Subjective Verification of Numerical Models as a Component of a Broader Interaction between Research and Operations

JOHN S. KAIN

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma and NOAA/National Severe Storms Laboratory, Norman, Oklahoma

MICHAEL E. BALDWIN

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma and NOAA/National Severe Storms Laboratory, and NOAA/NWS/Storm Prediction Center, Norman, Oklahoma

PAUL R. JANISH AND STEVEN J. WEISS

NOAA/NWS/Storm Prediction Center, Norman, Oklahoma

MICHAEL P. KAY

Cooperative Institute for Research in Environmental Sciences, University of Colorado and NOAA/Forecast Systems Laboratory, Boulder, Colorado

GREGORY W. CARBIN

NOAA/NWS/Storm Prediction Center, Norman, Oklahoma

(Manuscript received 9 June 2002, in final form 31 January 2003)

ABSTRACT

Systematic subjective verification of precipitation forecasts from two numerical models is presented and discussed. The subjective verification effort was carried out as part of the 2001 Spring Program, a seven-week collaborative experiment conducted at the NOAA/National Severe Storms Laboratory (NSSL) and the NWS/Storm Prediction Center, with participation from the NCEP/Environmental Modeling Center, the NOAA/Forecast Systems Laboratory, the Norman, Oklahoma, National Weather Service Forecast Office, and Iowa State University. This paper focuses on a comparison of the operational Eta Model and an experimental version of this model run at NSSL; results are limited to precipitation forecasts, although other models and model output fields were verified and evaluated during the program.

By comparing forecaster confidence in model solutions to next-day assessments of model performance, this study yields unique information about the utility of models for human forecasters. It is shown that, when averaged over many forecasts, subjective verification ratings of model performance were consistent with preevent confidence levels. In particular, models that earned higher *average* confidence ratings were also assigned higher *average* subjective verification scores. However, confidence and verification scores for *individual* forecasts were very poorly correlated, that is, forecast teams showed little skill in assessing how ''good'' individual model forecasts would be. Furthermore, the teams were unable to choose reliably which model, or which initialization of the same model, would produce the ''best'' forecast for a given period.

The subjective verification methodology used in the 2001 Spring Program is presented as a prototype for more refined and focused subjective verification efforts in the future. The results demonstrate that this approach can provide valuable insight into how forecasters use numerical models. It has great potential as a complement to objective verification scores and can have a significant positive impact on model development strategies.

1. Introduction

Since the Storm Prediction Center (SPC) began full operations at the National Severe Storms Laboratory (NSSL) facility in early 1997, close proximity and a mutual interest in operationally relevant research problems have cultivated a strong working relationship between the two organizations. Informal daily map discussions and collaborative research projects (e.g., Baldwin et al. 2002; Craven et al. 2002; Evans and Doswell 2001; Kain et al. 2000, 2003a; Stensrud and Weiss 2002)

Corresponding author address: Dr. John S. Kain, NSSL, 1313 Halley Circle, Norman, OK 73069.
E-mail: jack.kain@noaa.gov

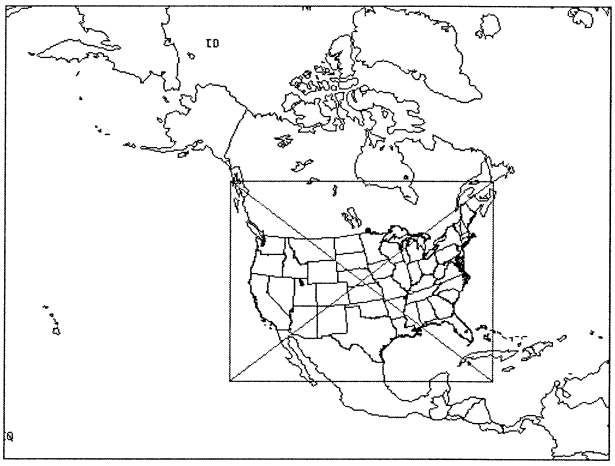

FIG. 1. Horizontal domain of the operational Eta Model during the 2001 Spring Program, with the inset indicating the domain of the EtaKF forecasts.

are among the organized interactions, but the cornerstone of this collaboration in the last several years has been intensive multiweek research programs conducted during each spring severe convective weather season. This effort has become known as the NSSL/SPC Spring Program (Kain et al. 2003b, hereafter KBAMS).

Forecasters at the SPC rely on a variety of observational data and mesoscale model guidance in preparation of convective outlooks, severe thunderstorm and tornado watches, and other operational forecast products. Two models routinely used at the SPC include the operational Eta Model (Black 1994) from the National Centers for Environmental Prediction (NCEP) and an experimental version of the Eta Model run in parallel at NSSL. The configuration of the NSSL version of the Eta (hereafter EtaKF) differs from the operational version in only three ways: 1) it contains the Kain–Fritsch convective parameterization (Kain and Fritsch 1993, hereafter KF) in place of the operational Betts–Miller–Janjić scheme (Janjić 1994, hereafter BMJ), 2) it uses fourth-order horizontal diffusion (with a 90% reduction in the diffusion coefficient) rather than the second-order algorithm used operationally, and 3) it runs over only a subset (about one-fifth) of the operational domain (Fig. 1). The alternative convective scheme and the more scale-selective horizontal diffusion both favor the development of circulations and features that are smaller in scale and higher in amplitude than corresponding structures in the operational model (Baldwin and Wandishin 2002). Computational resources at NSSL dictate the limited domain size. For the results discussed in this study, horizontal grid-spacing, terrain, and surface characteristics were identical to the operational Eta.

A key element in the 2001 Spring Program was subjective evaluation and verification of numerical weather prediction models. This element was inspired by empirical comparisons of Eta and EtaKF output over the past several years. Operational forecasters and collaborating research scientists have noted that although the EtaKF does not outperform the operational Eta every day, it complements the Eta well. Its performance is consistently comparable in skill, yet different in character, often providing unique information or a different perspective that is not available in operational forecasts. For example, forecasters report being intrigued by the EtaKF output because of the following:

- Unique output fields are produced from EtaKF runs, including parameterized updraft mass flux and the updraft source layer (Kain et al. 2003a). These fields have proven to be quite useful and have become an important ingredient in the routine forecast-preparation process for many forecasters.
- The KF scheme conforms closely to the general concepts of parcel theory and convective triggering associated with low-level airmass boundaries, consistent with most forecasters' conceptual model of convection. In contrast, the BMJ scheme is more fundamentally based on bulk tropospheric properties (Baldwin et al. 2002), making its behavior more difficult to link to specific observed physical processes.
- Operational experience suggests that the KF scheme provides more accurate timing of convective initiation in certain types of severe weather environments, in particular those with so-called ''loaded gun'' (e.g., Moller et al. 1994) soundings.
- Preconvective forecast soundings generated by the EtaKF are often more realistic and verify better than those of the operational Eta in the lowest 2–3 km due to differences in parameterized shallow convection (Baldwin et al. 2002).
- The structure and configuration of model-generated convective precipitation fields from the EtaKF often ''look'' more like observed radar imagery than output from the operational Eta. In particular, the EtaKF tends to produce more small-scale, high-amplitude precipitation structures than the operational Eta.

The EtaKF output has become ''popular'' with SPC forecasters, yet this appeal is not reflected in widely used objective verification measures. For example, a bellwether metric used by NCEP's Environmental Modeling Center (EMC) is the equitable threat (ET) score (Mesinger 1996). This measure rewards diffusive, smoothly varying forecasts over solutions with relatively small-scale, high-amplitude structures (Baldwin et al. 2001). Yet, forecasters at the SPC (and elsewhere) clearly appreciate having access to more detailed (though not necessarily higher resolution) model output. Our experience has revealed that the ET score often rewards Eta solutions with higher scores than the EtaKF, even on days when the EtaKF solution (with more ''structure'') is preferred by forecasters.

Realization of this contradiction has heightened our sensitivity to a more general problem with model verification: current verification metrics do not necessarily reflect the value of model forecasts to human forecast-

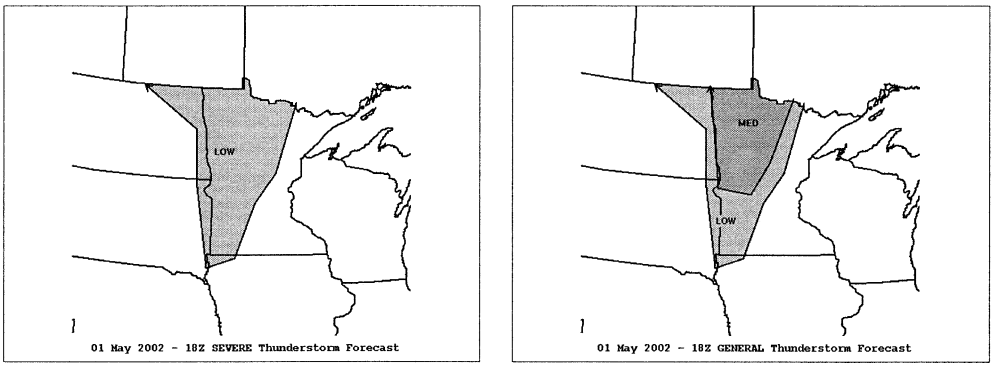

FIG. 2. An experimental forecast product issued at 1700 UTC for the 1800–2100 UTC time period on 1 May 2001.

ers. This warrants serious consideration because verification scores directly influence trends in model development. In recent years, newer generations of operational models have tended to favor diffusive representations of atmospheric processes in spite of the fact that the primary end users of model guidance (human forecasters) often prefer more realistic-looking detail.

The subjective verification component of the 2001 Spring Program was designed to address several aspects of this problem. The primary goal of this effort was to determine whether subjective interpretation and evaluation of numerical model output provides a valid measure of model performance when it is done using systematic and quantitative procedures. As corollary objectives, this study seeks to 1) document the disparity between widely used objective verification measures and human judgments of model performance, 2) develop a database of subjective verification statistics that could be used to calibrate new objective verification techniques that are being developed at NSSL and SPC (Baldwin et al. 2001), and 3) develop a better understanding of how forecasters are using model guidance.

The objectives of this paper are to document the procedures used to obtain subjective verification statistics during the 2001 NSSL/SPC Spring Program, to report on our progress in achieving the goals outlined above, and to offer recommendations for future subjective verification efforts. Section 2 provides a fairly detailed description of the methodology, followed by a summary of results in section 3, focusing on subjective verification of Eta and EtaKF precipitation forecasts. Then, section 4 contains a discussion of the results and methodology, followed by a summary in section 5.

2. Methodology

A thorough overview of the 2001 Spring Program is provided in KBAMS. This program brought operational forecasters from the SPC and the Norman, Oklahoma, National Weather Service Forecast Office together with numerical modeling experts from NSSL, EMC, the Forecast Systems Laboratory, and Iowa State University. The overriding goal of the program was to evaluate whether mesoscale model output could be used more effectively to predict convective initiation and severe convective weather development.

Each day forecast teams first performed a subjective verification of the previous day's forecasts and model guidance (see below). By about 1600 UTC, the teams completed the verification and moved on to an examination of the current day's weather. Although there was considerable interaction between team members during this period, operational forecasters naturally focused more on observational data for short-term convective forecasting, while those with numerical modeling expertise concentrated on a detailed examination of output from numerical models. After about an hour and a half, team members convened and formulated consensus probabilistic forecasts of general thunderstorms and severe convective weather over a limited domain (approximately 10° latitude by 10° longitude). This process was repeated for a second forecast later in the day. By following preset procedures, forecasters used familiar routines for much of the forecast-preparation process. Two important differences from operational convective watch procedures were that 1) they were required to issue forecast products at specified times, as opposed to normal event-driven responses, and 2) they infused information gleaned from expert interpretation of numerical guidance by model developers into the forecast-preparation process.

a. Model evaluation

After the team reached a consensus prediction, the lead forecaster began preparing graphics and text for the formal experimental forecast products (e.g., Fig. 2). At the same time, other team members began a systematic evaluation of models used in the preparation of the forecast. The lead forecaster joined in this task after the forecast product was issued.

A relatively small subset of the forecast domain was typically chosen for the model evaluation. Figure 3 shows an example of an evaluation area within the fore-

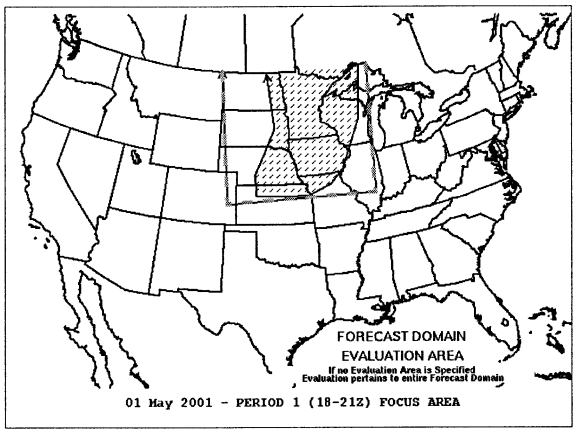

FIG. 3. Forecast domain, enclosed by the square box over the upper Midwest, for the time period indicated in Fig. 2. The hatched area indicates where the models were evaluated and verified.

cast domain. In this case, the evaluation area was larger than normal because the forcing for convective activity was widespread along a frontal boundary that extended through a large part of the forecast domain. Oftentimes, however, the evaluation area was made quite small in the interest of isolating a portion of the domain where the forecasting challenges and meteorological parameters were nearly homogeneous. This allowed a more distinct classification of convective events for later study, based on the meteorological regime within the evaluation area. For example, all dryline cases or warm frontal events could be grouped together. Placement of this evaluation area within the forecast domain was determined by the heightened interest of forecast teams. This interest was partly of a meteorological nature, but forecasters also were specifically instructed to consider disparity in model solutions as a primary criterion for choosing the evaluation area.

The Web-based evaluation form was fairly lengthy. It first queried team members for information about numerous meteorological factors, including the uncertainty associated with each one. These factors involved measures of moisture, instability, wind shear, and dynamic forcing—four fundamental ingredients used by forecasters to assess the likelihood of severe convection (e.g., Johns and Doswell 1992). Participants were then asked to provide a detailed evaluation of each model that was used in preparing the forecast, based on two fundamental characteristics of the forecasts (e.g., Fig. 4).

First, they were asked to provide an assessment of how favorable each factor was (in the specific model guidance) for the development of severe convection, using a scale from 0 to 10. The purpose of gathering data on this characteristic was to provide subjective

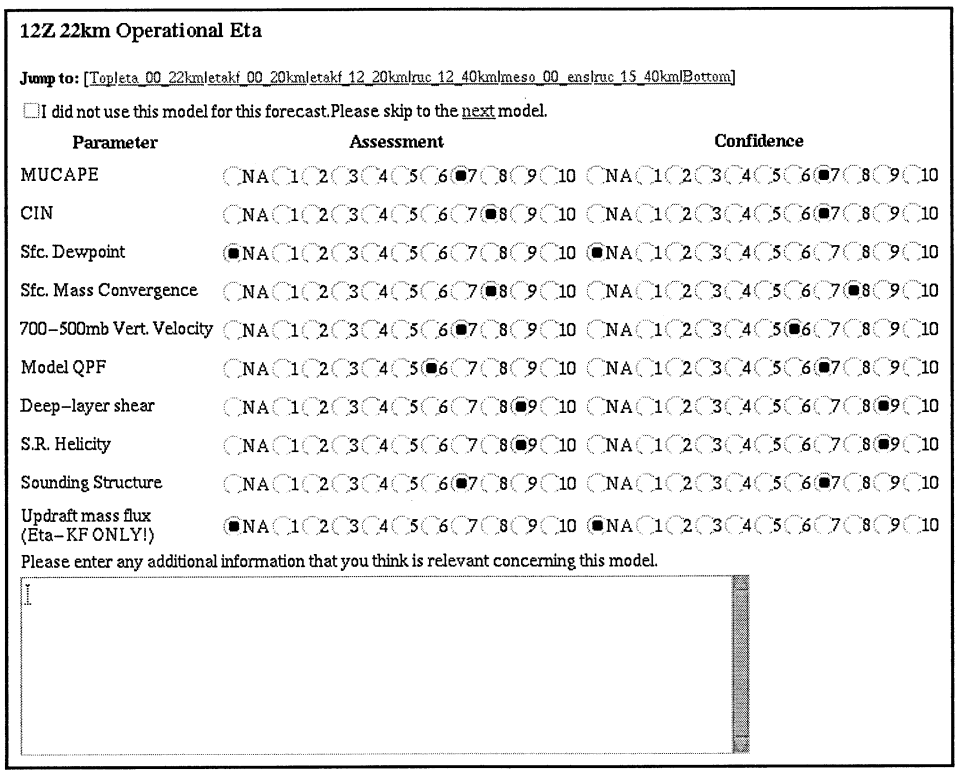

FIG. 4. A portion of the Web-based evaluation form showing the ''assessment'' and ''confidence'' values assigned to the 1200 UTC initialization of the Eta Model used to formulate the experimental forecast shown in Fig. 2.

comparisons of the model forecasts of certain fields. Such information can be valuable to forecasters, for example, to show that one model systematically predicts higher convective available potential energy (CAPE) values compared to another.

Next, team members were asked to express their *confidence* in each model's prediction of the same fields (i.e., whether they believed the forecast was accurate), again on a scale of 0 to 10. This query allowed us to document a fundamental element of the decision-making process involved in forecast preparation; namely, how consistent is the model's prediction with the forecaster's best judgment? This is fundamentally a measure of the goodness of a forecast in Murphy's (1993) type I sense. When combined with next-day verification data (see below), this evaluation enabled program organizers to ascertain whether forecasters were making the correct decisions when incorporating model guidance into their formulation of convective forecasts.

After all of the models used in the forecast process were evaluated with this detailed approach, participants were instructed to rate the overall utility of each model as a basis for the forecast. Finally, team members were asked to provide comments on short-range ensemble products and their utility in preparing the short-term convective forecast.

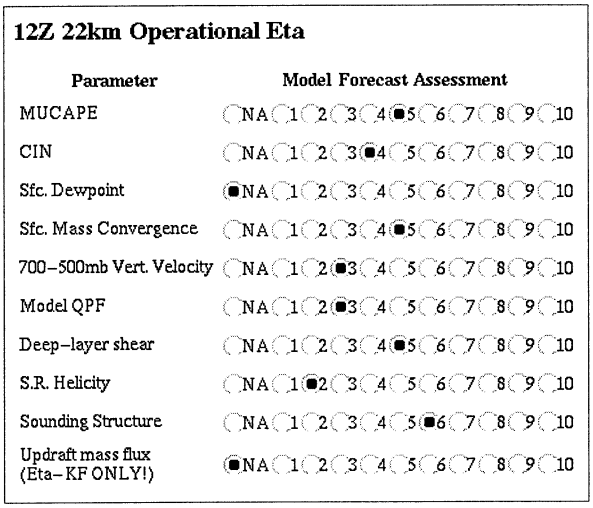

FIG. 5. A portion of the Web-based, next-day verification form showing the values assigned to the 1200 UTC initialization of the Eta Model used to formulate the experimental forecast shown in Fig. 2.

b. Model verification

As mentioned above, verification of the previous day's human forecasts and model guidance was the first order of business each day. The *same team* that both issued the forecasts and evaluated the model guidance performed the verification. This provided continuity in understanding the decision-making and interpretation processes that went into the use of model guidance and formulation of forecasts. Here, the focus is on verification of the models rather than the human forecasts. Each model field that was evaluated for confidence and utility at the time of forecast preparation was verified for accuracy the following day. As with the evaluation, subjective verification statistics were recorded on a Web-based form using a rating scale of 0–10 (e.g., Fig. 5).

Guidelines were given for assigning all numerical ratings. For example, forecast teams were instructed to assign verification ratings based on the following reference values:

- rating 0 = poor forecast: model missed all significant features and provided misleading guidance;
- rating 5 = fair forecast: model captured some significant features and provided fair guidance;
- rating 10 = excellent forecast: model captured all significant features and provided excellent guidance.

These guidelines were intentionally nonspecific because no two days were alike in terms of the character of meteorological features and the difficulty of the forecasting challenges. Forecast teams were encouraged to assign a numeric value to the ''best'' model forecast for each period first, and then assign lower numbers to other forecasts based on the assessment of each relative to the most accurate forecast.

Multiple observational data sources were used to infer ''ground truth'' for the verification process. Since there were modest levels of uncertainty in the verification of many fields, especially those involving three-dimensional structures, this paper focuses on the model quantitative precipitation forecast (QPF) output. However, it is emphasized that numerical guidance for precipitation is only one small ingredient in a multifaceted decision-making process for the SPC and other convective weather forecasters.

Model precipitation forecasts were compared to radar data and hourly surface reports, with occasional corroboration from satellite data. The combination of these sources provided very reliable estimates of the areal coverage of precipitation. Since convective initiation was the predominant interest of the forecast teams, the quantitative aspects of the precipitation field were not emphasized as much as timing and coverage in the evaluation and verification process (despite the label of quantitative precipitation forecast!). Note that forecast teams were not given specific instructions to differentiate between errors in coverage, displacement, amount, orientation, etc., although such instructions could be useful in a more refined approach.

3. Results

The results presented herein are limited to the Eta and EtaKF models. The purpose of this study is to examine the viability of subjective evaluation procedures, not to infer a definitive judgment in favor of one model

or another. Yet, comparison of the Eta and EtaKF output is of particular interest to us because the EtaKF is run and objectively verified locally. Furthermore, by limiting the comparison to just two different models, the sample size (number of forecast periods) is increased, enhancing the statistical significance of the results. Sample size tends to increase as the number of included models decreases because a particular forecast period had to be excluded if any model in the set to be compared was missing. Finally, attention is focused on measures of forecaster confidence during the forecast stage and verification of the same data the following day.

In this section, a particular event is presented in order to illustrate the judgments that were inherent in evaluating and verifying the models. Note that forecast teams did not document their justification for the ratings assigned in this case (or in most other cases). Consequently, the discussion that accompanies the presentation below is based largely on the authors' interpretation of the specific ratings. The case study is followed by statistical results covering the entire seven-week period of the Spring Program.

a. Case study

As described in Fig. 2, Spring Program forecasters were concerned about the possibility for severe convective weather over the northern plains on 1 May 2001. A deep tropospheric short-wave trough was moving rapidly into this area and a surface cold front was expected to be the focus for convective development (Fig. 6). The experimental forecast issued by the team at 1700 UTC, valid during the 1800–2100 UTC time period, is considered. For the purpose of this illustration, the only model outputs shown are the 3-h precipitation forecasts for this period.

The most significant difference between the Eta and EtaKF precipitation guidance for this event was along the southern portion of the cold front, from southwestern Minnesota to northeastern Nebraska. Over this region, the EtaKF triggered convection before 2100 UTC (Figs. 7b and 7d), whereas the operational Eta did not (Figs. 7a and 7c). This was true for both the 0000 and 1200 UTC initializations of these models. Forecasters expressed higher confidence in operational Eta solutions, with confidence ratings of 8 and 7 (out of 10) for the 0000 and 1200 UTC runs, respectively (Figs. 7a and c). The EtaKF received confidence ratings of 6 for both runs (Figs. 7b and 7d). So, the forecast teams did not reject the possibility of earlier initiation, as predicted by the EtaKF forecast, but they thought that this solution was somewhat less likely.

During the forecast period, deep convection did initiate along the southern half of the front, with the first cells forming near Sioux City, Iowa, between 1900 and 2000 UTC (Fig. 8a). By the end of the period (2100 UTC), the area of active convection had expanded along the front in both directions (Fig. 8b).

Next-day verification scores favored the EtaKF for correctly predicting the early initiation. The 1200 UTC EtaKF scored a 9 out of 10, likely because it clearly predicted two separate areas of convective rainfall, corresponding very well in timing and location with observed activity during the period (Fig. 7d). The 21-h forecast from the 0000 UTC EtaKF received a verification rating of 7, apparently being rewarded for indicating convective initiation along the southern half of the boundary, but losing points for failing to indicate two separate areas of activity (Fig. 7b). The 0000 UTC run of the operational Eta received a 5, apparently viewed favorably because it provided some indication that convection would activate toward the southern half of the frontal boundary, but penalized for not extending convection far enough south (Fig. 7a). The 1200 UTC forecast from the Eta was given only 3 out of 10 possible points. This run was likely penalized for limiting convective activity to northern Minnesota before 2100 UTC (Fig. 7c).

This example provides a cursory overview of the model evaluation and verification process. It is important to recognize that the convective precipitation fields shown here were evaluated and verified within the context of a much broader examination of model output and meteorological parameters, including time continuity within a specific model. For example, forecasters may have expressed more confidence in the Eta's QPF because it seemed to be providing more realistic predictions of related fields, such as CAPE and convective inhibition. In this type of experiment, it is impossible to remove these complicating factors. Yet, the goal of this experiment is to emulate the decision-making processes that forecasters face on a daily basis and to quantify the judgments that they make. Although one could quibble about the absolute values assigned the various model runs in this case, the relative rankings are readily justifiable. Over many different cases and with many different forecast teams, statistical analysis of these data can yield uniquely valuable information.

b. Statistical analysis

Data collected in the manner described above were compiled for statistical analysis. Results from this analysis are expressed in two different ways. First, mean values based on the raw ratings are computed. These values provide useful information about subjective impressions from the forecast teams, including inferences about how much better or worse one forecast is perceived to be compared to another (on average). These results can be misleading, however, because the benchmarks used to gauge model performance vary from forecast to forecast. For example, a perfect forecast for one event might turn out to be a prediction of no precipitation, while the next event may require extremely realistic timing and evolution of complex mesoscale convective structures for perfection.

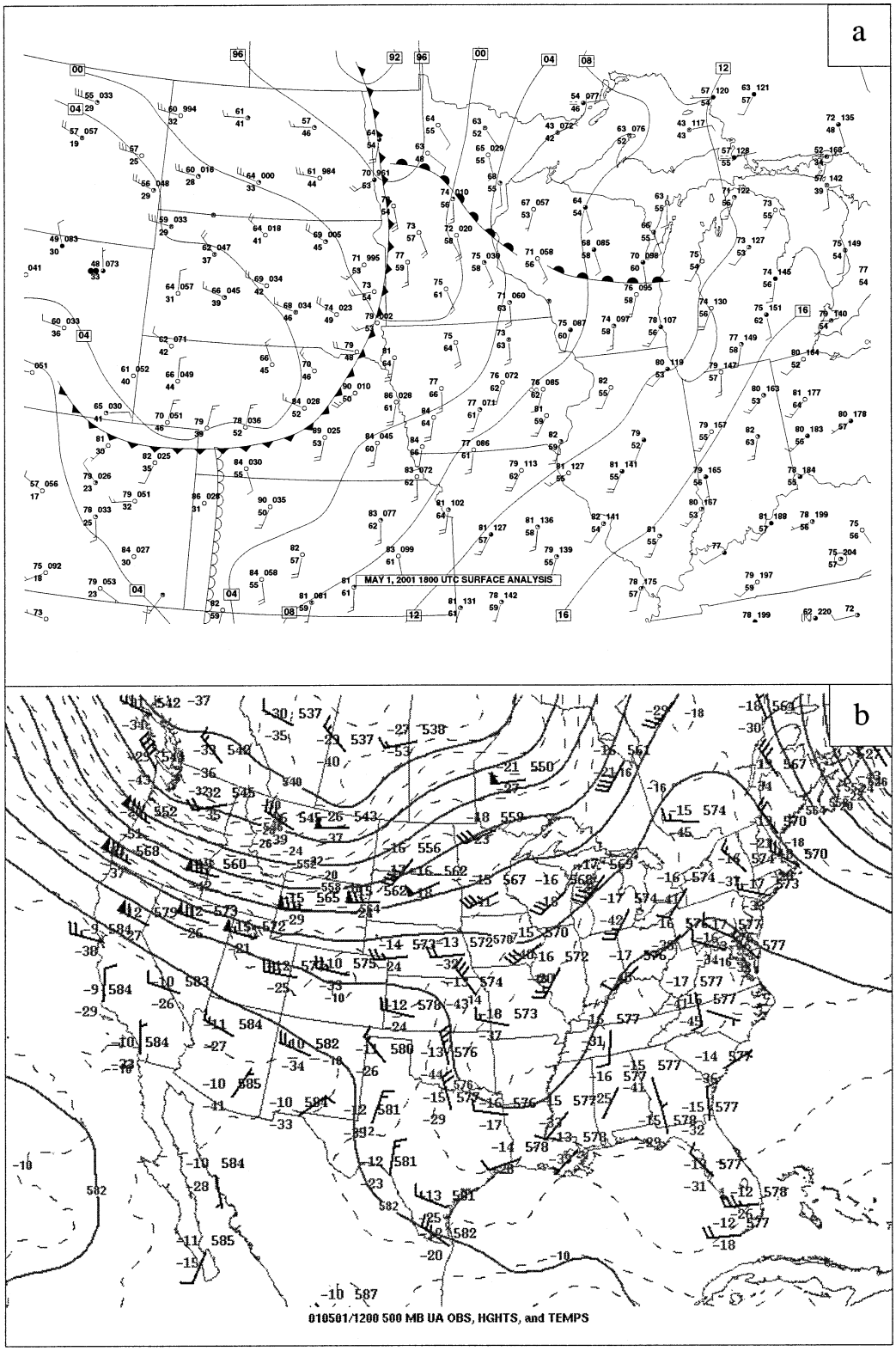

FIG. 6. Large-scale meteorological conditions for 1 May 2001. (a) Sea level pressure (4-hPa interval) and surface frontal analysis valid 1800 UTC, and (b) 500-hPa geopotential (solid lines, 6-dm interval) and temperature (dashed lines, 2°C interval) valid 1200 UTC. Station models and frontal symbols are standard.

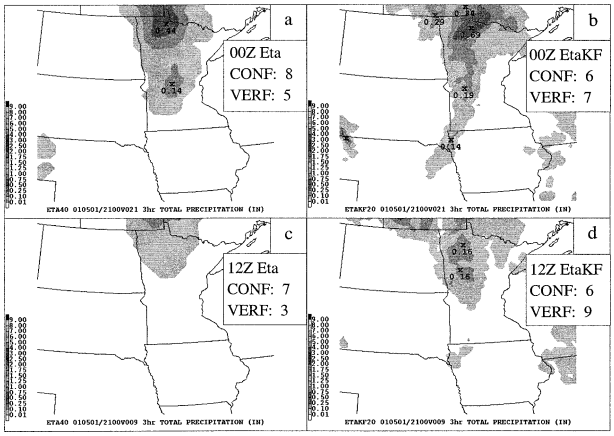

FIG. 7. Total precipitation forecasts for (a), (c) the Eta and (b), (d) the EtaKF for the 3-h time period 1800–2100 UTC 1 May 2001. Top panels show 21-h forecasts from the 0000 UTC initializations, while bottom panels show 9-h forecasts from the 1200 UTC initializations.

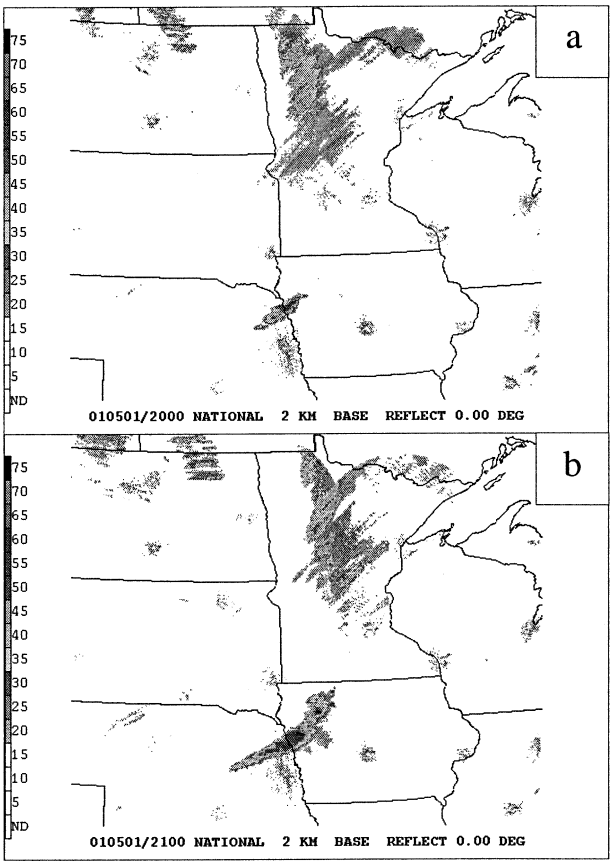

FIG. 8. Hourly maximum base reflectivity (dBZ) from the national Weather Surveillance Radar-1988 Doppler (WSR-88D) radar network valid at (a) 2000 and (b) 2100 UTC 1 May 2001.

To compensate for this inconsistency in absolute scale, a second analysis that is based on the relative rankings only is provided. These numbers are generated by ranking raw scores for each forecast period according to highest (rank value equal to the number of model forecasts in the comparison), second highest (rank value equal to number of forecasts minus 1), etc. In the case of ties, a mean number is assigned. For example, if for a particular forecast period one model out of four was given a rating of 8, two received 6s, and one received a 3, the relative rankings would be 4, 2.5, 2.5, and 1, respectively.

For each method, paired *t*-test scores (e.g., Wilks 1995) were computed in order to assess the statistical significance of any differences. A *t*-test score of 0.05 indicates that differences are significant at a 95% confidence level, and this value is often used as a threshold to distinguish between significance and nonsignificance. This threshold is used as a reference point, but a more general usage of *t*-test scores is emphasized, such that lower values imply a greater probability that differences are real and higher values suggest differences may not be real (see Nicholls 2001).

There were a total of 23 forecast periods from which the 0000 and 1200 UTC precipitation forecasts from the Eta and EtaKF were all evaluated and verified. Considering forecast-team *confidence* at the time forecasts were issued, statistical analysis of the *raw scores* shows that, on average, forecasters expressed the highest confidence in forecasts from the 1200 UTC run of the EtaKF, followed by the 0000 UTC EtaKF, the 1200 UTC Eta, and the 0000 UTC Eta (Fig. 9a). Very low *t*-test scores in pairings of either initialization of the EtaKF with either run of the Eta imply that confidence in both of the EtaKF runs was significantly higher than confidence in the Eta. Paired *t*-test scores were still quite low when the 1200 and 0000 UTC runs of the EtaKF were compared, suggesting that the graphical difference between these two initializations is significant. In contrast, the *t*-test score was close to the maximum value of 1 when confidence ratings for the two initializations of the Eta were compared, implying that there is little discernible difference in forecaster confidence for these two runs. When the analysis was based on mean rank, rather than the raw rating, some subtle changes occurred in *t*-test scores, although the order of the models from highest to lowest did not change (Fig. 9b).

Next-day *verification* scores followed the same order from highest to lowest when the raw ratings were considered (Fig. 9c). This result is encouraging because it seems to suggest that, on average, forecast teams were making good decisions in choosing which runs to favor and which to discount. However, this comparison does not reveal whether the confidence and verification ratings are correlated, that is, whether high (low) confidence ratings are associated with high (low) verification ratings for individual forecasts.

In general, *t*-test scores are slightly higher for these verification ratings than they were for confidence, indicating a somewhat lower probability that differences between individual pairings are real. The 1200 UTC EtaKF appears to verify with a significantly higher av-

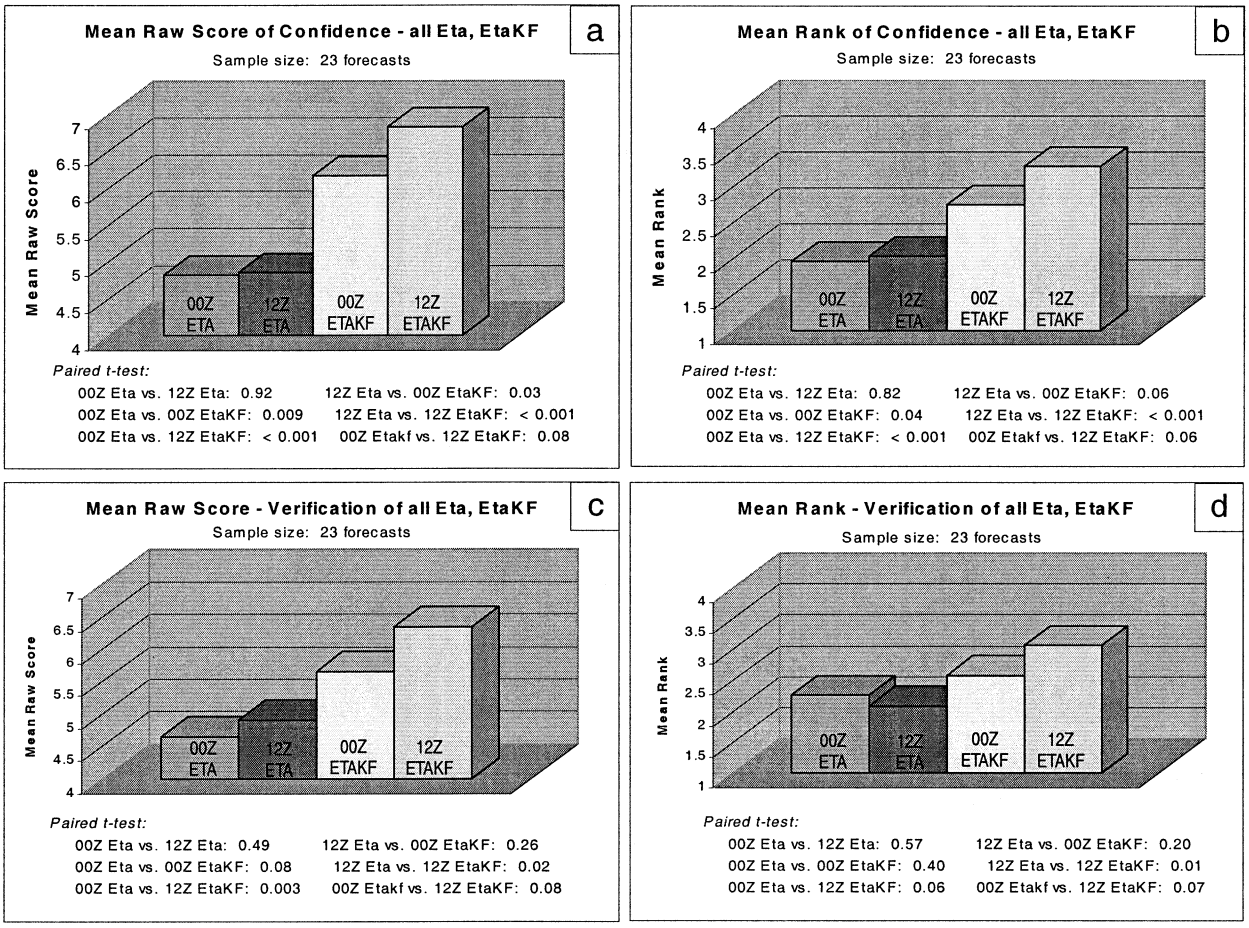

FIG. 9. Statistical results from surveys of (a), (b) day-1 forecaster confidence and (c), (d) day-2 verification for Eta and EtaKF forecasts. Mean raw scores are shown on the left and mean rankings are shown on the right.

erage rating than both the 0000 and 1200 UTC Eta runs, as was the case with forecaster confidence in this run. Yet, there is less certainty about differences between the 0000 UTC initialization of the EtaKF and the two runs of the Eta than there is for confidence ratings. Most notably, a pairing of the 1200 UTC Eta and the 0000 UTC EtaKF yielded a *t*-test score of 0.26, leaving considerable doubt about the significance of this difference.

When the verification data were transformed based on ranking rather than raw numeric ratings, an interesting change occurred. The 0000 UTC Eta earned better verification numbers than the 1200 UTC run (Fig. 9d). Paired *t*-test scores remained quite high, but the lack of distinction in itself yields the surprising result that 6–9-h convective rainfall forecasts from the Eta are often less skillful than 18–21-h forecasts. When the two EtaKF initializations are compared, a paired *t*-test score of 0.07 still inspires a fairly high degree of certainty that these two runs are different. The *t*-test scores for the 1200 UTC Eta–EtaKF pairing remain quite low at a value of 0.01.

These comparisons can be refined and the sample size increased by focusing on individual pairings of model runs. Comparisons of like-time Eta and EtaKF initializations and different initializations for the same model configuration are summarized in Table 1, revealing that the fundamental relationships depicted in Fig. 9 are largely unchanged with the slightly larger sample sizes.

Additional insight into the data can be gained by examining scatterplots. A plot of the verification ratings for the 1200 UTC run of the Eta against the same initialization of the EtaKF (Fig. 10a) clearly reveals the highest density of data points favoring the EtaKF. Yet, it also shows that the Eta verified much better than the EtaKF on some days, in spite of the lower average rating. A similar plot for the two different initializations of the Eta (Fig. 10b) shows a more evenly distributed density of data points, although the 0000 UTC run is favored.

A different picture emerges when forecaster confidence is plotted against verification score for individual forecast periods. Such plots show little, if any, correlation between confidence and verification, regardless of which model or initialization time is considered. For example, Figs. 11a and 11b show data from the 0000 UTC initializations of the Eta and EtaKF, respectively.

TABLE 1. Summary of forecaster confidence and subjective verification statistics for selected individual pairings of Eta and EtaKF forecasts during the 2001 Spring Program.

Model runs	No. of forecasts	Forecaster confidence				Subjective verification			
		Raw	*t* test	Rank	*t* test	Raw	*t* test	Rank	*t* test
1200 UTC Eta	34	4.76	<0.001	1.22	<0.001	4.91	0.005	1.28	0.007
1200 UTC EtaKF		6.38		1.78		6.35		1.72	
0000 UTC Eta	30	4.73	0.001	1.32	0.025	5.03	0.3	1.48	0.85
0000 UTC EtaKF		6.17		1.68		5.57		1.52	
0000 UTC Eta	35	4.6	0.62	1.46	0.57	5.26	0.38	1.61	0.15
1200 UTC Eta		4.77		1.54		4.94		1.39	
0000 UTC EtaKF	25	6.16	0.23	1.46	0.6	5.76	0.26	1.4	0.2
1200 UTC EtaKF		6.6		1.54		6.24		1.6	

These plots are consistent with the relative distribution of mean scores discussed above; for example, points on the EtaKF scatterplot are concentrated at comparatively high values of both confidence and verification, but *the wide scatter of individual points in both plots suggests that forecast teams had little or no skill in predicting how accurate a particular model forecast would be.* Essentially the same result is obtained if individual model rankings are analyzed instead of the raw ratings (not shown).

Before passing further judgment on these results, it is important to emphasize the context from which they were derived. During the 2001 Spring Program, forecast teams were instructed to focus on areas where there was disagreement among model forecasts and significant meteorological unknowns as well. Thus, by design they chose evaluation areas where the forecasting challenges were the greatest and the inherent uncertainty was maximized. This strategy was adopted as a means of encouraging forecasters to aggressively interrogate numerical model output for guidance when they faced tough forecasting scenarios, consistent with the broad objectives of the 2001 Spring Program. Notwithstanding this challenging forecasting environment, the scatterplots of confidence versus verification suggest that forecast teams were not able to identify a ''model of the day'' in advance, at least not with any consistency.

In summary, statistical analysis suggests that forecast-team confidence was consistent with the perceived skill of model forecasts when averaged over all events. However, forecasters appeared to have little or no skill in discerning which model(s) would provide the best guidance for a particular event.

c. Comparison with an objective measure

As stated earlier, a corollary objective of the 2001 Spring Program was to compare forecaster impressions of model performance with objective verification measures of the same model forecasts. To this end, ET scores for Eta and EtaKF were calculated for the same forecast periods and spatial domains used to generate Fig. 9. Results show that the two models score similarly, but the Eta Model rates somewhat higher, especially at the rain–no rain threshold (Fig. 12). It is worth noting that the EtaKF output has a higher-frequency bias than the Eta at all precipitation thresholds (not shown). Since higher bias values can inflate ET scores (Hamill 1999), it is possible that the amplitude of the EtaKF curve is somewhat exaggerated in this figure. Yet, it still lies well below the Eta curve at the lowest threshold, in distinct contrast to the relative positions of the Eta and EtaKF bars in Figs. 9c and 9d. This inconsistency substantiates the concern that prominent objective verification measures used at operational centers often fail to provide a judgment that is consistent with forecasters' impressions of forecast value. This problem is well known and other alternatives to ET and bias scores are currently being investigated by scientists in the National Weather Service (National Oceanic and Atmospheric Administration 1999; Charba et al. 2003).

4. Discussion

As stated in the introduction, the purpose of subjective verification for numerical models is to provide quantitative verification measures that are consistent with end users' (e.g., forecasters') impressions of model performance. These measures are badly needed because currently used objective verification techniques are only partially successful (at best) in this regard. Yet, systematic subjective verification is rare, perhaps because it is a challenging and tedious process into which uncertainty creeps from numerous sources. Some of the challenges that were faced in the Spring Program 2001 are highlighted below. These can be used as guidelines for future subjective verification efforts.

Every participant in the Spring Program was predisposed to personal bias in some way. For example, many forecasters were most comfortable or familiar with certain models. Likewise, modeling experts typically had research interests in only one or two of the models. However, these personal inclinations do not invalidate the results. All participants were encouraged to be as objective as possible and individual ratings were assigned by consensus of at least two team members. Furthermore, teams comprised a diverse group of individuals, with at least one modeling expert and one fore-

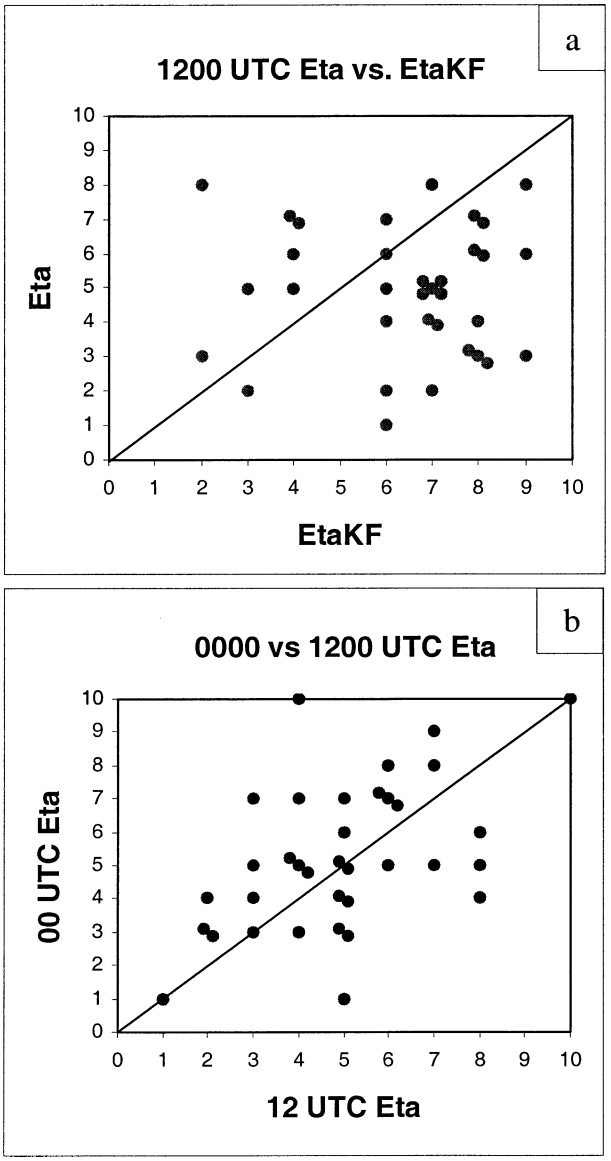

FIG. 10. Scatterplots for subjective verification raw ratings for (a) 1200 UTC Eta vs 1200 UTC EtaKF and (b) 0000 UTC Eta vs 1200 UTC Eta.

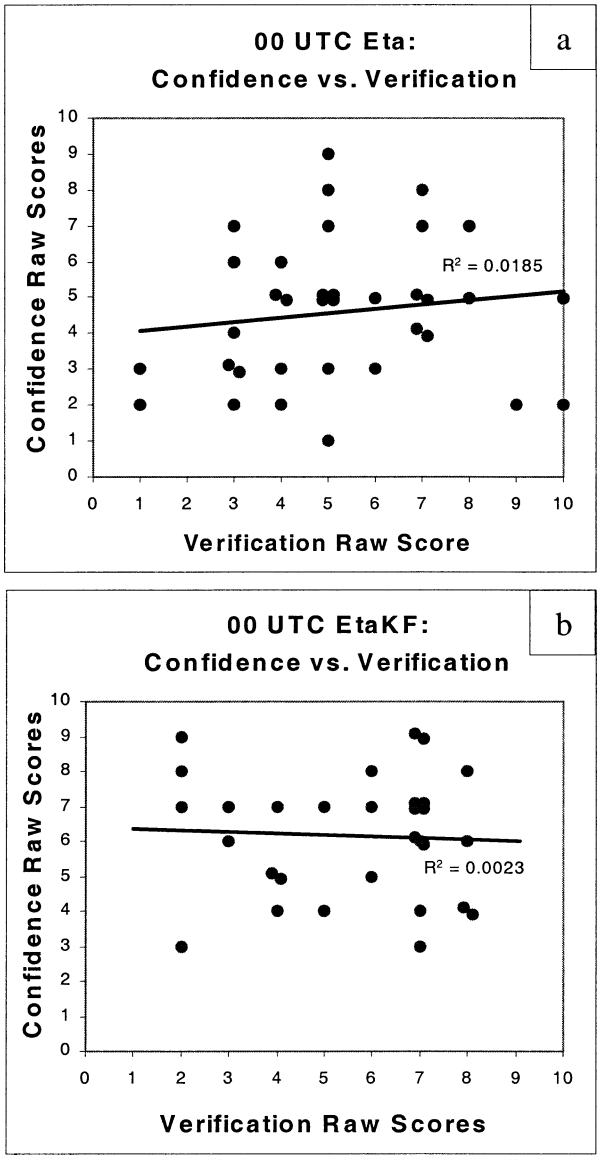

FIG. 11. Scatterplots of confidence vs verification for the raw ratings of individual forecasts for (a) 0000 UTC Eta and (b) 0000 UTC EtaKF. Linear best-fit lines (least squares method) are shown in each plot and the coefficients of determination (R^2) are indicated.

caster in each group; teams members came from different modeling and forecast centers and they were encouraged to rotate through the various tasks associated with model evaluation and verification over the course of each week. Personal biases cannot be eliminated, but we believe their impact was minimized by the procedures used.

It is possible that evaluation and verification procedures may have been influenced by inconsistencies in model output displays. Ideally, all displays of a given output field should use identical map backgrounds, contour intervals, color fill patterns, etc.; if not, comparisons can become quite difficult. Early in the 2001 Spring Program, there was some inconsistency in the various output displays from several models. Before these variations were eliminated, the use of different color fills, contour intervals, etc. complicated the comparisons of some output fields. This was not an issue for the QPF fields, the focus of this paper, yet it was a problem for some other fields and it is an important consideration for future studies.

Another potential factor was mental fatigue. Despite efforts to streamline the process as much as possible, evaluation and verification forms were lengthy and rather tedious. These factors could facilitate human error or simply diminish the level of effort contributed to the task.

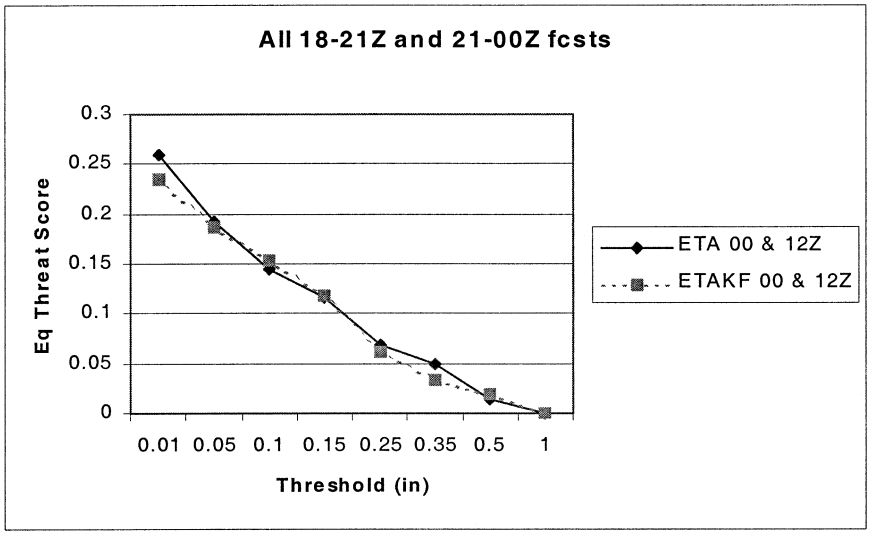

FIG. 12. Equitable threat scores as a function of precipitation threshold over the same time and space domains (evaluation areas) included in the subjective verification results shown in Fig. 9.

In spite of these factors, we believe that the dataset from the 2001 Spring Program is quite robust. Some may argue that these issues might weaken, or even invalidate, the findings. Although a number of logistical and/or ''human'' issues may have influenced the statistical results, we believe that in the end the diversity of the participants and their unique backgrounds and interests are actually a major strength of the experiment. In particular, this diversity is valuable because it allows us to receive input from experts across a spectrum of areas that are critical to the development and application of numerical models.

Results from a survey of all participants seem to corroborate this assessment. Excluding the four organizers of the program, 19 out of 26 participants responded to the postprogram survey. As part of the survey, they were asked two questions related to the validity of the evaluation and verification components. In the spirit of the program, they were asked to respond on a scale of 0–10, with 10 representing excellent and 0 poor, as follows:

- *Evaluate how well the process of evaluating* model utility/confidence *at the time of the forecast provided a fair and accurate assessment of the relative role played by each model in creating the forecast and extracting feedback from the forecast team—average rating, 7.1.*
- *Evaluate how well the process of* subjective verification *in postanalysis led to a fair and accurate assessment of model forecasts of fields relevant to convective initiation and severe convection—average rating, 7.1.*

Comments solicited from this section of the survey suggest that when lower ratings were given, it was because team members sometimes expressed initial disagreement over the appropriate ranking. Some compromise and/or discussion were often necessary before a team could settle on a final numeric value in the rating process. This was sometimes frustrating, but we believe it reflects a strength of the process rather than a weakness. It is not surprising that operational forecasters and model developers, or even different forecasters, view model output from different perspectives in some, if not all, cases. When this happens, a compromise opinion is appropriate, though it may not be particularly satisfying to any of the participants. In the final analysis, *this consensus approach is usually better than relying on only forecasters or research scientists to conduct an evaluation/verification.*

We believe that the difficulties and uncertainties associated with subjective verification have discouraged efforts like this in the past. However, we strongly advocate the use of subjective verification because current objective verification metrics are not consistent with the notion of measuring the value of mesoscale model forecasts to human forecasters. The statistics obtained in this study will be used to help calibrate new verification metrics that are currently being developed at the NSSL and SPC (Baldwin et al. 2001).

5. Summary

A seven-week program of model evaluation and experimental forecasting took place at the NSSL and SPC during the spring of 2001. The 2001 Spring Program was one of a continuing series of collaborative efforts at the NSSL/SPC facility that have been characterized by rare synergies of operational and research meteorologists (KBAMS). The primary objective of this program was to evaluate whether convective initiation could be predicted more accurately if forecasters could

use mesoscale models more effectively. The program was specifically designed to take SPC forecasters ''out of their comfort zone'' by introducing two important changes to their routine forecast-preparation process. First, they were assigned to forecast teams with at least one numerical modeler, so that expert interpretation and interrogation of model output would be integrated into the forecast-preparation process. Second, forecast teams were required to issue an experimental product similar to a convective watch (i.e., severe thunderstorm or tornado watch), but the product was issued at scheduled times rather than being driven by developing weather (see KBAMS). In particular, forecasters were unable to wait until satellite or radar data indicated that convective development was imminent. Probability forecasts for convective development were often issued hours before observations revealed the precise location and timing of activity. Forecast teams were encouraged to utilize model guidance to compensate for the lack of timely observational evidence for convective development.

Subjective evaluation and verification of the model guidance were central components of the 2001 program and the focus of this paper. The evaluation component included a survey of forecast-team confidence in various model solutions. In particular, at the time that experimental forecasts were issued, forecast teams were instructed to assess their confidence that a particular model solution would be correct, on a scale from 0 to 10. The same forecast team performed a subjective verification of model forecasts the following day, using a similar rating scale. This methodology was particularly useful because it allowed a comparison of confidence and verification statistics, providing an assessment of whether forecasters were making good decisions in their use of model data. In essence, the confidence assessment provided a measure of the ''goodness'' of a forecast in Murphy's (1993) type 1 sense (i.e., a measure of its consistency with a forecaster's best judgment), while the verification component yielded a unique quantification of goodness in Murphy's type 2 sense (a measure of its correspondence to observations).

In this paper, a relatively small subset of the evaluation and verification statistics is presented. Specifically, this paper focuses on precipitation forecasts (primarily convective initiation) from the 0000 and 1200 UTC Eta and EtaKF model runs. Averaged over many forecast periods during the 2001 Spring Program, the relative rankings of these four model runs were the same for both confidence and verification. In particular, the 1200 UTC initialization of the EtaKF had the highest ranking, followed by the 0000 UTC EtaKF, the 1200 UTC Eta, and the 0000 UTC Eta. The consistency in trends indicates that forecasters favored particular model solutions during the prediction stage about as frequently as those solutions were favored in postevent subjective comparisons with observed data. Surprisingly, however, when confidence and verification were compared for individual forecasts, there was little or no correlation between the two. It seems that *forecast teams demonstrated little skill in predicting how accurate individual model forecasts of convective initiation would be.* Further examination also revealed that forecasters were unable to discern which model output would be more accurate on a given day, that is, what the relative accuracy of the model runs would be.

These results suggest that increased reliance on precipitation forecasts from mesoscale models (at least from those models available during the spring of 2001) would be unlikely to improve predictions of convective initiation by SPC forecasters. Although most of the models that were tested produced very good predictions (with respect to timing and location) of deep convection some of the time, forecast teams were unable to discern when model forecasts could be trusted. *Interpretation and assessment of model forecasts remains a very challenging task, even when convective forecasting specialists and model experts combine efforts as part of a forecast team.* These results support anecdotal evidence from the operational community that forecasters have few tools enabling them to know how much confidence to place in a particular mesoscale model solution on a given day, especially when small temporal and spatial scales are considered. Work is under way to determine how forecast accuracy might vary as a function of larger-scale meteorological conditions and other factors (e.g., Kay and Baldwin 2002).

The subjective evaluation and verification approach described in this paper provides unique insight into the ways that forecasters use model data and it allows investigators to focus on a particular element of model forecasts that is important to certain groups of users. For example, the precipitation field highlighted in this paper was used more as a measure of convective initiation rather than as QPF guidance, the former being more relevant to the needs of the SPC as opposed to, say, the Hydrometeorological Prediction Center (HPC). This approach could be substantially refined without much additional effort. For example, with regard to the precipitation field, verification teams could be asked to elaborate on comparisons between predicted and observed fields, separately quantifying errors in timing, displacement, and areal coverage of specific meteorological features. Data of this type would have significant value for model developers and would be consistent with the intended end result of new objective verification procedures currently being developed at NSSL and SPC (Baldwin et al. 2001). Our ultimate goal is to use subjective verification to provide ''insight into what is right and what is wrong about the forecasts, rather than the mere production of verification statistics for ranking of relative performance'' (Doswell and Flueck 1989). Thus, the subjective verification procedures described here are viewed as a foundation upon which future verification efforts can build.

Even though the data gathered during the 2001 Spring Program were relatively crude, they clearly provide in-

formation that cannot be inferred from the equitable threat score, a bellwether metric at NCEP. When this score was computed for the same spatial domain and time periods as our subjective verification, it showed distinctly different results. This disparity is consistent with anecdotal evidence supplied by SPC forecasters. It is important that many different types of verification metrics be used to guide numerical model development, as emphasized by others (McDonald 1998; National Oceanic and Atmospheric Administration 1999; Charba et al. 2003). Carefully and systematically gathered subjective verification data appear to have an important role to play in this process.

Acknowledgments. Special thanks and appreciation are extended to all participants and staff for assisting in the preparations/planning, programming, and data flow issues associated with the 2001 Spring Program. In particular, special thanks to John Hart (SPC) for software support and development; Phillip Bothwell (SPC) and Gregg Grosshans (SPC) for providing access to model and verification data; Dave Stensrud (NSSL) for experimental MM5 data access; Jay Liang (SPC), Doug Rhue (SPC), Steve Fletcher (NSSL), and Brett Morrow (NSSL) for assistance in configuring hardware and software; and Charlie Crisp (NSSL) for his expert meteorological analysis and many other contributions. We further wish to recognize the full support of SPC and NSSL management and enthusiasm by participants from the Forecast Systems Laboratory (FSL), Environmental Modeling Center (NCEP/EMC); Norman, Oklahoma, National Weather Service Forecast Office; and the Iowa State University who helped make this undertaking a positive experience for everyone. Thanks to Harold Brooks (NSSL) for his review and guidance of this work and to Nita Fullerton (FSL) for a thorough editorial review of this manuscript. Critical comments and suggestions from three anonymous reviewers and Matt Wandishin of the University of Arizona/NSSL significantly improved this manuscript. Our interpretation of results benefited from discussions with Chuck Doswell. This work was partially funded by NOAA–OU Cooperative Agreement NA17RJ1227 and COMET Cooperative Project 099-15805.

REFERENCES

Baldwin, M. E., and M. S. Wandishin, 2002: Determining the resolved spatial scales of Eta Model precipitation forecasts. Preprints, *19th Conf. on Weather Analysis and Forecasting,* San Antonio, TX, Amer. Meteor. Soc., 85–88.

——, S. Lakshmivarahan, and J. S. Kain, 2001: Verification of mesoscale features in NWP models. Preprints, *Ninth Conf. on Mesoscale Processes,* Fort Lauderdale, FL, Amer. Meteor. Soc., 255–258.

——, J. S. Kain, and M. P. Kay, 2002: Properties of the convection scheme in NCEP's Eta Model that affect forecast sounding interpretation. *Wea. Forecasting,* **17,** 1063–1079.

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

Charba, J. P., D. W. Reynolds, B. E. McDonald, and G. M. Carter, 2003: Comparative verification of recent quantitative precipitation forecasts in the National Weather Service: A simple approach for scoring forecast accuracy. *Wea. Forecasting,* **18,** 161–183.

Craven, J. P., R. E. Jewell, and H. E. Brooks, 2002: Comparisons between observed convective cloud-base heights and lifting condensation level for two different lifted parcels. *Wea. Forecasting,* **17,** 885–890.

Doswell, C. A., and J. A. Flueck, 1989: Forecasting and verifying in a field research project: DOPLIGHT '87. *Wea. Forecasting,* **4,** 97–109.

Evans, J. S., and C. A. Doswell III, 2001: Examination of derecho environments using proximity soundings. *Wea. Forecasting,* **16,** 329–342.

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

Janjić, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.,* **122,** 927–945.

Johns, R. H., and C. A. Doswell III, 1992: Severe local storms forecasting. *Wea. Forecasting,* **7,** 588–612.

Kain, J. S., and J. M. Fritsch, 1993: Convective parameterization for mesoscale models: The Kain–Fritsch scheme. *The Representation of Cumulus Convection in Numerical Models, Meteor. Monogr.,* No. 46, Amer. Meteor. Soc., 165–170.

——, S. M. Goss, and M. E. Baldwin, 2000: The melting effect as a factor in precipitation-type forecasting. *Wea. Forecasting,* **15,** 700–714.

——, M. E. Baldwin, and S. J. Weiss, 2003a: Parameterized updraft mass flux as a predictor of convective intensity. *Wea. Forecasting,* **18,** 106–116.

——, P. R. Janish, S. J. Weiss, M. E. Baldwin, R. Schneider, and H. E. Brooks, 2003b: Collaboration between forecasters and research scientists at the NSSL and SPC: The Spring Program. *Bull. Amer. Meteor. Soc.,* in press.

Kay, M. P., and M. E. Baldwin, 2002: Combining objective and subjective information to improve forecast evaluation. Preprints, *19th Conf. on Weather Analysis and Forecasting,* San Antonio, TX, Amer. Meteor. Soc., 411–414.

McDonald, B. E., 1998: Sensitivity of precipitation forecast skill to horizontal resolution. Ph.D. dissertation, University of Utah, 135 pp. [Available online at ftp://ftp.hpc.ncep.noaa.gov/brett/diss/.]

Mesinger, F., 1996: Improvements in quantitative precipitation forecasts with the Eta regional model at the National Centers for Environmental Prediction: The 48-km upgrade. *Bull. Amer. Meteor. Soc.,* **77,** 2637–2650.

Moller, A. R., C. A. Doswell III, M. P. Foster, and G. R. Woodall, 1994: The operational recognition of supercell thunderstorm environments and storm structures. *Wea. Forecasting,* **9,** 327–347.

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

National Oceanic and Atmospheric Administration, 1999: The modernized end-to-end forecast process for quantitative precipitation information: Hydrometeorological requirements, scientific issues, and service concepts. National Weather Service/Office of Climate, Water, and Weather Services, 187 pp. [Available from Office of Climate, Water, and Weather Services, W/OS, 1325 East West Hwy., Silver Spring, MD 20910.]

Nicholls, N., 2001: Commentary and analysis: The insignificance of significance testing. *Bull. Amer. Meteor. Soc.,* **81,** 981–986.

Stensrud, D. J., and S. J. Weiss, 2002: Mesoscale model ensemble forecasts of the 3 May 1999 tornado outbreak. *Wea. Forecasting,* **17,** 526–543.

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