9.3 SUBJECTIVE VERIFICATION OF DETERMINISTIC MODELS DURING THE 2003 SPC/NSSL SPRING PROGRAM

John S. Kain[*,**], Steven J. Weiss[***], David R. Bright[***], Michael E. Baldwin[*,**,***], Marc Dahmer[****], and Jason Levit[***]

*Cooperative Institute for Mesoscale Meteorological Studies, Norman, OK
**NOAA/OAR/National Severe Storms Laboratory, Norman, OK
***NOAA/NWS/NCEP/Storm Prediction Center, Norman, OK
****University of Missouri, Columbia, MO

1. INTRODUCTION

The SPC/NSSL (Storm Prediction Center/National Severe Storms Laboratory) Spring Program is a collaborative exercise held in Norman, OK during the peak severe convective weather season. It brings together a variety of meteorologists from research and operational communities to investigate specific applied research problems and to promote interactions between the two communities (Kain et al. 2003a). The 2003 Spring Program was anchored by SPC forecasters and NSSL/CIMMS (Cooperative Institute for Mesoscale Meteorological Studies) researchers and rounded out with visiting scientists from numerous institutions, including the Environmental Modeling Center (NCEP/EMC), the Forecast Systems Laboratory, the Norman, OK and White Lake, MI NWS Forecast Offices, the University of Arizona, the University of Oklahoma, the University of Washington, Iowa State University, the Massachusetts Institute of Technology, the United Kingdom Meteorological Office, and the Meteorological Service of Canada. In addition, observers from COMET and USWRP participated.

Subjective verification of numerical weather prediction models has been an integral part of the Spring Program for several years (Kain et al. 2003b) and 2003 was no exception. This year subjective verification methods were used to evaluate two promising applications of numerical models in forecasting severe weather: 1) the use of Short-Range Ensemble Forecast (SREF) prediction systems, and 2) the use of high-resolution deterministic models. This paper will focus on the latter application as the former is discussed by Levit et al. (2004).

The objective of the deterministic model evaluation during the 2003 Spring Program was to compare the performance of three experimental model configurations to two "benchmark" models that are used regularly at the SPC. It is hoped that the results of this subjective verification exercise will provide insight into the utility of these models for severe weather forecasting and also provide guidance for numerical model developers.

2. METHODOLOGY

Methods used during the subjective verification of deterministic models in the 2003 Spring Program were very similar those documented by Kain et al. (2003b). In particular, model output fields were subjectively compared to observations of convective initiation and evolution using a rating scale from 1 to 10, with 1 being a very poor forecast and 10 being excellent. Ratings were assigned using a web-based form.

Corresponding author address: Jack Kain, NSSL, 1313 Halley Circle, Norman, OK 73069. e-mail: jack.kain@noaa.gov

During this year's program, the rating process was limited to precipitation forecasts. Although precipitation is often the last model output field to be examined by SPC forecasters, it is well suited to subjective verification because its general character is readily verifiable with radar data. In assigning forecast ratings, teams were instructed to focus on this character, including the timing, location, orientation/configuration, movement, etc. of precipitation fields, rather than on accumulated amounts. Radar data used in this process were composited over the same time period that the precipitation was accumulated in the model, with the radar composite displayed as the maximum reflectivity at each display pixel over the period in question. For most of the ratings, a 3 hour period was used, although a subset of the data (see below) was collected using 1 hour intervals.

The operational Eta model (e.g., Black 1994; Janjic 1994) and the EtaKF (Kain et al. 2003c) were considered benchmarks in the 2003 Spring Program. These models are used routinely at the SPC and have been part of subjective verification efforts in the recent past (Kain et al. 2003b). Output from these models was compared to forecasts from NCEP's NMM (Janjic et al. 2001), a non-hydrostatic derivative of the Eta model that is currently being run operationally with 8 km grid spacing in EMC's "HiRes Window" production slot (G. DiMego 2003, personal communication) and two configurations of the WRF (Weather Research and Forecasting) model (Michalakes et al. 2001). The first WRF configuration used 12 km grid spacing over a CONUS domain (hereafter WRF12), with convection parameterized by the Kain-Fritsch scheme (Kain 2004). The second WRF run used a re-locatable domain, 1200 km on a side with 3 km grid spacing and no parameterized convection (hereafter WRF03). This high resolution WRF domain was centered over the area of greatest concern for the SPC, determined each morning in consultation with the SPC Lead Forecaster on duty. The subjective assessment of all models was limited to this regional domain, using display software to zoom in on the common area for the models with larger areal coverage.

All models were based on the same initial conditions, namely those of the 1200 UTC operational Eta model, but the effective resolution of the initial data varied significantly from model to model. The NMM grid was populated by interpolating directly from the Eta's native 12 km grid, but the other three runs were initialized from standard NCEP output grids. Specifically, the EtaKF (with 22 km spacing) was initialized from the 212 grid (40 km spacing), while both configurations of the WRF were populated by interpolating from the 211 grid (80 km spacing). Obviously, these procedures for generating initial conditions were not optimal, especially for the WRF runs in which initial data were defined on much coarser scales than the model configurations were capable of resolving. This initial condition problem likely handicaps model performance, but the magnitude of this disadvantage has not been quantified.

3. RESULTS

Subjective verification exercises were planned for a total of thirty days during the 2003 Spring Program (six weeks at five days per week). Of these thirty days, forecasts from all five models were available on twenty-one days, or 70% of the time. The initial focus of the statistical analysis is on these twenty-one days. Two forecast periods were evaluated each day; the 1800-2100 UTC period and the 2100-0000 UTC period. Thus, a total of forty-two forecast periods were surveyed for the complete set of five different model forecasts.

Results from this survey 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.

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 two received a 3s, the relative rankings would be 5, 3.5, 3.5, 1.5, and 1.5, respectively.

For each method, paired t-test scores (e.g.,

	18 - 21 UTC					21 - 00 UTC				
DATE	**ETA**	**ETAKF**	**NMM**	**WRF12**	**WRF03**	**ETA**	**ETAKF**	**NMM**	**WRF12**	**WRF03**
29-Apr	7	7	7	6	4	6	3	7	2	4
30-Apr	7	6	7	7	6	6	5	5	3	7
1-May	4	2	4	4	1	1	1	1	2	0
4-May	4	6	3	5	5	5	5	2	7	7
5-May	5	5	5	5	4	4	6	4	3	6
6-May	5	6	6	6	5	6	3	5	3	4
7-May	2	2	2	2	1	5	5	3	4	4
11-May	2	4	2	4	3	4	6	3	7	5
12-May	4	4	5	4	3	5	4	6	3	3
13-May	4	5	4	5	2	3	5	4	6	2
14-May	7	8	6	7	2	7	8	9	6	2
15-May	8	8	8	6	7	8	7	8	6	6
20-May	5	6	5	7	3	2	4	2	6	2
21-May	5	6	4	4	1	7	8	6	5	1
22-May	3	6	2	7	7	1	4	1	6	7
28-May	5	7	6	6	9	6	7	5	6	8
1-Jun	4	6	5	4	1	4	7	6	4	1
2-Jun	3	7	4	8	2	6	3	7	4	2
3-Jun	4	5	3	6	2	3	7	5	3	1
4-Jun	7	3	5	6	2	4	6	3	5	1
5-Jun	5	8	7	3	4	6	5	6	3	3
Number	21	21	21	21	21	21	21	21	21	21
Average	4.7619	5.57143	4.7619	5.33333	3.52381	4.71429	5.19048	4.66667	4.47619	3.61905

Table 1. Subjective verification ratings for the 21 days on which all 5 models were available.

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).

3.1 Three-hourly forecasts: All five models.

Ratings from the days when all five models were available can be seen in Table 1. All models earned a wide range of ratings and no model was consistently rated worse or better than the others, but certain trends are clearly discernible. When averaged over all rating periods, the EtaKF stands out with higher scores while the WRF03 appears to be an outlier on the lower end of the scale (Fig. 1a). Results from paired t-tests (Table 2) confirm the significance of these differences. Pairings between various combinations of the WRF12, Eta, and NMM produce t-test values greater than 0.6, strongly suggesting that none of these models was significantly better or worse than the others, on average. In contrast, when EtaKF or WRF03 output is paired with any of the other forecasts, t-test scores are quite low, indicating a high likelihood that the differences are real. When the comparison is done in terms of rank instead of rating, the results are qualitatively unchanged. In particular, the order from highest to lowest is the same (Fig. 1b) and t-test scores indicate comparable levels of significance.

Additional insight into these results comes from examining the correlation between different elements. Although all models used the 1200 UTC Eta for initial conditions, they clearly followed different paths in predicting the evolution of precipitation fields over the ensuing 12 h. Inter-model correlation was strongest between the Eta and NMM, with a correlation coefficient of 0.822 (Table 3). Forecasts from these two models received the same rating on many days and differed by as

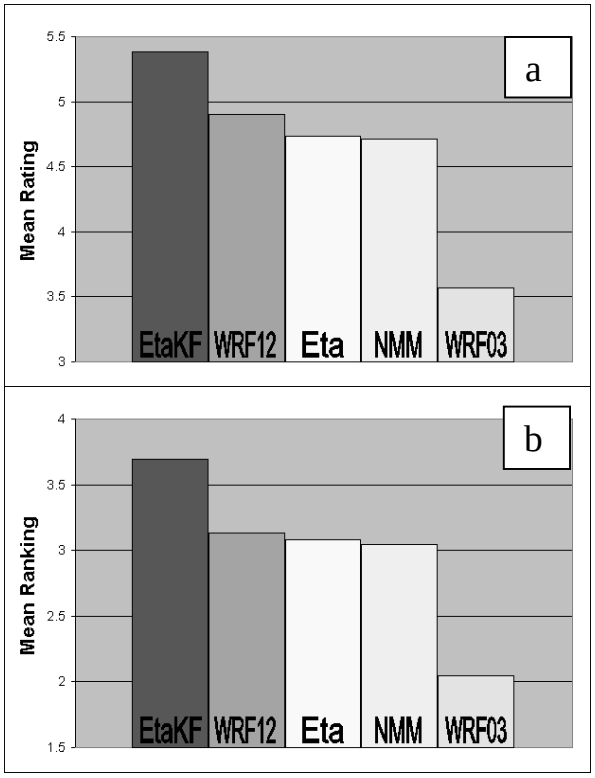

Fig. 1. Mean a) ratings and b) rankings for all times shown in Table 1.

Model Pairings	Paired t-tests	
	Raw Ratings	Rankings
Eta - EtaKF	0.034	0.033
Eta – NMM	0.893	0.854
Eta – WRF12	0.622	0.883
Eta – WRF03	0.005	0.002
EtaKF - NMM	0.030	0.035
EtaKF – WRF12	0.086	0.042
EtaKF – WRF03	<0.001	<0.001
NMM – WRF12	0.628	0.816
NMM – WRF03	0.015	0.010
WRF12 – WRF03	0.001	0.002

Table 2. Paired t-test values for the data shown in Table 1

Correlation between Forecast Ratings	
Model Runs Compared	**Correl. Coeff.**
Eta - EtaKF	0.437
Eta – NMM	0.822
Eta – WRF12	0.186
Eta – WRF03	0.255
EtaKF – NMM	0.485
EtaKF – WRF12	0.474
EtaKF – WRF03	0.262
NMM – WRF12	0.022
NMM – WRF03	0.094
WRF12 – WRF03	0.319
Eta: 1st vs. 2nd fcst pd	0.490
EtaKF: 1st vs. 2nd fcst pd	0.303
NMM: 1st vs. 2nd fcst pd	0.632
WRF12: 1st vs. 2nd fcst pd	0.174
WRF03: 1st vs. 2nd fcst pd	0.868

Table 3. Correlation for both inter-model and intra-model comparisons

many as 3 rating points on only one day (Table 1 and Fig. 2a). In comparison, Eta and EtaKF forecasts were given very different ratings on numerous days and showed only a moderate correlation (Tables 1, 3; Fig. 2b). Other correlations were generally weak to moderate.

These differing correlations appear to be related to variations in model physics and, perhaps to a lesser extent, model dynamics. For example, the NMM and Eta models use essentially the same physics package (Janjic et al. 2001), consistent with their strong correlation in ratings. The EtaKF, configured with a different convective parameterization but otherwise the same physical parameterizations as Eta and NMM, shows a moderate correlation to both of these models. The WRF12, using a physics package that is completely different from Eta and NMM, is only weakly correlated with these models, but it is moderately correlated with EtaKF, with which it has the KF convective parameterization in common. The WRF03, with no convective parameterization, is weakly correlated with all other models. It shows a somewhat stronger correlation with WRF12, consistent with the similarity in physics and dynamics in these two configurations of WRF.

Intra-model changes between consecutive forecast periods are also revealing. The WRF12 showed the least consistency between first and second periods, with a correlation coefficient of only 0.174 (Table 3). Moreover the average rating of the WRF12 forecasts dropped by nearly a full point from one period to the next. This appears to reflect a problem that was frequently noted with the WRF12, namely that the areal coverage of WRF12 precipitation shrunk dramatically and unrealistically as solar heating in the model waned late in the day. Work is underway to understand the reasons for this excessive response to the diurnal cycle. The EtaKF ratings followed similar, but less dramatic trends from first to second period, dropping by

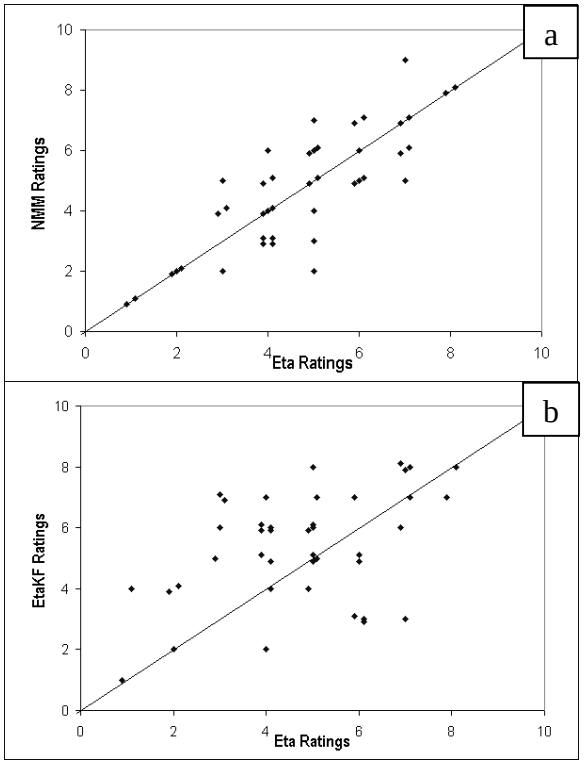

Fig. 2. Scatterplots of a) NMM vs. Eta and b) EtaKF vs. Eta for all times shown in Table 1

about 0.4 points and showing only weak to moderate correlation. Since the EtaKF and WRF12 both use the KF convective parameterization, these inconsistencies could be related to this scheme. Ratings for the Eta and NMM also dropped for the later forecast period, but this decline was minimal. These two models appeared to have more temporal consistency than the EtaKF and WRF12, but still only a moderate correlation between first and second periods.

Despite having the lowest ratings overall, the WRF03 demonstrated the greatest consistency between forecast periods. In addition, it was the only model to have a positive trend with time. These results suggest that the processes discussed by Warner and Hsu (2000) may be operative in the other forecasts, i.e., in the model runs using parameterized convection. In particular, Warner and Hsu (2000) showed that feedbacks from parameterized convection can disrupt subsequent convective initiation over regional scales within a few hours of the initial activation. Although they examined this effect in the context of parameterized convection on large grids affecting explicitly resolved convection within embedded high-resolution grids, the same principles are likely to hold on a single grid with parameterized convection. Parameterized convective feedbacks introduce local imbalances on the model grid, spawning gravity waves that propagate into the surrounding environment quite rapidly. The deepest gravity wave mode induces subsidence over a deep layer and propagates at ~30 m s^{-1} (e.g., Mapes 1993). As shown by Warner and Hsu, these waves can strongly affect the surrounding environmental stability and humidity, two important factors in both the BMJ and KF trigger functions (Baldwin et al. 2002; Kain et al. 2003c), as well as the vertical velocity, which is a key element in the KF trigger.

The impact of this process is difficult to quantify, or even confirm. Nonetheless, we speculate that it is operative in daily forecasts with parameterized convection. Ironically, the KF scheme seems to be most susceptible to this problem (Warner and Hsu 2000), yet it is associated with the best forecasts of convective initiation and evolution in this and other studies (Kain et al. 2003b). Moreover, while model forecasts without parameterized convection may be immune to this problem, they performed significantly worse, on average, than the other forecasts in this study.

On a final note for this dataset, it is worth considering exceptional, rather than average performances by the models. For many forecast periods, two or more models shared the distinction of earning the highest or lowest rating (Table 4). That is, there was a tie for first or last place in the ratings. When ties are included in computing frequency, the numbers of highest ratings follow essentially the same pattern as the average ratings. The WRF03 is clearly an outlier on the lower end of the highest ratings and on the higher end of the lowest ratings. However, the more interesting results emerge when ties are not included, highlighting the events when one model was distinctly better or worse than all the others. With this criterion, the EtaKF stands out with nine highest ratings and zero lowest ratings. The NMM and WRF12 both have fairly large numbers of highest ratings, but they also earned the lowest ratings on a comparable

Frequency of High and Low Ratings					
	Eta	**EtaKF**	**NMM**	**WRF12**	**WRF03**
Highest Rating (including ties)	11	18	14	16	7
Lowest Rating (including ties)	5	2	9	9	27
Highest Rating (NOT including ties)	2	9	5	7	4
Lowest Rating (NOT including ties)	1	0	6	5	21

Table 4. Frequency of high and low ratings for the data shown in Table 1

number of forecasts. The operational Eta received very few highest or lowest ratings. It tends to produce "conservative" precipitation forecasts with relatively smooth, low amplitude features, but it rarely fails to provide a signal for convective precipitation associated with meso and larger-scale disturbances. Finally, it is noteworthy that the WRF03 earned the lowest ratings more than twice as often as all other models combined. Yet it also earned the highest rating for four different forecasts. These numbers highlight the importance of considering many different meteorological scenarios before passing judgement on the comparative performance of different models in forecasting convection.

3.2 One-hourly forecasts: Comparison of WRF12 and WRF03

As can be deduced from the 3-hourly results discussed above, WRF03 forecasts were quite unimpressive for the most part. Comparison of 1-hourly precipitation fields from the WRF12 and WRF03 yielded more of the same - a significant disadvantage for the WRF03 in terms of timing, location, and evolution of convective activity. However, the 3 km configuration often provided useful information about convective mode (e.g., isolated convective cells vs. a squall line), whereas such indications were generally lacking in all of the coarser resolution models (see also Done et al. 2004). Furthermore, in spite of the generally poor performance of our "no-CP" (no convective parameterization) version of WRF, we are encouraged about the future of higher resolution WRF forecasts by the results obtained using a 4 km version of the model in support of BAMEX (Bow echo and MCV experiment, also carried out during the spring and early summer of 2003). The BAMEX 4 km WRF often performed well in terms of timing, evolution, and mode of mesoscale convective systems (see Weisman et al. 2004 and a thorough on-line dataset at http://www.joss.ucar.edu/bamex/catalog/). The BAMEX configuration included a much larger domain, somewhat different physical parameterizations, and the Eta 212 grid (40 km grid spacing) for initial conditions. These differences apparently enhanced the WRF forecasts for BAMEX significantly (compared to our forecasts) and they paint a more optimistic picture of the future of convection–resolving numerical forecasts.

4. SUMMARY

Forecast teams subjectively compared predictions of convective initiation and evolution from five different deterministic forecast models during the 2003 SPC/NSSL Spring Program. Results substantiated the validity and utility of the systematic subjective verification process and they provided valuable information about the comparative performance of the models.

Relative ratings of two benchmark models, the operational Eta and the EtaKF, were consistent with subjective verification results from a similar experiment in 2001 (Kain et al. 2003b). In particular, the EtaKF was rated considerably higher than the Eta on average and the difference was statistically significant. As in 2001, however, there were many individual forecast periods for which the Eta received higher ratings. This result suggests that forecasters are wise to consider solutions from both models in preparing their forecasts.

In contrast, forecasts from the Eta and NMM displayed similar characteristics and their individual subjective verification ratings were strongly correlated. This result should not be surprising, considering that the Eta and NMM use very similar physical parameterizations. It bodes well for a smooth transition from the Eta to the NMM as the primary 1-3 day deterministic forecast model, yet it suggests that the higher resolution of the NMM will not necessarily translate into more

detailed or "better" forecasts of convective systems.

Mesoscale forecasts with the WRF model, using the Kain-Fritsch convective parameterization (Kain 2004) and 12 km grid spacing, were quite good on some days, especially considering that this model was initialized from a smoothed version of Eta initial conditions on a grid with 80 km spacing. On average, WRF12 forecasts earned ratings that were lower than EtaKF scores, but higher than both the Eta and NMM ratings (though not in a statistically significant sense). The WRF12 often provided much better forecasts during the first half of the rating period (1800-2100 UTC), than during the 2100-0000 UTC period. The cause of this inconsistency is under investigation.

A high-resolution configuration of WRF (3 km grid spacing) received the highest rating about 10% of the time, but it was rated considerably worse than all other models on many days. This version of the model was handicapped by very coarse initial conditions, lateral boundary effects exacerbated by a small domain (1200 km on a side), and possibly other factors. Forecast teams frequently noted that explicitly predicted convection in the WRF03 was just beginning to "spin up" by the end of the forecast period. Spin up and lateral boundary problems appear to have been much less severe in realtime WRF forecasts from NCAR, generated during an overlapping time period in support of BAMEX (Chris Davis 2003, personal Communication; Weisman et al. 2004). Although the BAMEX forecasts used slightly coarser horizontal resolution (4 km grid spacing) they benefited from a considerably larger domain and higher resolution initial conditions. Obviously, much more work is needed to optimize model configurations for realtime forecasts without parameterized convection.

Finally, it is worth emphasizing that the internal consistency of these subjective verification results validates the methods used to obtain them. In particular, the fact that differences in mean ratings are completely consistent with differences in model physics provides strong evidence that the subjective verification methods used in the Spring Program are efficacious. As emphasized by Murphy (1993), the "goodness" of a forecast is a multifaceted concept that can be nebulous and very difficult to express. Standardized objective verification metrics may sample some aspects of goodness, but they are likely to fall short of the holistic assessment of human analysts and forecasters. Systematic subjective verification provides unique information that is inherently lacking in traditional verification metrics.

Acknowledgements

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 2003 Spring Program. In particular, we appreciate Brett Morrow (CIMMS/NSSL) and Steve Fletcher (CIMMS/NSSL) for hardware support that enabled local runs of the WRF model; Phillip Bothwell (SPC), Gregg Grosshans (SPC), Jay Liang (SPC), and Doug Rhue (SPC) for assistance in configuring hardware and software. We further wish to recognize the full support of SPC and NSSL management and enthusiasm by participants from NCEP/EMC, FSL, the Norman, OK and White Lake, MI NWS Forecast Offices, the University of Arizona, the University of Oklahoma, the University of Washington, Iowa State University, the Massachusetts Institute of Technology, the United Kingdom Meteorological Office, the Meteorological Service of Canada, COMET, and USWRP who helped make this undertaking a positive experience for everyone. We are also grateful to Chris Davis, Bill Skamarock, and Morris Weisman, all of NCAR, who provided input to

help us configure our hi-resolution WRF runs, as well as output and interpretation from their BAMEX forecasts. This work was partially funded by NOAA-OU Cooperative Agreement #NA17RJ1227 and COMET Cooperative Project S01-32796. COMET Partners Project S03-38671 provided travel support for the Spring Program.

5. References

Baldwin, M. E., 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.

Done, J., C. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the weather research and forecast (WRF) model. Submitted to *Atmospheric Science Letters.*

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.

Janjic, Z. I., J. P. Gerrity, and S. Nickovic, 2001: An alternative approach to nonhydrostatic modeling. *Mon. Wea. Rev.,*: **129**, 1164–1178.

Kain, J. S., 2004: The Kain-Fritsch convective parameterization: An update. *J. Appl. Meteor.*, in press.

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

Kain, J. S., M. E. Baldwin, S. J. Weiss, P. R. Janish, M. P. Kay, and G. Carbin, 2003b: Subjective verification of numerical models as a component of a broader interaction between research and operations. *Wea. Forecasting*, **18**, 847-860.

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

Levit, J. J., D. J. Stensrud, D. R. Bright, and S. J. Weiss, 2004: Evaluation of Short-Range Ensemble Forecasts during the SPC/NSSL 2003 Spring Program. This volume.

Mapes, B.E., 1993: Gregarious tropical convection. *J. Atmos. Sci.*, **50**, 2026-2037.

Michalakes, J., S. Chen, J. Dudhia, L. Hart, J. Klemp, J. Middlecoff, and W. Skamarock, 2001: Development of a next-generation regional weather research and forecast model. *Developments in Teracomputing: Proceedings of the Ninth ECMWF Workshop on the Use of High Performance Computing in Meteorology*, W. Zwieflhofer and N. Kreitz, Eds., World Scientific, 269–276.

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

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

Warner, T. T., H.-M. Hsu, 2000: Nested-model simulation of moist convection: The impact of coarse-grid parameterized convection on fine-grid resolved convection. *Mon. Wea. Rev.*, **128**, 2211–2231.

Weisman, M. L., C. Davis, J. Done, W. Wang, and J. Bresch, 2004: Real-time explicit convective forecasts using the WRF model during the BAMEX field program. *This volume*.

Wilks, D. S., 1995: *Statistical methods in the atmospheric sciences: An introduction.* Academic Press, 467 pp.