Evaluation of WRF model output for severe-weather forecasting from the 2008 NOAA Hazardous Weather Testbed Spring Experiment

Michael C. Coniglio*, Kimberly L. Elmore, John S. Kain

NOAA/OAR/National Severe Storms Laboratory, Norman, OK

Steven J. Weiss

NOAA/NCEP/Storm Prediction Center, Norman, OK

Ming Xue

Center for the Analysis and Prediction of Storms and School of Meteorology, University of Oklahoma, Norman, OK

Morris L. Weisman

National Center for Atmospheric Research

Boulder, CO

Revised manuscript submitted to *Weather and Forecasting*

June 23, 2009

* *Corresponding author address*: Michael Coniglio, NSSL, 120 David L. Boren Blvd., Norman, OK 73072; e-mail: Michael.Coniglio@noaa.gov

Abstract

This study assesses forecasts of the pre-convective and near-storm environments from the convection-allowing models run for the 2008 National Oceanic and Atmospheric Administration (NOAA)/Hazardous Weather Testbed (HWT) Spring Experiment. Evaluating the performance of convection-allowing models (CAMs) is important for encouraging their appropriate use and development for both research and operations. Systematic errors in the CAM forecasts included a cold bias in mean 2-m and 850-hPa temperatures over most of the United States and smaller than observed vertical wind shear and 850-hPa moisture over the High Plains. The placement of airmass boundaries was similar in forecasts from the CAMs and the operational North American Mesoscale (NAM) model that provided the initial and boundary conditions. This correspondence contributed to similar characteristics for spatial and temporal mean error patterns. However, substantial errors were found in the CAM forecasts away from airmass boundaries. The result is that the deterministic CAMs do not predict the environment as well as the NAM. It is suggested that parameterized processes used at convection-allowing grid length, particularly in the boundary layer, may be contributing to these errors.

It is also shown that mean forecasts from an ensemble of CAMs were substantially more accurate than forecasts from deterministic CAMs. If the improvement seen in the CAM forecasts when going from a deterministic framework to an ensemble framework is comparable to improvements in mesoscale model forecasts when going from a deterministic to an ensemble framework, then an ensemble of mesoscale model forecasts could predict the environment even better than an ensemble of CAMs. Therefore, it is suggested that the combination of mesoscale (convection-parameterizing) and CAM configurations is an appropriate avenue to explore for optimizing the use of limited computer resources for severe-weather forecasting applications.

1. Introduction

The Storm Prediction Center (SPC) and the National Severe Storms Laboratory (NSSL) conducted the 2008 Spring Experiment (SE08) over a seven-week period during the peak severe convective season, from mid April through early June, as part of the activities for the National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT). As in past Spring Experiments, a vital component to its success was the active participation by diverse members of the meteorological community (e.g., forecasters, research scientists, model developers) who have a passion for operationally relevant meteorological challenges (Kain et al. 2008). As in recent years, the primary focus in 2008 was on the examination of convection-allowing configurations of the Weather Research and Forecasting (WRF) model (Skamarock et al. 2005) covering approximately the eastern three-fourths of the continental U. S. (CONUS) in a simulated severe-weather-forecasting environment. As in previous experiments, WRF model forecasts were evaluated based on their ability to predict the location and timing of thunderstorm initiation and evolution, and offer operationally relevant and useful information on thunderstorm morphology. In addition, the experiment continued to evaluate a real-time, large domain 10-member convection-allowing ensemble forecasting system to gauge the potential benefits of uncertainty information at convection-allowing model (CAM) resolutions (Kong et al. 2008; Xue et al. 2008).

Several studies have examined the effect of model horizontal grid length on forecast accuracy. As described in the review by Mass et al. (2002), increasing resolution produces better-defined and more realistic structures in general (evaluated subjectively), but few studies have demonstrated that forecast accuracy, measured objectively over an extended period of time,

increases as grid spacing decreases below approximately 10-15 km. More recently however, several modeling studies using convection-allowing configurations of the WRF model with horizontal grid spacings of ~4 km have suggested that these high-resolution models add value as weather forecast guidance tools. For example, Done et al. (2004) demonstrated that convection-allowing 4-km WRF forecasts predict the frequency and structures of convective systems better than 10-km WRF forecasts that used convective parameterization. Similarly, Kain et al. (2005), Weisman et al. (2008), and Schwartz et al. (2009) all found that 4-km WRF forecasts with explicitly resolved convection yielded better guidance for precipitation forecasts than the 12-km North American Mesoscale (NAM) model that uses convective parameterization. Additionally, these experiments revealed that running the WRF model at 4 km without convective parameterization does not result in grossly unrealistic precipitation forecasts, even though a 4-km grid is too coarse to fully capture convective scale circulations (Bryan et al. 2003). This is consistent with the findings of Weisman et al. (1997) that 4-km grid length is sufficient to capture much of the mesoscale structure and evolution of convective systems.

Despite the successes of CAM forecasts noted above, studies on the accuracy of model forecasts of the mesoscale and larger scale fields within the CAM forecasts and how they relate to forecasts of the environment from models that do not allow convection at the grid scale are scarce (Weisman et al. 2008). The studies mentioned in the previous paragraph focus largely on the quantitative precipitation forecasting (QPF) problem. The lack of investigation into model-environment forecasts led the developers of the SE08 to assemble a daily evaluation of the relationship between model forecasts of convective storms and model predictions of the pre-convective and near-storm environment. This endeavor was driven by recent subjective impressions that large errors in 18-30 h explicit forecasts of convection may be controlled

frequently by errors in the fields that are introduced by the initial conditions (ICs) and lateral boundary conditions (LBCs) (Kain et al. 2005; Weisman et al. 2008).

SE08 participants were asked to spend ~1 h each day analyzing the 18-30 h forecasts of the mesoscale and larger scale environments from convection-allowing models and identifying differences with the associated analyses from the Rapid Update Cycle (RUC) model (Benjamin et al. 2004b). Indeed, on some days, it was possible to identify errors in phase and amplitude within the larger scale fields that clearly had a negative impact on the CAM forecasts of convective initiation and evolution. The errors were apparently inherited from the IC/LBCs provided by the forecasts from the operational version of the NAM, and on some days it was noted that NAM forecasts of convective precipitation showed biases that were very similar to those from the CAMs.

Although the subjective impressions on the quality of the model forecasts suggested many days in which the CAM forecasts were driven largely by the NAM ICs and LBCs, it is not clear if the increased resolution improved the forecasts of the environment overall compared to the NAM forecasts. It is important for model developers to know the performance characteristics of the forecasts of the larger scale fields within the convection-allowing model forecasts, which can determine the location and timing of the explicit forecasts of convection. In addition, forecasters continue to rely heavily on relationships between the environment and convective mode and evolution (Weiss et al. 2008); so it is important to maintain quality in the environmental forecasts as resolution increases.

The purpose of this paper is to produce a quantitative comparison of forecasts of the pre-convective and near-storm environments from the convection-allowing models and forecasts of the environments from operational mesoscale models. Most of the model evaluations are

restricted to the regions of the U.S. that have the potential for severe weather on a given day, which provides a potentially more meaningful analysis than if the evaluations were performed over the entire model domains, which cover at least the eastern two-thirds CONUS. More specifically, this paper addresses the following questions:

1. What are the model biases and how quickly do forecast errors in the pre-convective and near-storm environmental fields grow in the CAMs and in the lower resolution models?
2. How do forecasts of the environment from the CAMs compare to those from lower resolution operational models?
3. Do ensemble mean CAM forecasts of the environment improve upon those from the deterministic CAMs composing the ensemble?

Knowledge gained from answers to these questions can provide specific information to model developers that can guide efforts to improve various components of the WRF model and aid in continued development of operational CAM systems. Furthermore, this knowledge may help forecasters assess how much confidence to have in model guidance for severe-weather forecasting applications.

2. SE08 models

a. Deterministic WRF models

The model evaluation includes output from various configurations of the WRF model provided by the Center for Analysis and Prediction of Storms (CAPS) at the University of

Oklahoma, the National Centers for Environmental Prediction Environmental Modeling Center[1] (EMC), the National Center for Atmospheric Research (NCAR), and NSSL (see Table 1 for a description of each model). The forecasts were initialized at 0000 UTC and 1-hrly forecast output was provided out to at least 30 h for each day of the experiment. Although the size of the domains varied among the models, all of the models covered roughly the eastern three-fourths of the CONUS.

The EMC and NSSL models used the 0000 UTC analyses of the operational NAM model (Janjić 2003) for the ICs and the corresponding NAM forecasts for the LBCs. The EMC model used the 12-km operational NAM while the NSSL model used the operational NAM fields interpolated to a 40-km domain. The NCAR model used the WRF 3DVAR assimilation system on a 9-km outer domain, cycling every 3 hours from 1200 UTC to 0000 UTC, to provide ICs for the NCAR one-way nested inner domain with 3-km grid spacing, which was then initialized at 0000 UTC. Both the 9-km and 3-km forecasts extended for an additional 36 h, with the LBCs for the 3-km nest coming from the parallel 9-km run. ICs and LBCs for the 9-km run were supplied from the operational EMC Global Forecast System (GFS) model (Environmental Modeling Center 2003). Although the NCAR model is not the primary focus of this paper, an evaluation of output from the 3-km domain is presented because this system provided the first extensive real-time test of the WRF 3DVAR system used with a convection-allowing configuration.

b. WRF-ARW 10-member ensemble

[1] This run is similar to the operational "high-resolution window" deterministic forecasts run at ~4-km horizontal resolution produced by the Environmental Modeling Center (EMC), which are nested within the 12 km NAM domain. For more details on these forecasts, see URL http://www.emc.ncep.noaa.gov/mmb/mmbpll/nestpage/.

The 10-member WRF ensemble was produced by CAPS (Table 2) and run at the Pittsburgh Supercomputing Center. It used physics and IC/LBC diversity in nine out of 10 members, with one "control" member (identified as "CAPS-CN" in Table 1 and "CN" in Table 2) and eight perturbed members. The control member uses a real-time analysis using the ARPS 3DVAR system for the ICs (Gao et al. 2004; Hu et al. 2006), in conjunction with the 12-km NAM 0000 UTC analysis that was used as the background field. Additional data from the wind profiler and surface observation networks, as well as radial velocity and reflectivity data from the WSR-88D radar network (Xue et al. 2008) were used in the assimilation. The forecasts from the 12-km NAM 0000 UTC cycle were used to provide the LBCs. Mesoscale atmospheric perturbations were introduced in the IC/LBCs of these eight members by extracting four pairs of positive/negative bred perturbations from EMC's operational Short Range Ensemble Forecast (SREF) system (Du et al. 2006) and applying them separately to the eight members. Furthermore, radar data (3D reflectivity and radial velocity) were assimilated into the control and eight perturbed members, using the ARPS 3DVAR data assimilation system. The tenth member was configured identically to the control member, except that it used the 12-km NAM 0000 UTC analysis directly, i.e., no data assimilation was conducted (identified as C0 in Table 2).

3. Model evaluation method

An objective evaluation of model output was performed to supplement the subjective findings of the SE08 participants. Select model forecast fields and 20-km RUC analyses were filtered to a common domain using a Gaussian distance-dependant weighting function. The common domain covers roughly the eastern two-thirds of the CONUS (Fig. 1) with a horizontal grid spacing of roughly (1/3) by (1/3) degree latitude and longitude, giving a zonal grid length of

about 24 km at the southern boundary and about 33 km at the northern boundary. The filtering of the model fields allows for a more direct comparison of the mesoscale and larger scale features from the various models to the resolvable scales of the associated RUC analyses than if the native grids were used and reduces greatly the influence of the convective-scale perturbations on the analysis (see the Appendix for more details on the filtering procedure).

Six fields were selected for the evaluation, including the 2-m and 850-hPa temperature (TMP2m and TMP850) and dewpoint (DPT2m and DPT850), the convective available potential energy of a surface parcel (CAPE[2]), and the magnitude of the vector difference of the winds (wind shear) between 10 m and 500 hPa (WSHR). These fields from the four daily WRF model forecasts (described in Table 1) and the mean of the CAPS ensemble (CAPS-ENSMEAN) are filtered to the common grid every three hours from 0 to 30 h. In addition, the accuracy of the CAM forecasts of the pre-convective and near-storm environments compared to the lower resolution models is examined by filtering the forecasts from the operational versions of the NAM and GFS[3] models to the common grid using the same method for filtering the CAM forecasts described above. The associated analyses are obtained by filtering the 20-km RUC analyses to the common grid using the same method. We refer the interested reader to Benjamin et al. (2004a) and Benjamin et al. (2007) for a detailed description of the RUC model and data assimilation system and to Benjamin et al. (2004b) for a detailed comparison of the RUC analyses to observations.

---

[2] It should be noted that the "surface" CAPE used in this study is not computed identically among the models. The parcel that is used to calculate the RUC "surface" CAPE is an average of the thermodynamic conditions of the lowest seven model levels (~50 hPa deep), whereas the "surface" CAPE from the NAM and the CAMs use the parcel with the maximum equivalent potential temperature (θ_e) in the lowest 70 hPa. Although these differences limit the ability to interpret any individual model/analysis comparisons, the results are presented to illustrate the magnitude of the typical differences in CAPE between the model outputs and the analysis that are currently available routinely in operations.

[3] The interpolation of the GFS forecast fields required different parameters in the filtering procedure because the resolution of the GFS input grids is lower than that of the common grid. This resulted in smoother fields compared to the filtered fields from the other models.

A daily task for the SE08 participants was the development of an experimental severe-weather forecast within a regional domain selected by the participants. The size of the regional domain was held fixed at roughly 1200 by 900 km, but was moved daily to a region deemed likely to experience convective weather during the forecast period (see Fig. 2 for an example of the regional domain and an experimental forecast). The objective evaluation of the model output discussed herein focuses on the area encompassed by these regional domains and to the 31 days on which an experimental forecast was issued (see Fig. 1 for the center points used for the regional domains during the SE08). This focuses the quantitative evaluation to specific regions known to be relatively favorable for severe convective weather throughout the period examined. The objective measures for each forecast field include the mean error (bias) and the root mean squared error (RMSE) between the filtered model fields and the filtered RUC analyses over all grid points within the regional domain at a given time or over a specified time period, depending upon the particular analysis.

4. Mean RMSE and bias time trends

a. Model RMSE averaged over the regional domains

Of interest to model developers and users of 18-30 h model forecasts for severe weather forecasting applications is the growth of the errors starting from the initial model state (represented here as the difference between model forecasts and the associated RUC analysis). The mean RMSEs for each variable, averaged over the regional domains for all days of the experiment, are found to have different growth characteristics (Fig. 3). Nearly all of the TMP850 errors grow steadily through 12 h, then are steady or decrease in the 12-21 h period

before exhibiting growth again from 18-30 h (Fig. 3c). For the next-day convective forecasts (18-30 h), the TMP850 errors are 1.4-1.8 times larger compared to the initial errors. Similarly, WSHR errors for most of the models show initial growth out to 9 h, then are steady or decrease during the 9-15 h period, before increasing once again from 15- 30 h (Fig. 3f). As for TMP850, the WSHR errors are 1.4-1.8 times larger for the 21-30 h forecasts compared to the initial errors, which is slightly smaller than the 1.5-2.5-day doubling times for errors in model fields dominated by larger-scale flows found in Simmons et al. (1995).

The TMP2m, DPT2m, and DPT850 errors (Figs. 3a, 3b, and 3d) seem to be affected more by the diurnal cycle over the 30 h period than the TMP850 and WSHR errors. This dependence on the diurnal cycle is clearly manifest in the CAPE errors, which are 2-3 times larger in the late afternoon (f21-f24) than in the early morning (f06-f12) (Fig. 3e). Although the error growth for CAPE contains a strong diurnal signal, the removal of this signal still reveals an increasing trend; the CAPE errors at 24 h (valid at 0000 UTC the following day; the same nominal time of the initial analysis) are 1.2-1.8 times larger compared to the initial time, which is smaller than the error growth for the TMP850 and WSHR (Figs. 3c and 3f). This relatively slow growth in model error of fields that are directly impacted by the planetary boundary layer (PBL) evolution compared to the typical growth of errors in larger-scale flows in the free atmosphere (Simmons et al. 1995, Wandishin et al. 2001) is found in Mass et al. (2002), Eckel and Mass (2005), and Kong et al. (2007). Mass et al. (2002) suggest that orography, surface forcing for heat and moisture fluxes, and the greater influence of the physical parameterizations of the PBL at the mesoscales add a more predictable, deterministic component to the deterioration of the larger-scale forecasts that acts to constrain the growth of the initial errors.

b. Model bias averaged over the regional domains

Some insight into the contributions to the mean RMSE values is gained by examining the mean biases (Fig. 4). The 2-m temperatures tend to be too warm at night (0300-1200 UTC) and too cool during the afternoon and evening (1800-0000 UTC), although a few do not exhibit this tendency (Fig. 4a). For example, the NCAR run tended to be too cool at the surface at all hours except at the 0000 UTC start time and at 1800 UTC, and the CAPS-CN and NAM models maintain a slight warm bias throughout the daytime hours.

The diurnal variation seen in the mean RMSE for TMP2m, DPT850, and CAPE (Figs. 3a, 3d, and 3e) is likely a reflection of the diurnal variation in the mean biases among the various models (Figs. 4a, 4d, and 4e). Interestingly, the two lower-resolution models (NAM and GFS) showed very little bias for DPT850 (Fig. 4d) and their RMSE values were among the smallest of all the models (Fig. 3d), whereas all of the CAMs show a distinct dry bias for DPT850 (Fig. 4d), which is discussed further in section 5. The TMP850 errors tend to be small, but show a slight warm bias early in the diurnal cycle (Fig. 4c).

Finally, it is noteworthy that all of the models under-forecast the magnitude of the wind shear during the day and the negative biases peak in the period when most severe convection tends to occur (Fig. 4f). The reasons for this low bias are examined further in section 5. This low WSHR bias could partly explain the inability of the CAMs to maintain strong convective structures in the early nighttime hours noted by the SE08 participants and which is shown in Fig. 4 of Schwartz et al. (2009b) and Fig. 3 of Clark et al. (2009) for the convection-allowing models run for the 2007 Spring Experiment.

5. Comparison of operational mesoscale models and CAM forecasts

A primary goal of this study is to compare the forecasts of the environment from the deterministic CAMs to the lower resolution models that provide the ICs and LBCs. An assessment of the environment forecasts remains an essential part of the severe convective weather forecast process, and these results can allow forecasters to gauge the performance of the CAMs against the NAM and GFS forecasts that are used widely in operations.

a. *Spatial distribution of model bias*

The model biases averaged over the regional domains (Fig. 4) are explored further by calculating the spatial distribution of the biases over all days and all times for which model output was available. The technique described in Elmore et al. (2006) to determine the statistical significance of biases in the face of naturally occurring spatial correlation is used for this comparison. It is found that the CAM forecasts tend to produce larger biases than the NAM forecasts over much of the CONUS and for many variables, particularly for low-level temperature. The NSSL model is used in an illustration of this larger bias (Figs. 5 and 6) because the low-level thermodynamic error characteristics from similar models have been examined extensively in previous Spring Experiments through subjective comparisons of forecast and observed soundings (Kain et al. 2008). These comparisons made in past Spring Experiments revealed systematic biases that are confirmed in this study and discussed below.

The NAM 24-h forecasts of TMP2m show regions that tend to be too cool and too warm in regional corridors (Fig. 5a), but a common systematic bias noticed by the Spring Experiment participants is illustrated by the NSSL 24-h forecasts that clearly show a cool bias over most of the domain (Fig. 5b). This led Kain et al. (2005) to suggest that the Mellor-Yamada-Janjić

(MYJ) scheme in the WRF-ARW core, working with the NAM ICs and LBCs, is largely responsible for creating these conditions that are usually too cool near the surface during the late afternoon and early evening hours. The regions that show the largest TMP2m cold biases in the NSSL model (southern Iowa and Missouri and over the Appalachian states) are regions that show only a slight cool bias in the NAM (Figs. 5a and 5b). The biases in TMP850 show similar characteristics (Figs. 5c and 5d). In particular, the NSSL model has a significant cold bias at 850 hPa along the high Plains, where there is no significant bias in the NAM model. Likewise, the regions that show a small to insignificant cool bias at 850 hPa in the NSSL forecasts (central to northern High Plains and eastern Texas) are regions that show a warm bias in the NAM model (Figs. 5c and 5d). This suggests that the physical parameterizations used by the NSSL model are systematically adding a cold bias to forecasts that use the NAM ICs and LBCs.

Kain et al. (2005) and Weisman et al. (2008) also suggested that the MYJ scheme run with convection-allowing resolutions produces a moist bias in the PBL. The present results for the NSSL DPT2m forecasts confirm their findings, particularly over the Midwest and Great Lakes regions (Fig. 6b). Note that the 2-m moist bias is even worse for the NAM forecasts over this same region (Fig. 6a). However, the improvement in the NSSL 2-m dewpoint forecast bias over the NAM model over the Midwest and Great Lakes region is not evident in the bias time trends (Fig. 4b) because the regional domains were usually not located in this area (Fig. 1). Over the regional domains, which were usually centered in the central and southern Plains (Fig. 1), the DPT2m biases were relatively small for both the NSSL and NAM forecasts (Figs. 4a and 4b) and there was no substantial difference between the NSSL and NAM forecasts seen in the mean spatial bias (Figs. 6a and 6b).

The moist bias at 2 m for the NAM and NSSL forecasts is not as widespread at 850 hPa (Figs. 6c and 6d), but there are differences in the biases between the NAM and NSSL forecasts. The significant mean errors for DPT850 tend to be too moist for the NAM forecasts (except over western Texas), but the sign of the dewpoint bias for the NSSL model forecasts depends on location (Fig. 6d). The NSSL forecasts tend to be too moist over the northern High Plains and Midwest and too dry from Texas into the lower Ohio Valley. Thus, a dominant dry bias exists in the time trend at 850 hPa for the NSSL forecasts (Fig. 4d) because the regional domains were usually located over the central and southern Plains (Fig. 1) where the NSSL model tends to be too dry at 850 hPa (Fig. 6d). Note that the NAM DPT850 forecasts were nearly unbiased over the regional domains (Fig. 4d), indicating again that the NSSL model is producing a bias that is not evident in the NAM over the regional domains.

Output from the NAM model, two deterministic models (NSSL and NCAR) and the CAPS-ENSMEAN are used to illustrate an interesting regional bias in vertical wind shear common to most of the models (Fig. 7). Overall, the WSHR forecasts show little significant bias over much of the domain at 21 h, but a significant negative bias of 2-3 m s^{-1} is found over much of the central and southern High Plains. This creates the negative WSHR biases seen in the time trends averaged over the regional domains (Fig. 4f). Southeastern Colorado is a region with low WSHR forecasts in particular, in which the mean negative bias exceeds 4 m s^{-1} in the NAM (Fig. 7a) and CAPS-ENSMEAN (Fig. 7d). Included on Fig. 7 is the mean vector error in the 10-m winds at the locations with a statistically significant mean WSHR bias. The westerly direction of most of the error vectors illustrates a tendency of the models to veer the 10-m winds too much over this region, resulting in an overall decrease in the mean wind shear. Although the reasons for this tendency of the models to veer the winds over this region are not clear, the slight dry bias

in low levels, especially at 850 hPa (Fig. 6d), suggests that the models tend to push the dryline too far east. Indeed, this was a tendency noted by SE08 participants on several occasions.

b. RMSE ranks

The previous section illustrated that the CAMs tend to produce larger biases over much of the CONUS compared to the NAM. Likewise, there is no consistent improvement in the forecasts of the CAMs over those of NAM and GFS models in terms of the mean RMSE and bias averaged over the regional domains (Figs. 3 and 4). This is consistent with the results of Mass et al. (2002), who showed little to no consistent improvement in objective verification results for model fields at 4-km grid spacing compared to 12-km grid spacing over a much more limited region of the Pacific Northwest (although the fast LBC update in their two-way nesting procedure ensures more influence from the LBCs in their case).

The lack of a consistent improvement in the CAM forecasts versus the GFS, and the NAM forecasts especially, is shown further by viewing the frequency distribution of the relative RMSE ranks of the models for each day (Fig. 8). The models are ranked using the 24 days for which the model output was available for all of the models over the period examined. The 15-21 h period is examined to allow the next day's diurnal cycle to be represented in the statistics, while focusing on the time period that typically precedes convective initiation or represents the early evolution of the daily convective activity (Schwartz et al. 2009b; Clark et al. 2009). The NAM forecasts of DPT850 and CAPE over the 15-21 h forecast period are ranked in the top three almost 80% of the time (Figs. 8d and 8e). There does not appear to be any substantial improvement in the performance of the CAMs compared to the NAM and GFS, and in fact, the CAM forecasts tend to be slightly worse overall according to the mean RMSE ranks.

c. Similarity in large-scale boundaries and CAM errors away from the boundaries

Although the use of a mean RMSE over a regional domain cannot separate objectively the contribution of timing, location, and magnitude errors, inspection of the forecasts by the SE08 participants and by the authors revealed that the placement of mesoscale and synoptic scale features in the NAM forecasts was very similar to the location of the features in the CAMs, as noted similarly in Weisman et al. (2008). This is most apparent by focusing on the EMC model, which uses the same dynamic core (WRF-NMM; Janjíc et al. 2005) as the NAM. The EMC model used in this study is essentially a higher-resolution version of the NAM used without cumulus parameterization (there are other minor differences, however, which prevents a complete isolation of the effects of higher resolution).

A comparison of the NAM and EMC TMP2m forecasts on a day in which the NAM forecast of the pre-convective environment showed one of the largest improvements in RMSE over the CAMs is shown in Fig. 9. The only significant large-scale airmass boundary in the region was a cold front positioned from the Nebraska/Iowa border into the Oklahoma panhandle (Figs. 9a and 9d). Both the NAM and EMC forecasted the cold front too far to the north and west. The difference fields of both models compared to the associated RUC analysis (Figs. 9c and 9f) show that the incorrect placement of the front resulted in positive temperature errors over Kansas and central Nebraska and negative temperature errors to the east. Although there are differences in the placement of the cold front between the NAM and EMC forecasts, which results in different error magnitudes in this region (Figs. 9c and 9f), the models are clearly more like each other than the associated RUC analysis. Based on the author's experience, operational forecasters at the SPC have recognized this similarity the last few years.

The example in Fig. 9 illustrates the typical similarities between the NAM and CAM forecasts. Although not *always* the case, the errors in the placement of airmass boundaries in the CAMs often appear to be driven largely by the LBCs, resulting in similar error patterns in the placement of mesoscale and synoptic scale features. However, an important point to make is that *significant differences in the errors between the NAM and CAMs also are present away from the airmass boundaries*. For example, the negative temperature errors found in the NAM forecast in eastern Oklahoma and Missouri (Fig. 9c) are exacerbated in the EMC forecast (Fig. 9f). An inspection of model soundings by the SE08 participants revealed the most noticeable difference between the observations and the model forecasts from eastern Oklahoma to central and northern Missouri is the presence of widespread low-level clouds in both the NAM and EMC forecasts. Morning surface observations and subsequent visible satellite observations show that much of this region was indeed covered by low-level clouds during the morning hours, but this cloud cover had thinned and become much more isolated by the afternoon hours (not shown). This analysis suggests that the large negative temperature errors in this region are likely due to the inability of the model physical parameterizations to properly predict the evolution of the low-level clouds and their effect on the temperature and moisture profiles in the PBL during the day, which was a common analysis made by the SE08 participants.

It is recognized that a comparison of the RMSE time trends cannot separate the contributions of timing, location, and spatial-structure errors. Although the use of a verification technique that could separate these sources of error (e.g., Davis et al. 2006) is beyond the scope of this paper, the influence of location errors can be removed through an inspection of the distributions of forecast and observed variables over the regional domains. An example is shown in Fig. 10, which reveals that the EMC forecasts also have a cold bias overall, as found for the

NSSL (Figs. 4a and 5b) and NCAR models (Fig. 4a). Furthermore, Fig. 10 adds that the EMC forecasts are too cold more frequently than the NAM forecasts. This is revealed by the fact that the distribution of EMC temperatures is shifted to the left toward colder values compared to both the NAM forecasts and the RUC analyses (Fig. 10). This conclusion that the forecast values tend to have a distribution displaced further away from the RUC analyses than the NAM forecasts is valid for many of the CAMs and the other variables (not shown). This is further evidence that the larger errors noted in the EMC forecasts overall and in the other CAMs (Fig. 3) have a significant contribution from errors present over much of the domain and aren't simply due to location errors of the airmass boundaries.

The modulating effect of errors that occur away from the airmass boundaries is also suggested for many variables in Fig. 3 (not just TMP2m), in which the changes in the mean RMSE for the NAM forecasts with time (black dashed line) follow the changes in the forecasts from the deterministic CAM models that use the NAM for the ICs and LBCs. Again, this comparison is best made between the EMC and NAM models, since these models share the same WRF dynamic core and model physics, aside from the lack of cumulus parameterization in the EMC model (compare the purple line to the black dashed lines in Fig. 3).

To summarize this last point, the placement of airmass boundaries is often similar between the CAM forecasts and the NAM forecasts that provide the ICs and LBCs. Significant location errors can be the result, as exemplified in Figs. 9c and 9f. However, as also exemplified in Fig. 9, a portion of the errors in the pre-convective and near-storm environments in the CAMs appears to emanate from inaccuracies in the forecasts away from these boundaries. The NAM/EMC comparisons (Figs. 9 and 10) and the trends and rankings of the RMSEs (Figs. 3 and 8) suggest that these errors away from the airmass boundaries tend to be worse in the CAMs as

compared to the NAM forecasts. Although the specific reasons for this result are not clear, this suggests that the configuration of the physics used at coarser resolutions (and the attendant simplifications/assumptions) are not sufficient at higher resolutions. This insufficiency may be leading to degradation of the CAM forecasts of the environment, which is discussed further in section 6.

d. Comparison of CAPS-ENSMEAN to other models

The final goal of this study is to examine forecasts of the environment provided by the CAPS ensemble compared to the other model forecasts. As found with mesoscale ensembles (Wandishin et al. 2001; Homar et al. 2006; Jones et al. 2007), the present results show a substantial improvement in the model forecasts of the environment with the convection-allowing ensemble compared to the deterministic convection-allowing model forecasts. It is clear from Fig. 3 that the CAPS-ENSMEAN forecasts almost always improved upon the CAPS-CN forecast (examination of daily error comparisons, not shown, confirms this result), and from Fig. 8 that the CAPS-ENSMEAN forecasts tend to be ranked higher than the other models for all of the variables, except for CAPE[4]. It is intriguing that for most of the variables, the CAPS-ENSMEAN forecasts are ranked in the top three on at least 70% of the days. In fact, the forecasts of WSHR are ranked first almost 80% of the time and are ranked in the top three on all 24 of the days (Fig. 8f).

To ensure that these improvements were not simply the result of the smoothing inherent in producing the mean fields, the relative ranks shown in Fig. 8 are calculated again, but only for the 15- and 18-h forecasts (valid at 1500 and 1800 UTC) over the 16 days with a "clean-slate"

[4] The CAPS-CN forecasts of CAPE are often ranked the worst among all the models, for reasons that are not clear, but that the CAPS-ENSMEAN forecast still almost always is ranked higher than the CAPS-CN forecasts, even if the forecasts are not accurate compared to the other models.

pre-convective environment, which are defined to be days with little to no deep convection within the regional domain from 0900 UTC through 1900 UTC. This prevents any convective feedback from unfairly handicapping the deterministic CAMs in a comparison to the ensemble mean fields. It is seen that the CAPS-ENSMEAN forecasts still improve upon the CAPS-CN forecasts for all variables except CAPE when only the undisturbed pre-convective environments are examined (Fig. 11). For the WSHR forecasts, although the percentage of number-1 rankings for the CAPS-ENSMEAN declines (Figs. 8f and 11f), it still ranks first on almost 60% of the days (10 out of 16) and the percentage of rankings in the top two remains nearly the same for all days.

The above analysis gives further support that the filter produces an equitable quantitative comparison of the environments and damps the convective-scale features, since the long-term error statistics on all days and on those days with little to no convection is very similar. In addition, it strengthens the conclusions that the CAM forecasts are typically worse overall than the NAM and GFS forecasts for environmental parameters; yet, the CAPS-ENSMEAN forecasts are almost always better than the control and are frequently better than the other models, regardless of their native grid lengths and effective resolution. Implications for this result are discussed in section 6.

6. Summary and discussion

This study examines the quality of the pre-convective and near-storm environment forecasts from the convection-allowing WRF models (CAMs) run for the 2008 NOAA/HWT Spring Experiment (SE08). The goal of this paper is to present the typical error characteristics of CAMs in severe weather-relevant sub domains and to compare them to the lower resolution

models used frequently in operations and to the mean of an ensemble of CAMs. Motivation came from recent studies (Weisman et al. 2008; Weiss et al. 2008; Schwartz et al. 2009a) and experiences during the SE08 that show a strong correspondence of CAM model forecasts of convection to the precipitation forecasts of the lower-resolution models that provide the initial conditions (ICs) and the lateral boundary conditions (LBCs). Indeed, SE08 participants noted errors in the pre-convective environment that had a large influence on the timing and location of convection for 18-30 h forecasts on several days. However, it was not clear how the forecasts of the environment with increased resolution compared overall to the forecasts of the environment from the lower resolution model output. Understanding the nature of the errors in the prediction of the environments by the CAMs is important to the continued development of the models at such high resolution and in the use of CAM-based systems at operational forecasting centers.

Although the use of summary statistics in this study (mean RMSE and bias) cannot separate the contribution of timing, location, and magnitude errors objectively, inspection of the forecasts reveals that the placement of larger-scale airmass boundaries in the NAM forecasts was often very similar to the location of the same boundaries represented in the CAM models, which resulted in similar timing and placement of convective features for next-day convective forecasts (18-30 h). This correspondence contributed to similar temporal trends in the mean RMSE values among several variables examined. However, despite these similarities in the mean error patterns in the CAM and NAM forecasts resulting from location errors, substantial errors were also found away from the airmass boundaries that were often exacerbated in the CAMs. For example, the tendency for the CAMs to display a systematic cold bias was examined in detail.

Although this study does not isolate specific reasons for these error characteristics, we offer some hypotheses for some of the more pertinent results. Although the influence of the

initialization and the strategy for the lateral boundary conditions cannot be ruled out (Warner et al. 1997), experience with a subjective evaluation of CAM forecasts during NOAA/HWT Spring Experiments (Kain et al. 2005; Kain et al. 2008), and the results presented in this study showing the model biases in the CAMs (Figs. 4-7 and Fig. 10), suggests that these errors away from airmass boundaries are linked to physical parameterizations used at model grid lengths that allow convection, especially parameterizations for boundary layer processes and shallow (non-precipitating) clouds at or near the top of the boundary layer. All of the fields examined in this study are strongly modulated by these two parameterizations because each field is sensitive to the growth and decay of the planetary boundary layer (PBL), in addition to other parameterized processes such as exchanges with the underlying surface below and the free atmosphere above and the degree of cloud-cover. The MYJ PBL scheme, and the methods for how it interacts with other parameterized processes, was used in all of the deterministic forecasts examined here, yet this scheme was developed and calibrated using grid-spacing of ~O(10 km). While it may be one of the better performing PBL schemes tested in the WRF model, the use of the MYJ scheme, or any PBL scheme, at ~4-km grid spacing could be impinging upon a "grey area" in grid spacing, in which resolved large eddies or horizontal convective rolls begins to blend with the parameterized mixing from the PBL scheme (Stensrud 2007). This PBL grey area could be analogous to the use of convective parameterization with grid lengths ~O(10 km), in which the parameterized convective processes mixes with grid-scale convection (Molinari and Dudek 1992). The fact that the model fields that are strongly influenced by the PBL showed relatively slow error growth with time and a strong diurnal trend in the objective statistics (Fig. 3) suggests further a dominant role of the surface and PBL physics at these resolutions compared to other

sources of model error (Mass et al. 2002), such as boundary conditions and error growth in the background larger-scale flows.

Another contributor to the relatively high amplitude CAM errors may be the absence of a shallow-convection (SC) parameterization in the CAM configurations. The BMJ (Betts-Miller-Janjić) SC parameterization (Janjić 1994) has been documented to produce unrealistic vertical structures at times (Baldwin et al. 2002), but in the NAM it provides a critically important transport mechanism between the MYJ PBL and the free atmosphere, modulating boundary layer growth processes and the exchange of mass between the PBL and the atmosphere just above it. There is no parameterization for this process in the CAMs.

Although the CAMs appear to suffer from relatively high amplitude errors in predictions of the convective environment, we stress that these errors do not prevent the CAMs from outperforming the NAM in terms of QPF from both deterministic (Schwartz et al. 2009a) and ensemble perspectives (Clark et al. 2009). It seems likely that this QPF advantage could be further enhanced if the WRF model's physical parameterizations could be refined for higher resolution applications. Again, this study does not isolate the specific causes of the error characteristics of the CAMs, including the reasons offered above. However, it is hoped that this study provides an impetus for further research to identify the true cause for the poorer environmental forecasts in absence of significant convective activity so that forecasts of both the environment and convection, and the convective feedback to the environment, can be improved.

Finally, it is intriguing that a CAM-based ensemble mean provides better forecasts of the pre-convective and near-storm environment than the parent NAM, even though the high-resolution ensemble is anchored by the CAPS-CN configuration, which does not predict the environment better than the NAM. Since the NAM appears to provide a deterministic

framework for predicting the mesoscale environment as good or a better than the CAMs, it seems reasonable to ask, could a mesoscale ensemble provide even better forecasts of the environment than a CAM ensemble? Furthermore, can we then drive individual CAMs with members from the mesoscale ensemble to derive an even better deterministic CAM forecast? If these suppositions are true, a potentially useful strategy, given current capabilities of operational models, is to employ a combination of mesoscale and convection-allowing configurations to provide guidance for severe weather forecasters. Ensembles based on mesoscale models may provide optimal forecasts for the background convective environment and such ensembles could then be used to launch CAMs in either deterministic or ensemble frameworks for explicit predictions of convective storms. This strategy may be prudent currently for a number of reasons, not the least of which is cost: computational expenses for a 4-km grid are at least 27 times higher than those for the same domain with 12-km grid spacing. This underscores the need to develop ensemble systems that are appropriate for explicit predictions of convection. Recent studies have begun to examine the creation of CAM forecasts driven by an ensemble of mesoscale forecasts (Kong et al. 2006, Dowell and Stensrud 2008) and research is ongoing at the NSSL and SPC to configure such a system for testing in upcoming NOAA/HWT Spring Experiments (Stensrud et al. 2008).

Acknowledgments

Jay Liang, Gregg Grosshans, and Joe Byerly of the SPC make the daily operations of the NOAA/HWT Spring Experiment and dataflow possible. The creative web-based and diagnostic applications of Greg Carbin, John Hart, David Bright, and Jason Levit of the SPC makes it possible to evaluate a wide array of model output in an efficient manner. Linda Crank provides exceptional attention to detail in making arrangements for local and out of town participants.

At the NSSL, Brad Sagowitz, Brett Morrow, Jeff Horn, Steve Fletcher, James Murnan, Vicki Farmer, and Bob Staples provided invaluable technical support to ensure the smooth operation of the Spring Experiment. Linda Foster and Kelly Lynn contribute valuable help with travel arrangements and budgetary concerns.

CAPS scientists have been invaluable contributors to multiple HWT Spring Experiments. We appreciate the expert development and planning by Fanyou Kong and Keith Brewster, and the incredibly dedicated execution by Kevin Thomas. We also are grateful to Craig Schwartz, whose analysis of the data and associated manuscripts have made this endeavor easier.

NCAR scientists have been scientific leaders and valuable partners in numerous Spring Experiments and we are especially grateful to Wei Wang.

We have benefitted greatly from a long and productive working relationship with NCEP/EMC, especially Matt Pyle and Zavisa Janjić. Jun Du and Zoltan Toth have been valuable collaborators and supporters of ensemble-based scientific efforts and Geoff DiMego has provided support at all levels for this collaboration.

Finally, the paper was greatly improved by reviews by Matt Bunkers, the Science and Operations Officer at the National Weather Service office at Rapid City, SD, James Correia Jr., and one anonymous reviewer.

Appendix

A Gaussian distance-dependent weighting function was used to interpolate the model fields to the evaluation domains and was designed to remove convective-scale details completely while retaining meso-β and larger scale features; the filter removes over 99% of the amplitude of 30-km wavelength features while retaining over two-thirds of the amplitude of 100-km wavelength features (see Fig. A1 for the filter response function). The example shown in Fig. A2 illustrates that the filtered forecast fields from the CAMs contain spatial structures much closer to the structures seen in the 20-km RUC analyses that from the fields on the native CAM grids.

It should be noted that, although the filter is effective at removing convective-scale details up to seven times the grid length of the CAMs (Fig. A2), feedback of convective systems to the larger scale environment, is retained to some degree, especially for larger convective systems. However, the scale and amplitude of the convectively induced features that are retained in the filtered CAM fields are similar to the scale and amplitude of the convectively induced features that can be resolved by the 12-km NAM forecasts and the 20-km RUC analyses. The resulting comparison is then largely a comparison of the mesoscale and larger scale synoptic fields, with a small contribution from the feedback from convection. We believe the contribution of the convective feedback to the errors is relatively small, however. As shown in Section 5, a comparison of the long-term error statistics on all days and on those days with little to no convection in the real atmosphere or the model environments is very similar.

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.

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

Benjamin, S. G., D. Devenyi, S. S. Weygandt, K. J. Brundage, J. M. Brown, G. A. Grell, D. Kim, B. E. Schwartz, T. G. Smirnova, T. L. Smith, and G. S. Manikin, 2004b: An hourly assimilation-forecast cycle: The RUC. *Mon. Wea. Rev*., **132**, 495-518.

Benjamin, S.G., S.S. Weygandt, J.M. Brown, T.G. Smirnova, D. Devenyi, K. Brundage, G.A. Grell, S. Peckham, T. Schlatter, T.L. Smith, and G. Manikin, 2007: From the radar-enhanced RUC to the WRF-based Rapid Refresh. Preprints, *22nd Conf. Wea. Analysis Forecasting / 18th Conf. Num. Wea. Pred*., Park City, UT, Amer. Meteor. Soc., paper J3.4 [http://ams.confex.com/ams/pdfpapers/124827.pdf]

Bryan, G. H., J.C. Wyngaard, and J.M. Fritsch, 2003: Resolution requirements for the simulation of deep moist convection**.** *Mon. Wea. Rev.,* 131, 2394–2416.

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small near-convection-permitting and large convection-parameterizing ensembles. Accepted to *Wea. Forecasting*., January 2009.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev*., **134**, 1772-1784.

Done J., C. A. Davis, and M. L. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) model. *Atmos. Sci. Lett.*, **5**, 110–117, doi:10.1002/asl.72.

Dowell, D. C. and D. J. Stensrud, 2008: Ensemble forecasts of severe convective storms. Preprints, *24th Conf. on Severe Local Storms*, Amer. Meteor. Soc., Savannah, GA, paper 13A.5, [Available online at http://ams.confex.com/ams/pdfpapers/141628.pdf.]

Du, J., J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H. Chuang, 2006: New dimension of NCEP Short-Range Ensemble Forecasting (SREF) system: Inclusion of WRF members, *Preprint,WMO Expert Team Meeting on Ensemble Prediction System*, Exeter, UK, Feb. 6-10, 2006, 5 pp.

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.

Eckel, F.A., and C.F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350.

Elmore, K. L., M. E. Baldwin, and D. M. Schultz, 2006: Field Significance revisted: Spatial bias errors in forecasts as applied to the Eta model. *Mon. Wea. Rev*., **134**, 519-531.

Environmental Modeling Center, 2003: The GFS atmospheric model. NCEP Office Note 442, NCEP/NWS, 14 pp. [Available online at http://www.emc.ncep.noaa.gov/officenotes/newernotes/on442.pdf.]

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.

Gao, J.-D., M. Xue, K. Brewster, and K. K. Droegemeier, 2004: A three-dimensional variational data analysis method with recursive filter for Doppler radars. *J. Atmos. Ocean. Tech.*, **21,** 457-469.

Hall, W. D., R. M. Rasmussen, and G. Thompson, 2005: The new Thompson microphysical scheme in WRF. Preprints, 2005 WRF/MM5 User's Workshop, 27-30 June 2005, Boulder, CO. CD-ROM, paper 6.1.

Homar, V, D. J. Stensrud, J. J. Levit, and D. R. Bright, 2006: Value of human generated perturbations in short-range ensemble forecasts of severe weather. *Wea. Forecasting*, **21**, 347-363.

Hu, M., M. Xue, and K. Brewster, 2006: 3DVAR and cloud analysis with WSR-88D level-II data for the prediction of the Fort Worth, Texas, tornadic thunderstorms. Part I: Cloud analysis and its impact. *Mon. Wea. Rev*., **134**, 675-698.

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

Janjić, 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.

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

Janjić, Z.I., T. L. Black, M. E. Pyle, H.- Y. Chuang, E. Rogers, and G. J. DiMego, 2005: The NCEP WRF NMM core. Preprints, 2005 WRF/MM5 User's Workshop, 27-30 June 2005, Boulder, CO. CD-ROM, paper 2.9.

Jones, M. S., B. A. Colle, and J. S. Tongue, 2007: Evaluation of a mesoscale short-range

ensemble forecast system over the northeast United States. *Wea. Forecasting*, **22**, 36-55.

Kain, J. S., S. J. Weiss, M. E. Baldwin, G. W. Carbin J. J. Levit, D. R. Bright, and J. A. Hart, 2005: Evaluating high-resolution configurations of the WRF model that are used to forecast severe convective weather: The 2005 SPC/NSSL Spring Program. Preprints, *21st Conf. on Weather Analysis and Forecasting and 17th Conf. on Numerical Weather Prediction,* Washington, D.C., Amer. Meteor., Soc., 2A.5. [Available online at http://ams.confex.com/ams/pdfpapers/94843.pdf.]

Kain, J. S., S. J. Weiss, S. R. Dembeck, J. J. Levit, D. R. Bright, J. L. Case, M. C. Coniglio, A. R. Dean, R. Sobash, and C. S. Schwartz, 2008: Severe-weather forecast guidance from the first generation of large domain convection-allowing models: Challenges and opportunities. Preprints, *24th Conf. on Severe Local Storms*, Amer. Meteor. Soc., Savannah, GA, paper 12.1, [Available online at http://ams.confex.com/ams/pdfpapers/141723.pdf]

Kong F., K. K. Droegemeier, and N. L. Hickmon, 2006: Multiresolution ensemble forecasts of an observed tornadic thunderstorm system. Part I: Comparison of coarse- and fine-grid experiments. *Mon. Wea. Rev*., **134**, 807-833.

Kong, F., M. Xue, K. K. Droegemeier, D. Bright, M. C. Coniglio, K. W. Thomas, Y. Wang, D. Weber, J. S. Kain, S. J. Weiss, and J. Du, 2007: Preliminary analysis on the real-time storm-scale ensemble forecasts produced as a part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment. Preprints*, 22nd Conference On Weather Analysis and*

Forecasting/18th Conference on Numerical Weather Prediction, Park City, UT, Amer. Meteor. Soc., CD-ROM, 3B.2.

Kong, F., M. Xue, K.W. Thomas, K. K. Droegemeier, Y. Wang, K. Brewster, J. Gao, J. S. Kain, S. J. Weiss, D. R. Bright, M.C. Coniglio and J. Du, 2008: Real-time storm-scale ensemble forecast experiment: Analysis of 2008 Spring Experiment Data. Preprints, *24th Conf. on Severe Local Storms*, Amer. Meteor. Soc., Savannah, GA, paper 12.3, [Available online at http://ams.confex.com/ams/pdfpapers/141827.pdf]

Mass, C. F., D. Owens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts? *Bull. Amer. Meteor. Soc*., **83**, 407-430.

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. and M. Dudek, 1992: Parameterization of convective precipitation in mesoscale numerical models: A critical review. *Mon. Wea. Rev.*, 120, 326-344.

Powers, J. G., and J. B. Klemp, 2004: The advanced research WRF effort at NCAR. Preprints, *5th WRF/14th MM5 Users Workshop*, 22-25 June 2004, Boulder, CO, pp. 157-160.

Schwartz, C. S., J. S. Kain, S. J. Weiss, M. Xue, D. R. Bright, F. Kong, K. W. Thomas, J. J.

Levit and M. C. Coniglio, 2009a: Next-day convection-allowing WRF model guidance: A second look at 2 vs. 4 km grid spacing. Accepted for publication in *Mon. Wea. Rev.*.

Schwartz, C., J. S. Kain, S. J. Weiss, M. Xue, D. R. Bright, F. Kong, K. Thomas, J. J. Levit, M. C. Coniglio, and M.S. Wandishin, 2009b: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing guidance with small ensemble membership. Accepted for publication in *Wea. Forecasting.*

Skamarock, W. C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang and J. G. Powers, 2005: A description of the Advanced Research WRF Version 2, NCAR Tech Note, NCAR/TN–468+STR, 88 pp. [Available from UCAR Communications, P.O. Box 3000, Boulder, CO, 80307]. Also at: http://box.mmm.ucar.edu/wrf/users/docs/arw_v2.pdf)

Simmons, A. J., R. Mureau, and T. Petroliagis, 1995: Error growth and estimates of predictability from the ECMWF forecasting system. *Quart J. Roy Meteor. Soc*., **121**, 1739-1771.

Stensrud, D. J., 2007: Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models. Cambridge University Press, 459 pp.

Stensrud, D. J., N. Yussouf, D. C. Dowell, and M. C. Coniglio, 2008: Assimilating surface data into a mesoscale model ensemble: Cold pool analyses from Spring 2007. *Atmos. Research*, accepted.

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

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev*., **129**, 729-747.

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

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

--------, C. Davis, W. Wang, K.W. Manning, and J.B. Klemp, 2008: Experiences with 0-36 h explicit convective forecasts with the WRF-ARW model. *Wea. Forecasting*, **23**, 407-437.

Weiss, S. J., M. E. Pyle, Z. Janjić, D. R. Bright, J. S. Kain, and G. J. DiMego, 2008: The operational high resolution window WRF model runs at NCEP: Advantages of multiple model runs for severe convective weather forecasting. Preprints, *24th Conf. Several Local Storms*, Savannah, GA, Ameri. Meteor. Soc., Paper P10.8. [Available online at http://ams.confex.com/ams/pdfpapers/142192.pdf.]

Xue, M., F. Kong, K. W. Thomas, J. Gao, Y. Wang, K. Brewster, K. K. Droegemeier, J. Kain, S.

Weiss, D. Bright, M. Coniglio, and J. Du, 2008: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2008 Spring Experiment. Preprints, *24th Conf. Several Local Storms*, Savannah, GA, Ameri. Meteor. Soc., Paper 12.2. [Available online at http://ams.confex.com/ams/pdfpapers/142036.pdf.]

Figure Captions

Figure 1. The common domain that contains the filtered RUC analysis and model forecast fields. The 3-letter station identifiers indicate the center points for the regional domains used to make experimental forecasts during the SE08 and to evaluate the models in this study (see Fig. 2 for an example of the size of the regional domains). A total of 31 domains were used during the experiment (RSL and HLC were chosen twice).

Figure 2. Example of a regional domain selected by the SE08 participants for the experimental forecast. The size of this domain remained the same throughout the experiment and only varied by center point, which was Clinton, OK (CSM) on this day. The contours depict areas expected to receive severe weather from 2100 UTC 5 May 2008 to 0500 UTC 6 May 2008; the percentages are the chance of receiving severe weather within 25 miles of any point.

Figure 3. RMSEs from the model runs, averaged over the regional domains for all days versus forecast hour for the six model fields described in the text.

Figure 4. As in Fig. 3, except for the mean error (bias).

Figure 5. A comparison of the spatial distribution of the 24-h forecast mean temperature errors (C) (biases, or model minus analyses) between the NAM and NSSL models at 2 m and at 850 hPa. The areas filled in brown for the 850 hPa plots indicate a mask used for data below ground or data outside of the native grid for the NSSL plot and the light blue areas indicate a mask used for grid points over the Gulf of Mexico and the Atlantic Ocean. All

colored areas outside of the masks indicate significant errors at a 95% confidence level (see Elmore et al. 2006 for details on how the confidence levels are calculated).

Figure 6. As in Fig. 5, except for the mean 24-h forecast dewpoint errors.

Figure 7. As in Fig. 5, except for a comparison of the mean 21-h forecast 10-m to 500-hPa wind shear errors (m s^{-1}) for the NAM, NSSL, NCAR, and CAPS-ENSMEAN. The vectors indicate the mean 10-m wind vector errors and are shown only where the wind shear errors are statistically significant.

Figure 8. The relative ranks of the RMSE, ranked from lowest RMSE to highest RMSE, averaged over the 15 - 21 h forecast period and averaged over the regional domains for each model and for the six model fields described in the text. For example, the upper-left panel shows that the CAPS ensemble mean forecasts of 2-m temperature had the lowest RMSE on 35% of the days and the second lowest RMSE on 25% of the days. Only those days for which model output was available for all the models were used to calculate the rankings for each variable. Note that the TMP850 and DPT850 were not available for the EMC model and the CAPE was not available for the GFS model.

Figure 9. A comparison of the TMP2m RUC analysis at 2100 UTC 1 May 2008 (upper-most panels) and the NAM (middle-left panel) and EMC (middle-right panel) 21 h forecasts valid at the same time within the regional domains. The frontal positions are determined subjectively through inspection of the filtered temperature and wind fields. The difference in TMP2m between the NAM forecast and the RUC analysis is shown in the bottom-left panel

and the difference between the EMC forecast and the RUC analysis is shown in the bottom-right panel.

Figure 10. The distribution of 21-h forecasts and RUC analyses of TMP2m over the regional domains on 10 "clean-slate" convective days, in which there was little to no deep convection in the model forecasts or in the real atmosphere within the regional domains in the period leading up to the forecasts (see Section 5b). The lines connect the frequency values for each bin for the NAM (thick black line) and the EMC (thick grey line) forecasts and the RUC analyses (thin black line).

Figure 11. As in Fig. 8, except for the 16 "clean-slate" convective days, in which there was little to no deep convection in the model forecasts or in the real atmosphere within the regional domains in the period from 0900 UTC to 1900 UTC. The EMC and NCAR models are not shown because output was available on only 10 of the 16 clean-slate days for these models.

Figure A1. Response function of the Gaussian weighting function used to interpolate the model fields to the common domain.

Figure A2. The 20-km RUC analysis of TMP2m valid at 2100 UTC on 1 May 2008 and the corresponding 21-h forecasts from the filtered and unfiltered TMP2m fields from the EMC 4-km model on the regional domain.

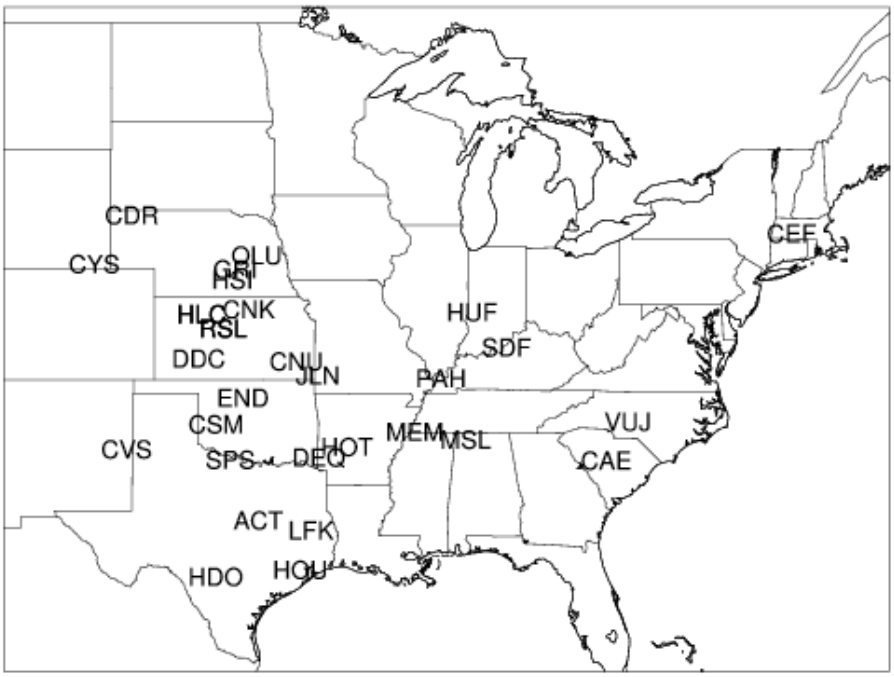

Figure 1. The common domain that contains the filtered RUC analysis and model forecast fields. The 3-letter station identifiers indicate the center points for the regional domains used to make experimental forecasts during the SE08 and to evaluate the models in this study (see Fig. 2 for an example of the size of the regional domains). A total of 31 domains were used during the experiment (RSL and HLC were chosen twice).

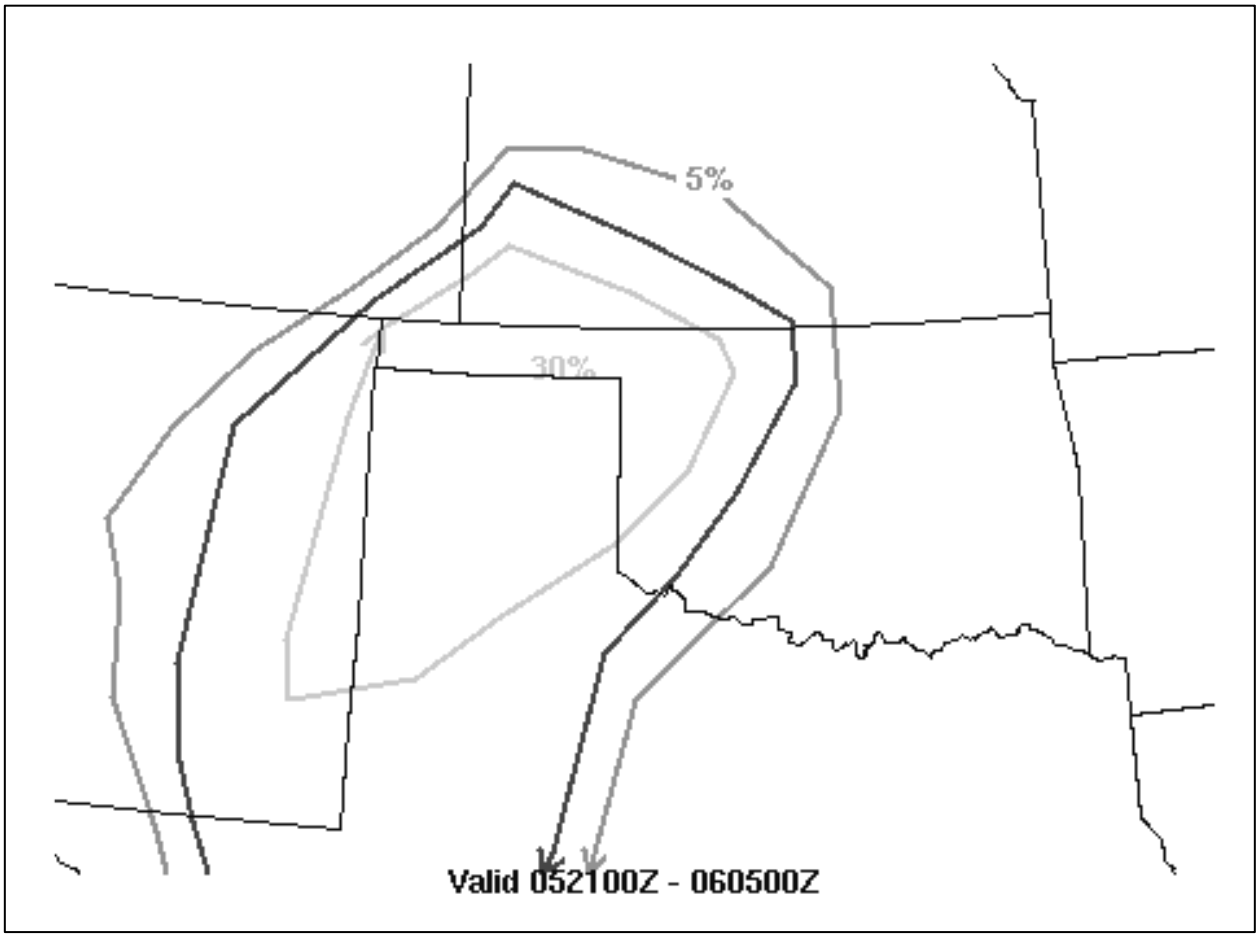

Figure 2. Example of a regional domain selected by the SE08 participants for the experimental forecast. The size of this domain remained the same throughout the experiment and only varied by center point, which was Clinton, OK (CSM) on this day. The contours depict areas expected to receive severe weather from 2100 UTC 5 May 2008 to 0500 UTC 6 May 2008; the percentages are the chance of receiving severe weather within 25 miles of any point.

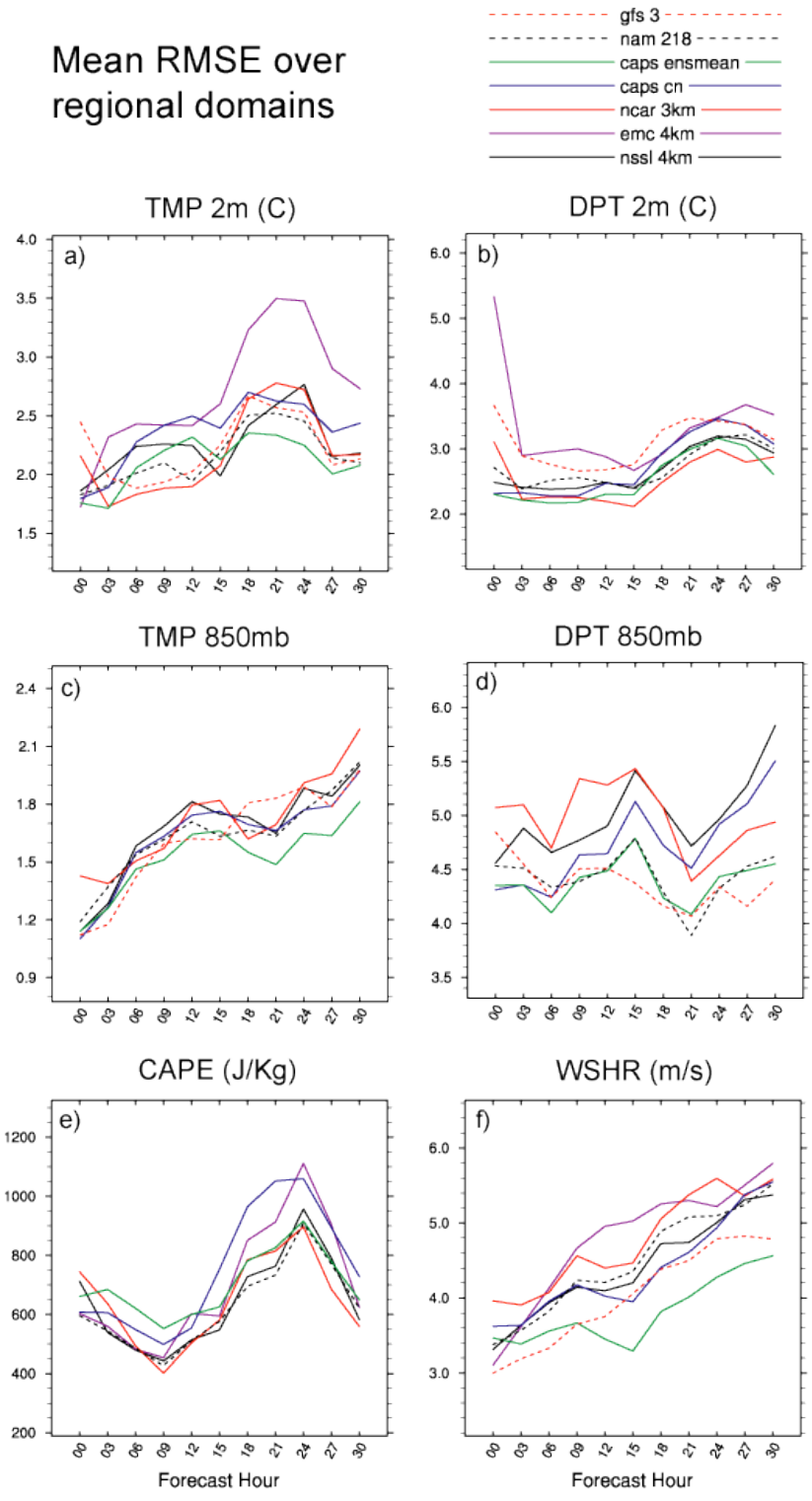

Figure 3. RMSEs from the model runs, averaged over the regional domains for all days versus forecast hour for the six model fields described in the text.

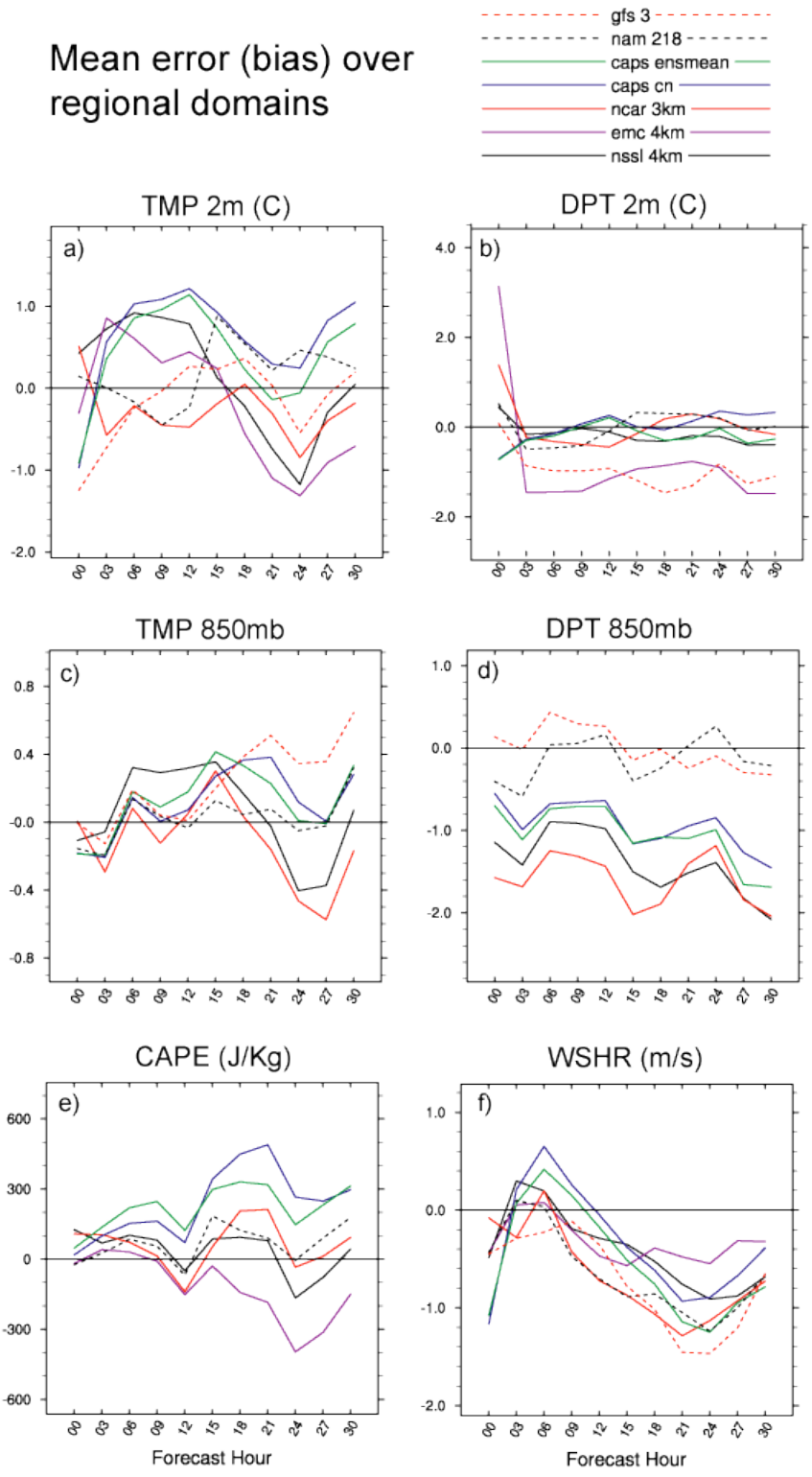

Figure 4. As in Fig. 3, except for the mean error (bias).

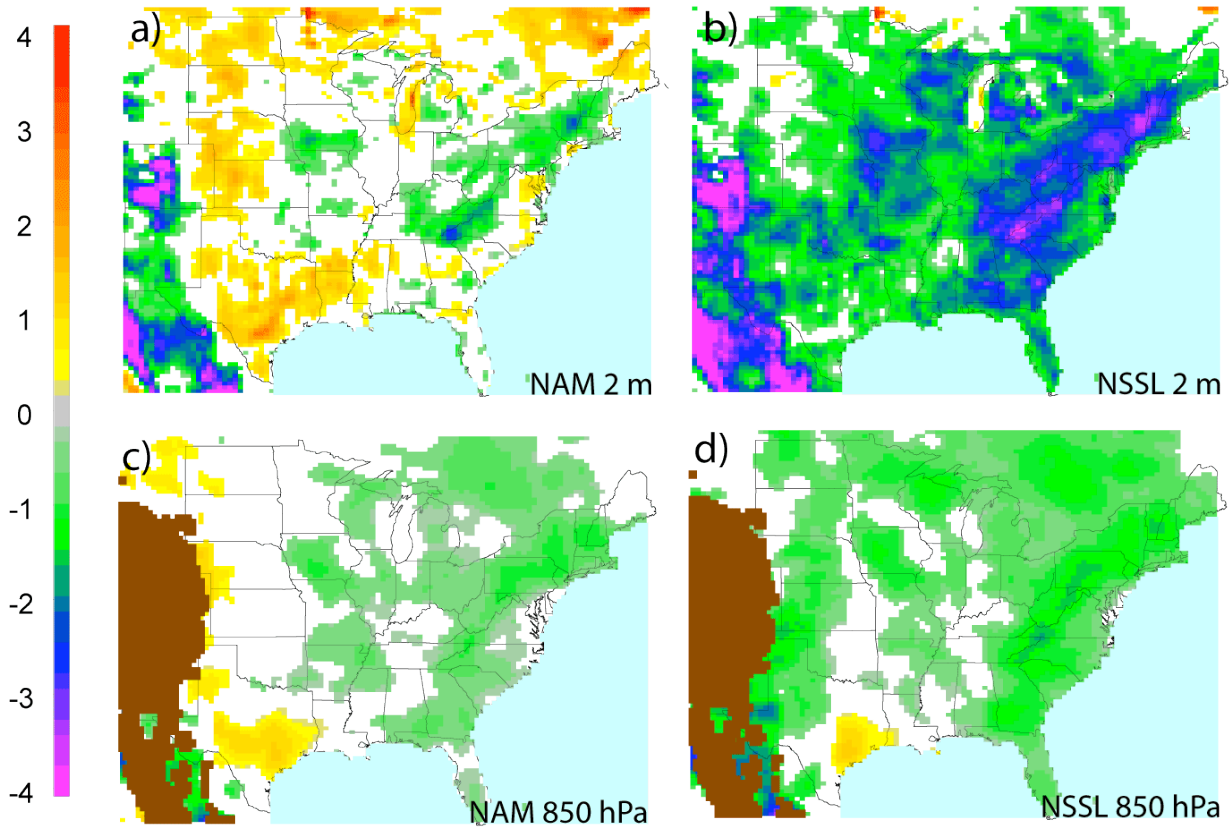

Figure 5. A comparison of the spatial distribution of the 24-h forecast mean temperature errors (C) (biases, or model minus analyses) between the NAM and NSSL models at 2 m and at 850 hPa. The areas filled in brown for the 850 hPa plots indicate a mask used for data below ground or data outside of the native grid for the NSSL plot and the light blue areas indicate a mask used for grid points over the Gulf of Mexico and the Atlantic Ocean. All colored areas outside of the masks indicate significant errors at a 95% confidence level (see Elmore et al. 2006 for details on how the confidence levels are calculated).

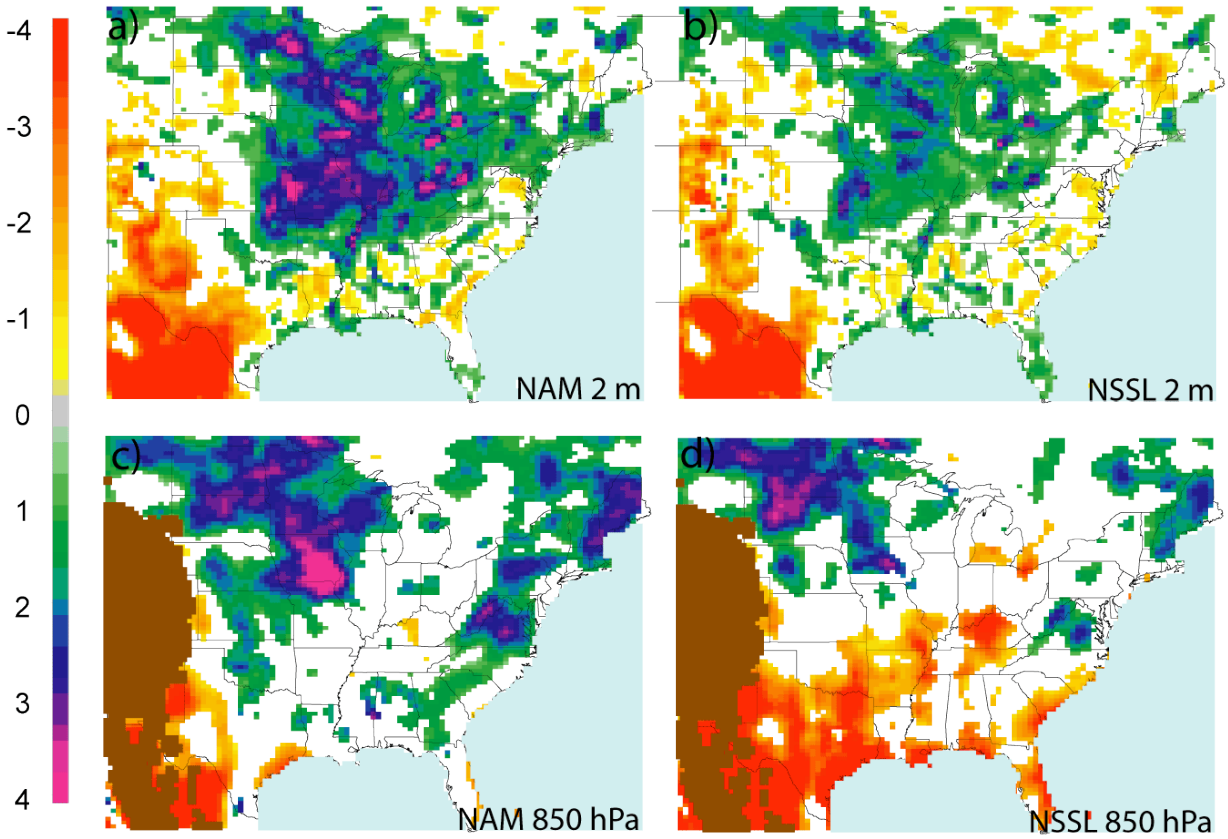

Figure 6. As in Fig. 5, except for the mean 24-h forecast dewpoint errors.

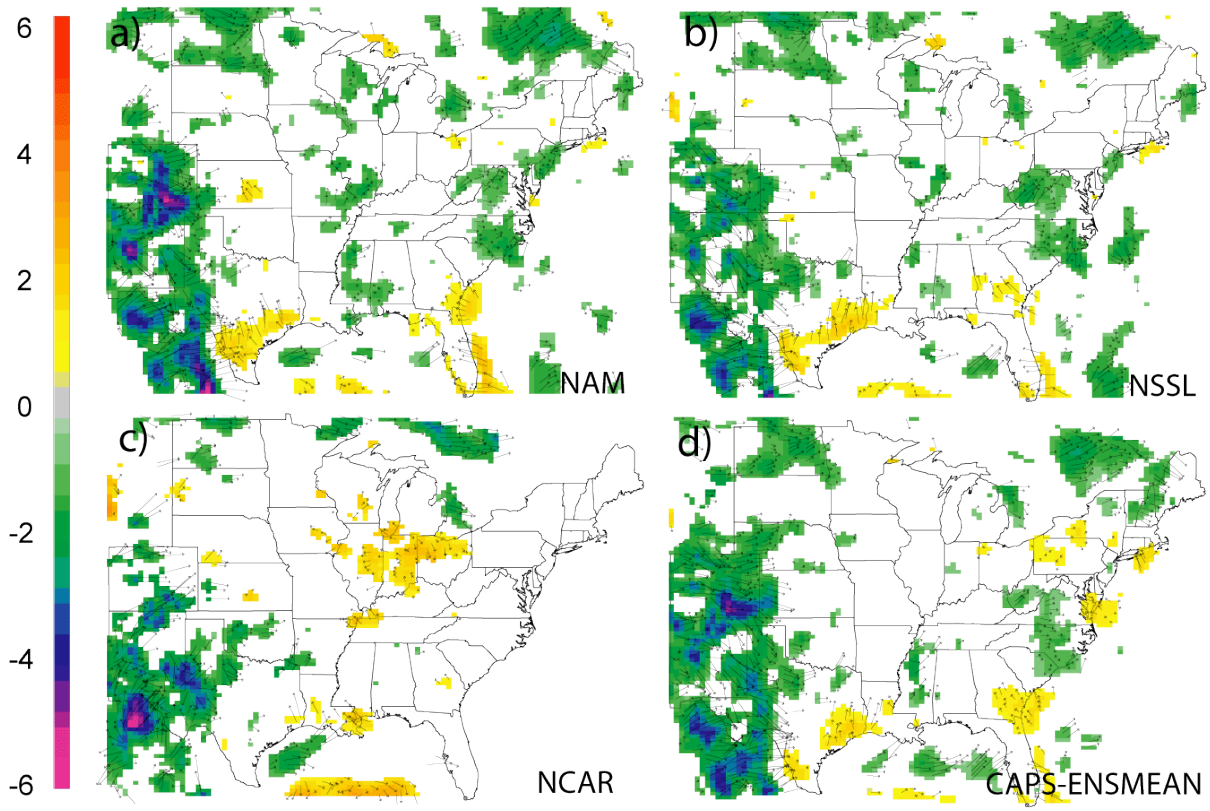

Figure 7. As in Fig. 5, except for a comparison of the mean 21-h forecast 10-m to 500-hPa wind shear errors (m s^{-1}) for the NAM, NSSL, NCAR, and CAPS-ENSMEAN. The vectors indicate the mean 10-m wind vector errors and are shown only where the wind shear errors are statistically significant.

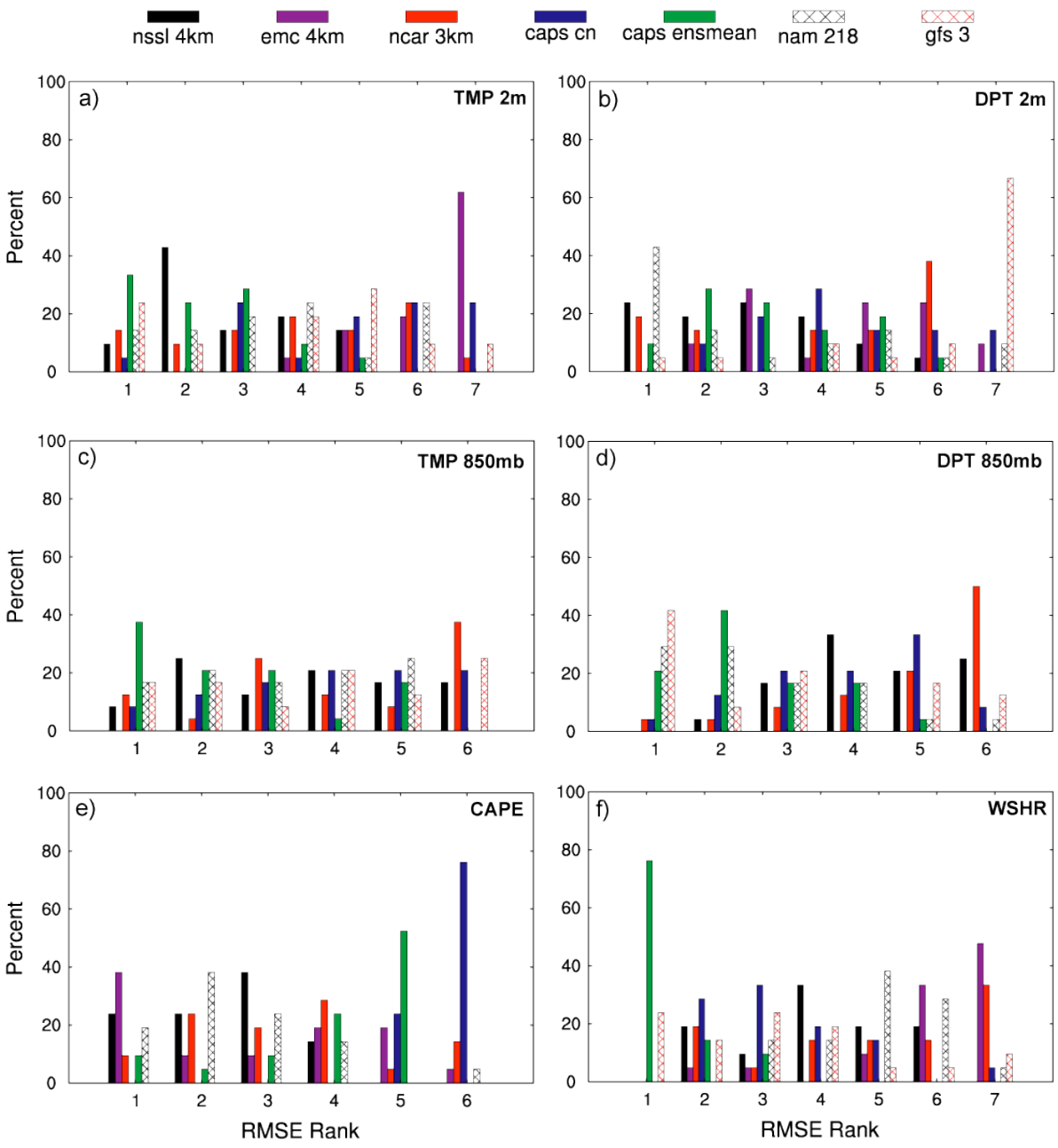

Figure 8. The relative ranks of the RMSE, ranked from lowest RMSE to highest RMSE, averaged over the 15 - 21 h forecast period and averaged over the regional domains for each model and for the six model fields described in the text. For example, the upper-left panel shows that the CAPS ensemble mean forecasts of 2-m temperature had the lowest RMSE on 35% of the days and the second lowest RMSE on 25% of the days. Only those days for which model output was available for all the models were used to calculate the rankings for each variable. Note that the TMP850 and DPT850 were not available for the EMC model and the CAPE was not available for the GFS model.

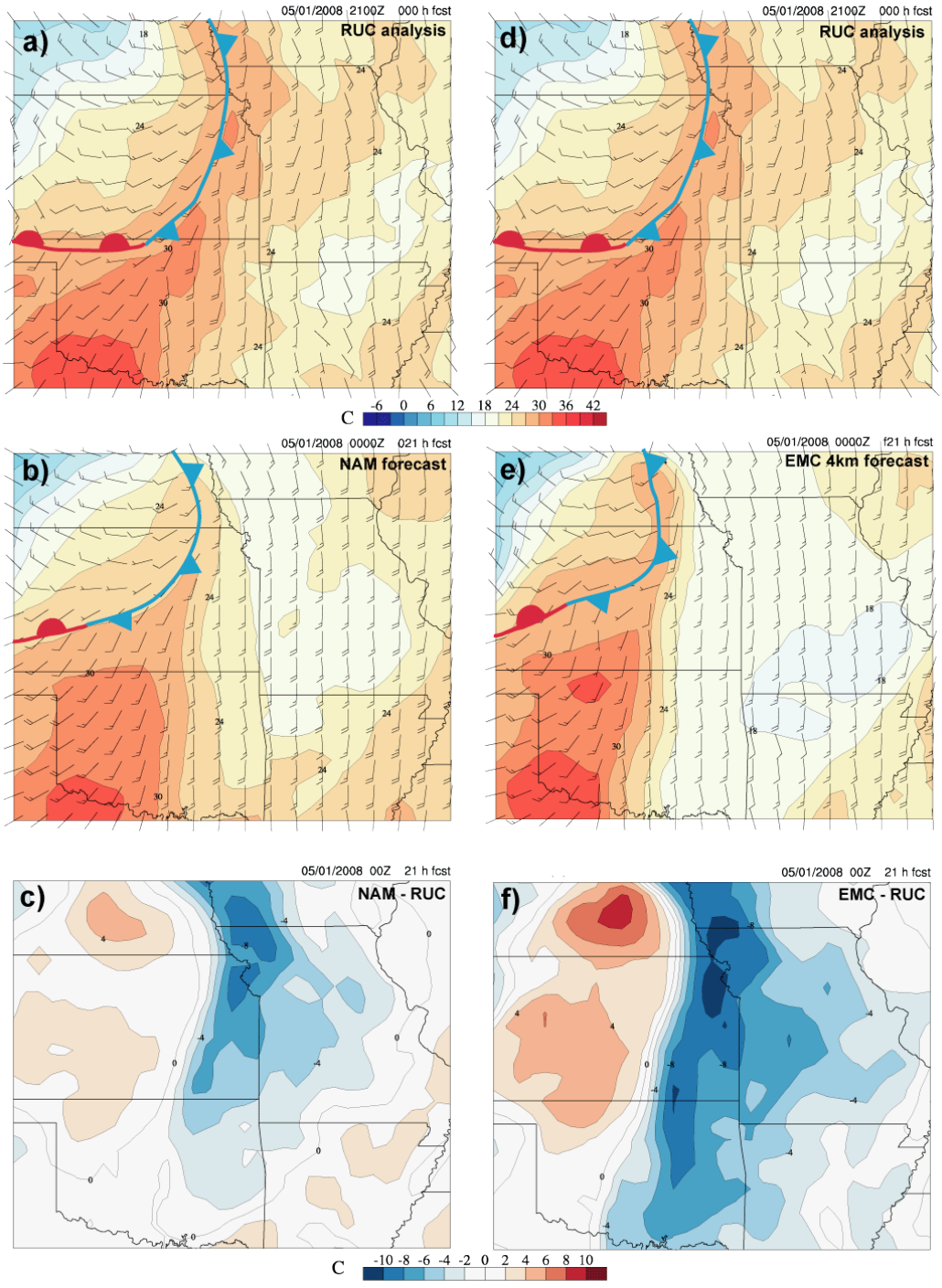

Figure 9. A comparison of the TMP2m RUC analysis at 2100 UTC 1 May 2008 (upper-most panels) and the NAM (middle-left panel) and EMC (middle-right panel) 21 h forecasts valid at the same time within the regional domains. The frontal positions are determined subjectively through inspection of the filtered temperature and wind fields. The difference in TMP2m between the NAM forecast and the RUC analysis is shown in the bottom-left panel and the difference between the EMC forecast and the RUC analysis is shown in the bottom-right panel.

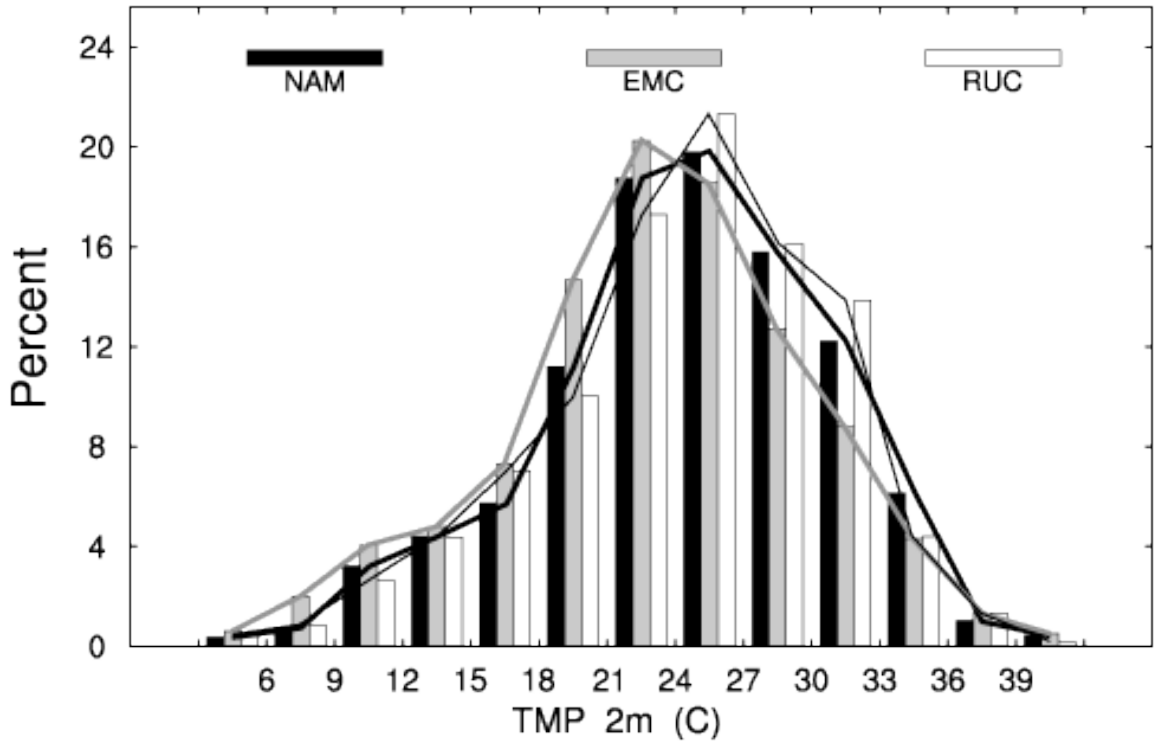

Figure 10. The distribution of 21-h forecasts and RUC analyses of TMP2m over the regional domains on 10 "clean-slate" convective days, in which there was little to no deep convection in the model forecasts or in the real atmosphere within the regional domains in the period leading up to the forecasts (see Section 5b). The lines connect the frequency values for each bin for the NAM (thick black line) and the EMC (thick grey line) forecasts and the RUC analyses (thin black line).

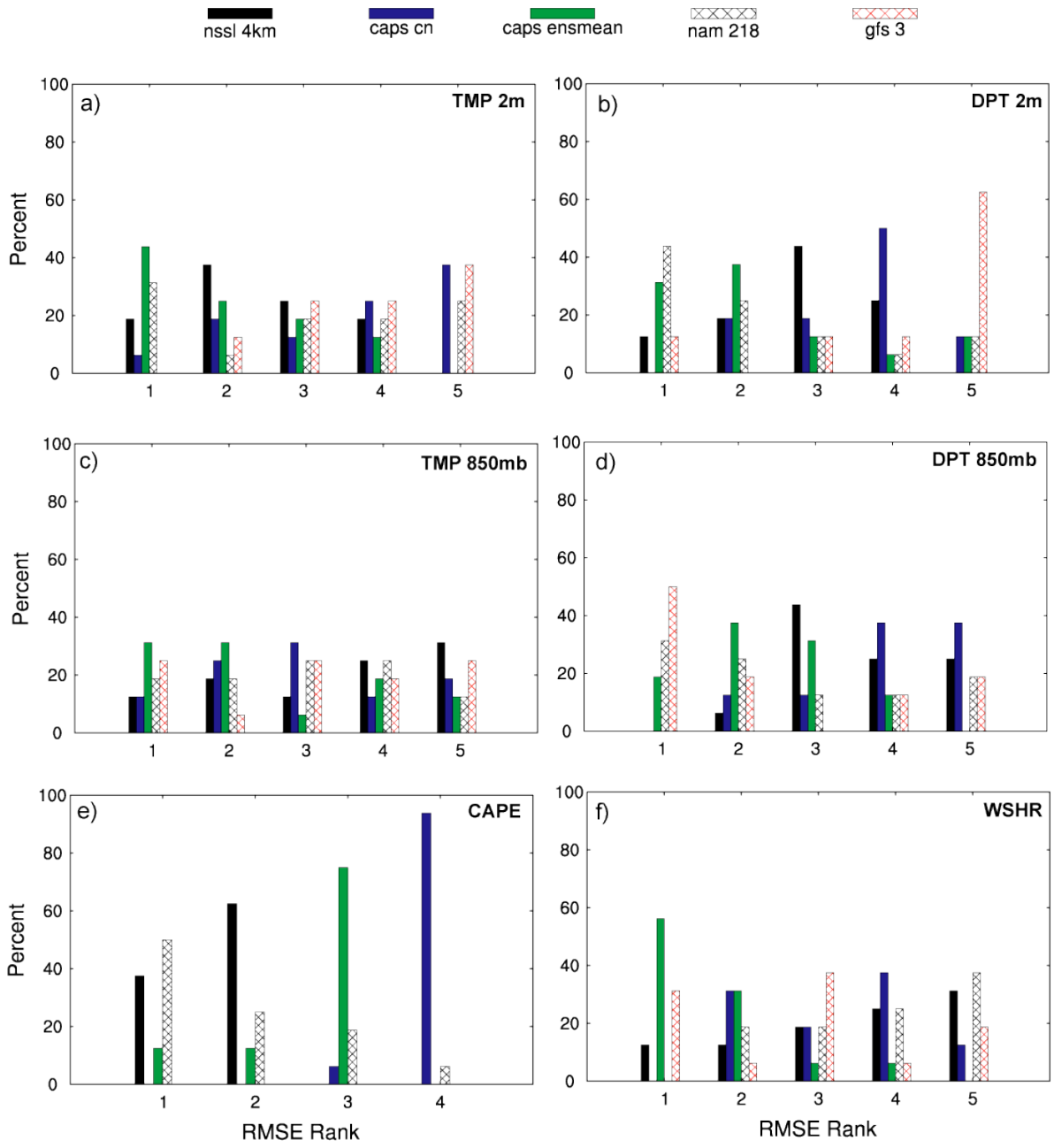

Figure 11. As in Fig. 8, except for the 16 "clean-slate" convective days, in which there was little to no deep convection in the model forecasts or in the real atmosphere within the regional domains in the period from 0900 UTC to 1900 UTC. The EMC and NCAR models are not shown because output was available on only 10 of the 16 clean-slate days for these models.

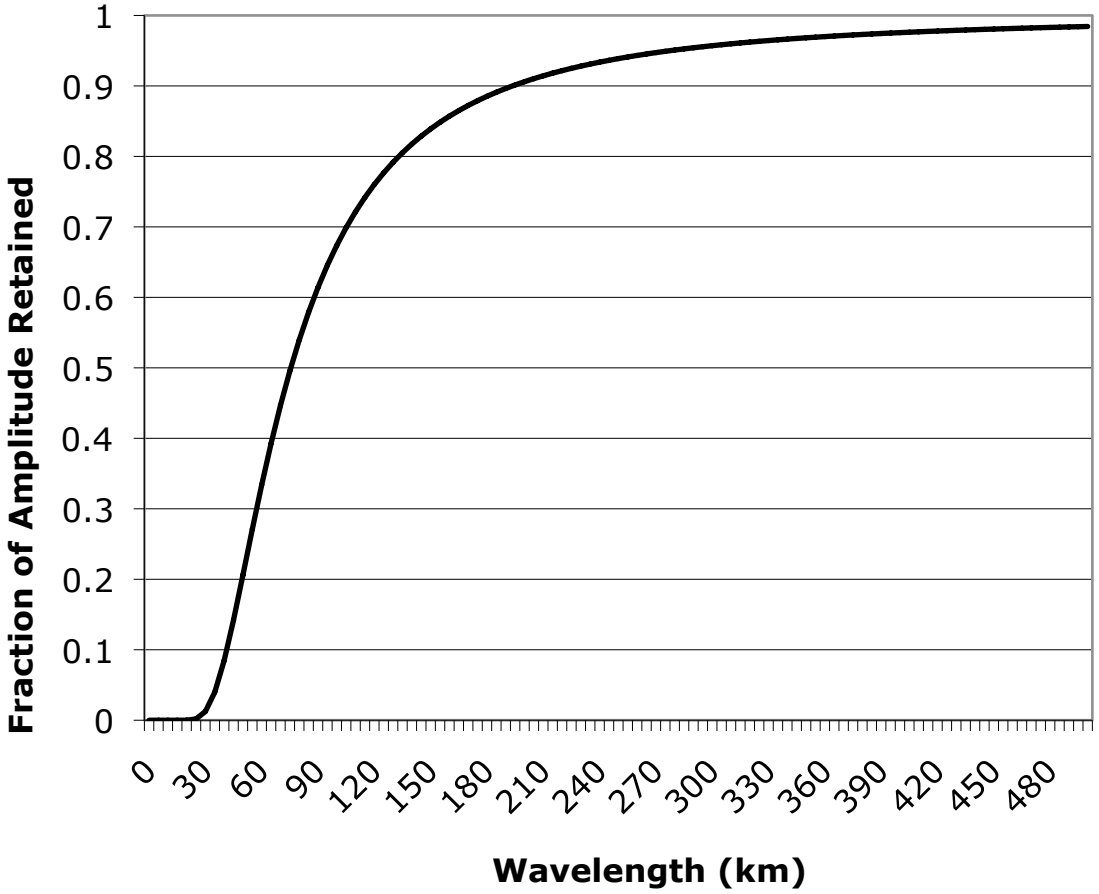

Figure A1. Response function of the Gaussian weighting function used to interpolate the model fields to the common domain.

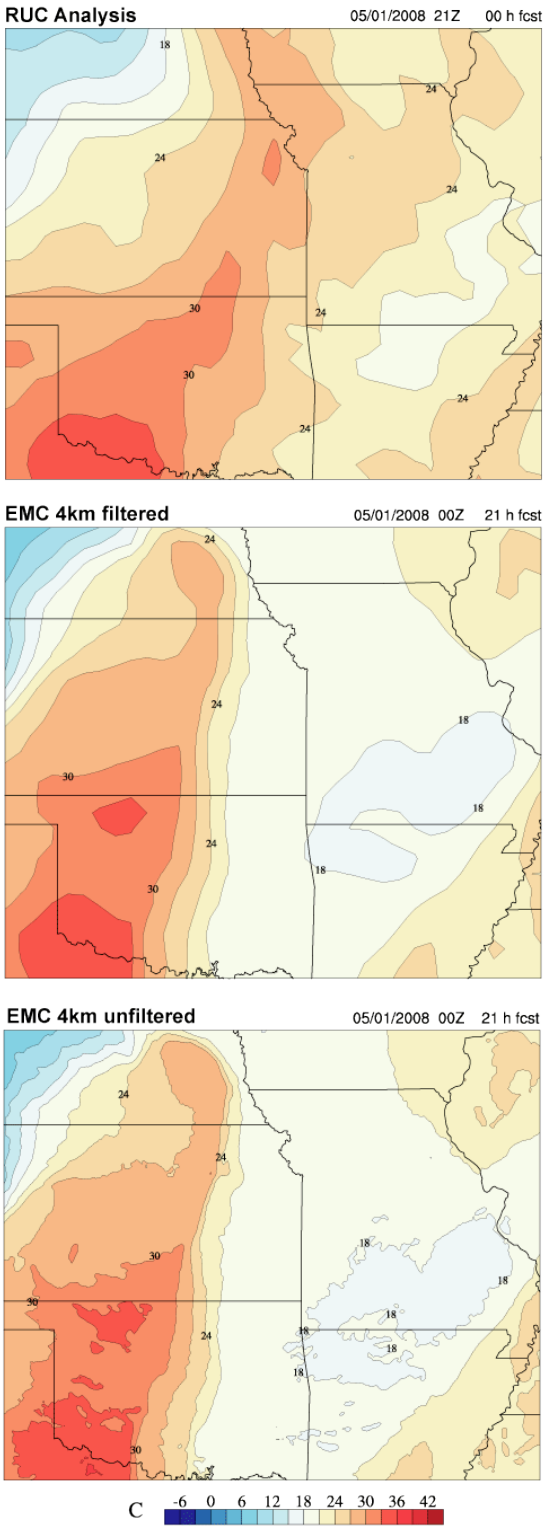

Figure A2. The 20-km RUC analysis of TMP2m valid at 2100 UTC on 1 May 2008 and the corresponding 21-h forecasts from the filtered and unfiltered TMP2m fields from the EMC 4-km model on the regional domain.

Table 1. Configurations for the four deterministic WRF models examined in this study. MYJ: Mellor-Yamada-Janjić turbulence parameterization scheme (Janjić 2001), NMM: Nonhydrostatic Mesoscale Model (Janjić et al. 2005), ARW: Advanced Research WRF (Powers and Klemp 2004), Ferrier: Ferrier et al. (2002); WSM6: WRF single moment, 6 class microphysics; Thompson: Hall et al. (2005), GFDL: Geophysical Fluid Dynamics Laboratory (Tuleya 1994); Dudhia: Dudhia (1989); RRTM: Rapid Radiative Transfer Model (Mlawer et al. 1997).

WRF name	**NSSL**	**EMC**	**NCAR**	**CAPS-CN**
Horizontal Grid Spacing (km)	4.0	4.0	3.0	4.0
Vertical Levels	37	43	40	51
Boundary Layer/Turbulence	MYJ	MYJ	MYJ	MYJ
Microphysics	WSM6	Ferrier	Thompson	Thompson
Radiation (shortwave/longwave)	Dudhia/ RRTM	GFDL/GFDL	Dudhia/ RRTM	Goddard/RRTM
Initial Conditions	40 km NAM	32 km NAM	parallel 9 km WRF/GFS	3DVAR analysis using 12 km NAM analysis background
Dynamic Core	ARW v2.2	NMM v2.2	ARW v.3	ARW v2.2

Table 2. Variations in initial condidtions, lateral boundary conditions, microphysics, shortwave radiation, and planetary boundary layer physics for the 2008 CAPS 4-km WRF-ARW ensemble forecasts. NAMa –12km NAM analysis; NAMf – 12km NAM forecast. All members used the RRTM longwave radiation scheme, and the Noah land-surface scheme. YSU refers to the Yonsei University PBL scheme. Additional details about the initial and boundary conditions and the perturbations can be found at URL http://www.emc.ncep.noaa.gov/mmb/SREF/SREF.html and in Xue et al. (2008).

Member	IC	LBC	Radar	Microphysics	Shortwave Radiation	PBL Scheme
CN	ARPS 3DVAR Analysis	00Z NAMf	yes	Thompson	Goddard	MYJ
C0	00Z NAMa	00Z NAMf	no	Thompson	Goddard	MYJ
N1	CN – arw_pert	21Z SREF arw_n1	yes	Ferrier	Goddard	YSU
P1	CN + arw_pert	21Z SREF arw_p1	yes	WSM6	Dudhia	MYJ
N2	CN – nmm_pert	21Z SREF nmm_n1	yes	Thompson	Goddard	MYJ
P2	CN + nmm_pert	21Z SREF nmm_p1	yes	WSM6	Dudhia	YSU
N3	CN – etaKF_pert	21Z SREF etaKF_n1	yes	Thompson	Dudhia	YSU
P3	CN + etaKF_pert	21Z SREF etaKF_n1	yes	Ferrier	Dudhia	MYJ
N3	CN – etaBMJ_pert	21Z SREF etaBMJ_n1	yes	WSM6	Goddard	MYJ
P4	CN + etaBMJ_pert	21Z SREF etaBMJ_n1	yes	Thompson	Goddard	YSU