17.1 EXAMINATION OF SEVERAL DIFFERENT VERSIONS OF THE WRF MODEL FOR THE PREDICTION OF SEVERE CONVECTIVE WEATHER: THE SPC/NSSL SPRING PROGRAM 2004

Steven J. Weiss*,[1], J. S. Kain[2], J. J. Levit[1], M. E. Baldwin[2], and D. R. Bright[1]

[1]NOAA/NWS/Storm Prediction Center
[2]University of Oklahoma/CIMMS/National Severe Storms Laboratory

1. INTRODUCTION

Co-location of the Storm Prediction Center (SPC) with the National Severe Storms Laboratory (NSSL), the Cooperative Institute for Mesoscale Meteorological Studies (CIMMS), and other agencies in the Norman, OK Weather Center has facilitated considerable interaction and collaboration on a variety of operationally relevant research problems (e.g., Baldwin et al. 2002; Craven et al. 2002; Robbins and Cortinas 2002; Stensrud and Weiss 2002; Elmore et al. 2003; Kain et al. 2003a; Thompson et al. 2003). Over the past five years, the most visible component of this collaboration has been an annual event held during the peak severe convective weather season. This event, known as the Spring Program (Kain et al. 2003b), has attracted a wide cross section of local and visiting forecasters and researchers. The specific emphasis of this program varies each year, but the underlying structure allows forecasters to evaluate new tools or concepts that emanate from the research community, while immersing research scientists in the challenges, needs, and constraints of the operational forecasting environment. This approach promotes forecast improvements by accelerating the transfer of science and technology into forecast operations at the SPC and by providing researchers with the knowledge to formulate research strategies that will directly benefit operational forecasting.

The SPC is responsible for the prediction of severe convective weather over the contiguous United States on time scales ranging from several hours to three days. To meet these responsibilities, the SPC issues Convective Outlooks for the Day 1, Day 2 and Day 3 periods to highlight regions with enhanced potential for severe thunderstorms (defined as thunderstorms producing $\mathrm{hail} \geq 0.75$ inch in diameter, wind $\mathrm{gusts} \geq 50$ kt or thunderstorm induced wind damage, or tornadoes). These outlooks are issued in both categorical (slight, moderate, or high risk) and probabilistic formats, and at set times each day, regardless of the meteorological situation. When the threat level becomes elevated over a mesoscale area in time and space, the SPC may issue a Mesoscale Discussion (MD) product to notify local National Weather Service (NWS) offices that SPC forecasters are closely monitoring the area and to articulate the reasons for their concern. MDs are often issued as a precursor to convective Watches, the cornerstone of SPC forecast products. Watches are issued when conditions become favorable for the development of severe thunderstorms or tornadoes in a specific location over the next several hours. They are designed to alert a wide variety of users, including NWS and private meteorologists, the public, emergency managers, broadcast media, and aviation interests, of the threatening environmental conditions.

In current practice at the SPC, the issuance of both MDs and convective Watches is driven primarily by observational data, which are monitored diligently by forecasters. The transition from MD to Watch often is triggered by a specific observation, such as the first sign of deep convective clouds in satellite or radar data. By waiting for such an observation, SPC forecasters undoubtedly increase the accuracy in their placement of convective Watches. The drawback to this approach is that it limits the lead time that forecasters can provide between watch issuance and the development of severe thunderstorms. Over the last several years, the SPC has been making a concerted effort to find ways to increase Watch lead time

---

*Corresponding author address: Steven J. Weiss, Storm Prediction Center, 1313 Halley Circle, Norman, OK 73069 without sacrificing placement accuracy or increasing size.

Another pressing challenge for the SPC is to improve predictions of convective mode, or morphology. In recent years it has become evident that the type of severe weather that occurs (tornadoes, hail, or damaging winds) is often closely related to the convective mode that storms exhibit. For example, SPC forecasters expect distinctly different severe weather threats from storms that form and remain as discrete cells, as opposed to those that organize into quasi-linear structures (such as squall lines) or multi-cellular clusters. In addition, some severe storms develop as dynamically unique classes of thunderstorms such as supercells and bow echoes, which are believed to produce a disproportionate number of tornado and widespread straight-line wind damage events, respectively. Currently, the prediction of convective mode is based on assessments of key physical properties (e.g., estimates of CAPE, convective inhibition, and magnitudes of vertical motion) that are difficult to gauge accurately, as well as concepts derived from cloud scale model results and observational studies (e.g., orientation of surface boundaries relative to mean wind and vertical shear vectors – see Weisman et al. 1988, Bluestein and Weisman 2000, and Dial and Racy 2004). If forecasters could find ways to anticipate convective mode more accurately, the specificity and value of all three convective guidance products - Watches, MDs, and Outlooks – would likely increase as well.

One way to address these forecasting challenges is to explore new methods in Numerical Weather Prediction (NWP). In current practice at the SPC, both deterministic and ensemble configurations of numerical models are used routinely, but their primary function is to define the larger-scale backdrop for convective activity, rather than provide specific details about convective storms in time and/or space. However, both computer power and modeling expertise continue to grow, and in recent years it has become feasible to run deterministic models over large domains (i.e., CONUS size) and grid spacing of about 4 km in a semi-operational environment. This grid spacing is best described as a near-convection-resolving compromise – the equivalent resolution appears to be fine enough to obviate parameterization of deep convection, but it is close to the upper limit of grid spacing needed to resolve essential aspects of organized convective systems (Weisman et al. 1997). Some studies have suggested that models with this resolution can be used to predict skillfully the initiation and mode[1] of convective systems as much as 36-48 h in advance (Fowle and Roebber 2003; Done et al. 2004). Since this may very well be the class of model used in the next generation of NWP, it behooves us to investigate the value of this type of modeling system for severe convective weather forecasting at the SPC.

Such an investigation was the theme of the 2004 SPC/NSSL Spring Program. In this year's program, three different configurations of the Weather Research and Forecast (WRF) model, each configured with ~ 4 km grid spacing and no convective parameterization, were used to predict convective activity over near-CONUS domains each day. Output from these runs was used to generate a daily experimental forecast of severe weather, but only after a baseline had been established by preparing a control forecast using routine observations and model guidance. Specifically, two probabilistic forecasts of severe weather were prepared, over a regional spatial domain and a Watch-like time frame. The first was a control forecast, designed to emulate current operational practice, with data access restricted to operational data streams. The second was the experimental forecast, prepared with access to high-resolution output, after the first forecast was submitted. Differences between these two forecasts were measured to gauge the impact of the high-resolution output. In addition, numerous aspects of the individual high-resolution forecasts were systematically evaluated and compared to the same characteristics of current operational models.

---

[1] It should be emphasized that the term "convective mode" is a bit of a misnomer in the context of this study. By focusing on the three categories of 1) isolated cells, 2) multi-cellular structures, and 3) quasi-linear structures, convective mode is synonymous with the meso-γ-scale organizational characteristics of convection in this study.

A key component of the program was the participation of operational SPC forecasters, whose insights and experience impart a real-world severe weather forecasting perspective when assessing the usefulness of high-resolution WRF models. The primary goal of this study is to use data from the Spring Program to assess whether SPC forecasters can make better predictions of severe convective weather when their current data stream of observational and model data is supplemented with output from near-cloud-resolving forecast models. Further, we aim to identify specific characteristics of the high resolution output that provide added value, as well as those that might have a detrimental or misleading impact. Finally, we compare the performance of the different WRF configurations in order to provide feedback to model developers. The specific methods used in the 2004 Spring Program are outlined in the next section, followed by a summary of results and a discussion of their implications.

2. METHODS USED IN THE 2004 SPRING PROGRAM

As has been the case in several previous Spring Programs, the 2004 effort had two primary components: 1) experimental human forecasts for severe convective weather and 2) an evaluation of experimental numerical forecast models. Each of these is described below, following a description of the evaluation methods.

2.1 *Subjective Evaluation*

A compelling objective of the Spring Program is to facilitate engaging discussion and lively interaction between forecasters and researchers. One of the ways that we promote this activity is through a subjective evaluation process in which all participants become members of a panel of experts. A new panel is constructed each week in the form of a forecast team, consisting of a minimum of one SPC forecaster, one NSSL or CIMMS modeling expert, and one other forecaster or research scientist. On most days in 2004, there were five or six panel members with a wide variety of backgrounds. Subjective ratings of both human forecasts and model predictions were obtained by means of consensus among all panel members. Achieving consensus was not always easy, but the deliberation process was very effective in soliciting input from all team members. Consensus ratings were assigned on a scale from 0 to 10, with 10 being a superior rating and 0 corresponding to the lowest possible assessment. All ratings were entered on a web-based form, similar to that described in Kain et al. (2003a).

Diversity of viewpoint is essential for a credible subjective evaluation process. For 2004 such diversity was characteristic of the Spring Program. In all, there were about 50 participants over a seven-week period (April 19 – June 4), including contributors from numerous NOAA research and forecasting organizations, ten major universities, the Air Force Weather Agency, NCAR, and international visitors from Canada and Finland (Appendix A). The variety of backgrounds and perspectives in this group was viewed as a key to minimizing the impact of any personal predispositions that could bias subjective assessments.

2.2 *Human Forecasts*

Morning activities during the 2004 program revolved around preparation of the control and experimental forecasts. The specific forecast product was designed to be a hybrid between the current operational SPC Watch and Outlook products, in that it was issued at scheduled times for fixed time periods (like an Outlook), but was valid for shorter time frames (like a Watch). It consisted of a probabilistic forecast of severe convective weather, including graphic and text components. The graphic depicted probability contours of all severe storms; that is, it did not distinguish between severe weather types such as large hail, damaging winds, and tornadoes, although the attendant discussion alluded to various distinctions, including a prediction of the likelihood of three possible convective modes: discrete cells, quasi-linear systems, and multi-cellular clusters (see the Spring Program Operations Plan at http://www.spc.noaa.gov/exper/Spring_2004/sp04opsplan.pdf for additional details). In addition, the graphic delineated areas where a 10% or greater probability existed for significant severe events (F2 or greater tornado, hail 2 inches or larger, or wind gusts 65 kt or greater). The forecast covered a regional domain (approximately 14^o longitude by 8^o latitude) and a six-hour time frame, centered in both time and space on the greatest threat for severe weather. Threat severity was determined by examining the operational 1300 UTC SPC Day 1 Outlook, observational data, deterministic model forecasts from the Eta (Black 1994) and RUC (Benjamin et al. 2004) models, short-range ensemble forecasts from NCEP (National Centers for Environmental Prediction) (Du et al. 2004), and by consultation with operational SPC forecasters. On most days the forecasts were valid for the 1800-0000 UTC period, in order to capture the time of anticipated afternoon convective initiation. However, on some days when convective development was not expected to occur before late afternoon or evening, the six-hour forecast time period was shifted to 2100-0300 or 0000-0600 UTC. Finally, we were most interested in focusing on initiation of new storms, rather than continuation of existing storms, so this rationale was factored into the domain selection process.

In addition to the graphic and text product, forecast teams were solicited for probabilistic forecasts of individual convective modes over the forecast domain during the valid period, and for a prediction of a 2-hour time window of the first severe storm report. The former documented the likelihood of occurrence of the three basic convective modes, while the latter addressed issues of severe storm timing. Both are important considerations that affect the accuracy of SPC severe weather products.

As part of the forecast experimental design, it is important to emphasize that high-resolution model output was deliberately and uncompromisingly excluded from the control forecast, in order to more directly determine the impact of the WRF models on the severe weather forecasting process. Once this first product was issued, high-resolution model output was then introduced (and real-time observational data updates were disabled to exclude other influences on the subsequent final forecast decisions). The experimental forecast was prepared using the full suite of operational and experimental model data, coupled with the identical observational data used in preparation of the control forecast. Through these procedures, the impact of the high-resolution data could be measured by comparing the two forecasts, both subjectively and objectively.

2.3 *Experimental Forecast Models*

The suite of high-resolution models used for the 2004 program was comprised of different configurations of the Weather Research and Forecasting (WRF) model, including two configurations based on the so-called Eulerian mass core (Michalakes et al. 2001; hereafter WRF-EM) and one based on the NCEP Nonhydrostatic Mesoscale Model (NMM) core (Janjic 2003; Janjic et al. 2004; hereafter WRF-NMM). Through partnerships with the NCEP's Environmental Modeling Center (EMC), the National Center for Atmospheric Research (NCAR), and the University of Oklahoma's Center for Analysis and Prediction of Storms (CAPS), these different versions of the WRF model were run and the output was post-processed at remote locations. A subset of the full hourly output was transferred to the SPC and provided to forecast teams. All versions of the model used approximately 4 km grid spacing and domains that covered ~ 2/3 or more of the CONUS (Fig. 1). Details of the configurations can be found in Table 1, but several similarities and differences are noted for emphasis here: 1) The NCAR and CAPS runs (hereafter WRF-EM-NCAR and WRF-

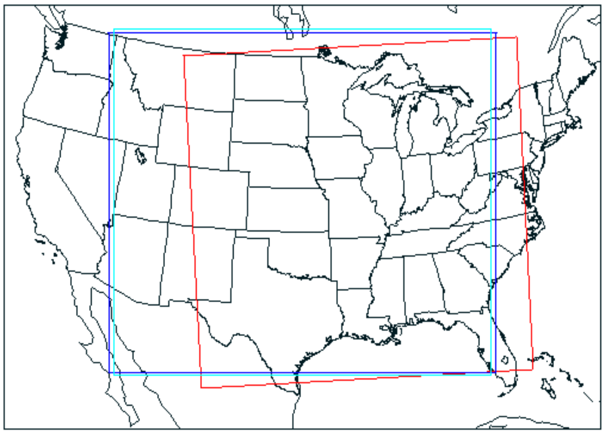

Fig. 1. Model domains for the WRF-NMM-EMC (red), WRF-EM-NCAR (blue), and WRF-EM-CAPS (cyan) forecasts.

EM-CAPS, respectively) both used the Eulerian mass core while the EMC forecast (hereafter WRF-NMM-EMC) used the NMM dynamic core; 2) all runs were initialized using 0000 UTC initial conditions from the operational Eta model (40 km 212 grid), but this relatively coarse resolution initial condition was enhanced in the WRF-EM-CAPS forecast using an experimental data assimilation procedure (ADAS) that also introduced hydrometeor fields derived from radar data (Brewster 1996); 3) the WRF-EM-NCAR and WRF-EM-CAPS runs used nearly identical domains (Fig. 1), physics options, and parameter settings, differing most significantly in their numbers of vertical levels and initialization procedures.

It is also important to emphasize that the high-resolution WRF model forecasts were all experimental. Each of these model runs included elements that were relatively untested and still under development. For example, the data assimilation system used by CAPS to enhance initial conditions was originally developed for convection resolving grids (grid spacing 1-2 km) and 0-12 h forecasts using the Advanced Regional Prediction System (ARPS) model. The code was adapted to the WRF model and a 4 km grid length in the weeks before the start of the program, with the intent of testing the utility of this adaptation as part of the Spring Program. Model forecasts early in the program suggested that the assimilation cycle was having a negative impact on the skill of 12-36 h forecasts. Nonetheless, CAPS scientists agreed to retain the original procedure through the end of the experiment in order to maintain the integrity of the dataset. Likewise, the WRF-EM-NCAR forecast was generated using a beta test version of WRF that had not been released for general users at the start of the program. Scientists at NCAR detected systematic anomalies in output from this run and they reverted back to a previous release of the model for their ongoing real-time forecasts after the end of the Spring Program. They later discovered that the anomalies were caused by lack of horizontal numerical diffusion, which had been disabled inadvertently. Finally, the WRF-NMM-EMC run was also quickly configured over a matter of a few weeks prior to that start of the program. However, it is likely that this run benefited from a relatively robust physics package, derived directly from the well-calibrated, inter-dependent group of physical parameterizations used in the operational Eta model.

As indicated above, only a subset of the full high-resolution output from these runs was ingested into the data stream for the program, focusing on selected output fields that are commonly examined by severe convection forecasters. Specific output fields included instantaneous, one-hourly, and three-hourly rainfall rates, low-level wind and moisture fields (including their derivatives such as mass and moisture convergence), CAPE (Convective Available Potential Energy), CIN (Con-

	WRF-NMM-EMC	**WRF-EM-NCAR**	**WRF-EM-CAPS**
Horiz. Grid Spacing (km)	4.5	4.0	4.0
Vertical Levels	35	35	51
PBL/Turb. Param.	MYJ	YSU	YSU
Microphysical Param.	Ferrier	Lin et al.	Lin et al.
Radiation Param. (SW/LW)	GFDL/GFDL	Dudhia/RRTM	Dudhia/RRTM
Initial Conditions	40 km Eta	40 km Eta	40 km Eta +ADAS +Level II Radar

Table 1. Model configurations used for the high resolution forecasts. MYJ: Mellor-Yamada-Janjic (Janjic 2001); YSU: Yonsei University (Noh et al. 2001); Ferrier: Ferrier et al. (2002); Lin et al. (1983); GFDL: Geophysical Fluid Dynamics Laboratory (Tuleya 1994); Dudhia: Dudhia (1989); RRTM: Rapid Radiative Transfer Model (Mlawer et al. 1997; Iacono et al. 2000); ADAS: ARPS (Advanced Regional Prediction System) Data Assimilation System (Brewster 1996).

vective INhibition), and several vertical shear and storm-relative helicity parameters. Model evaluation ratings were based primarily on the hourly rainfall field and its comparison with an hourly radar mosaic base-reflectivity field (displayed as the maximum reflectivity at each pixel over the previous hour). Ideally, a more direct comparison with radar observations could be made by computing an equivalent reflectivity field from instantaneous model hydrometeor fields, but the necessary model output and post-processing algorithms for equivalent reflectivity were not available for all models at the start of the program. One-hour precipitation fields proved to be quite adequate for assessing the general characteristics of model-predicted convective initiation, evolution, and mode.

Because of the substantial time required to integrate the WRF models and generate the high-resolution output, we were forced to initialize the experimental models with 0000 UTC data. However, it should be noted that 1200 UTC initializations of the operational models (Eta and RUC) were used both for forecast guidance and for subjective comparison. Later runs of the Eta and RUC models were used for two reasons: 1) the desire to emulate operational routines in preparing the control forecast - SPC forecasters typically focus on the most recently updated model guidance in preparing forecasts operationally, and 2) operational RUC forecasts are only 12 h in length and guidance was needed for the afternoon to evening time period. Although the latest updates of model forecasts are not necessarily the most skillful (e.g., see Kain et al. 2003a), it is recognized that this approach may handicap the high-resolution models in a direct comparison with the 1200 UTC RUC and Eta runs. Note that, as with the higher resolution models, output from the Eta and RUC models was displayed on native model grids (spacing of 12 and 20 km, respectively).

3. RESULTS

Results from the 2004 program are first illustrated by using one day's forecast as an example, followed by an overall assessment of human forecasts, and an overview of all model forecasts.

3.1 *Example of a Severe Weather Forecast*

High-resolution model guidance had a significant positive impact on human forecasters on 28 May. As forecast teams assessed the meteorological scenario on this day, they noted that an upper level ridge was in place over the central and northern Plains while an embedded short-wave trough was passing over northeastern Wyoming (not shown). This trough was expected to move over the Dakotas by late in the day and, in conjunction with a lee trough and associated warm front at the surface, to trigger convection in this region. Wind fields and surface-based instability were judged sufficient to support severe thunderstorms, including isolated supercells.

Observational data and operational model guidance suggested that precipitation would develop over this region between 2100 and 0000 UTC and move eastward with the prevailing flow, thus the forecast team focused on the 2100 – 0300 UTC time frame and a regional domain centered on Sioux Falls, South Dakota for the two forecasts. Since the 1200 UTC RUC guidance was available only through 0000 UTC, forecast teams relied most heavily on the Eta model for deterministic guidance in this case. Between 0000 and 0100 UTC the Eta model predicted a broad swath of precipitation along the central Iowa-Minnesota border, with a lobe extending southwestward into northeastern Nebraska and an additional extension along the warm front towards the northwest (Fig. 2b). Based largely on this coverage pattern, the forecast team outlined a large area with 15% probability of severe weather, extending from west-central North Dakota southeastward into South Dakota and encompassing parts of Minnesota, Iowa, and Nebraska (Fig. 3a).

When the high-resolution model output was made available, forecast teams were presented with a different scenario. In particular, the WRF-NMM-EMC and the WRF-EM-NCAR runs developed intense convection over a much smaller area, concentrated in southeastern South Dakota (Figs. 2c and d), with little precipitation elsewhere. Forecasters accepted this scenario, inspired by the consistency between these two model forecasts, the reasonable behavior of these models during the first ~ 15h of integration, and the consistent evolution of other fields in the model guidance.

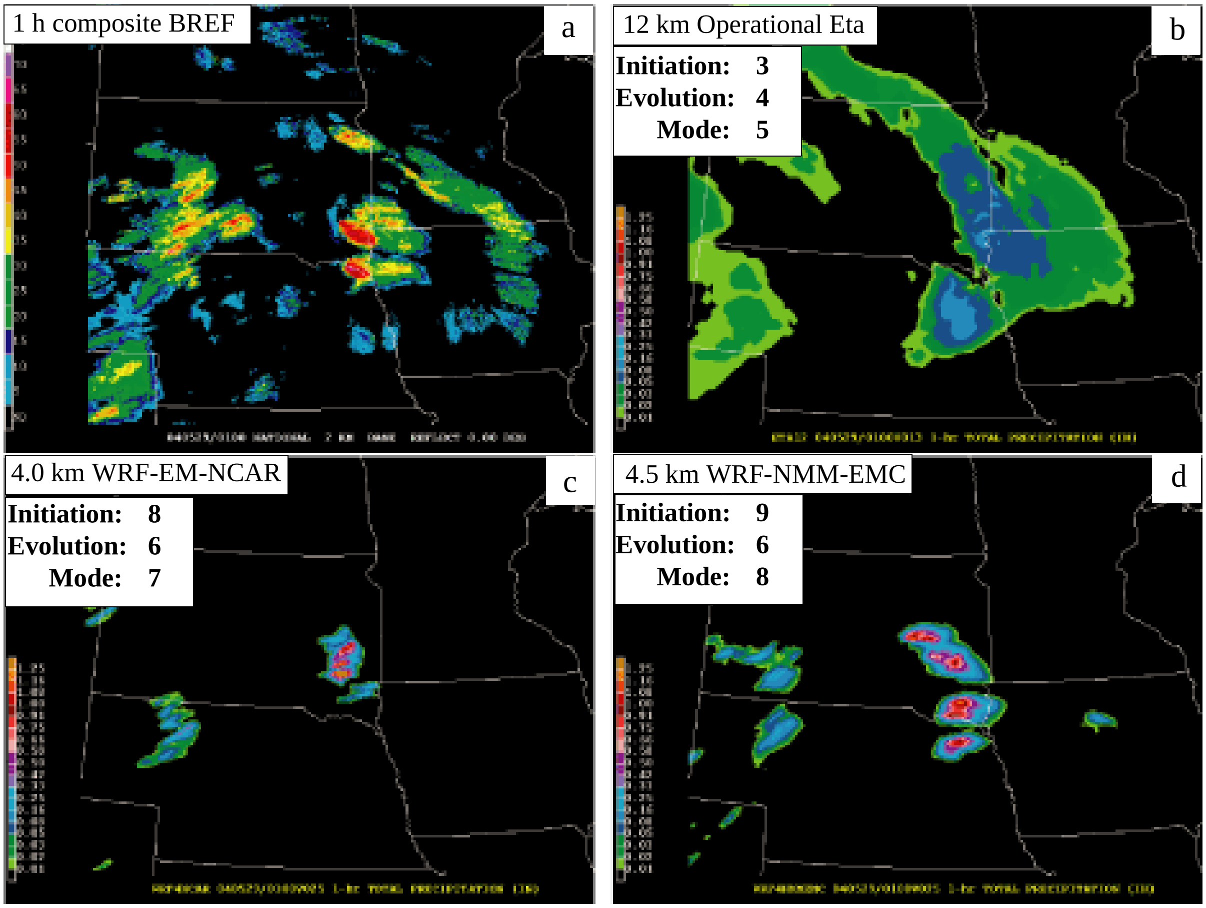

Fig. 2. Verifying radar data and model forecasts valid 0100 UTC 29 May 2004. Time composite of base reflectivity (a) is derived as maximum reflectivity at each pixel over the previous hour. Model forecasts show accumulated precipitation over the previous hour. Subjective verification ratings are indicated in the upper left of the model-forecast frames.

Their response was to reduce the areal coverage of the 15% probability contour substantially and add an area with 25% probability over southeastern South Dakota (Fig. 3b).

When severe weather reports were examined the next working day, it was quite obvious that the high-resolution models had a favorable impact (note the location of severe weather reports in figs. 3a, b). Since these forecasts were made on a Friday, they were evaluated by a new forecast team on Monday morning. The first forecast received 5 points out of 10, receiving credit for encompassing all reports within the 15% contour, but penalty points for the large area farther to the northwest where no reports were received. By comparison, the forecast that benefited from the high-resolution numerical guidance was given 8 points out of 10, as high as any forecast during the entire program. The improvement on this day (3 points) was higher than on any other day on which all three high-resolution models were available.

All NWP forecasts were also evaluated on Monday (note ratings in the upper left hand corners of Fig. 2b, c, d). Although Fig. 2 shows only one of the six hourly frames that were used to evaluate the models, it provides a good sense of the differences between the Eta, WRF-NMM-EMC, and WRF-EM-NCAR forecasts and their relative correspondence with radar data. Other model forecasts were also evaluated on this day, but for the sake of brevity, herein we limit discussion to the highest rated WRF configurations and the “benchmark” Eta model.

In terms of convective initiation, the WRF-NMM-EMC forecast received an exceptionally high score of 9 out of 10, showing excellent correspondence with observations in both timing and location of severe storm development in southeastern

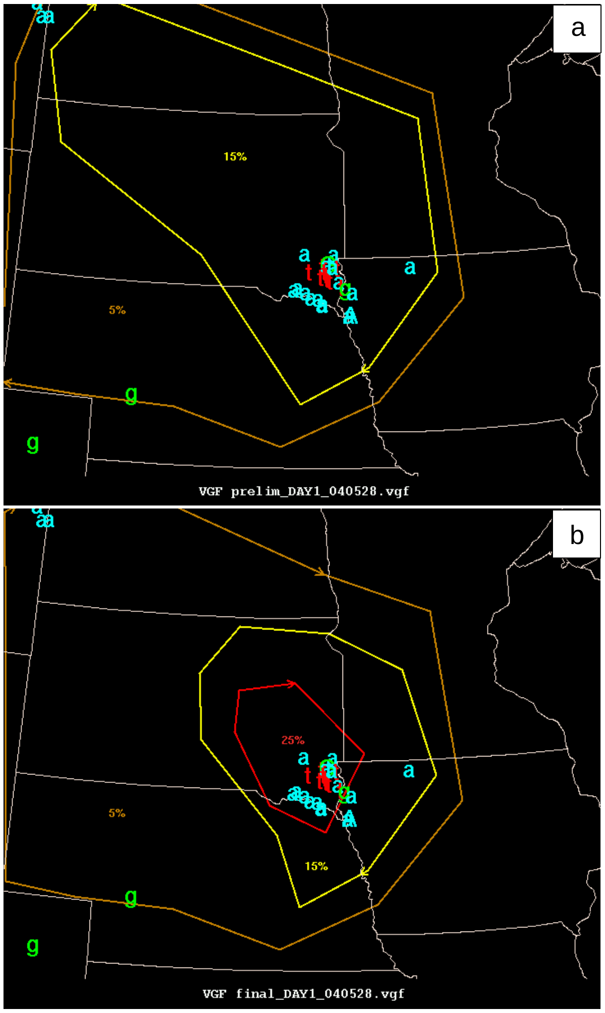

Fig. 3. a) control and b) experimental severe weather forecasts valid 2100 UTC 28 May - 0300 UTC 29 May 2004. Brown line denotes the outline of the 5% probability contour, yellow line 15%, red line 25%. Severe weather reports from this period are denoted by the letters **t** (tornado), **a** (hail ≥ 0.75 in.), and **g** (wind gusts ≥ 50 kts).

South Dakota. The WRF-EM-NCAR was only one point lower, as it was 1-2 h late in activation, but excellent in placement of intense convection. The Eta model was penalized quite heavily by forecast teams because it activated parameterized convection too early and over much too broad of an area, with no focal point for more intense activity (3 out of 10).

In evaluating convective evolution, forecast teams were instructed to focus on direction and speed of system movement, areal coverage, and configuration and orientation of mesoscale features. The nondescript structure of the Eta precipitation field left a negative impression with forecast teams in this category as well. It received a rating of 4 out of 10, with archived comments indicating credit for predicting the direction and speed of movement quite well, but penalty points for too much coverage and "obscured configuration of mesoscale structures". The two WRF configurations fared somewhat better, both receiving a 6 in the evolution category. Both of these forecasts were credited with forecasting the location, movement, configuration and orientation well, but they were criticized for under-predicting the areal coverage.

The two WRF forecasts received high ratings for convective mode (8 for the WRF-NMM-EMC and 7 for the WRF-EM-NCAR) because they correctly predicted intense isolated convective cells where most of the severe reports occurred, with a slight penalty for missing some non-severe multicellular convection that developed elsewhere in the forecast domain. The Eta model earned a 5 for convective mode, producing a "blobbish" precipitation field that was categorized as multi-cellular by the process of elimination, i.e., because neither quasi-linear structures nor isolated cells could be discerned.

3.2 *Human Forecasts: Areal Coverage of Severe Convection*

gusts ≥ 50

On most days, forecast teams made relatively minor updates to the control forecasts when they issued their experimental counterparts (in contrast to the significant adjustment shown in Fig. 3). This is consistent with routine practice at the SPC: Operational forecasters tend to make only incremental changes when updates are issued unless they discover compelling evidence that major modifications are needed. In SPC operations, this approach is prudent because every existing forecast contains a certain amount of inertia, having been systematically assembled from a large body of evidence, including observational surface and upper air data, multiple derived convective parameters, satellite and radar imagery, operational mesoscale and SREF modeling systems, and forecaster experience. During the Spring Program, the control forecasts carried the same weight of supporting evidence, seemingly insulating them from major adjustments. Furthermore, significant changes were difficult to justify because performance characteristics and systematic biases of the particular WRF configurations were relatively unknown to the new forecast teams that were assembled each week. Consequently, forecasters tended to proceed with caution in formulating experimental forecasts.

Control and experimental forecasts were verified using one subjective and two objective measures. In this paper, we focus on the period from 7 May to the end of the program (3 June), since at least two out of three of the high-resolution WRF forecasts were available every day during this period and data collection and archiving procedures were very robust. The sample size for this period is 20 days.

The subjective verification was based on next-day panel evaluations of the accuracy and usefulness of each forecast, focusing on areas with greater observed severe storm coverage or higher forecast probabilities within the regional domain. The evaluation teams also had access to radar signature and severe weather warning information, which was used to supplement the severe reports in regions where low population might affect the number of ground-truth reports of severe weather. Particular attention was given to the skill of the experimental forecast relative to the control (e.g, was it better, worse, or similar in accuracy and usefulness?). In this way, although the panel members varied from week to week and the raw rating numbers were not always uniformly calibrated, the difference between the forecasts could be used to assess relative skill.

Objective measures of forecast skill, verified exclusively against local severe weather reports, were computed using the Brier Score (Brier 1950) and the area under the Relative Operating Characteristic (ROC) curve (Mason 1982). The Brier Score is commonly used to verify probabilistic forecasts and ranges from 0 to 1, with 0 being perfect (lower scores indicate better forecasts). The ROC is also useful for verifying probabilistic forecasts and their ability to discriminate occurrences from non–occurrences. If the area under the ROC-curve is integrated, values range from a perfect score of 1 to a useless value of $\leq$0.5, with an area of > 0.7 considered to represent reasonable discriminating capability. Severe weather reports and forecast probablities were both mapped to an 80 km grid for objective verification, roughly consistent with the concept of detecting severe weather within 25 miles of a point.

Results from the various verification measures are summarized in Table 2. Each of the three approaches provides a more favorable mean number for the experimental forecast, but the differences are very small for the objective measures. All three measures indicate that the experimental forecast was more skillful on about half of the days, but the objective measures suggest that the control forecast was better on nearly half of the days, while the subjective verification indicated a degraded forecast on only two days (10%). The nature of this discrepancy is not entirely clear, but it should be noted that many of the negative changes in the objective scores were quite small, perhaps proportionately less than the smallest increment available in subjective ratings (1 point).

Reliability charts for the control and experimental forecasts were nearly identical, and only the experimental forecast chart is shown (Fig. 4). This indicates excellent reliability for all probability values. For these values the appearance of slight under-forecasting is largely an artifact of the allowable forecast probability intervals (5, 15, 25, etc.),

	Subjective Ratings	**Brier Scores**	**ROC Curve Area**
Mean Forecast Rating (control)	5.4	0.0704 (0.0614)	0.723 (0.772)
Mean Forecast Rating (experimental)	5.9	0.0700 (0.0611)	0.733 (0.784)
Days With Forecast Improvement	11	10	10
Days With Forecast Degradation	2	9	8
Days with Unchanged Ratings/Scores	7	1	2

Table 2. Summary of subjective ratings and objective verification scores for the 19 human forecasts made during the period from 10 May - 3 June. Numbers in parentheses are the overall objective scores for the same period, which differ from the arithmetic mean.

such that a 25% value actually represents all probabilities in the 25-34% range. The slight over-forecast bias evident in 35% probability forecasts may reflect a very small sample size, as these forecasts included only 27 total grid blocks.

In general, these subjective and objective verification measures are consistent in indicating that the high-resolution model data had a small positive impact on the experimental severe weather forecasts. The subjective scores seem to provide a more favorable relative assessment, perhaps reflecting the consideration of additional data sources (radar signatures and warning information) in the subjective evaluation process. These initial experimental results should be regarded as evidence that experienced severe weather forecasters can gain useful and valuable information from near cloud-resolving models on some days.

3.3 ***Human Forecasts: Timing of First Severe Report***

The ability of the forecast teams to predict the time of first severe storm occurrence in two-hour time windows was also explored. For this assessment, the full dataset of 30 forecast days, dating back to April 19, was utilized. Severe storms did not occur within the forecast domain on three days, while severe storms continuing from the morning into the afternoon impacted three other days. For the remaining 24 forecast days, the experimental forecast correctly predicted the time of the initial severe report on 58% (14) of the days. The experimental forecast changed the control forecast time of the first report on only five days, moving to the correct time window on two of these days while trending in the proper direction on two other days. These findings suggest that specific prediction of the onset of severe weather remains very challenging, as details of convective initiation and subsequent intensification are dependent on mesoscale and storm-scale processes that are not well understood.

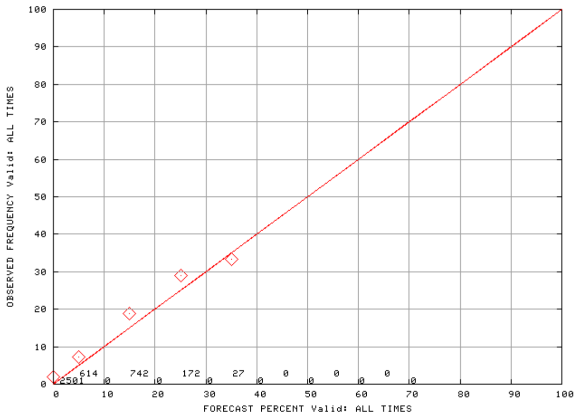

Fig. 4. Reliability diagram for the 20 experimental human forecasts made during the period 7 May - 3 June.

3.4 ***Overall Assessment of Model Forecasts***

Although up to eight different model forecasts were verified on some days, discussion herein is limited to a comparison of the Eta model (as an operational benchmark) and the three high-resolution WRF models. The RUC model would also be a viable operational benchmark, and it was commonly used in preparation of control forecasts. However, output from the 1200 UTC RUC was available in only 3 h intervals beyond the 3 h forecast time (1500 UTC), which greatly limited assessment of convective initiation and evolution. In addition, forecast guidance from the 1200 UTC RUC was only available through 0000 UTC, so it could not be determined if RUC forecasts exhibited possible timing delays. It also contained insufficient data for the selected days when the experimental forecast period extended beyond 0000 UTC. Thus, the RUC could not be evaluated by the same criteria that were used for the other models and it is not included in the statistical results.

Precipitation output from the Eta model and all three high-resolution WRF models was generated on the models' respective native grids every hour. Over the course of the program, there were fifteen days on which complete output from all four of these models was available (all from 7 May to 3 June, the same period considered for verification of human forecasts). The mean subjective ratings for categories of initiation, evolution, and mode on

---

[2] Statistical significance was assessed using the paired t-test (Wilks 1995). Differences that are described as “statistically significant” herein are characterized by a t-test score of 0.05 or lower, indicating that differences are significant at the 95% confidence level.

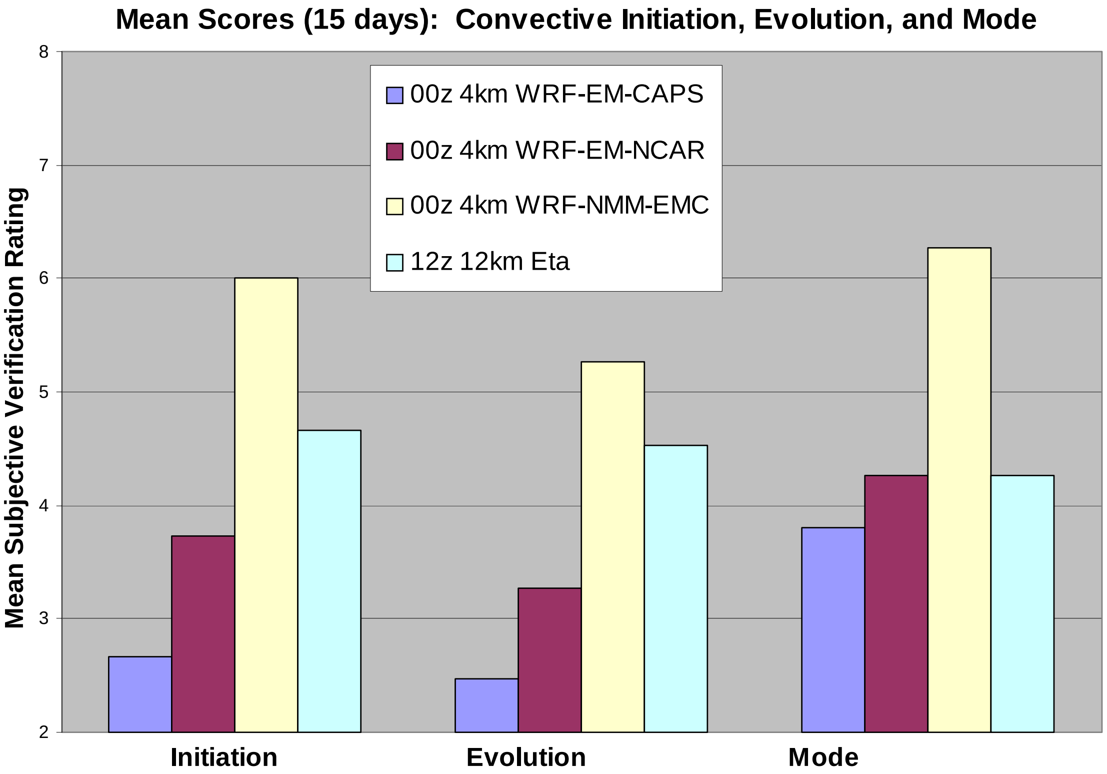

Fig. 5. Mean subjective verification ratings for the operational Eta model and the 3 high-resolution configurations of the WRF model, for categories of convective initiation, evolution, and mode, for the 15 days when all 4 models were available

these days are shown in Fig. 5. Interestingly, the Eta model was rated higher than 2 out of 3 of the high-resolution WRF models. However, one of the high-resolution models (WRF-NMM-EMC) earned the highest mean rating in all three categories. In general, the higher ratings for this model were statistically significant[2] when compared to all other models in all categories, except for the Eta model in the evolution category. The Eta forecasts were rated higher than those from both WRF-EM-NCAR and WRF-EM-CAPS in both initiation and evolution, though differences with the WRF-EM-NCAR run are not statistically significant. In the convective mode forecasts, the WRF-EM-NCAR, WRF-EM-CAPS, and Eta ratings were essentially the same, while the WRF-NMM-EMC rating was significantly better.

The statistical significance of some differences was limited because of the relatively small sample size. Sample size (i.e., the number of days on which all compared models were available) tended to increase as the number of models considered decreased, so numerous direct comparisons were made in addition to those plotted in Fig. 5. In some cases, when differences between two models were already substantial, they became statistically significant in head-to-head comparisons. In general, however, changes in individual model ratings were consistently small when the sample size was increased, such that the results in Fig. 5 remain viable measures of model inter-comparisons.

The relative ratings are somewhat surprising. For example, previous studies have shown that convective initiation is delayed in model simulations with coarsely, but explicitly resolved convection compared to analogous runs using parameterized convection (e.g., Molinari and Dudek 1986; Weisman et al. 1997). Although this finding is consistent with the relative scores of the WRF-EM-NCAR, WRF-EM-CAPS, and Eta forecasts (with Eta earning the higher rating), the comparatively high initiation rating of WRF-NMM-EMC is contrary to previous (albeit limited) work in this area.

Perhaps the most surprising result is in the mode category, which can relate directly to the type, coverage, and intensity of severe convective weather. The WRF-EM-NCAR and WRF-EM-CAPS runs, like the WRF-NMM-EMC, are certainly capable of resolving the meso-γ-scale (~2-20 km) organizational mode of convection much more faithfully than the Eta model. Furthermore, one would expect these two WRF configurations to depict more realistic convective structures because they are not parameterizing convection, yet the convective mode ratings for these two high-resolution models were not significantly different from those for the operational Eta model.

Considering this surprising result, one may be tempted to ask: How credible are these subjective ratings? Fortunately, the convective mode ratings can be substantiated with additional data that was collected during the program. This data is in the form of specific assessments of convective mode from both model forecasts and observations. Mode information from each model forecast was collected during the forecast preparation process, i.e., in the morning before the anticipated convective event occurred. Specifically, based on the 6 one-hourly model precipitation images that were available, forecast teams were asked to partition model-predicted convective activity over the regional forecast domain into different convective modes (discrete cells, multicellular clusters, quasi-linear structures) for each model. In other words, they were asked to partition convective activity by mode in both time and space (with individual modes estimated in 10% increments and a total of 100% required). This is quite different from hierarchical mode classifications that have been done in previous studies (e.g., Fowle and Roebber 2003; Done et al. 2004). Furthermore, it differs in that partitioning of model forecasts was done before the event occurred, precluding any bias that could be associated with knowing “the answer” ahead of time. The following day, the observed mode was derived from radar data using the same classification method. In particular, observed convection during the previous afternoon/evening’s convective event was partitioned into the same 3 mode categories using the 6 one-hour composites of base reflectivity corresponding to the valid forecast period.

A quasi-independent measure of model skill in forecasting convective mode can be obtained by comparing the partition levels estimated from model forecasts to those from radar observations. If the relative skill levels are comparable to those obtained from the average subjective ratings, the subjective ratings will have corroborating support. Several different methods of comparing predicted and observed modes are reflected by the data in Table 3. (Note that the WRF-EM-CAPS runs were excluded here to bring the sample size up to 21

	Mean Daily Error	**Std. Deviation Of Mean Daily Error**	**Mean Absolute Daily Error**	**Corr. Coeff.: predicted vs. observed**
Eta				
Discrete cells	-28.6	24.0	30.5	0.160
Multi-cellular	5.2	26.7	22.4	0.373
Quasi-linear	23.3	30.3	31.0	0.141
WRF-NMM-EMC				
Discrete cells	-11.4	22.9	18.1	0.416
Multi-cellular	-1.0	24.5	20.0	0.340
Quasi-linear	12.4	18.2	15.2	0.510
WRF-EM-NCAR				
Discrete cells	-7.6	35.3	28.6	0.205
Multi-cellular	-4.3	36.5	30.0	-0.018
Quasi-linear	11.9	25.9	18.6	0.381

Table 3. Error statistics related to model forecasts vs. radar observations of convective mode for the 21 days when all 3 models were available. Mode was determined by subjectively partitioning convective activity in both time and space over the 6 h period of each forecast. See text for additonal information.

forecasts.) The second column in Table 3 shows the mean error (difference between predicted and observed) for all modes. The two WRF models show a tendency to under-forecast the occurrence of discrete cells and multicell clusters, while over-forecasting convective lines. The over-forecasting bias evident in the line forecasts is nearly balanced by the under-forecasting of discrete cells. This tendency to produce linear convection in the WRF-EM-NCAR was also found by Done et al. (2004).

The Eta displays the largest under-forecast (over-forecast) bias when cells (lines) are considered. This likely reflects the Eta model characteristic to produce smooth precipitation fields with little in the way of mesoscale structure (Baldwin and Wandishin 2002), which makes it difficult to ascertain discrete cells when the precipitation fields are interpreted in a literal sense. By default, the Eta output gives the appearance of generating multicell clusters and lines more frequently than they are observed. This over-forecasting bias is especially evident for lines, possibly reflecting the tendency of the Eta's Betts-Miller-Janjic convective parameterization (Janjic 1994) to produce gravity-wavelike-phenomena that can appear as bow-like structures in the precipitation field (Bukovsky and Kain 2004).

Interestingly, the WRF-EM-NCAR forecasts have the smallest mean error, which is not consistent with the subjective mode ratings (Fig. 5). However, the WRF-EM-NCAR also shows the largest standard deviation values for mean error, suggesting that the mean value may not provide the best indication of skill. Indeed, mean absolute error values (fourth column) are quite large for the WRF-EM-NCAR suggesting that the small mean error is the result of averaging relatively large positive and negative errors. The WRF-NMM-EMC has the smallest mean absolute errors for all three mode categories, which is consistent with the mean ratings presented in Fig. 5. The correlation coefficient between predicted and observed modes (fifth column) substantiates this finding, with the WRF-NMM-EMC showing moderate correlation for the two higher order structures of discrete cells and lines. Conversely, the Eta displayed the lowest correlation for cells and lines, likely reflecting its tendency to produce smooth, low amplitude precipitation fields that were most often interpreted to be multicell clusters. Since multicell clusters tend to occur more frequently in the atmosphere compared to intense discrete cells and lines, it is reasonable that Eta forecasts exhibited the highest correlation for this mode (slightly higher than the WRF-NMM-EMC).

In general, this specific data on convective mode is consistent with the subjective ratings of mode. Mean absolute error values and correlation coefficients (Table 3) are quite consistent with the relative ratings of mode shown in Fig. 5. In particular, the WRF-NMM-EMC forecast clearly earned the most favorable assessments, while the WRF-EM-NCAR and Eta forecasts were rated similarly, but somewhat lower. These results lend credibility to the subjective ratings, not only those for convective mode, but, by association, for initiation and evolution as well.

4. DISCUSSION

Based on data collected and subjective impressions gained during the program, it appears that the WRF model configured with near storm scale resolution can provide valuable supplemental guidance to severe weather forecasters. In particular, it can provide information on specific topics of convective initiation, evolution, and mode that directly impact the issuance of SPC Mesoscale Discussion and Severe Thunderstorm/Tornado Watch products. On some days, the 0000 UTC WRF models exhibited remarkably skillful and detailed convective forecasts during the 18-30h forecast period, including proper delineation and transition between different convective modes (discrete cells and linear structures). Overall, the WRF-NMM-EMC version showed better skill compared to the other two WRF versions and the benchmark Eta for all three convective categories of initiation, evolution, and mode. The use of the NMM core and its well-calibrated physics packages derived from the Eta model may have in part contributed to the performance of the WRF-NMM-EMC. On the other hand, none of the WRF versions had been previously tested and evaluated for near storm scale forecasting applications (although the WRF-EM-NCAR was utilized for mesoscale convective system forecasting during the 2003 BAMEX program). And the WRF-EM-CAPS model was configured to test the impact of a state-of-the-science data assimilation system (ADAS) that had been originally optimized for use with the ARPS model. The incorporation of Level II radar data into the start of the WRF-EM-CAPS run appeared to be quite effective, with very high correspondence between the model precipitation fields and radar images at one hour into the model run. However, the positive impact of the data assimilation procedure typically ended by 3 hours into the model run, such that the subsequent forecast into the next day was no better (and often worse) than that produced by the two "cold start" WRF runs. These types of ground-breaking experiments are absolutely necessary in order to assess initial performance characteristics and potential modeling system improvements, and as such these results should not be considered representative of future WRF model skill.

Not surprisingly, precipitation forecasts from the WRF model sometimes corresponded poorly to observed radar data when the two were compared the next day. Yet, even on many of these days the high-resolution output looked very similar to detailed precipitation structures and patterns that are commonly seen in images of radar reflectivity. That is, rather than looking like traditional mesoscale model output consisting of relatively smooth and less structured patterns, the WRF output (especially 1-hour and instantaneous precipitation fields) often closely resembled weather phenomena that appeared to be realistic and plausible. As a result, it was imperative to examine other WRF model output relevant to convective forecasting (e.g, CAPE, CIN, surface temperature, dewpoint, and wind fields, vertical shear, etc.). These fields were not only useful in linking the model precipitation to appropriate physical mechanisms, but it was also important to compare these basic fields with observed data to determine if the first 12-15 hours of the model forecast appeared reasonable. Otherwise, there was a risk of uncritically accepting the WRF precipitation output simply because it more readily "looked like" typically observed radar data.

The incorporation of any model guidance into the forecast decision-making process also requires forecaster determination of the likelihood that model guidance will provide useful information. The introduction of the WRF models into the experimental forecast process proved to be no different. This weighing of information is typically done by applying knowledge of model configuration and physics (e.g., Baldwin et al. 2002, Kain et al 2003c) and known model performance characteristics. However, this process was much less well-defined for the WRF models given the lack of a previous experience base to draw upon (although very short-term performance characteristics were sometimes used by the weekly forecast teams to identify possible model biases). Since model forecasts are not always correct or always incorrect, the ability of forecasters to know "how much to believe" about specific model guidance is a vitally important component in forecasting. During the program, the forecast teams determined their confidence level in the model precipitation solutions during each daily forecast preparation phase (i.e., they were asked to express their confidence on a scale from 0 to 10 that the model precipitation forecasts would correspond well to the observed convective weather). This allowed a determination of how much weight was given to each model solution and how it was factored into the resultant forecast. Since verification data on model performance were collected the next day, a comparison of forecaster confidence with model performance can be directly determined, allowing an assessment of how well the forecast teams knew when to believe the model and when to discount the guidance.

The results in Table 4 show that the mean forecaster confidence ratings for categories of initiation, evolution, and mode are highest for the WRF-NMM-EMC, followed by the Eta and WRF-EM-NCAR (the values for the WRF-EM-CAPS are not listed because of the smaller sample size). The order of ranking is identical to the mean model verification ratings, suggesting that the forecast teams were, on average, able to discern relative levels of skill in the different models during the forecast preparation process. Note that the mean confidence scores tended to be slightly higher than the verification scores. Given the large uncertainty in day-to-day forecasting of convective details, it is understandable that forecasters may place more weight on model guidance in an effort to compensate for the absence of other definitive convective forecast information, especially during more weakly-forced situations.

	Mean Confidence Rating	Mean Verification Rating	Conf.-Verif. Correlation
Eta			
Initiation	5.9	4.7	0.42
Evolution	5.6	4.7	0.43
Mode	5.3	4.3	-0.03
WRF-NMM-EMC			
Initiation	6.6	5.9	0.68
Evolution	5.9	5.1	0.27
Mode	6.4	6.4	0.07
WRF-EM-NCAR			
Initiation	4.6	4.1	0.16
Evolution	4.3	3.4	0.55
Mode	4.3	4.3	0.38

Table 4. Forecast-team confidence in model forecasts, assessed at the time human forecasts were issued, compared to next-day verification ratings.

A somewhat different picture emerges when the correlation between forecaster confidence and model verification is examined. While there is modest correlation (usually less than 0.6) between confidence and verification for categories of initiation and evolution, there is virtually no correlation for the mode category when the Eta and WRF-NMM-EMC models are considered. Higher mode correlation was found with the WRF-EM-NCAR model (around 0.38), but this may be related to the lower confidence scores assigned to this model on average. These results suggest that it remains quite difficult to know ahead of time how much confidence to place in high resolution model forecasts of precipitation on a day-to-day basis, which is reasonable given the complexity of the convective forecasting problem. Some of these results were undoubtedly influenced by the lack of known skill levels and performance characteristics for the experimental models. However, it is also somewhat encouraging that moderate levels of confidence-verification correlation were exhibited for initiation (WRF-NMM-EMC), and evolution and mode (WRF-EM-NCAR), while the operational Eta showed respectable levels of correlation for both initiation and evolution.

Since the forecast teams received guidance from multiple WRF models, they were required to deal with new types of forecast uncertainty, especially as it related to prediction of small scale convective details. It was not uncommon to see considerable variability in the temporal and spatial development of model generated convection, which immediately raised questions about how to incorporate this added level of uncertainty into the forecasting process. This step will require model users to think more carefully about the amount of detail they need from a model, including whether there can actually be too much detail in the guidance. Interestingly, this assessment can vary even within a single organization such as the SPC. For example, the time and space detail of model guidance needed by a forecaster issuing Tornado Watches is greater than that required by a forecaster issuing a Convective Outlook, although this difference is diminishing as the requirements to provide Outlook details have increased in recent years. And while this may suggest that high resolution model output may have less margin for error compared to the traditional smooth, less detailed precipitation output from operational mesoscale models, we found that forecasters can still gain valuable information about potential convective development even when timing and/or location details are in error. For example, experienced forecasters often can make adjustments for model placement errors while still incorporating specific information on evolution and/or mode. Recognition of these aspects of forecast uncertainty also highlights the impact of observational and model physics errors on any NWP forecast, suggesting that it may be more fruitful to pursue high resolution WRF ensemble prediction approaches (e.g, Levit et al. 2004).

The explicit portrayal of detailed convective patterns and structures from the high resolution models was routinely viewed by program participants as being more useful than output from current operational mesoscale models. We found that details from all models can be enhanced by generation of higher temporal resolution output grids (i.e., 1-hour output fields instead of 3-hour fields), which is especially valuable when attempting to discern characteristics of convective initiation, evolution, and mode. Although this initial testing of large domain, high resolution WRF models was, on average, favorably received by program participants, there remain many questions and issues that need to be addressed in order for these types of WRF models to be implemented in an operational forecast environment, most notably related to intensive computational, communications, storage, and workstation display requirements. Our experience during the program suggests that specialized severe weather forecasters working closely with model developers and research scientists can combine their complementary knowledge bases and begin to identify ways of applying experimental high resolution model forecasts to address operational forecasting needs within relatively short periods of time. Thus, continued collaboration and dialogue between the operational forecasting and model development communities is essential in order to expedite the development of operationally relevant NWP systems and subsequent transfer of these systems from research to operations.

5. SUMMARY

The 2004 SPC/NSSL Spring Program focused on testing three experimental configurations of the WRF model to determine: 1) if there is new and useful information for an operational forecaster perspective, and, 2) whether severe weather forecasts can be improved when forecasters have access to near-stormscale models using explicit precipitation physics, compared to mesoscale models with parameterized convection. During the seven week experiment conducted during the prime severe weather season, weekly panels consisting of a mix of forecasters, model developers, and research scientists participated in a series of daily tasks. These included both forecasting and evaluation activities aimed at exploring the value of near-cloud scale resolving experimental versions of the WRF model for severe weather forecasting purposes. Convective categories of initiation, evolution, and mode were individually examined, in order to assess forecast skill in these important prediction issues. The prediction of convective mode is an essential aspect of improving severe weather forecasts, given the relationship between severe weather type and intensity and mode (i.e., discrete cells, linear structures, and multicellular clusters), with tornadoes and damaging winds more closely correlated to discrete cells and line segments, respectively.

A comparison of precipitation forecast guidance from the benchmark operational Eta model and the WRF models with observed radar reflectivity found that the WRF-NMM-EMC model scored higher than other models for convective initiation, evolution, and mode. The Eta model was ranked equal to or higher than the WRF-EM-NCAR and WRF-EM-CAPS versions for these three convective categories, although it should be stressed that all models performed well on individual days. Overall, the WRF models demonstrated respectable skill, especially considering they were initialized 12 hours earlier than the Eta model and, unlike the Eta, were required to predict the evolution of nocturnal convection prior to the start of the next day's diurnal cycle. In particular, it was found that the high resolution WRF models were substantially more capable of resolving convective structure compared to the Eta model, and they showed promise of providing new and unique guidance for severe weather forecasters. The severe weather forecasts that incorporated output from the WRF models displayed small but positive improvement on the majority of days using both subjective and objective measures, suggesting that the forecast teams were able to extract useful information from these models. Given the lack of prior experience and knowledge about performance characteristics of the WRF models, this result is quite meaningful.

It is also important to acknowledge that the near-stormscale resolving models tested during the Spring Program were early versions of the WRF model, and they did not have the benefit of a long period of "fine-tuning" to ensure that all of the model physics, parameterization schemes, and initialization procedures were well-calibrated for each particular model configuration. For example, the WRF-EM-CAPS run used a state-of-the-science data assimilation system, but this system was developed and calibrated in the ARPS model and was relatively untested with the WRF package; scientists at NCAR discovered after the program that a horizontal diffusion parameter had been set incorrectly in the WRF-EM-NCAR configuration. Model performance during the Spring Program certainly did not reflect the best we can hope to see from the WRF model. Rather, the Spring Program provided an early benchmark for the performance of this model in severe convective weather situations. Results from the program provide valuable and unique feedback to model developers as they continue to evaluate the WRF model and optimize its performance.

Finally, before the start of the Spring Program we were uncertain about the value and potential impact of high resolution models in the operational severe weather forecasting arena. The experiences of the numerous forecasters and research scientists who participated in the experiment and the documented findings indicate that high resolution WRF models demonstrated remarkable forecast skill on some days. These results strongly suggest that continued development efforts in high resolution WRF modeling will be beneficial to operational forecasting, and they have clear potential to improve severe weather forecasts.

Acknowledgements

We are very grateful to numerous individuals at CAPS, EMC, and NCAR for their valuable collaboration and extraordinary efforts to generate and disseminate high resolution model output on a daily basis. In particular, we thank Kevin Thomas, Dan Weber, and Kelvin Droegemeier from CAPS; Matt Pyle, Zavisa Janjic, and Geoff DiMego from EMC; and Wei Wang, Morris Weisman, and Joe Klemp from NCAR. We thank Adam French, a student from Valparaiso University who visited the SPC under the ORISE program and compiled much of the data discussed herein. In addition, we thank all participants in the 2004 Spring Program (see Appendix A). This program would not be possible without the technical support of many people at the NSSL/SPC/CIMMS facility in Norman, especially Phillip Bothwell, Gregg Grosshans, Jay Liang, and Doug Rhue from the SPC; and Brett Morrow and Steve Fletcher from Indus/NSSL. Linda Crank (SAIC/SPC), Tracy Reinke (CIMMS/NSSL), and Peggy Stogsdill (SPC) are deeply appreciated for making complex travel arrangements as simple as possible. Finally, we would like to thank COMET, an organization that has supported this and previous collaborative activities at the Oklahoma Weather Center. This year's Spring Program was partially supported by COMET Cooperative Project S03-44678 and COMET Partner's Project S04-44693.

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.

Baldwin, M. E., and M. S. Wandishin, 2002: Determining the resolved spatial scales of Eta model precipitation forecasts. *Preprints, 15th Conference on Numerical Weather Prediction*, San Antonio, TX, Amer. Meteor. Soc., 85-88.

Benjamin, S. G., G. A. Grell, J. M. Brown, T. G. Smirnova, and R. Bleck, Rainer, 2004: Mesoscale weather prediction with the RUC hybrid isentropic–terrain-following coordinate model. *Mon. Wea. Rev.,* **132**, 473–494.

Black, T. L. 1994: The New NMC Mesoscale Eta Model: Description and Forecast Examples. *Wea. Forecasting*, **9**, 265–284.

Bluestein, H. B. and M. L. Weisman, 2000: The interaction of numerically simulated supercells initiated along lines. *Mon. Wea. Rev.*, **128**, 3128-3149.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.

Brewster, K., 1996: Application of a Bratseth analysis scheme including Doppler radar data. *Preprints, 15th Conf. on Weather Analysis and Forecasting,* Norfolk, VA, Amer. Meteor. Soc., 92–95.

Bukovsky, M. S., and J. S. Kain, 2004: Predicting propagating convective systems using operational forecast models. *Preprints, 22nd Conference on Severe Local Storms*, Hyannis, MA, Amer. Meteor. Soc., CD-ROM, paper 17.7.

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

Dial, G. L. and J. P. Racy, 2004: Forecasting short term convective mode and evolution for severe storms initiated along synoptic boundaries. *Preprints, 22nd Conference on Severe Local Storms*, Hyannis, MA, Amer. Meteor. Soc., CD-ROM, paper 11A.2

Done, J., C. Davis, and C. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) model. *Atmos. Sci. Let.*, accepted for publication.

Du, J., J. McQueen, G. J. DiMego, T. L. Black, H. Juang, E. Rogers, B. S. Ferrier, B. Zhou, Z. Toth, and Steve Tracton, 2004: The NOAA/ NWS/NCEP Short Range Ensemble Forecast (SREF) system: Evaluation of an Initial Condition vs Multiple Model Physics Ensemble Approach. *Preprints, 20th Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., CD-ROM, 21.3

Dudhia, J., 1989: Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.,* **46**, 3077–3107.

Elmore, K. L., S. J. Weiss, and P. C. Banacos, 2003: Operational ensemble cloud model forecasts: Some preliminary results. *Wea. Forecasting*, **18**, 953–964.

Ferrier, B. S., Y. Jin, Y. Lin, T. Black, E. Rogers, and G. DiMego, 2002: Implementation of a new grid-scale cloud and precipitation scheme in the NCEP Eta model. *Preprints, 15th Conf. on Numerical Weather Prediction*, Amer. Meteor. Soc., San Antonio, TX, 280-283.

Fowle, M. A., and P. J. Roebber, 2003: Short-range (0-48 h) numerical prediction of convective occurrence, mode and location. *Wea. Forecasting*. **18**, 782-794.

Iacono, M. J., E. J. Mlawer, S. A. Clough, and J.-J. Morcrette, 2000: Impact of an improved longwave radiation model, RRTM, on the energy budget and thermodynamic properties of the NCAR Community Climate Model, CCM3. *J. Geophys. Res.,* **105**, 14873–14890.

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., 2001: Nonsingular Implementation of the Mellor-Yamada Level 2.5 Scheme in the NCEP Meso Model. NOAA/NWS/NCEP Office Note #437, 61 pp.

Janjic, Z. I., 2003: A nonhydrostatic model based on a new approach. *Meteor. Atmos. Phys.*, **82**, 271-285.

Janjic, Z. I., T. L. Black, M. E. Pyle, H.-Y. Chuang, E. Rogers, and G. J. DiMego, 2004: The NCEP WRF core. *Preprints, 5th WRF/14th MM5 User's Workshop*, 22-25 June, Boulder, CO, 184-187.

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

Kain, J. S., P. R. Janish, S. J. Weiss, M. E. Baldwin, R. S. Schneider, and H. E. Brooks, 2003b: Collaboration between forecasters and research scientists at the NSSL and SPC: The Spring Program. *Bull. Amer. Meteor. Soc.*, **84**, 1797-1806.

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, N. L., K. K. Droegemeier, and F. Kong, 2004: High-Resolution Storm-Scale Ensemble Forecasts of the March 28, 2000 Fort Worth Tornadic Storms, *Preprints, 20th Conference on Weather Analysis and Forecasting/16th Conference on Numerical Weather Prediction*, Seattle, WA, Amer. Meteor. Soc., CD-ROM, paper 23.6.

Lin, Y.-L, R. D. Farley, and H. D. Orville, 1983: Bulk parameterization of the snow field in a cloud model. *J. Clim. Appl. Met*., **22**, 1065-1092.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Met. Mag*., **30**, 291-303.

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.

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.,* **102**, 16663–16682.

Molinari, J.M., and M. Dudek, 1986: Implicit versus explicit convective heating in numerical weather prediction models. *Mon. Wea. Rev.*, **120**, 326-344.

Noh, Y., W. G. Cheon, and S. Raasch, 2001: The improvement of the K-profile model for the PBL using LES. *Preprints of the International Workshop of Next Generation NWP Model*, Seoul, South Korea, 65-66.

Robbins, C. C., J. V. Cortinas, 2002: Local and synoptic environments associated with freezing rain in the contiguous United States. *Wea. Forecasting*, **17**, 47–65.

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

Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243–1261.

Tuleya, R. E., 1994: Tropical storm development and decay: Sensitivity to surface boundary conditions. *Mon. Wea. Rev.,* **122**, 291–304.

Weisman, M. L., J. B. Klemp, and R. Rotunno, 1988: Structure and evolution of numerically simulated squall lines. *J. Atmos. Sci*., **45**, 1990-2013.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.*, **125**, 527-548

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

APPENDIX A. Participants in the 2004 Spring Program

NOAA Agencies:
- **NWS/NCEP/SPC:**
  - D. Bright
  - G. Carbin
  - G. Dial
  - R. Edwards
  - S. Goss
  - J. Levit
  - C. Mead
  - J. Peters
  - J. Racy
  - S. Weiss
- **NWS/NCEP/EMC:**
  - J. Du
  - B. Ferrier
  - Z. Janjic
  - G. Manikin
  - Z. Toth
- **NWS/NCEP/HPC:**
  - M. Eckert
  - P. Manousos
- **NWS/DTX:**
  - G. Mann
- **OAR/FSL:**
  - J. Brown

Cooperative Institutes:
- **CIMMS/OU/NSSL:**
  - M. Baldwin
  - D. Burgess
  - J. Gourley
  - J. Kain
  - D. Schultz
  - P. Spencer
  - M. Wandishin
- **CIRES/CU/CDC:**
  - B. Mapes

Universities:

- Arizona:
  - S. Mullen
- Iowa State:
  - B. Gallus
- Missouri-Columbia:
  - P. Market
- Naval Postgraduate School:
  - R. Elsberry
- Oklahoma:
  - M. Bukovsky
- SUNY-Albany:
  - L. Bosart
  - T. Galarneau
  - A. Wasula
- Utah:
  - J. Steenburgh
- Valparaiso:
  - A. French
- Wisconsin - Madison:
  - G. Tripoli
- Wisconsin - Milwaukee:
  - P. Roebber

NRC/NSSL:

- V. Homar

Air Force Weather Agency:

- R. Curran
- S. Rugg

CAPS/OU:

- D. Weber

NCAR:

- M. Weisman

Meteorological Service of Canada:

- B. Bukoski
- A. Erfani
- P. King
- D. Patrick
- P. Vaillancourt

Finnish Meteorological Institute:

- J. Teittinen

Koch/Entergy Corporation:

- P. Janish