Summary of the 1998 Workshop on Mesoscale Model Verification

Christopher Davis* and Frederick Carr[†]

ABSTRACT

The major results and discussion items presented at the 1998 Workshop on mesoscale model verification, held 18–19 June in Boulder, Colorado, are summarized. This forum represents perhaps the first attempt to bring together the mesoscale modeling and statistical communities in an attempt to discuss the most challenging issues related to verifying mesoscale forecasts. Pervading discussion was the issue of uncertainty in predictions and observations and how to account for this when performing verification. This article discusses techniques to verify both deterministic and probabilistic predictions and provides recommendations for approaches to future endeavors in mesoscale model verification.

1. Introduction

With the recent emphasis on locally run real-time forecasts, especially those approaching cloud-resolving scales, the new operational datasets available and the use of experimental forecasting techniques such as ensembles, the issue of mesoscale model verification has taken on new significance. Verification here refers primarily to the comparison of forecasts with observations of the state of the atmosphere, although it can include comparison of model results with analytic solutions or with simulations of idealized flows performed by other, established models. The comparison with observations is perhaps the most challenging aspect of verification, for the issue of imperfect data (instrument error, partial coverage in space or time) is central to all verification efforts.

Verification, if done properly, can serve several purposes. First, it can indicate model performance. This is particularly important as research and opera-

tional prediction enters the “cloud-scale” regime. In this regime, traditional methods of verification are likely to fail because phenomena such as convection, clear-air turbulence, clouds, and fog are localized and episodic. These weather elements may be on the edge of predictability or entirely stochastic from the perspective of current numerical weather prediction (NWP) models. As such, traditional verification methods, based on standard observations (soundings and surface), or based on the instantaneous comparison of analyzed and predicted fields, may not yield useful information.

It was felt that an initial workshop on model verification would help assess the current state of the science and bring together the statistics and modeling communities. Such a meeting was held at the National Center for Atmospheric Research in Boulder, Colorado, on 18 and 19 June 1998. This article reports on the major results of that meeting and makes general recommendations regarding future verification practices.

2. Joint session with real-time NWP workshop

The meeting was structured so as to first provide a summary of current verification efforts and discuss some of the difficulties of such efforts. This was done in a joint session with the 1998 Real-Time Numerical

*National Center for Atmospheric Research,[#] Boulder, Colorado.

[†]University of Oklahoma, Norman, Oklahoma.

Corresponding author address: Christopher A. Davis, NCAR, P.O. Box 3000, Boulder, CO 80307.

E-mail: cdavis@ucar.edu

[#]The National Center for Atmospheric Research is sponsored by the National Science Foundation.

In final form 13 August 1999.

©2000 American Meteorological Society

Weather Prediction Workshop. The advantage of discussing verification of real-time forecasts is that usually a large sample of cases, in a variety of flow regimes, are available. Thus, the results are relatively unbiased and often contain important information about systematic model errors. This avoids some of the problems associated with case studies, though it is recognized that detailed verification using specialized, case-dependent datasets is also necessary. The examination of large samples of forecasts also offers some promise of understanding the limits of mesoscale predictability.

A topic that permeated most of the presentations was the verification of relatively high-resolution forecasts of quantities such as precipitation (J. Horel, UU;¹ and L. Bernadet, FSL), icing (J. Mahoney, FSL), mesoscale circulations (N. Seaman, PSU; M. Ramamurthy, UI), trace chemical constituents (W. Lyons, FMA), and terrain-induced circulations (T. Holt, NRL; P. Haines, ARL). Standard means of verification, typically applied to large scales, gave some useful information, but most of these investigators used other verification tools to obtain their results.

Other talks examined verification on larger scales, but with some new twists. S. Silberberg (NIU) discussed verification of parameters associated with quasigeostrophic ascent and the spatial distribution of large errors. J. Hacker (UBC) discussed the construction and verification of ensembles where the initial perturbations were coherent structures whose location was varied. Since it is not possible to discuss results from all presentations in detail, the remainder of this section will review the topics upon which significant discussion was focussed.

J. Horel (UU) performed analyses of precipitation over the western United States and attempted an intercomparison among models run by the National Centers for Environmental Prediction (NCEP). Within the suite of models were the Nested Grid Model (NGM) and the Eta Model (Black 1994) run at 10-, 29-, and 48-km resolution. At issue was how to meaningfully intercompare the precipitation forecasts from these models in regions of complex terrain. The suggestion was to use climatological information to postprocess the precipitation forecasts from each model and map the results onto a common grid before the equitable threat score (ETS) was computed (Fig. 1).

FIG. 1. Method to disaggregate coarse-resolution precipitation forecast (top left) by applying finescale climatological information (top right) to repartition the forecasted precipitation on a higher-resolution domain (bottom). The total precipitation, integrated over the coarse grid box, remains the same.

Here

$$ETS = \frac{C - \epsilon}{F + O + C - \epsilon}, \quad (1)$$

where C is the number of grid boxes with a correct forecast (i.e., precipitation greater than a certain threshold), F is the number of forecasted occurrences, O is the number of observed occurrences, and ϵ refers to the expected number of correct predictions using a random guess for the precipitation category. In the example presented there was a dramatic improvement in the skill of the NGM at high precipitation thresholds resulting from the addition of higher-resolution climatological information. For 24-h precipitation totals exceeding 1.5 in., the ETS increased from 0.03 to 0.12. Another notable result was the appearance of a peak in skill at precipitation amounts between 0.25 and 0.5 in. for 24-h forecasts, though it was unclear why this peak existed.

The situation is much different over the eastern United States, where complex terrain plays a smaller

¹See the appendix for meaning of abbreviated affiliations.

role in focusing precipitation. L. Bernadet (FSL) presented results from using the Regional Atmospheric Modeling System (RAMS) over the southeastern United States during the summer of 1996. The ETS for forecasts out to 20 h was between 0.05 and 0.1 for most precipitation thresholds. Similar results were obtained from the Eta Model during the period.

The contrast between the Horel and Bernadet studies demonstrated the large difference in predictability between the winter season near complex terrain and the summer season with weak terrain forcing. Results indicate that warm-season precipitation over the southeast United States is generally a stochastic process, and future verification efforts would be better off concentrating solely on the statistical properties of precipitation for forecasts longer than perhaps an hour. For example, Bernadet presented statistics on the number of points where precipitation above a certain threshold was predicted versus observed and showed that RAMS had a significant bias of underforecasting heavy precipitation. This result was subsequently tied to a diurnal cycle that was too weak and hence to the surface energy balance. Thus, the absence of predictability, in the strictly deterministic sense, is not grounds to abandon verification efforts, for results from such efforts can still be quite informative about model behavior.

In general, verification of mesoscale forecasts must rely on data sources that are asynoptic and measure quantities other than basic atmospheric variables (wind, temperature, pressure, humidity). W. Lyons discussed verification using trajectory analysis aided by the observed concentrations of trace constituents. With source regions presumed known (for instance, in a deliberate release), the general airflow can be estimated and compared with model results. Lyons also emphasized the importance of nonstandard data sources, including boundary layer profilers, sodars, and surface mesonet data for detailed verification of planetary boundary layer structure in models. Examples were shown in connection with the Lake Michigan Ozone Study (Lyons et al. 1995).

Along these lines, an interesting proposal was made by T. Schlatter (FSL), who suggested that all the observations currently used by national centers for 3D- and 4D-variational data assimilation could be used for model verification. This prospect is made simpler by the fact that differences between model and observations, in observation space, already are computed to evaluate the cost function that is minimized in such procedures. This means that satellite radiances,

Doppler velocity, and reflectivity from radars and other data sources can be incorporated into model verification routinely. The main requirement is storing the differences for subsequent analysis. The advantage of computing the differences in observation space is that the forward operator will already be coded for the data assimilation, and a forward operation avoids errors associated with retrievals of the model variables (i.e., the inverse calculation).

3. Invited presentations

a. B. Brown (NCAR): Coping with poor observational data

As noted in the presentation by J. Mahoney, efforts to verify many nonstandard, but meteorologically very significant, derived quantities such as icing, fog, turbulence, and severe convection are hampered by poor observations. In this context poor observations refers to data with one of the following deficiencies: 1) Forecasts do not have matching observations, 2) observations are not representative, 3) Observations are not systematic, 4) Observations have large errors or biases. Poor observations can lead to verification results that are incorrect, biased, unrepresentative, or unreliable.

Examples of forecasts with poorly matched observations are severe weather reports and pilot reports (PIREPS) of aircraft icing or turbulence. In both cases, many events go unobserved, so there is poor matching. The data are also of questionable accuracy. Pilot reports are often issued at a location other than where the event occurred and may inaccurately report icing or turbulence severity.

One way to cope with poor observational data is to develop appropriate methods to at least partially compensate for the data deficiencies. Brown presented examples of the types of verification one can perform with imperfect data, PIREPS in this case (Brown et al. 1997). Icing is typically verified in terms of a “yes–no” (dichotomous) forecast. The skill of these forecasts can be represented by a contingency table (Fig. 2). In the case of PIREPS, it is not possible to verify null forecasts because PIREPS are typically issued only when an event is encountered and, even then, not systematically. Thus, only the probability of detection can be computed for forecasts of icing, because it is the only statistic that does not depend on the forecast distribution. Hence, the false alarm rate and the bias cannot be computed. To compensate for this limitation,

Forecast	Observation		Total
	Yes	No	
Yes	YY	YN	YY+YN
No	NY	NN	NY+NN
Total	YY+NY	YN+NN	N

$POD = YY / (YY + NY)$	Observations-based
$POD-No = NN / (YN + NN)$	Observations-based
$FAR = YN / (YY + NY)$	Forecast-based
$Bias = (YY + YN) / (YY + NY)$	Obs and Forecast Distributions

FIG. 2. Contingency table for dichotomous forecasts.

the areal and volumetric extent of the forecast regions of icing are also computed. These variables are reasonable because, given two forecasts with the same probability of detection, the one that predicts icing over a smaller area will be of more use.

Other approaches to avoid some of the issues related to poor data are 1) to restrict the verification in space or time such that the observations are “locally” satisfactory; 2) to use other variables that approximate the true observations, but that offer better coverage in space or time; and 3) to understand that comparative verification (i.e., comparing the performance of two or more verification systems) may be appropriate even if absolute verification is not. The second approach, known as a “covariate” approach, utilizes known relationships between an observable field (i.e., satellite observations) and the unobservable field being forecast (i.e., icing) to estimate standard verification measures.

b. H. Brooks: Issues associated with forecast uncertainty

In this presentation, the issue of forecasting rare (but important!) events was discussed. Here, severe convection was given as an example (e.g., that which produces tornadoes, large hail, or damaging straight-line winds). The main topic related to these rare events is spatial uncertainty, that is, the events themselves are highly localized and have little explicit predictability. However, on any given day, forecasters may identify areas (much larger than the event) that exhibit favorable conditions for severe weather. How, then, does one verify a forecast that a small-scale event will oc-

cur somewhere in a much larger-scale area? Another way to phrase the question is, what is a reasonable upper limit of skill in this situation?

This reasonable upper limit of skill was assessed by considering a probability density function computed from point reports of severe weather on a single day. For this purpose, we show a hypothetical set of severe weather reports (+’s in Fig. 3) for some arbitrary geographical region. The associated probability density distribution is computed as in Brooks et al. (1998) and may be thought of as a Gaussian smoothing operation, with the smoothing parameter σ chosen such that larger values of σ correspond to smoother distributions. The thresholds of different probability represent the chance that a severe weather event occurred in a given grid box, subject to a particular choice of σ (Fig. 3). Obviously, the more clustered the events, the higher the probability. Note that this distribution is computed entirely from the observations.

In performing this analysis, the σ is an arbitrary parameter. Since we are trying to assess the maximum reasonable skill of a severe weather forecast, the σ will be related to how detailed the forecast would be and, hence, to forecast lead time. Forecasts of longer lead times would naturally involve larger area coverage, hence more smoothing, just due to uncertainty on the mesoscale and synoptic scales. The particular smoothing here might be appropriate for forecasts of 12–24 h and a horizontal resolution of a few tens of kilometers.

For a choice of σ , one can compute a critical success index (CSI) assuming that all grid points where the probability density exceeds a certain threshold are counted as events.² For this case, it turns out that the maximum CSI for a smoothing parameter representative of severe weather outlooks is about 0.24 for a probability threshold of 15%. This can be interpreted as the best a forecaster would hope to do when issuing a severe weather outlook. In the example shown, a hypothetical convective outlook (area covered by asterisks) exhibits a CSI of about 0.09 for the same threshold probability. This forecast skill is therefore roughly 38% of the upper bound, implying that despite a low CSI in absolute terms, the hypothetical convective outlook exhibits considerable relative skill in this case. By this relative measure of skill it turns out that forecasters will perform better on days when severe

²CSI and threat score are identical. The threat score is ETS with $\epsilon = 0$.

weather events are clustered in space than when they occur more randomly.

Another issue related to spatial uncertainty is the fact that numerical forecasts of highly variable quantities such as precipitation are often not entirely deterministic. In fact, at any grid point, the predicted amount of precipitation might fall into one of many, almost equally likely, precipitation categories. Thus, a way of interpreting the numerical model output is to effectively smooth it, as was done with the severe weather forecasts, by taking into account whether a given precipitation event was predicted at adjacent grid points. The simplest threshold to consider is that of measurable precipitation. If we consider that all grid points within a radius R from a central grid point are equally likely forecasts of the event for the central grid point, then the probability is just the number of points at which precipitation greater than 0.01 in. was predicted, divided by the number of grid points within a distance R . Using an artificial dataset, it was shown that the best forecast, measured by the Brier skill score,³ was not for the raw numerical forecast, but for a forecast created with $R = 2$ grid points. That is, the best forecast, on average, was obtained when all grid points within $R = 2$ were considered equally likely representations of the forecast for the central grid point. This means that a model will generally show higher skill for forecasts of highly variable fields like precipitation if the output is smoothed. However, there is clearly an opti-

³The Brier skill score (Wilks 1995, p. 260) is defined as the variance of the forecast error in the case where observations assume values of only 0 or 1, depending on whether or not an event occurs at a given point, and forecasts are constrained on the interval (0, 1).

FIG. 3. (a) Hypothetical severe weather reports (+s) indicating the occurrence of an event within box of length dx on each side (see scale). Shaded are probability values representing the “practically perfect” forecast generated by smoothing the individual reports with a Gaussian filter with half-width $3 dx$ (Brooks et al. 1998). The probability of a severe weather report lying in any grid box is shaded in increments of about 5.6%. (b) Asterisks mark the area encompassed by a hypothetical severe weather outlook, with the shaded field the same as in (a); (c) CSI for practically perfect forecasts made assuming different probability thresholds (%) and different smoothing of the severe weather reports (σ is in units of dx).

mum value of R , beyond which the forecast rapidly deteriorates.

c. T. Hamill: Some thoughts on the statistical evaluation of precipitation forecasts

The first portion of this talk concerned verification of probabilistic forecasts, of which ensemble forecasts are an example. Ensembles were originally designed as a set of forecasts from an identical model and a variety of plausible initial conditions. In recent years, the notion of ensembles has also been extended to include variations in model physical parameterizations,

with identical initial conditions, or with variations of both physics and initial conditions. The intent of the ensemble is to (a) estimate an envelope of uncertainty in forecasts and (b) sample the probability distribution associated with weather events.

Regardless of the design of the forecasts that populate current operational ensembles, too often the real state of the real atmosphere lies outside the ensemble envelope. This may be due to insufficient growth of differences among ensemble members, or to model error that causes some or all ensemble members to be biased.

1) RANK HISTOGRAMS

A visualization of this problem may be seen by examining “rank histograms” (Hamill and Colucci 1997). A rank histogram, also known as a Talagrand diagram, displays the frequency with which the observation occurs relative to the sorted ensemble. The rank is just the position of the observation relative to the ensemble forecasts (1 to $N+1$ ranks for N ensemble members), ordered in some systematic way at each grid point, say, from coldest to warmest, or driest to wettest. For instance, if for a particular time interval and grid point, the precipitation predictions from a 10-member ensemble were 0.0, 0.0, 0.2, 0.4, 0.8, 1.5, 1.9, 2.8, 3.8, and 5.1 mm, and the verification was 2.6, the rank of the verification would be 7 out of the 11 possible ranks. Rank statistics are accumulated over many grid points and many cases to compose a histogram depicting the frequency with which each rank is occupied. For a properly constructed ensemble (such as would occur with a perfect model and properly sampled initial conditions), the rank histogram should be approximately uniform, meaning that an observation is equally likely to occur near any ensemble member.

Most operational ensembles depart significantly from a uniform ranked histogram. Precipitation prediction is a good example. Coarse-resolution models do not capture heavy precipitation events accurately, partly because of the resolution and partly because of physics that is scale dependent. Typically there are many precipitation values outside the ensemble spread. Small-scale variability also means that there will be many points that are dry in reality, when the model forecasts large areas of light precipitation in all members. Hence, there is a large nonuniformity of the ranked histogram (Fig. 4).

Several methods to more appropriately model the forecast uncertainty in ensemble forecasts have been

FIG. 4. Examples of nearly uniform (top) and nonuniform (bottom) ranked histograms based on a hypothetical ensemble of 20 members.

suggested. Some of the more promising are the incorporation of stochastic physical parameterizations (Stensrud and Fritsch 1994; Buizza et al. 1997) and designing perturbations consistent with analysis uncertainty (Houtekamer 1995; Barkmeijer et al. 1998, 1999; Houtekamer and Mitchell 1998; Hamill et al. 2000). In addition, increasing resolution, domain size (moving boundaries out) (Warner et al. 1997) and including perturbations to the land surface initial condition (Houtekamer et al. 1996; Hamill and Colucci 1998) all should increase the spread of ensembles.

2) RELIABILITY DIAGRAMS

Another verification tool to use for inherently probabilistic forecasts is the reliability diagram (Wilks 1995). If we consider a set of precipitation probability forecasts that predict a 30% chance of measurable rain, an important question to ask is whether measur-

able rain is observed 30% of the time. A reliability diagram, depicting observed frequency versus predicted probability, allows one to assess whether forecasts are unbiased in their probability assignments, that is, whether they are properly calibrated. For instance, when a 30% chance of rain is predicted, an observed frequency of less than 30% yields a point in the diagram below the 1:1 line and is interpreted as an “overforecast” bias.

A shortcoming of reliability diagrams is that one needs a large number of forecasts to generate a meaningful diagram (see Wilks 1995, p. 266). A different diagram was discussed here (Hamill 1997), known as the multicategory reliability diagram (MCRD). Simultaneous forecasts of probabilities in several categories (of precipitation, for example) are allowed, not just dichotomous predictions. The MCRD was designed to evaluate the average probability that the observation will fall below certain quantiles (percentiles) of a probability distribution. For example, if a set of forecasts is calibrated, then 30% of the time an observation can be expected to be below the 30th percentile of the probability distribution.

Figure 5 illustrates the mechanics of constructing the MCRD. Here, model output statistics (MOS) from NCEP’s NGM assigns probabilities to each of six categories of precipitation, 0–trace, 0.01–0.09 in., 0.1–0.24 in., 0.25–0.49 in., 0.5–1.0 in., and > 1.0 in. Let us assume that the probabilities in each category, rounded to the nearest tenth, are 0.2, 0.3, 0.3, 0.2, 0, and 0, respectively. Now we assign a forecast category to 10 evenly spaced quantiles in the probability distribution. The forecast category for each quantile is obtained by mapping the forecast probabilities for the six precipitation categories into these 10 quantile bins, such that the assigned category for a given quantile corresponds to the accumulated probability in the original forecast. This can be thought of as a remapping of probability as a function of category, $\text{Pr}(\text{cat})$, to category as a function of accumulated probability. Hence, the 25th quantile is assigned category 2, because 0.25 is greater than the probability for cat-

Multicategory Reliability Diagrams

Hypothetical MOS Categorical Precipitation Forecast

Categories	0.0–Tr	0.01–0.09	0.10–0.24	0.25–0.49	0.50–1.00	> 1.00
Forecast	0.2	0.3	0.3	0.2	0.0	0.0

Verification: Category 2

Quantiles	0.05	0.15	0.25	0.35	0.45	0.55	0.65	0.75	0.85	0.95
Fcst category at each quantile	1	1	2	2	2	3	3	3	4	4
Verification category at each quantile	2	2	2	2	2	2	2	2	2	2
$P(\text{verif} < \text{fcst})$ at each quantile	0	0	1/6	1/2	5/6	1	1	1	1	1

FIG. 5. Sample methodology for construction of an MCRD. See text for explanation.

egory 1 (0.2) but less than the accumulated probability of categories 1 and 2 (0.5).

Assume now that the observed precipitation falls into category 2 (0.01–0.09 in.). Thus, the probability that the verification category is less than the forecast category is 0 in quantiles assigned to category 1, and 1 for quantiles assigned to categories greater than 2. For quantiles assigned category 2, the probability is forced to vary linearly from 0 to 1 (Hamill 1997). Next, the procedure is repeated over and over using other forecast probability distributions and the associated observations. A relationship between quantiles and the probability that the observation is less than that quantile is established.

A benefit of the multicategory reliability diagram is that each quantile is populated from a single forecast, reducing the minimum sample size needed for a smooth, meaningful curve. MCRDs yield information similar to that contained in standard reliability diagrams but are interpreted differently. Forecasts above the 1:1 line signal overprediction of precipitation (probabilities in the higher precipitation categories are

too large) and the reverse for points below the line. The practical benefit of this diagram is that it tells one how to adjust probability distributions in order to improve forecasts.

3) HYPOTHESIS TESTING

Often it is difficult to tell whether changes in a physical parameterization or initialization procedure lead to a meaningful improvement in model forecast skill. Similarly, it is often a subjective matter to state whether one model performs better than another. Here it was emphasized that hypothesis tests of differences between forecasts, either with different or the same models, could be used to validate claims of improvement or superiority. Hamill (1999) reviews the problems with such hypothesis tests. In most cases, a paired *t* test (Wilks 1995) provides an adequate assessment of statistical significance, but when evaluating rare events such as heavy precipitation, more careful resampling tests are required.

d. E. Kalnay: Validation of usefulness

1) ENSEMBLE VERIFICATION

Perhaps too much emphasis has been placed on the fact that the ensemble mean has more skill, on average, than any particular ensemble member. This fact is mostly a by-product of the goal of ensemble forecasting. Problems with assessing the skill of the ensemble mean can easily be seen when precipitation events are considered. The ensemble mean will naturally reveal lighter precipitation, covering a broader area, than individual ensemble members. Typically, the threat score for the ensemble mean is better than for single, high-resolution models, but the ensemble mean has a tremendous bias, forecasting a great deal of light precipitation but almost no heavy precipitation.

A more useful measure of the quality of precipitation forecasts by the ensemble is the probabilistic threat score (PTS):

$$\text{PTS} = \frac{\sum_{i=1}^N F_i \cap O_i}{\sum_{i=1}^N F_i \cup O_i}. \quad (2)$$

The difference between this and the traditional threat score is that the PTS computes the overlap between the precipitation forecasts for each ensemble member,

then performs a summation over the ensemble. It then normalizes by the summed union between forecasts and observations (Fig. 6). The result can be interpreted as the probability that any ensemble member will correctly predict the exceedence of a specified precipitation threshold. Note that even if the ensemble mean may never exceed the specified precipitation threshold anywhere, the PTS may still be greater than zero.

2) VALIDATION OF USEFULNESS

With the advent of “storm-scale NWP” (Lilly 1990) comes the issue of whether a forecast that is not strictly accurate in the deterministic sense is still useful. While not only an issue on the storm scale, the extreme local variability of certain phenomena such as convection makes usefulness of imperfect forecasts a more important issue. The first question to be answered is, useful to whom? In the present context, it is forecasters and scientists wishing to utilize and improve numerical prediction. A prediction of a single supercell thunderstorm might be off by 40 km, but the fact that a model predicts a supercell is potentially useful information. It is difficult to quantify *how* useful.

One strategy was introduced by R. Hoffman and collaborators (e.g., Hoffman et al. 1995), which is an object-oriented verification. If an object, such as an area of precipitation, can be uniquely identified, then a usefulness threat score can be defined that uses a translation of the predicted object in space and time to define a maximum threat score. This score is then penalized (exponentially) by the distance in time and space one had to move the object. Thus the right thunderstorm in the wrong place does not receive a threat score of zero, but rather, some nonzero number that becomes larger as the prediction becomes more accurate in position and time. Typically, the exponential decay length scales and timescales are calibrated using forecasts from the wrong day (i.e., useless by definition). This helps overcome the issue of making sure that the object in the forecast is indeed the object that verifies.

4. New approaches and recommended verification strategies

a. New directions

The clearest theme emerging from the invited presentations is one of reconciling uncertainty, either in

FIG. 6. Illustration of the advantage of the probabilistic threat score (see text) for forecasts of the area of accumulated precipitation greater than 1 in. F1 and F2 refer to two ensemble forecast members. The intersection (double hatching) is the area of greater than 1 in. in the ensemble mean. The traditional threat score for the ensemble mean is zero.

observations or forecasts or both, when attempting forecast verification. As models attempt to push the theoretical limits of predictability, whatever they may be, we are constantly confronted with uncertainty and are continually exploring ways to quantify it.

There are several tools available to assess the skill of inherently probability based forecasts or only partially deterministic forecasts. The question of what is a reasonable upper limit on skill (Brooks et al. 1998) turns out to have an answer that is often much less than 100%. It is critical to address questions such as these. First of all, it improves our understanding of predictability on the cloud scale and mesoscales. Second, it provides a basis for the allocation of resources on different forecast problems.

Most verification methods for probabilistic forecasts discussed herein can be equally well applied to both numerical forecasts and model-aided forecasts made by forecasters. It is likely that a systematic application of some of these techniques (where not already done), would lead to further improvements in forecasts issued to the public even with little additional technological advance. Furthermore, some of the techniques we have discussed would be beneficial to those building and improving new, state-of-the-art forecast systems. These techniques can help the effort to produce ensembles with the correct dispersion and hence a better estimate of predictability. The quantification of predictability limits is crucial for determining the optimal allocation of resources for returning the greatest improvement in forecasts.

b. Recommendations

We close with some specific recommendations concerning the use of some of the relatively new tools we have discussed for mesoscale model verification.

- 1) We stress the need for efforts to verify large samples of forecasts rather than individual cases. We recognize that it is not always practical to consider samples of hundreds or thousands of cases, as might be done at operational centers, when comparing models against each other or evaluating changes to a given model. To make the evaluation tractable, we recommend choosing cases in different flow regimes or seasons that are believed to highlight model sensitivity. Often, considerable insight can be gained from sample sizes as small as $O(10)$. For example, one might consider weak and strong flows, moist and dry conditions, warm and cool seasons, a total of eight cases. It can be useful to verify even a single forecast, so that more detailed analyses and special datasets can be applied to the verification and so that a greater number of sensitivity simulations can be executed. Detailed verification of a single forecast will be most insightful if the forecast errors appear representative of recurring model deficiencies.
- 2) Operational centers that already compute the difference between forecasts and observations (in observation space) should routinely archive the difference fields for model verification. Those engaging in model verification need to work closely with those developing variational data assimilation systems because of the need for model error covariances in calculating the background term.
- 3) For ensembles, statistics (such as threat score) of quantities that are highly variable in space and time should be computed for the individual ensemble members, then averaged, rather than be computed on the ensemble mean.
- 4) When considering the skill of forecasts of rare events, care must be taken to estimate the maximum reasonable skill score that could be expected, which is sometimes much less than 1.
- 5) Verify against another, cleverly chosen forecast. If the other forecast is from a different model, perform verification on a common grid. This may involve downscaling the coarser-resolution forecast. Conditional climatologies and lagged persistence forecasts are examples of often worthy competitors. Providing the sample size is adequate,

one should apply hypothesis tests of the differences between forecasts to assess statistical significance.

- 6) Use model verification to attempt to understand mesoscale predictability. While general theoretical estimates exist for predictability, the results of model verification are necessary to understand what limits our predictability in practical situations.
- 7) Some standard verification statistics should be computed for all mesoscale models. Though this point was agreed upon, the details are the subject of continuing discussion. This list was compiled with two criteria: (a) statistics that would tend to point toward a particular part of the model code as the source of error, despite the strong interdependency that exists within models, and (b) statistics that can be calculated by regularly available operational datasets today. Statistics in (b)–(e) (see below) may be computed on a single forecast or a set of forecasts. A preliminary list is as follows.

(a) Basic model dynamics (“one-time” calculations): tests for which analytic solutions, or converged numerical solutions exist:

- Zero initial flow; no forcing; no gradients; steep (slope > 0.1) topography
- Long’s solution
- Low Froude number flow and lee vortices
- Splitting supercell (minimal microphysics)
- 2D squall line

(b) Precipitation physics:

Datasets:

- Stage IV precipitation analysis from NCEP (4-km resolution).⁴
- National Lightning network

Statistics:

- Bias for different thresholds; computed on a grid of resolution 20–30 km
- Threat scores for multiple precipitation categories

- Contingency diagram for prediction of deep convection.⁵

(c) Boundary layer winds and temperature:

Datasets:

- Aviation Routine Weather Reports (METAR), local mesonets

Statistics:

- Diurnal cycle error (bias) (separate out clear days)
- Biases in wind direction and speed (separate into strong and weak flow regimes)
- Bias in water vapor mixing ratio

(d) Upper air:

Datasets:

- Soundings, profilers and operational analyses

Statistics:

- Bias in temperature, wind and humidity
- Errors in strength/location of jet maxima at about 900 and 300 mb
- rmse of geopotential heights, winds and temperatures

(e) Clouds:

Datasets:

- Geostationary Operational Environmental Satellite (GOES); IR, visible and water vapor (perhaps other channels)

Statistics:

- Biases in total cloud cover for low, middle, and high clouds

⁴Although the quality of this data set is improving, there are some notable deficiencies especially over the western United States. Hence we recommend that these data be used with caution.

⁵This statistic would perhaps make use of both the lightning and stage IV analysis and require some assumptions about what constitutes deep convection in a mesoscale model. It could also be more instructive if applied using the methodology of section 3b.

- Biases in cloud-top temperature
- Biases in fog/stratus coverage, especially as a function of time of day

As always, the particular verification performed will depend somewhat on particular forecast problems and data quality. These recommendations are meant to (a) add more commonality and quantitateness to the assessment of mesoscale forecasts, and (b) allow forecast model verification to serve a purpose that goes beyond the determination of whether a single forecast is good or bad.

Acknowledgments. The authors would like to thank Ingrid Moore of NCAR for her tireless effort in the organization of this workshop.

Appendix: Affiliations

ARL	Army Research Laboratory, White Sands, New Mexico
FMA	FMA Research Inc., Fort Collins, Colorado
FSL	Forecast Systems Laboratory (NOAA), Boulder, Colorado
NIU	Northern Illinois University, Dekalb, Illinois
NRL	Navy Research Laboratory, Monterey, California
NSSL	National Severe Storms Laboratory, Norman, Oklahoma
OU	Oklahoma University, Norman, Oklahoma
PSU	The Pennsylvania State University, State College, Pennsylvania
UBC	University of British Columbia, Vancouver, British Columbia, Canada
UI	University of Illinois, Urbana-Champaign, Urbana, Illinois
UU	University of Utah, Salt Lake City, Utah

References

Barkmeijer, J., M. van Gijzen, and F. Bouttier, 1998: Singular vectors and estimates of the analysis-error covariance metric. *Quart. J. Roy. Meteor. Soc.*, **124**, 1695–1713.

—, R. Buizza, and T. N. Palmer, 1999: 3D-Var Hessian singular vectors and their use in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor.*, **51**, 1–79.

Black, T., 1994: The new NMC mesoscale Eta model: Description and forecast examples. *Wea. Forecasting*, **9**, 265–278.

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

Brown, B. G., G. Thompson, R. T. Brintjes, R. Bullock, and T. Kane, 1997: Intercomparison of in-flight icing algorithms. Part II: Statistical verification results. *Wea. Forecasting*, **12**, 890–914.

Buizza, R., M. Miller, and T. N. Palmer, 1997: Stochastic simulation of model uncertainty in the ECMWF ensemble prediction system. *Proc. ECMWF Workshop on Predictability*, Shinfield Park, United Kingdom, ECMWF, 215–243.

Hamill, T.M., 1997: Reliability diagrams for multicategory probabilistic forecasts. *Wea. Forecasting*, **12**, 736–741.

—, 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167.

—, and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327.

—, and —, 1998: Perturbations to the land surface condition in short-range ensemble forecasts. Preprints, *12th Conf. on Numerical Weather Prediction*, Phoenix, AZ, Amer. Meteor. Soc., 273–276.

—, C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular vector and perturbed observation ensembles. *Mon. Wea. Rev.*, in press.

Hoffman, R. N., L. Zheng, and C. Grassotti, 1995: Distortion representation of forecast errors. *Mon. Wea. Rev.*, **123**, 2758–2770.

Houtekamer, P. L., 1995: Construction of optimal perturbations. *Mon. Wea. Rev.*, **123**, 2888–2898.

—, and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811.

—, and Coauthors, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225–1242.

Lilly, D. K., 1990: Numerical prediction of thunderstorms—Has its time come? *Quart. J. Roy. Meteor. Soc.*, **116**, 779–798.

Lyons, W. A., C. J. Tremback, and R. A. Pielke, 1995: Applications of the Regional Atmospheric Modeling System (RAMS) to provide input to photochemical grid models for the Lake Michigan Ozone Study (LMOS). *J. Appl. Meteor.*, **34**, 1762–1786.

Stensrud, D. J., and J. M. Fritsch, 1994: Mesoscale convective systems in weakly forced environments. Part III: Numerical simulations and implications for operational forecasting. *Mon. Wea. Rev.*, **122**, 2084–2104.

Warner, T. T., R. A. Peterson, and R. E. Treadon, 1997: A tutorial on lateral boundary conditions as a basic and potentially serious limitation to regional numerical weather prediction. *Bull. Amer. Meteor. Soc.*, **78**, 2599–2617.

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

HANDS-ON

METEOROLOGY

*Stories, Theories, and
Simple Experiments*

Zbigniew Sorbjan

Hands-on Meteorology presents the subject of atmospheric science in an exciting, intriguing, and entertaining way. Utilizing a "hands-on" approach, readers perform simple experiments that demonstrate concepts of basic meteorology: air pressure, heat, temperature, moisture, and wind. The experiments use common, everyday, inexpensive materials such as candles, bottles, cans, and balloons.

Hands-on Meteorology also contains meteorological anecdotes, historical narratives, references to important cultural events, and biographical sketches of prominent scientists.

Hands-on Meteorology is addressed to all those who are enthusiastic about learning and teaching about the atmosphere, weather, and climate and should prove entertaining for readers of all ages.

©1996 American Meteorological Society. Softcover, B&W, 306 pp., \$21.95 list/\$16.00 member (including shipping and handling). **Please send prepaid orders to:** Order Department, American Meteorological Society, 45 Beacon St., Boston, MA 02108-3693.

Project ATMOSPHERE is the educational initiative of the American Meteorological Society to foster the teaching of atmospheric topics across the curriculum grades K-12. It is a unique partnership between scientists and teachers with the ultimate goal of attracting young people to further studies in science, mathematics, and technology.

Project ATMOSPHERE
AMERICAN METEOROLOGICAL SOCIETY