant" verification measures has become widespread; and (e) perhaps most importantly, the science of verification is undergoing major changes and development, as standard methods have been found to not meet the needs associated with high-resolution gridded forecasts. As a result, more groups and individuals were interested in the workshop than had been anticipated. The approximately 90 participants (see Fig. 1) included meteorologists, hydrologists, statisticians, mathematicians,

researchers, and operational staff members from several countries, from weather services, universities, and research institutes. In addition to many areas of the United States, participants came from Canada, England, Scotland, Australia, and Finland.

The workshop included several components: invited speakers, contributed talks, a poster session, working group meetings and reports, and a panel discussion.

2. Presentations and posters

The workshop included nine invited presentations, twenty contributed oral presentations, and ten poster presentations. Most of the presentations and posters are available on the workshop web site at http://www.rap.ucar.edu/research/verification/pres.html. The invited speakers were selected to provide three perspectives on each of the three main areas of interest: *User and Operational Issues, Scaling and Observations, and Advanced Methods*. Many new ideas were presented in both contributed and invited talks; however, this summary will focus primarily on the invited talks; the web page can be used to refer to the contributed presentations.

2.1 Opening remarks

The opening remarks (by Barbara Brown) were designed to provide motivation for the workshop. Some of the issues in aviation weather forecast verification that were identified include problems associated with non-systematic, biased observations; working with three-dimensional gridded forecasts; scaling questions and appropriate approaches for matching forecasts and observations; and the need for verification approaches that are operationally meaningful. The related issues to be considered at the workshop included

- How to cope with various attributes of observations (good, bad; gridded, points…)
- Can more "meaningful" approaches be developed? (For Modelers? Forecasters? End-users?)
- Verification of finer scale models and ensemble forecasts – how should methods change?

Some motivational quotations regarding forecast verification and its role in science and forecasting were also included:

As A.H. Murphy (personal communication, April 1990) pointed out,

> *Forecast verification (i.e., the quantitative assessment of forecast quality) is an essential component of any weather forecasting system. Information concerning the quality of forecasts is needed… to monitor forecast performance and by so-called "end-users" to make the best possible decisions. Yet, forecast verification procedures currently in place… are quite primitive (from a methodological point of view) and generally fail to meet these fundamental information needs in a satisfactory manner."*

In a preface to a 1951 panel discussion on forecast verification, R.A. Allen remarked that

> "*Making a weather forecast is like conducting an experiment. One measures the initial conditions and sets up a hypothesis as to the outcome. To omit verification is comparable to conducting an experiment without learning the result."*

and that

> "*…there is a pressing need for verification systems which compare forecasts with subsequent weather in a useful way.*" (Allen et al. 1952)

Finally, in the same 1951 panel discussion, G.W Brier made the important point that

> "*…the 'best' forecast according to an accepted system of arbitrary scores may not be the most useful forecast.*" (Allen et al. 1952)

With these words of wisdom as guidance, the workshop was underway.

2.2 User and operational issues

Harold Brooks opened his talk with a defining quote from the past, which summarizes some of the problems associated with verification:

"*It is hazardous, and in many ways irrelevant to become entangled with any form or variety of verification schemes. No matter where one stands, one is always in the middle of a lusty controversy. The verification expert and the meteorologist are always ravishing each other."* (Holzman, 1947)

He then identified and considered the three aspects of forecast goodness that were defined by Murphy (1993):

a. *Consistency*: The correspondence between the forecaster's true beliefs and the actual forecast,
b. *Quality*: The correspondence between the forecast and the observations,
c. *Value*: the incremental benefit to users because of the use of the forecasts in making decisions.

As Brooks pointed out, measures of forecast quality are not equivalent to measures of forecast value. He also focused on the application of a "distributions-oriented approach" to verification, and demonstrated that it provides more meaningful information to the users of the verification information.

Ian Mason provided a discussion of the signal detection theory (SDT) approach to forecast verification, and economic interpretations of SDT results in the context of the cost-loss ratio decision-making model. A major conclusion of this work is that there is a large loss of information – and consequently, economic value – associated with use of non-probabilistic forecasts.

Finally, **Bruce Landsberg** provided a lively discussion of aviation weather and forecast verification from an aviation user's perspective. In particular, he discussed the needs by pilots and other aviation personnel for clearly defined weather information (i.e., it is not necessary for

the information to be completely accurate, but it must be well-defined). Landsberg also described a program implemented by the Aviation Operators and Pilots Association (AOPA) designed to collect more pilot reports, which will help to increase safety as well as contribute to research.

Contributed talks and posters on this general topic included discussions of the use of verification information to improve forecasts (P. Leftwich); various operational verification systems that either exist or are under development (P. Nurmi, J. Mahoney, N. Cajina, B. Glahn, and C. Kluepfel); methods that are appropriate for different types of users (F. Mosher and J. Evans); methods that can be applied to obtain the best information from probabilistic forecasts (I. Mason); and an application of the SDT model to estimate the value of probabilistic forecasts to aviation (R. Keith).

2.3 Scaling and observations

Efi Foufoula-Georgiou discussed a breadth of research that has been undertaken in hydrology regarding the representativeness error associated with the choice of approaches for matching gridded forecasts to point observations; scaling of precipitation fields (e.g., from radar or model output); and the impacts of the scaling effects on forecast verification results. These issues are of great concern in aviation weather forecast verification (e.g., for convective weather) and for numerical weather prediction model verification.

Mike Kay described the development and application of a method to define “practically perfect” hindcasts based on observations. For example, such an approach can be used to identify the best possible forecast that could be given, based on a sparse set of severe storm (i.e., tornado, hail, damaging wind) reports. The practically perfect forecasts are created by applying a statistical model to the spatial distribution of the observations. By comparing the practically perfect forecasts to the actual forecasts, it is possible to identify how far the actual forecasts are

from “practical perfection.” This approach makes it possible to provide meaningful feedback to forecasters and to determine objective limits to forecast skill, and it provides a useful visual aid for subjective forecast verification.

Ed Tollerud considered the impacts of observation error on forecast verification measures. Sources of uncertainty in verification data include random observations, systematic errors (e.g., due to observing practices), and natural variability and representativeness error. Comparisons of precipitation observations from neighboring stations – which could be considered an upper bound on the capability of precipitation forecasts – indicated that the maximum scores are quite limited, simply due to spatial variations in precipitation. Tollerud suggested that methods should be formulated to distinguish natural variability from truly observation-driven variability.

Contributed talks and posters related to this topic included discussions of the use of gridded observations for verification (A. Ghelli, M. Chapman); observation errors for space-based instruments (R. Frehlich); and observational issues that make convective forecast verification very complicated and difficult (C. Mueller).

2.4 Advanced methods

Beth Ebert considered various traditional and new approaches for verification of spatial forecasts. She identified the strengths and weaknesses of these approaches relative to a checklist of various forecast quality attributes that verification approaches should be able to evaluate. These attributes are the location, size, shape, mean value, maximum value, and spatial variability. The bottom-line approach to be mimicked objectively is visual verification (“Does the spatial forecast look right?”), which is able to consider all of the attributes of interest, but unfortunately is labor intensive and not objective. The verification approaches considered

include continuous statistics (e.g., mean absolute error, correlation); categorical statistics (e.g., POD, FAR, Bias); scale decomposition methods (e.g., discrete wavelet transforms); multiscale statistical organization methods, which examine how well the multi-scale attributes of the forecasts match those of the observations; entity-based methods, which use pattern matching to evaluate the displacement between forecast and observed entities and to decompose the error into various components (e.g., displacement, pattern, volume); object-based methods, which characterize forecast and observed regions in a natural way as geometric objects and consider errors associated with location, shape, size, and intensity; and event-oriented approaches, which verify the bulk properties of defined events (e.g., through compositing or classification of events). Ebert concluded that visual verification is the most effective approach, and that categorical statistics based on Yes-No forecasts and observations are the least informative. New methods (e.g., event- or object-based approaches) provide a more complete picture of forecast performance.

Mike Baldwin discussed the need to consider various viewpoints when determining approaches for verification of forecasts: the information needs of the forecaster, the modeler, and the weather information consumer are not the same. Traditional (measures-oriented) verification approaches cannot respond to these diverse needs. In addition, traditional approaches penalize forecasts that have a large amount of spatial detail, even though that spatial detail may be of considerable value to certain users. Baldwin proposed development of approaches to characterize the observed and forecast fields. Such characterization would provide users with more detailed information about the forecasts and their quality. Subjective characteristics of the forecasts (e.g., forecaster confidence) also could provide meaningful information to users. Measures of

similarity can be used to characterize differences between forecasts and observations. The challenge lies in distilling a large amount of information into useful "nuggets."

Laurie Wilson provided an overview of verification approaches available for evaluation of ensemble forecasts. The fundamental question here is how to compare a distribution to a single value. Ensemble forecast verification can include verification of the ensemble distribution, verification of individual ensemble members, and verification of probability forecasts based on the ensemble. Approaches for evaluation of the ensemble distribution include a distribution model comparison approach developed by Wilson; the rank probability score; the continuous rank probability score; and the rank histogram. Each approach measures different attributes and has different positive and negative aspects. Wilson suggested that the ensemble mean is not a meaningful forecast based on the ensemble distribution because it is not necessarily a trajectory of the model. Verification approaches for probability forecasts based on the ensemble distribution are the same as approaches for verification of other types of probability forecasts. In fact, Wilson suggested that reliability diagrams and the SDT relative operating characteristic (ROC) diagram provide a sufficient approach for verification of probability forecasts. Wilson concluded that the choice of verification approaches for ensemble forecast verification depends on how the ensemble forecast is to be used by the forecaster/user.

Contributed talks and posters in this session concerned a variety of topics, including statistical issues such as confidence intervals and power (T. Fowler and I. Jolliffe); ensemble forecast verification (K. Franz, A. Ghelli, and T. Gneiting); verification of precipitation and lightning (A. Loughe, C. Davis, W. Burrows, and W. Gallus); and development and implementation of new approaches (B. Brown, E. Ebert, B. Casati, S. Sandgathe, J. Nachamkin, and E. Grimit).

3. Working groups

Four working groups were convened during the workshop. Each group considered a general topic related to the three main themes of the workshop. Leaders for each group, in general, were selected from among the invited speakers[3]. The working groups met two times for 90 minutes.

3.1 User and operational aspects

Major recommendations from this working group include the following: (i) verification methods should be developed that are appropriate for each user group; (ii) the forecasts and the rules for verification should be clearly defined; (iii) the confidence in the forecasts should be included in the forecast (i.e., probabilistic forecasts are optimal) – if the phenomenon is hard to predict, the forecast should indicate that fact; (iv) verification statistics should reflect their value relative to perfect forecasts (note that this means we need to be able to define what a "perfect" forecast is); and (v) verifiers need to work closely with end users.

3.2 Advanced diagnostic methods

High- and low-resolution forecasts provide different types of information. A finer-scale model may perform better than a coarser scale model according to the attributes that are considered to be important, but standard verification techniques may suggest that it is worse. The benefit of going to higher resolution in numerical models may not be on the finest scales, at least in a deterministic sense.

Time scale issues also may mean that a high-resolution model gives a poorer forecast than the forecast provided by a coarse model later in the forecast period. Verification approaches

[3] The working group leaders were: (A) User and operational aspects – Ian Mason and Mike Kay; (B) Advanced diagnostic methods – Mike Baldwin and Beth Ebert; (C) Scaling and observations – Harold Brooks and Ed Tollerud; and (D) Ensemble verification methods – Laurie Wilson and Anna Ghelli.

are needed that do not penalize high-resolution forecasts of weather that occurs on unpredictable time scales.

Higher-resolution models need to be given a proper evaluation that measures their operational capabilities. It may be necessary to rely more on entity-based techniques, pattern recognition, and other scientific techniques, which are more appropriate for finer scales. It will be necessary to verify fine-scale forecasts using remotely sensed data, including satellite and radar observations, and to develop forward models that produce forecast "events" that are defined in the same way as the observed events (e.g., forecasts of reflectivity).

Another problem arises when forecasts on different spatial scales need to be compared. Translating them both to the same scale is not the best approach because it does not provide information about how well the finer-scale forecast performs on its own scale. Ideally, comparison of forecasts on different scales should be applied only to scales that are not very different from each other. One possibility would be to "ramp down" the scales to do comparisons. More advanced verification techniques (e.g., object or entity-based approaches) might not make sense for coarser scale models. Other approaches for moving between scales – such as wavelet transforms – might be more appropriate than simple averaging across a grid.

Measures of observational uncertainty could provide an idea of the possible precision of the verification. Three possible approaches for coping with observational uncertainty include (i) the practically perfect forecast approach (presented by Mike Kay); (ii) application of a fuzzy verification approach (e.g., Ebert poster); and (iii) defining a probability density function (pdf) for the observations. The practically perfect forecast approach makes it possible to move discrete observations to a better-behaved and more reasonable field. A pdf of observations could be re-sampled to obtain a distribution for the verification score. Another possibility would be to

employ Monte-Carlo simulations to test the possible existence of an event where observations are lacking.

Scale decomposition methods allow scale-related errors to be teased out of the overall error. Wavelets provide information on both averages and differences between points, and they allow reconstruction of the field at desired scales. However, “upscaling” is a mathematically simpler approach that may be more easily accepted and understood; for example, with this approach the threat score could be presented as a function of scale. Similarly, conditional rainfall amount could be verified as a function of scale (i.e., depth vs. area). Another possibility would be to develop a hybrid scale decomposition/entity approach; that is, a wavelet technique could be applied, and then objects at the various scales could be identified and verified.

New diagnostic approaches can and should be developed that address specific operational questions (e.g., flight path impacts; hydrologic forecasts). Some statistics can and should be tailored toward what particular users need to know. Verification “experts” should work with those users who have specific needs to jointly develop techniques that address their verification problems. However, methods that are developed through these specific applications should be made as general as possible so that they can be applied more widely. Graphical depictions should be very useful for many of these applications, with pictures and numbers presented together.

One approach to utilize advanced statistical models (e.g., spatial models) to improve verification approaches would be to take into account spatial correlation (e.g., a spatial structure function, correlogram, EOF analysis, etc.). These models also can be used for quality control (e.g., some decision trees like CART, and wavelets). Data mining methods – techniques that help to find useful patterns and knowledge out of a large database – could help reduce the large data

volume and find the meaningful information (however, users should be cautious about "black box" packages where it may be difficult to know what the data mining system is doing).

Typically, spatial models operate on a few specific scales – large scale such as global, and much smaller scales like rain cells (stochastic properties). For verification, it is necessary to access both the spatial structure and time evolution. One approach would be to stratify the error as a function of the properties of the object of interest. Scale decomposition methods could be useful, or it might be useful to apply mathematical/statistical models to describe and verify scale-dependent features.

The working group recommended that advanced education programs in meteorology should include courses on statistical methods. Web-based education also might be useful. A COMET course or additional AMS short courses on forecast verification would be valuable. In addition, a "toolbox" of approaches that is generally available would be very useful. This toolbox could be associated with a verification web site, which would include advice on which methods are appropriate for different applications, and a section for frequently asked questions (FAQs)[4]. The web site should consider the interpretation and meaning of a wide variety of verification methods. In addition, software might be included, as appropriate. The Weather Research and Forecasting (WRF) model web site for working group seven would be a possible site to host a verification toolbox, or provide links to it. The web site also should be linked from the American Meteorological Society (AMS) web page.

3.3 Scaling and observations

The basic question concerning observations, in the context of forecast verification, is "What is reality?" A related issue concerns determining/knowing the error characteristics of the

[4] Subsequent to the workshop, development of this website was begun as an activity of an *ad hoc* verification group (see http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html)

observations and observation systems. Another verification question relates to how we can make use of remotely sensed data for verification – what are the characteristics of the data that need to be taken into account, and how should the observations be compared to the forecasts?

Unfortunately, there frequently is a fundamental mis-match between forecasts and observations. In particular, forecast events frequently are not events that we observe. Additionally, disputes about what a model forecast actually represents are common. This situation is undesirable.

Evaluating the climatology of a model – i.e., measuring how well the model reproduces long-term weather conditions – can provide an upper bound on what can be expected of the forecasts. This comparison should be done in terms of the forecast and observed pdfs. Novel displays of forecasts and observations can shed light on the quality of the forecasts. Bounds on reasonable expectations for the maximum verification scores can/should take into account observational uncertainty.

3.4 Ensemble verification methods

The main verification-related issues in the area of ensemble forecasts involve predictability; use and verification of the ensemble mean as a predictor; and training. With regard to predictability, the questions concern at what projections it is desirable to switch between deterministic and probabilistic forecasts, and when it is best to switch to climatology. Skill scores can be used to aid in this determination; however, it is important to keep in mind that forecasts may have positive economic value even when they have negative skill. Some work is currently in progress to determine the relationship between ensemble spread and forecast error variance, which will help in these evaluations.

With regard to use of the ensemble mean as a predictor, the ensemble mean tends to score better than other ensemble-based predictors when evaluated using standard scores (due to its smoothing effect), but it is not likely to score as well when spatial or diagnostic methods are applied. Instead of the mean, it was suggested that forecasters could better use information about the ensemble distribution contained in a box and whisker plot. The extreme members of the ensemble also are useful because they provide information about the range of possible outcomes.

The training issue concerns the interpretation and use of verification methods: many users of the forecasts (including forecasters) may not be able to easily interpret information from the verification analyses. A related issue concerns understanding and applying appropriate approaches to make use of forecast pdfs in forecasting.

4. Panel discussion

The panel discussion was designed to provide a forum for a group discussion of difficult issues/questions related to verification, from a variety of perspectives. The panel was composed of individuals with a variety of backgrounds, and included Dave Pace, Barry Schwartz, Harold Brooks, Bill Gallus, and Jennifer Mahoney. A set of eighteen probing questions was created using suggestions from workshop participants and organizers. Each panel member selected one or more of the questions and led a discussion of that issue with all of the workshop participants. Due to time limitations, only six questions were actually discussed.

How can the goodness of a product be expressed simply and in a manner that is intuitive to users (e.g., as a "confidence level")?

The idea of the confidence level is to provide a simple scale that will tell users instantly how well the forecasts are performing. Historical information from past verification analyses could be used to provide this indicator; one possible common measure would be the d' measure

from SDT verification approaches, which could be transformed to a 0-10 scale. Information from ensemble forecasts (e.g., ensemble spread) could provide more specific information for particular forecast occasions. The use of probability forecasts was suggested as one possible approach, but some participants suggested that these forecasts are not likely to be used operationally in aviation and other application areas in the near future. Box and whisker plots also could be used to convey information about forecast quality.

Are statistical significance tests mis-used or under-used in verification studies?

Significance tests are often mis-used, due to questions regarding the validity of the underlying assumptions. However, significance tests also may not be used often enough. A paper by N. Nicholls, titled "The Insignificance of Significance Testing" (Nicholls 2001) suggests that it is more meaningful to use confidence intervals because these measures are much more informative than significance tests.

How can trust be built between forecasters/developers and verifiers?

First, useful information/feedback should be given to the forecasters, including information about how the forecasts can be improved. Verification methods should be transparent so that they can be easily understood, and forecasters should understand where the numbers came from. Ideally, the forecasters should be involved in the verification design process. Subjective information on the quality of the forecasts should be collected and compared to the objective results; if the two sets of results don't match, there is a problem.

How should verification of operational use of a forecast be related to meteorological verification of the forecast? Do users need both types of information?

Operational use of forecasts clearly should be related to the meteorological verification of the forecasts. One approach to improving the usefulness of verification information would be to stratify verification results according to subjective information (e.g., regarding the difficulty of the forecasting situation). Work is in progress at the University of Washington, regarding how forecasters use information to make decisions. Psychologists should be involved in these types of studies.

Allan Murphy once said that developing appropriate verification tools is just as hard as developing forecasting tools. Why, in the real world, is verification development always the "poor cousin" of forecast development?

Many scientists and forecasters believe that the verification problem is "solved" (e.g., that the RMSE provides adequate information) and that it is easy. In addition, many individuals have a fear of verification because its objective is to critique something that is of importance to the developer/forecaster. Developers should build verification into the development process; it is inappropriate to ignore this part of the process. However, it often *is* ignored in favor of other development activities. Finally, operational forecasts must be issued as expected, regardless of other factors, while verification results usually are not an operational requirement. In economic hard times this difference may be a factor.

There seems to be overkill in verification of precipitation. Why not look at state variables (p, T, etc.)? Is verification of precipitation forecasts useful? To whom?

Many participants agreed that there is **not** "overkill" in terms of precipitation verification: precipitation is very difficult to analyze and verify, and extensive improvements in the analyses and quality control of precipitation observations are needed. Moreover, precipitation integrates

all processes in a model. Other participants suggested that the user community is quite sensitive to other variables, and some of these variables (e.g., wind) are also difficult to verify.

5. Conclusion

A number of conclusions can be drawn from the workshop presentations and discussions. A few of those conclusions are listed here:

- Users of verification information should participate in the design of verification approaches and measures – each user may need a particular kind of information about forecast quality, which is likely to differ from other users' needs.
- Operationally relevant metrics are needed along with meteorologically relevant metrics; each type serves a particular, possibly distinct, purpose.
- Scale issues (e.g., observation scale vs. forecast scale) need to be taken into account in verification studies. Scale separation approaches are available and should be applied. However, this area still requires further research and testing.
- Current verification methods for spatial forecasts only provide limited information about the quality of these forecasts, especially as finer-scale models and forecasts have been developed. New object- or field-based approaches show promise for providing more useful information for at least some users. Further development of these approaches should be pursued.
- Observational uncertainty limits how well we can measure forecast quality; ideally, observational uncertainty should be taken into account in verification studies, but this objective is very difficult to actually accomplish and should be a subject of research.
- Additional educational opportunities regarding statistics and verification should be made available, through atmospheric science curricula, short courses, and web-based material.

- Future workshops on this topic would be desirable and were requested by many of the attendees.

In summary, this very successful workshop generated a great deal of interest and enthusiasm for new work in the forecast verification arena. A number of new ideas were shared, and new collaborations were developed. Many of these ideas will be incorporated into verification activities of the AWRP and also should be considered for other verification endeavors.

Acknowledgments

This research is in response to requirements and funding by the Federal Aviation Administration (FAA). The views expressed are those of the authors and do not necessarily represent the official policy and position of the U.S. Government.

We would like to thank Bruce Carmichael and Brant Foote for their support of this workshop, and NCAR for the support provided for invited speakers. Many individuals worked very hard to make the workshop happen and to keep it going – it would not have been possible without them. These individuals include Inger Gallo, who kept us organized before the workshop, and Carol Park who kept us on track during the workshop. A special thanks and acknowledgment of his fine work goes to Jamie Braid, who created and maintained the web page. We also would like to thank Randy Bullock, Mike Chapman, and Cherie Adams for their logistical support, and Anne-Marie Tarrant and Laura Kriho for providing computer support. Finally, we would like to thank all the workshop participants and especially those who gave presentations, for all the hard work that went into the preparations, and for their enthusiastic participation. The people made all the difference. NCAR is sponsored by the National Science Foundation.

References

Allen, R.A., G.W. Brier, I. Gringorton, J.C.S. McKillip, C.P. Mook, G.P. Wadsworth, and W.G. Leight, 1952: Panel discussion on forecast verification. *Bulletin of the American Meteorological Society*, **7**, 274-278.

Holzman, B.G., 1947: The separation of analysis and forecasting: A new basis for weather service operations. *Bulletin of the American Meteorological Society*, **28**, 281-293.

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.

Nicholls, N, 2001: Commentary and analysis: The insignificance of significance testing. *Bulletin of the American Meteorological Society*, **82**, 981–9

Figures

Figure 1. Workshop participants.

Figure 1. Workshop participants.