J4.2 SUBJECTIVE VERIFICATION OF NUMERICAL MODELS AS A COMPONENT OF A BROADER INTERACTION BETWEEN RESEARCH AND OPERATIONS

John S. Kain*,[1,2], Michael E. Baldwin[1,2,3], Steven J. Weiss[3], Paul R. Janish[3], Gregory W. Carbin[3], Michael P. Kay[4,5],
and Laurita Brown[6]

[1]Cooperative Institute for Mesoscale Meteorological Studies
[2]University of Oklahoma/NOAA Research - National Severe Storms Laboratory
[3]NOAA/NWS/Storm Prediction Center
[4]Cooperative Institute for Research in Environmental Sciences (CIRES)
[5]University of Colorado/NOAA Research – Forecast Systems Laboratory
[6]Jackson State University

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. 2002a; Kain et al. 2000, 2002) are among the organized interactions, but the cornerstone of this collaboration in the last several years has been intensive multi-week research programs conducted during each spring severe weather season. This effort has become known as the NSSL/SPC Spring Program (Kain et al. 2003).

Forecasters at the SPC rely on a variety of observational and mesoscale guidance in preparation of convective outlooks, severe thunderstorm and tornado outlooks, 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 et al. 2002, hereafter KF) in place of the operational Betts-Miller-Janjic scheme (Janjic 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 (http://www.nssl.noaa.gov/etakf). 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. 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 inspired by empirical comparisons of Eta and EtaKF output over the past several years. Operational forecasters and collaborating research scientists have noted that the EtaKF certainly does not outperform the operational Eta every day, but 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.

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. 2002b). 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 forecasters. 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, we sought 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

* *Corresponding author address:* Jack Kain, NSSL, 1313 Halley Circle, Norman, OK 73069

al. 2002b), and 3) develop a better understanding of how forecasters are using model guidance.

This paper is a condensed version of a full article that has been submitted to Weather and Forecasting, also available at http://www.nssl.noaa.gov/mag/subj_verf_paper.pdf. 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 brief description of the methodology, followed by results focusing on subjective verification of Eta and EtaKF precipitation forecasts, then a summary.

2. METHODOLOGY

An overview of the 2001 Spring Program is provided in Kain et al. (2003). This program brought operational forecasters from the SPC and the Norman, OK, 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 weather development.

Subjective evaluation and subjective verification were two key ingredients of the daily routine during the Spring Program. Both of these activities utilized web-based survey forms to query forecast teams about model forecasts. All forms were based on a rating scale 0-10, with 0 being the lowest possible score and 10 the highest.

Model evaluation occurred immediately after forecast products were issued and was designed to measure forecaster *confidence* in particular model solutions. Verification took place the next day and provided a subjective assessment of the consistency between each model forecast and observations. Comparison of confidence and verification ratings provides some insight in whether forecasters are making the right decisions in weighting some model solutions more heavily than others.

3. RESULTS

The surveys discussed above presented forecasters with opportunities to rate confidence and verification in regard to ten model output fields from up to ten different models (Kain et al. 2003), although on a typical day data were entered for only four or five models and about the same number of output fields. For this paper, we focus on the precipitation field from the Eta and EtaKF models. Our purpose is to examine the viability of subjective evaluation procedures, not to infer a definitive judgment in favor of one model or another.

3.1 *Statistical Analysis*

Data collected from the surveys 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. 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.

To compensate for this inconsistency in absolute scale, we provide a second measure that is based on the *relative* rankings only. 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 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. We use this threshold as a reference point, but emphasize a more general usage of t-test scores in which lower values imply a greater probability that differences are real and higher values suggest differences may not be real.

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. 1a). 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. 1b).

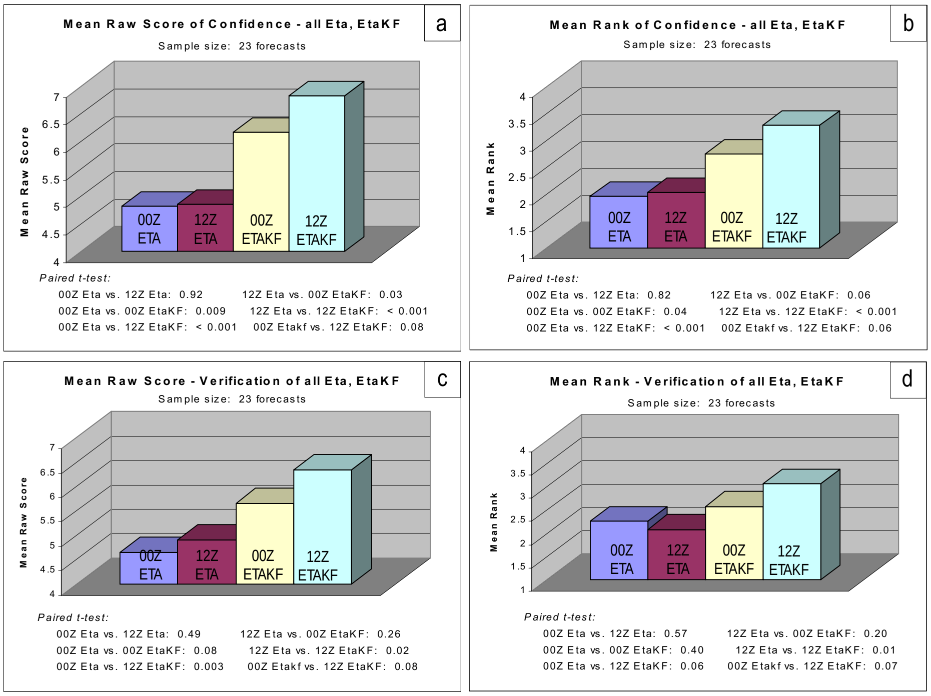

Fig. 1. Statistical results from surveys of day 1 forecaster confidence (a and b, top) and day 2 verification (c and d, bottom) for Eta and EtaKF forecasts. Mean raw scores are shown on the left and mean rankings are shown on the right.

Next-day *verification* scores followed the same order from highest to lowest when the raw ratings were considered (Fig. 1c). This result is encouraging because it suggests that forecast teams were making good decisions in choosing which runs to favor and which to discount. In general, t-test scores are slightly higher than they were for confidence ratings, indicating a somewhat lower probability that differences between individual pairings are real. The1200 UTC EtaKF appears to verify with a significantly higher average 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. 1d). 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. T-test scores for the 1200 UTC Eta-EtaKF pairing remain quite low at a value of 0.01.

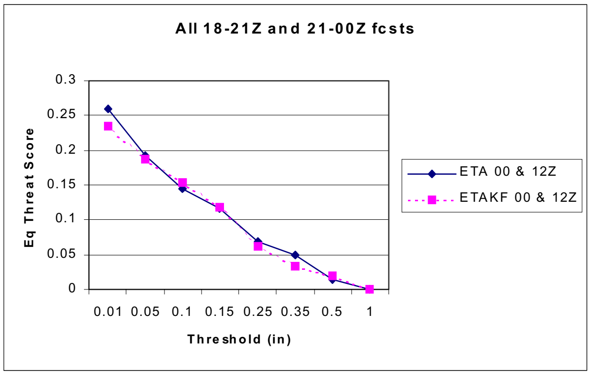

Fig. 2. Equitable threat scores as a function of precipitation threshold over the same time and space domains (evaluation areas) included in the subjective verification results shown in Fig. 1.

3.2 *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, equitable threat scores for Eta and EtaKF were calculated for the *same forecast periods and spatial domains* used to generate Fig. 1. Results show that the two models score similarly, but the Eta model is somewhat higher, especially at the rain/no rain threshold (Fig. 2). This result is distinctly different from the subjective comparison, which clearly favors the EtaKF, substantiating our 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.

4. SUMMARY

A seven-week long 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 in a continuing series of collaborative efforts at the NSSL/SPC facility that have been characterized by rare synergies of operational and research meteorologists. Subjective evaluation and verification of numerical models was a central component of this 2001 program and the focus of this paper. In essence, the evaluation component 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 part yielded a unique quantification of goodness in Murphy's type 2 sense (a measure of its correspondence to observations).

The results shown herein provide a sample of the utility of subjective, quantitative assessments of output from numerical weather prediction models. For example, by concentrating on precipitation forecasts from the 0000 and 1200 UTC Eta and EtaKF model runs, it has been shown that

- At the time when experimental forecasts were issued, forecast teams expressed the highest confidence in the 1200 UTC initialization of the EtaKF. This model also earned the highest subjective verification scores, but the margin of difference from the other models was smaller for verification than for initial confidence. These results suggest that forecast teams may have been overconfident about the EtaKF forecasts. For example, they may be predisposed to favor the KF scheme because it is based heavily on parcel theory, mirroring their conceptual view of convective development, whereas the BMJ algorithm is more strongly tied to bulk tropospheric properties and more difficult to relate to discrete convective initiation mechanisms.
- Forecast teams generally expressed higher levels of confidence in the most recent initialization of the models, *i.e.*, the 1200 UTC run. However, subjective verification statistics suggested that, on average, there was no significant difference between the two runs of the Eta in predicting convective initiation. Strictly speaking, the same could be said about the EtaKF, though paired t-test scores for the two runs of this model were much closer to meeting generally accepted criteria for statistical significance. These results suggest that forecasters should be very cautious in allowing updated model guidance to supercede previous guidance from the same model.

Additional conclusions could be inferred from the dataset, but our purpose is not to demonstrate the superiority of one model run over another or to draw definitive conclusions from this experiment, our first attempt at quantitative subjective assessment. Rather it is to establish the validity of the subjective verification approach and to demonstrate the potential utility of these methods. This approach 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.

The subjective verification approach used in this study 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, quantifying separately 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. 2002b). 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 information 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. 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. This work was partially funded by COMET Cooperative Project No. 099-15805.

REFERENCES

Baldwin, M. E., J. S. Kain, and M. P. Kay, 2002a: Properties of the convection scheme in NCEP's Eta model that affect forecast sounding interpretation. Accepted for publication in *Wea. Forecasting*.

Baldwin, M. E., S. Lakshmivarahan, and J. S. Kain, 2002b: Development of an "events-oriented" approach to forecast verification. *Preprints, Fifteenth Conference on Numerical Weather Prediction*, Amer. Meteor. Soc., San Antonio, TX, 12-16 August 2002.

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

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

Janjic, 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.

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

Kain, J. S., P. R. Janish, S. J. Weiss, M. E. Baldwin, R. Schneider, and H. E. Brooks, 2003: Collaboration between forecasters and research scientists at the NSSL and SPC. Submitted to *Bull. Amer. Meteor. Soc.*

Kain, J. S., M. E. Baldwin, and S. J. Weiss, 2002: Parameterized updraft mass flux as a predictor of convective intensity. Submitted to *Wea. Forecasting*.

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.

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