12.4 EVALUATION OF WRF MODEL OUTPUT FOR SEVERE-WEATHER FORECASTING FROM THE 2008 NOAA HAZARDOUS WEATHER TESTBED SPRING EXPERIMENT

Michael C. Coniglio[1*], Kimberly L. Elmore[1], John S. Kain[1], Steven J. Weiss[2], David R. Bright[2], Jason J. Levit[2], Gregory W. Carbin[2], Kevin Thomas[3], Fanyou Kong[3], Ming Xue[3], Morris Weisman[4], and Matt Pyle[5]

1. NOAA/OAR/National Severe Storms Laboratory, Norman OK
2. NOAA/NCEP/Storm Prediction Center, Norman OK
3. Center for the Analysis and Prediction of Storms, Univ. of Okla., Norman, OK
4. National Center for Atmospheric Research, Boulder, CO
5. NOAA/Environmental Modeling Center, Silver Spring, MD

1. INTRODUCTION

The NOAA Hazardous Weather Testbed (HWT) conducted the 2008 Spring Experiment (SE2008) over the seven-week period during the peak severe convective season, from mid April through early June. As in past Spring Experiments, a vital component to its success was the active participation by forecasters, model developers, and many others who have a passion for operationally relevant meteorological challenges (see Kain et al. 2008 for a detailed description of the SE2008). As in recent years, the primary focus in 2008 was on the examination of convection allowing (Δ = 2-4 km) configurations of the WRF model (referred to as CAMs hereafter) in a simulated severe-weather-forecasting experiment. These simulations are evaluated on their ability to predict the location and timing of thunderstorm initiation and evolution, and offer useful information on thunderstorm morphology. In addition, the experiment continued to evaluate a real-time, large domain 10-member convection-allowing storm scale ensemble forecast system to gauge the potential benefits of uncertainty information at these model resolutions.

A new endeavor for the SE2008 was a more detailed examination of the relationship between model forecasts of convective storms and model predictions of the environment, which is the focus of this paper. A daily task of SE2008 participants was to examine model predictions of the environment compared to the verifying analyses from the Rapid Update Cycle (RUC). This endeavor is driven by recent subjective impressions that significant errors in 18 – 36 h forecasts of explicit convection may be controlled frequently by larger-than-expected errors in the background fields that provide the initial and boundary conditions (Kain et al. 2005, Weisman et al. 2008).

This paper presents results from an ongoing quantitative evaluation of the models run for the SE2008 and addresses the following questions:

1. What are the model biases and how quickly do errors in the environment grow in the CAMs and in the lower resolution models?

2. How do forecasts of the environment from the CAMs compare to similar forecasts of the environment from lower resolution models?

3. Does the ensemble of CAMs provide forecasts of the environment that improve upon forecasts of the environment from the deterministic CAMs?

4. Is there a relationship between the subjective quality of forecasts of the environment from the CAMs and the subjective quality of the forecasts of explicit convection from the CAMs?

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 WRF modeling systems. Furthermore, it may be able to help operational forecasters assess how much confidence to have in model guidance in specific situations.

2. SE2008 MODELING SYSTEMS

a. Deterministic WRF models

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

The model evaluation focuses on output from various configurations of the WRF model provided for the SE2008 by The Center for Analysis and Prediction of Storms (CAPS), the Environmental Modeling Center (EMC), the National Center for Atmospheric Research (NCAR), and the National Severe Storms Laboratory (NSSL) (see Table 1 for a description of each model). The forecasts began at 0000 UTC and forecast output was provided out to at least 30h for each day of the experiment. Although the domains varied among the models, all of the models covered roughly the eastern 3/4^{ths} of the CONUS.

The CAPS, EMC, and NSSL models all used the operational NAM output in some form for the initial conditions (ICs) and lateral boundary conditions (LBCs), whereas the NCAR model used the WRF 3DVAR assimilation system to cycle continuously a relatively large domain forecast system with 9 km grid spacing. ICs for the domain with 3 km grid spacing are provided by a one way nesting within this domain every day at 0000 UTC. LBCs for the 3 km nest come from a parallel 9 km run initialized at 12 UTC, and LBCs for the 9 km run come from the operational GFS model. This system provided the first extensive real-time test of the WRF 3DVAR system.

b. WRF-ARW 10-member ensemble

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 9 out of 10 members, one "control" member (identified as CAPS-CN in Table 1 and "Cntl" in Table 2) and eight perturbed members. Mesoscale atmospheric perturbations were introduced in the initial and lateral-boundary conditions of these eight members by extracting four pairs of positive/negative perturbations from EMC's operational Short Range Ensemble Forecast (SREF) system and applying them separately to the eight members. Furthermore, radar data (reflectivity and radial velocity) were assimilated into all nine of these members, using the CAPS 3DVAR assimilation system (Hu et al. 2006) as a last step in the initialization process. The tenth member was configured identically to the control member, but it was not subjected to the final radar-data assimilation step (identified as C0 in Table 2).

3. MODEL EVALUATION METHODS

a. Subjective evaluation

During the SE2008, participants were asked to spend ~1 h each day analyzing guidance from daily 18-30 h WRF-NSSL and WRF-EMC forecasts and identifying differences with the verifying RUC analyses. Fields were displayed from model output that helped identify errors in boundary-layer thermodynamic structure, airmass boundaries, and sub-synoptic scale features in the free atmosphere. On some days, it was possible to identify errors in phase and amplitude within these fields that clearly had a negative impact on CAM forecasts of convective initiation and evolution. The errors were apparently inherited from initial conditions provided by the NAM, and on some days it was noted that NAM forecasts of convective precipitation showed biases similar to the CAMs. These subjective analyses were thoroughly documented and they provide a valuable reference for more detailed ongoing investigations and the more objective evaluation described next.

b. Objective evaluation

An objective comparison of model output was performed by interpolating select model forecast fields to a common evaluation grid that covers roughly the eastern 2/3^{rds} of the CONUS with a horizontal grid spacing of roughly (1/3) by (1/3) degree latitude and longitude (roughly 24-33 km horizontal resolution depending on latitude) (Fig. 1). The interpolation was designed to remove convective-scale details completely while retaining meso-β-scale and larger scale features; less than 1% of the amplitude of 30-km features is retained while over 2/3rds of the amplitude of 100-km features are retained (Fig. 2). This interpolation is done to allow for a more direct comparison of the mesoscale and larger scale features from the various models to the resolvable scales of the RUC analyses than if the native grids were used. Six fields were selected for the evaluation, including the 2-m temperature (TMP2m) and dewpoint (DPT2m), the 850-mb temperature (TMP850) and dewpoint (DPT850), the convective available

potential energy of a surface parcel (CAPE[1]), and the wind shear between 10 m and 500 mb (WSHR).

A daily task for the SE2008 participants was the development of an experimental severe-weather forecast within a domain selected by the participants in a region deemed likely to experience convective weather during the forecast period (see Fig. 3 for an example). The objective evaluation of the model output discussed herein is restricted roughly to the area encompassed by these regional forecast domains and to the 31 days on which an experimental forecast was issued (see Fig. 3 for the center points used for the forecast domains during the SE2008). The objective measures for each forecast field include the mean error (bias) and the root mean squared error (RMSE) between the interpolated model fields and the interpolated RUC analyses over all grid points within the regional domain over a specified period. Output from the four daily WRF model forecasts described in Table 1, output from the CAPS ensemble control member (CAPS-CN), and output calculated from the mean of the CAPS ensemble (CAPS- ENSMEAN) described in Table 2 is interpolated to the common evaluation grid every three hours from 0 to 30 h. In addition, question 1 from the introduction is addressed by interpolating forecasts from the operational versions of the NAM and GFS[2] models to the common grid. Finally, output from the 20-km RUC model was interpolated to this grid to provide the verifying analyses.

4. EVALUATION RESULTS

a. Model RMSE

The mean RMSEs for each variable, averaged over the forecast domains for all days of the experiment, have different growth characteristics (Fig. 4). The TMP850 and WSHR errors grow steadily through 30 h and are 1.4-1.8 times larger by the 21-27 h forecast period compared to the initial time. In contrast the TMP2m, DPT2m and DPT850mb errors do not show much growth with time and seem to be affected more by the diurnal cycle, particularly DPT850. This dependence on the diurnal cycle is clearly manifest in the CAPE errors, as the errors in the late afternoon (f21-f24) are 2-3 times larger than the errors in the early morning hours (f06-f12), likely because the surfaced-based CAPE is examined. However, the errors do show a growth with time outside of the diurnal signal- the CAPE errors at 24 h are 1.2-1.8 times larger compared to the initial time.

b. Model Biases

1) TEMPERATURE AND DEWPOINT

The mean biases within the forecast domains for the six variables are shown in Fig. 5. The 2-m temperatures tend to be too warm in the overnight hours and too cool in the daytime hours, although a few of the models break this tendency. For example, the NCAR model tended to be too cool at all hours except at the 0000 UTC start time, and the CAPS-CN and NAM maintain a slight warm bias through the daytime hours. The diurnal variation seen in the mean RMSE for TMP2m, DPT850, and CAPE (Fig. 4) is likely driven by the strong diurnal variation in the mean biases among the various models (Fig. 5). Interestingly, the two larger-resolution models showed very little bias for DPT850, but all of the CAMs show a distinct dry bias for DPT850. The TMP850 errors tend to be small, but show a slight warm bias early in the diurnal cycle.

A result from previous Spring Experiments is the bias of the Mellor-Yamada-Janjic (MYJ) planetary boundary layer (PBL) physical parameterization scheme, which tends to deepen the daytime boundary layer slowly at convection-allowing resolutions, resulting in PBL conditions that are often too cool and moist in the early evening (Kain et al. 2005). As shown in Table 1, all of the deterministic models used the MYJ scheme. The results from this study seem to contradict these findings if an assumption is made that the 850mb level resides within the PBL during the later period of the diurnal cycle. The reasons for this apparent contradiction are not clear. One possibility is that the comparisons that led to the conclusion that the MYJ scheme often produces boundary layers that are too cool and moist were made

[1] The results for the surface-based CAPE should be interpreted with caution since the WRF-based output does not use the virtual temperature correction (Doswell and Rasmussen 1994).

[2] The interpolation of the GFS forecast fields required different parameters in the Gaussian interpolation procedure that resulted in a smoother analysis because the resolution of the GFS input grids is lower than that of the evaluation grid.

with model soundings overlaid with observed soundings valid at either 1800 UTC or 0000 UTC that were usually located within warm sector air masses that originate over the Gulf of Mexico, whereas the comparison is this study is made at all locations within the forecast domain, regardless of the air mass regime.

This is examined further by calculating the mean biases over the entire evaluation domain (Fig. 2) over all days for which model output was available, to determine the spatial distribution of these biases. For this comparison, the technique described in Elmore et al. (2006) to determine the statistical significance of biases in the face of spatial correlation is used here. The NAM 24-h forecasts of TMP2m show regions that tend to be too cool and too warm, but the NSSL 24-h forecasts clearly show a cool bias over almost the entire domain (Fig. 6). The regions that show the largest cold biases in the NSSL model (southern IA/MO and over the Appalachian states) are regions that show a slight cool bias in the NAM. Likewise, the regions that show a small to insignificant cool bias in the NSSL forecasts (central to northern High Plains and eastern TX) are regions that show a warm bias in the NAM. This suggests that the physical parameterizations used by the NSSL model are systematically adding a cool bias on top of the bias provided by the NAM ICs and LBCs. This clearly shows the impact of the physical parameterizations used in the NSSL model and is consistent with Kain et al. (2005) that the MYJ scheme in the WRF-ARW core, working with the NAM ICs and LBCs, creates conditions that are usually too cool near the surface during the early evening hours.

The tendency for the MYJ scheme to produce a PBL that is too moist at higher resolutions is not as clear when viewing the mean 24-h forecast dewpoint errors (Fig. 7). Focusing on the errors over land, the significant mean errors at 850 mb tend to be too moist for the NAM 24-h forecasts (except over western TX), but the mean errors show a clear geographical bias for the NSSL model forecasts. The mean errors in the NSSL model tend to be too moist over the northern High Plains and Midwest and too dry from TX into the lower Ohio Valley. The clear dry bias for the NSSL 24 h forecast at 850 mb shown in Fig. 5 is likely due to the tendency for the forecast domains to be located over the central and southern Plains, where the NSSL model is too dry. Although a comparison of the model output to verifying soundings is preferred, these results suggest that the low-level moist bias associated with the MYJ scheme may not be apparent everywhere over the CONUS.

2) WIND SHEAR

The tendency for the afternoon and evening forecasts to show a low WSHR bias (bottom right panel in Fig. 5) is examined further in Fig. 8. Overall, the WSHR forecasts show little significant bias over much of the domain and show a slight low bias overall. The clear low biases found at 21 h averaged over the forecast domains (Fig. 5) appears to emanate from the central and southern High Plains regions, where low biases of 2-3 m s^{-1} cover much of the region. Southeastern CO seems to be a region with low WSHR forecasts in particular, in which the low bias exceeds 4 m s^{-1} in the NAM and CAPS-ENSMEAN. Although the reasons for this are not clear and are being investigated further, the dry bias in low levels over this region (Fig. 7) suggests that the models tend to push the dryline too far east and the associated veering of the winds could result in the reduced shear values, although the proximity of the low WSHR biases to the Rocky Mountains could also play a role in producing errors in the flow aloft.

b. Operational model forecasts versus deterministic CAM forecasts

One of the goals 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. Figs. 4 and 5 show that there appears to be no consistent improvement in the forecasts between the NAM and GFS forecasts and the higher resolution models (ignoring the CAPS ensemble mean for now), particularly for DPT850. The lack of a consistent improvement in the CAM environments forecasts versus the GFS, and the NAM forecasts especially, is examined further by viewing the frequency distribution of the relative RMSE ranks of the models for each day (Fig. 9). Fig. 9 is produced using the 24 days for which the model output was available for all of the models over the period examined, so that a comparison of the relative rankings using the RMSE could be made. The 15 - 21 h period is examined to allow the next day's diurnal cycle to be represented in the statistics, while reducing the

impact of convective feedback to the resolved scales on the evaluation domain.

Fig. 9 shows that the NAM forecasts of CAPE and DPT850 over the 15-21 h forecast period are ranked in the top three almost 80% of the time. The forecasts from the GFS tend not to rank as high as those from the NAM, but there does not appear to be any significant drop off in performance compared to the CAMs (except for DPT2m). This agrees with Weisman et al. (2008) that the ability to resolve convection does not necessarily have a large impact on resolved-scale mesoscale features, especially if the focus is on the pre-convective environment. As stated in Weisman et al. (2008), this is not necessarily a surprising result, since these models drive the CAMs through the ICs and LBCs. However, this provides evidence that the model physics and parameterizations used at the higher resolutions aren't necessarily providing an improved background environment over what could be garnered from the operational mesoscale forecast models, and in fact, the present results show that they could be having a slight detrimental impact overall.

Although the use of a mean RMSE over a regional domain can not separate objectively the contribution of timing, location, and magnitude errors, inspection of the forecasts reveals 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, which contributed to similar mean RMSE values. An example is shown in Fig. 10, which is a comparison of the NAM and CAPS-CN TMP2m forecasts on a day in which the NAM forecast of the pre-convective environment showed one of the largest improvements over the deterministic models. There was almost no convection in the preceding 21 hours in either the real atmosphere or in the models, so that the comparison is almost completely uncontaminated by convective feedback. Both the NAM and CAPS-CN forecast placed a cold front from south-central NE into western KS and a warm front from far eastern NE southward across the KS/MO border. The difference fields of both models compared to the verifying RUC analysis (Fig. 10, bottom panels) show that the warm front in particular was positioned at nearly the same location in both the NAM and CAPS-CN forecasts resulting in negative temperature errors over eastern KS and MO. The cold front was forecast too far to the north and west by both models resulting in positive temperature errors over KS and central NE. Compared to the RUC analysis, the position of the southern portion of the warm front was forecast well, but the position of the northern portion was forecast too far west over southeastern NE. This exemplifies the typical differences between the NAM forecasts and those from the CAMs. The errors in the placement of airmass boundaries seem to be driven largely by the NAM forecasts resulting in very similar errors in the timing and location of mesoscale and synoptic scale features. However, an important point to make is that errors also are present away from the airmass boundaries. For example, the negative temperature errors found in both the NAM and CAPS-CN forecasts in eastern OK and MO is exacerbated in the CAPS-CN forecasts, likely due to the physical parameterizations.

This modulating effect of the physical parameterizations can be seen in Fig. 4, in which the changes in the mean RMSE for the NAM forecasts with time (black dashed line) seem to follow the changes in the forecasts of the deterministic CAM models that use the NAM for the ICs and LBCs (NSSL, CAPS-CN, and EMC). This is most apparent by focusing on the two NAM-based models that use the WRF-ARW core (NSSL and CAPS-CN). As Fig. 10 shows, position errors of mesoscale features certainly contribute to the RMSE, but the RMSE for the NAM-based WRF-ARW models (NSSL and CAPS-CN) are not the same, mainly because of amplitude and not placement differences (Fig. 4) relating to the different physical parameterizations (see Table 1 for the differences).

Comparisons between the NAM and deterministic CAM fields were also made on the fewer days in which the NAM forecasts of the pre-convective environment were significantly worse than the CAMs. It is found that the NAM forecasts tend to place mesoscale precipitation regions incorrectly during the overnight or morning hours on these days. The most significant improvements in the statistics can be seen if the CAMs have a better representation of the convective systems. An example is shown in Fig. 11, in which the NAM forecast produced a swath of precipitation across SD over the early to late morning hours. A large shield of stratiform precipitation did indeed cover this area in the real atmosphere, but more significant deep convection was organized over north

central NE and moved to the southeast during the late morning. The CAPS-CN model forecast organized convection over this area and over this period similar to what was observed, and produced a good forecast of the convectively generated outflow over western NE during the afternoon that was not represented correctly in the NAM forecast (Fig. 11). This example shows that the few times the higher-resolution models showed a clear improvement upon the NAM forecasts were on days when the early morning convective activity was handled more accurately. However, this was not often the case, which helps explain why the NAM forecasts tended to rank higher than the deterministic CAMS over the period of the experiment.

c. Comparison of CAPS-ENSMEAN to other models

Another goal of this study is to examine forecasts of the environment provided by the CAPS ensemble compared to the other forecasts. It is clear from Fig. 4 that the CAPS-ENSMEAN forecasts almost always improves upon the CAPS-CN forecast and from Fig. 9 that the CAPS-ENSMEAN forecasts tend to be ranked higher than the other models for all of the variables, except for CAPE[3]. 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.

These improvements are larger than expected. Although the interpolation of the fields to the evaluation grid is designed to remove convective-scale features and to produce an analysis comparable to the interpolated RUC fields, it is possible that the large variability in the fields from the native grids of the CAMs on days with widespread convection could still be influencing the RMSE and falsely inflating the improvements provided by the CAPS-ENSMEAN. The potential influence of convective feedback is examined by limiting the evaluation of the relative ranks shown in Fig. 9 to the 16 days with a "clean-slate" pre-convective environment, which are defined to be days with little to no deep convection within the forecast domain from 0900 UTC through 1900 UTC. It is clear that the CAPS-ENSMEAN forecasts stil improve upon the CAPS-CN forecasts when only the nearly undisturbed pre-convective environments are examined (Fig. 12). For the WSHR forecast, although the percentage of number 1 rankings for the CAPS-ENSMEAN declines (Fig. 12), 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 as for all days.

Furthermore, it should be noted that the spatial structures of the CAPS-ENSMEAN fields on the interpolated grids is comparable to the spatial scales of the NAM, with or without convective feedback (not shown), yet the CAPS-ENSMEAN frequently improves upon the NAM (and GFS) forecasts (Figs. 9 and Fig. 12). This is further evidence that the CAPS-ENSMEAN forecasts are truly and improvement over the deterministic forecasts. Finally, the RMSE scores and rankings are examined for the CAPS-CN model output smoothed to produce spatial scales very similar to the CAPS-ENSMEAN fields. The mean RMSE for the smoothed CAPS-CN fields changes very little, and even becomes worse in a few instances compared to the original interpolated fields (not shown). This gives additional support that the experimental design minimizes the effects of convective feedback. In addition, it strengthens the conclusions that the CAMS forecasts do not show a clear improvement over the NAM and GFS forecasts and that the CAPS-ENSMEAN forecasts are frequently better than the other models (and almost always better than the CAPS-CN model).

d. RMSE versus subjective scores

On the day following the issuance of the experimental severe weather forecasts described in section 3 above, the SE2008 participants performed a subjective verification of 1 km AGL simulated reflectivity forecasts from several of the CAMS compared to verifying reflectivity observations similar to that described in Kain et al (2003). The subjective assessment includes how well model reflectivity forecasts corresponded to observed reflectivity, including convective initiation, direction and speed of system movement, areal coverage, the configuration and orientation of the convection,

[3] It should be noted that 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, even if the forecasts are not accurate compared to the other models.

and the convective mode, within the regional domain during the forecast period (Kain et al. 2008). The participants arrived at a consensus numerical rating of 0-10 for each CAM forecast (0 being poor, 10 being perfect).

The relationship between these numerical ratings and the quality of the predictions of the environment by the same models was explored through scatter plots of the mean RMSE of the six variables over a given time period versus the model numerical rating. Little, if any, correlations were found, except for the DPT2m forecasts over both the period of the forecast and, perhaps more importantly for forecasting purposes, at a time prior to the forecast period (Fig. 13). Although the linear relationship is not strong ($r \approx 0.40$ to 0.45), Fig. 13 suggests that the mean RMSE of the 18-h 2-m dewpoint forecast over the regional domain has some predictive value on the perceived quality of the predictions of convection by those models at later forecast times (f21-f30). This could be helpful to forecasters in an operational environment by steering their attention to the quality of the model forecasts of this field versus other fields, if this is indeed a robust relationship. In addition, this could provide justification to use the 2-m dewpoint as a variable to weigh more strongly than other variables in ensemble systems that require the selection of a "best member" (Vukicevic et al. 2008) or as a variable to focus on for initial perturbations for applied forecasting systems (Homar et al. 2004).

5. SUMMARY AND DISCUSSION

This study examines the quality of the predictions of the environment of the convection-allowing WRF models (CAMS) run for the 2008 NOAA/HWT Spring Experiment (SE2008). Motivation came from recent studies (Schwartz et al. 2008, Weisman et al. 2008) and experiences during the SE2008 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 and the lateral boundary conditions. SE2008 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. Understanding the nature of the errors in the specification of the environments by the CAMS is important to the continued development of the models at such high resolution and in the development of CAM-based systems to be used in operational forecasting centers.

The subjective assessment that errors in the CAMs are often tied to the larger scales resolved by the NAM ICs and LBCs is supported by the more objective assessment in this study. Although the use of a mean RMSE over a regional domain can not separate objectively the contribution of timing, location, and magnitude errors, inspection of the forecasts reveals that the placement of airmass boundaries in the NAM forecasts was very similar to the location of the same boundaries represented in the CAM models. This contributed to similar trends in the mean RMSE values among several variables examined. However, the RMSE values are modulated significantly by errors in the physical parameterization schemes among the models.

It is intriguing that the CAPS-CN model, which is forced by the 00Z NAM and does not tend to show an improvement over the NAM forecasts of the environment, but improvements over both the CAPS-CN and NAM forecasts are found in an ensemble framework that includes the CAPS-CN model. This result is not necessarily surprising since numerous studies have shown the benefits of ensemble forecasting for many applications. However, this result is relevant for practical applications of ensemble forecasting, in which it is very important to weigh computational cost versus the forecast benefits, especially if convective-scale forecasts are the goal. If, as these results suggest, little to no improvement in forecasts of the environment is gained by running a CAM versus the mesoscale model that provided its ICs and LBCs, but improvements are seen with an ensemble of these CAMs, then this suggests that little to no degradation in forecasts of the environment could be the result if an ensemble of *mesoscale* models are used to force the environment and to provide improved ICs and LBCs for a convection-allowing domain. The benefit of using an ensemble of mesoscale models to provide the environment versus running an ensemble in which all the members are at convection-allowing scales are obvious in that computational costs could be reduced significantly. The cost savings could be used to allow more frequent updating of the environment within a mesoscale ensemble, which could result in forecasters receiving CAM output at earlier

times with no sacrifice in quality. This highlights the importance of providing an accurate mesoscale environment to the CAMS and underscores the need to develop model perturbation strategies that are appropriate for convection-allowing ensembles. 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 (Stensrud et al. 2008) to configure such a system for testing in upcoming NOAA/HWT Spring Experiments.

6. REFERENCES

Doswell, C. A. and E. N. Rasmussen, 1994: The effect of neglecting the virtual temperature correction on CAPE calculations. *Wea. Forecasting*, **9**, 625-629.

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.]

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.

Homar, V, D. J. Stensrud, J. J. Levit, and D. R. Bright, 2004: 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.

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

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/24SLS/techprogram/paper_141723.htm.]

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

Schwartz, C., J. S. Kain, S. J. Weiss, D. R. Bright, M. Xue, F. Kong, K. Thomas, J. J. Levit, and M. C. Coniglio, 2008: Next-day convection-allowing WRF model guidance: A second look at 2 vs. 4 km grid spacing. Preprints, *24th Conf. on Severe Local Storms*, Amer. Meteor. Soc., Savannah, GA, paper 10.2, [Available online at http://ams.confex.com/ams/24SLS/techprogram/paper_142052.htm.]

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.

Vukicevic, T., I. Jankov, and J. McGinley, 2008: Diagnosis and optimization of ensemble forecasts. *Mon. Wea. Rev.*, **136**, 1054-1074.

Weisman, M.L., 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.

WRF version	NSSL	EMC	NCAR	CAPS-CN
Horiz. Grid (km)	4.0	4.0	3.0	4.0
PBL/Turb. Param.	MYJ	MYJ	MYJ	MYJ
Microphysics Param.	WSM6	Ferrier	Thompson	Thompson
Radiation (SW/LW)	Dudhia/ RRTM	GFDL/GFDL	?/?	Goddard/RRTM
Initial Conditions	40 km NAM	32 km NAM	parallel 9 km WRF/GFS	12 km NAM
Dynamic Core	ARW	NMM	ARW	ARW

Table 1. Configurations for the four deterministic WRF models examined in this study.

member	IC	LBC	Radar data	mp_phy	sw-phy	pbl_phy
Cntl	00Z NAMa	00Z NAMf	yes	Thompson	Goddard	MYJ
C0	00Z NAMa	00Z NAMf	no	Thompson	Goddard	MYJ
N1	Cntl – em_pert	21Z SREF em_n1	yes	Ferrier	Goddard	YSU
P1	Cntl + em_pert	21Z SREF em_p1	yes	WSM 6-class	Dudhia	MYJ
N2	Cntl – nmm_pert	21Z SREF nmm_n1	yes	Thompson	Goddard	MYJ
P2	Cntl + nmm_pert	21Z SREF nmm_p1	yes	WSM 6-class	Dudhia	YSU
N3	Cntl – etaKF_pert	21Z SREF etaKF_n1	yes	Thompson	Dudhia	YSU
P3	Cntl + etaKF_pert	21Z SREF etaKF_p1	yes	Ferrier	Goddard	MYJ
N4	Cntl – etaBMJ_pert	21Z SREF etaBMJ_n1	yes	WSM 6-class	Goddard	MYJ
P4	Cntl + etaBMJ_pert	21Z SREF etaBMJ_p1	yes	Thompson	Dudhia	YSU

Table 2. Variations in IC, LBC, microphysics (mp_phy), shortwave radiation (sw_phy), and planetary boundary layer physics (pbl_phy) for the 2008 CAPS WRF-ARW ensemble. NAMa –12km NAM analysis; NAMf – 12km NAM forecast. All members used the RRTM longwave radiation scheme, and the Noah land-surface scheme. Additional details about the IC and LBC perturbations can be found at URL http://www.emc.ncep.noaa.gov/mmb/SREF/SREF.html and in Xue et al. (2008).

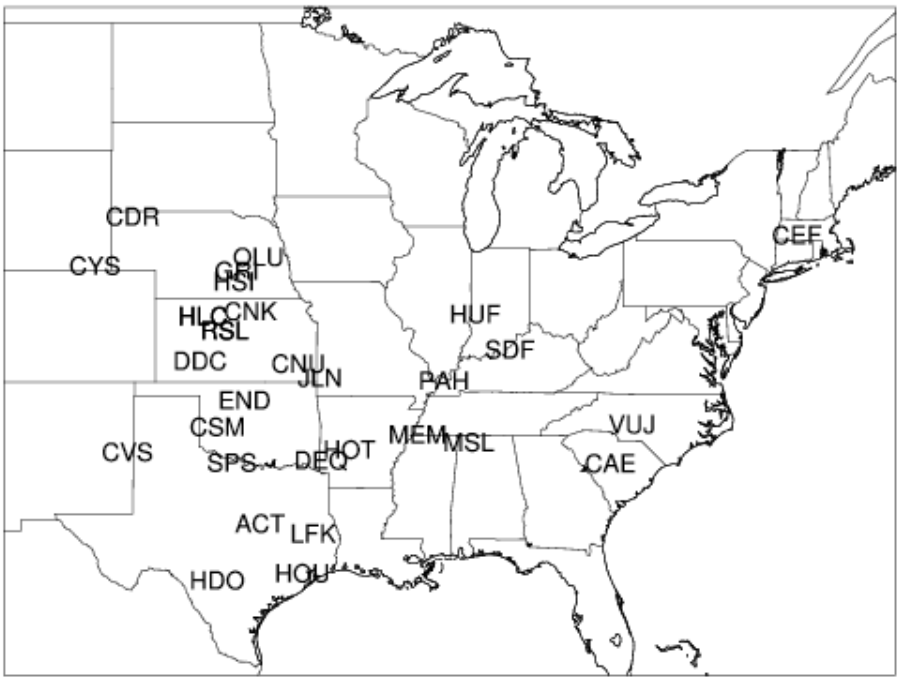

Figure 1. The evaluation domain that contains the interpolated model and analysis fields. The 3-letter station identifiers indicate the center points for the regional domains used to make experimental forecasts and to evaluate the models (see Fig. 3 for an example). A total of 31 domains were used during the experiment (RSL and HLC were chosen twice).

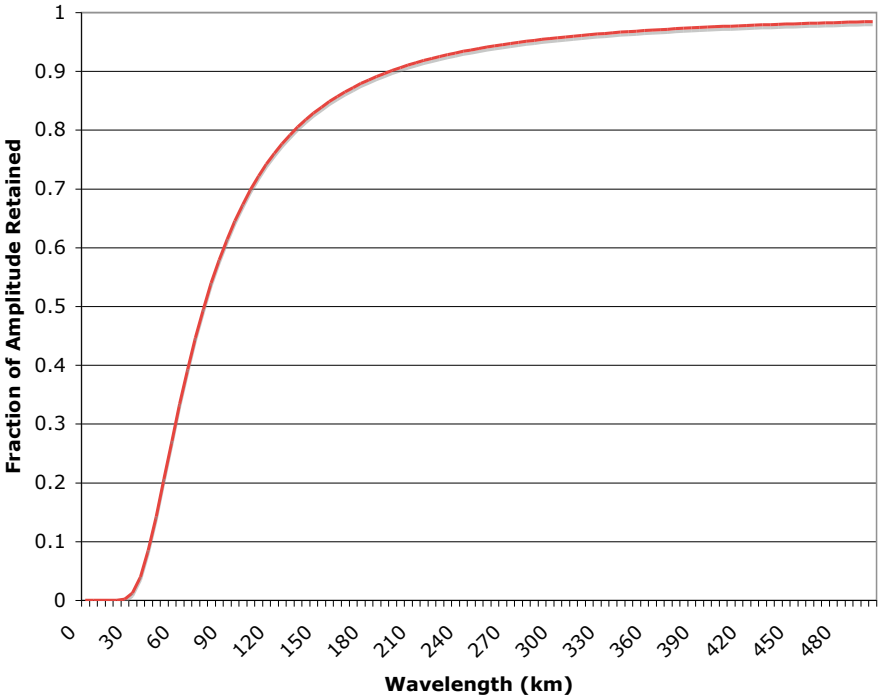

Figure 2. Response function of the Gaussian interpolation of the model fields to the common evaluation grid.

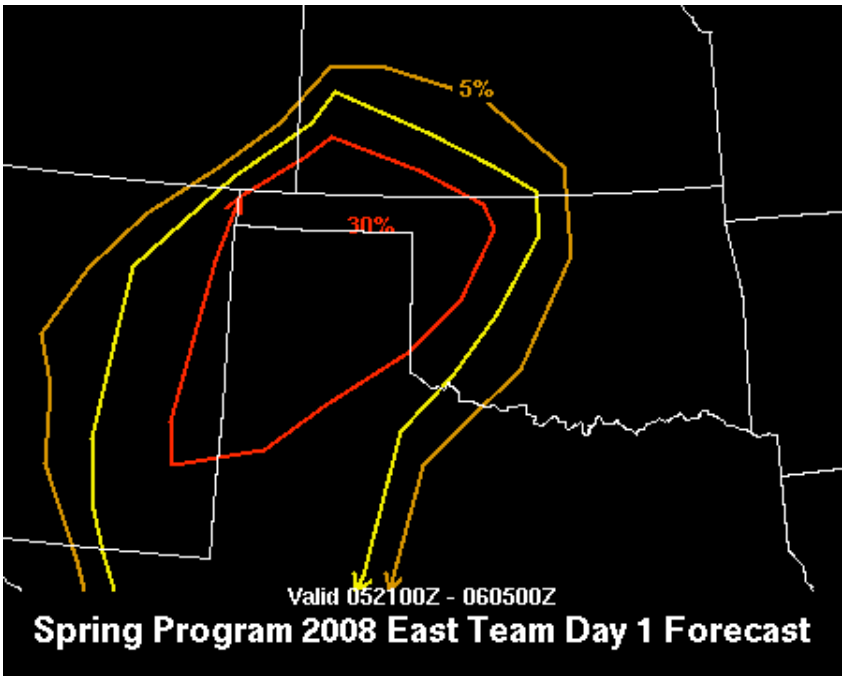

Figure 3. Example of a regional domain selected for the experimental forecast issued on 5 May 2008 by the SE2008 participants. The size of this domain remained the same throughout the experiment and only varied by center point.

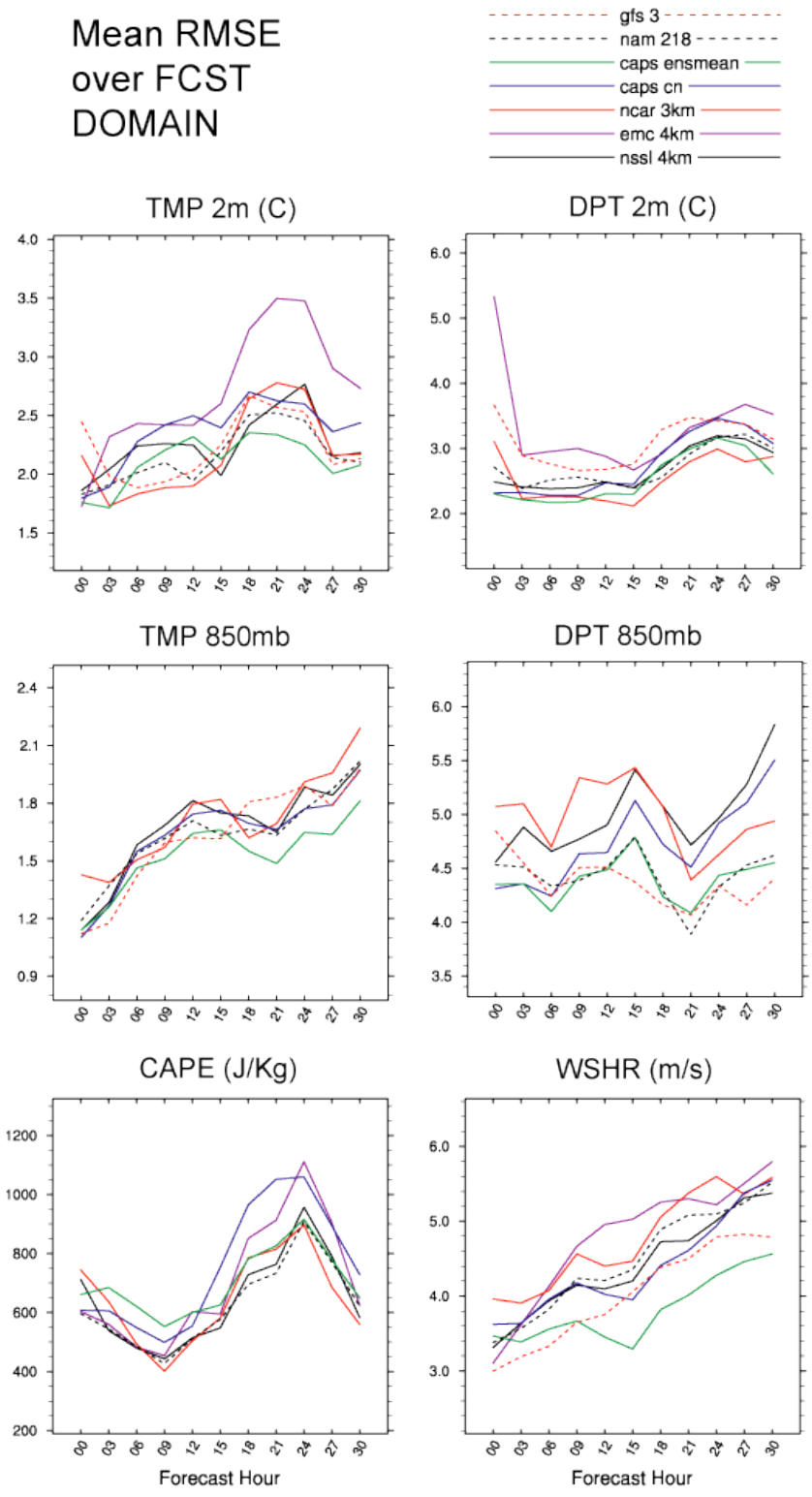

Figure 4. RMSE averaged over the forecast domain for all days versus forecast hour for six model fields.

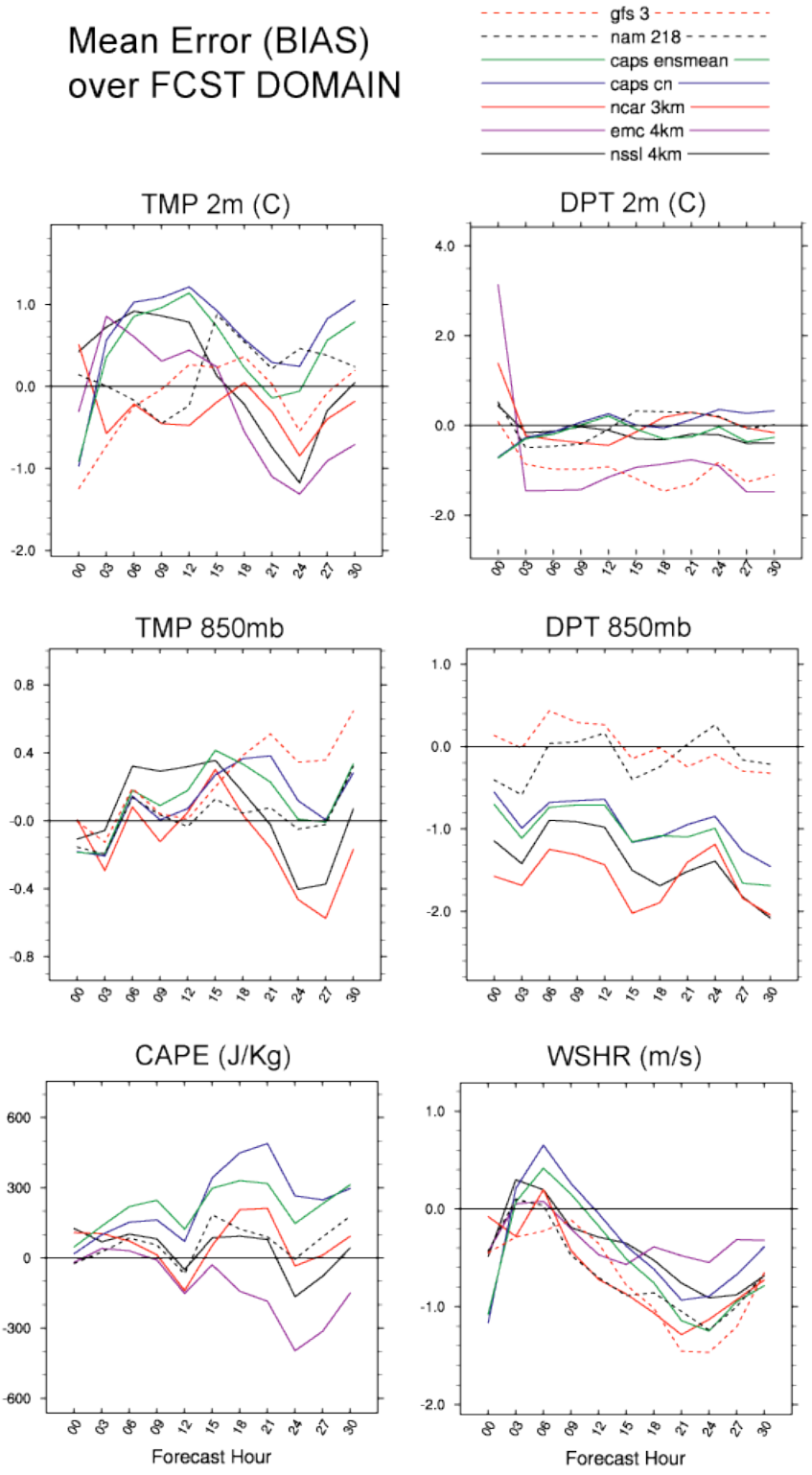

Figure 5. The mean error (bias) averaged over the forecast domain for all days versus forecast hour for six model fields.

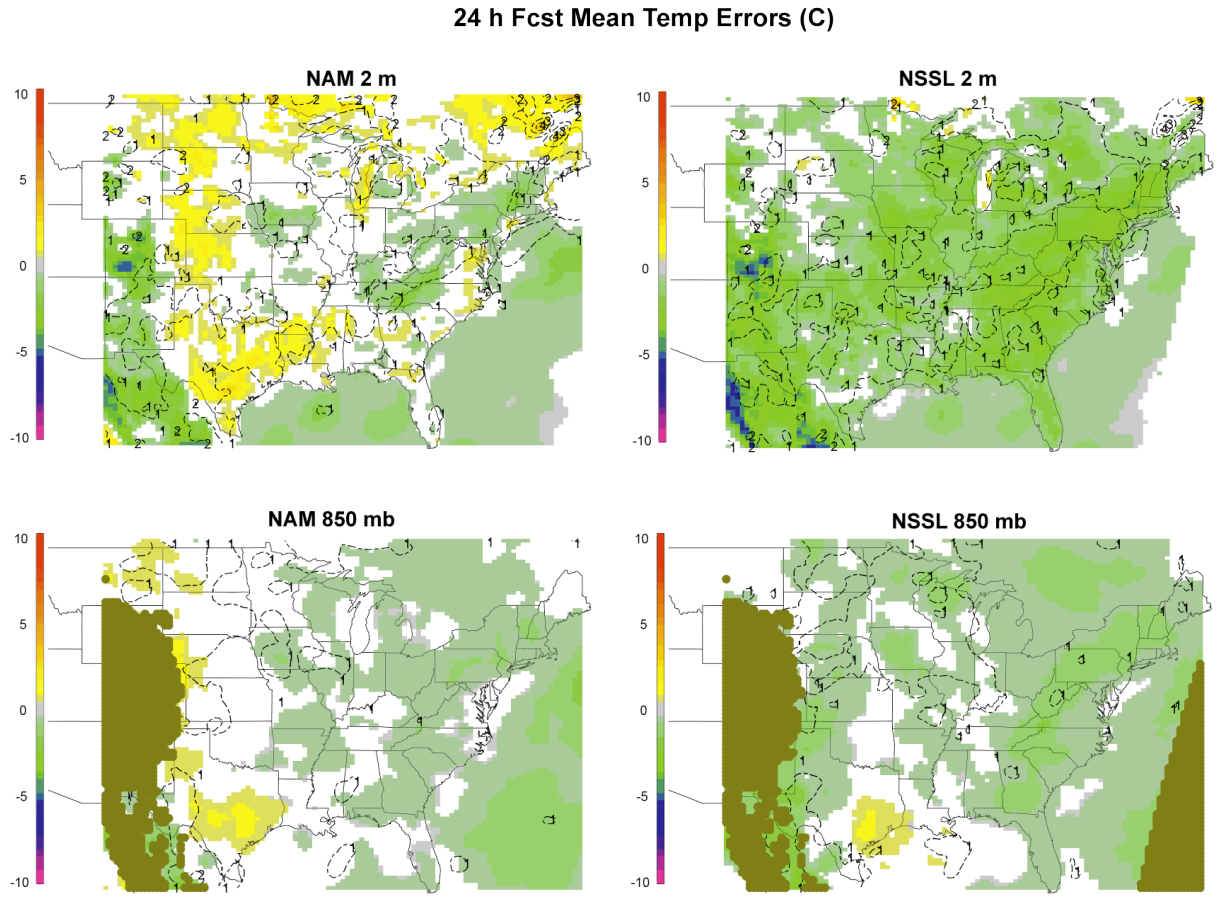

Figure 6. A comparison of the 24 h forecast mean temperature errors (C) (biases) between the NAM and NSSL models at 2 m (top row) and 850 mb (bottom row). The width of the 95% confidence intervals for differences between the model forecasts and the RUC analyses used as verification are contoured with the dashed lines (see Elmore et al. 2006 for details on how the confidence levels are calculated). All areas shaded in green and yellow indicate significant differences at the 95% confidence level. The areas filled in olive for the 850 mb plots indicate a mask used for data below ground or data outside of the native grid for the NSSL plot.

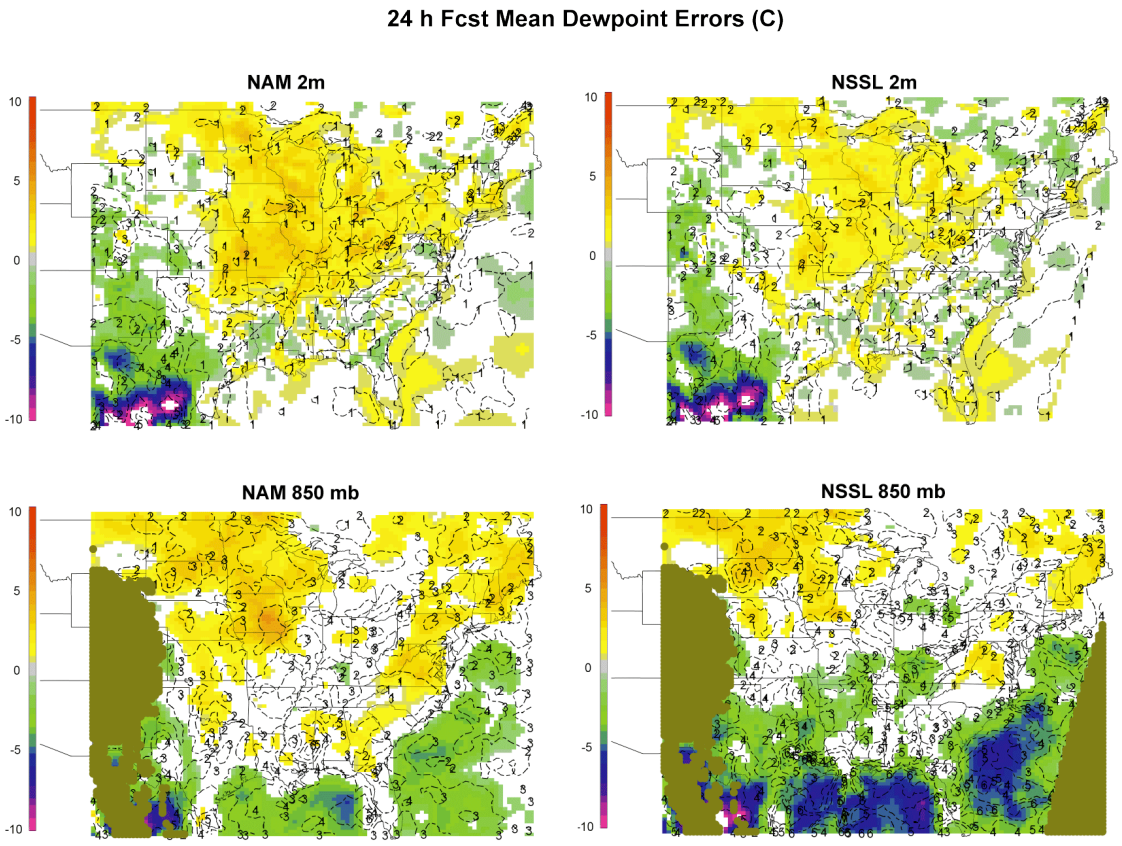

Figure 7. As in Fig. 6, except for the mean 24-h forecast dewpoint errors.

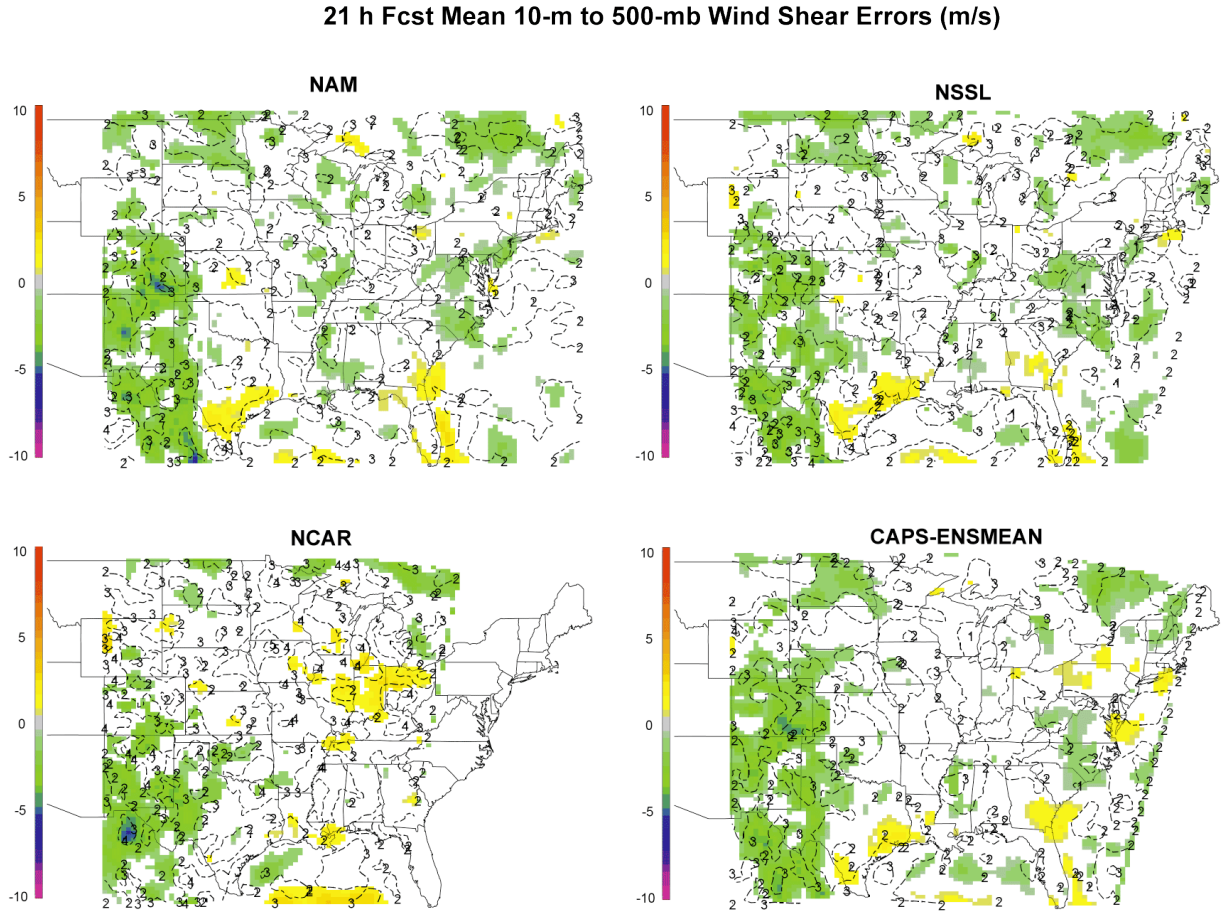

Figure 8. As in Fig. 6, except for a comparison of the 21 h forecast mean 10-m to 500-mb wind shear errors (m/s) (biases) for the NAM, NSSL, NCAR, and the CAPS-ENSMEAN output.

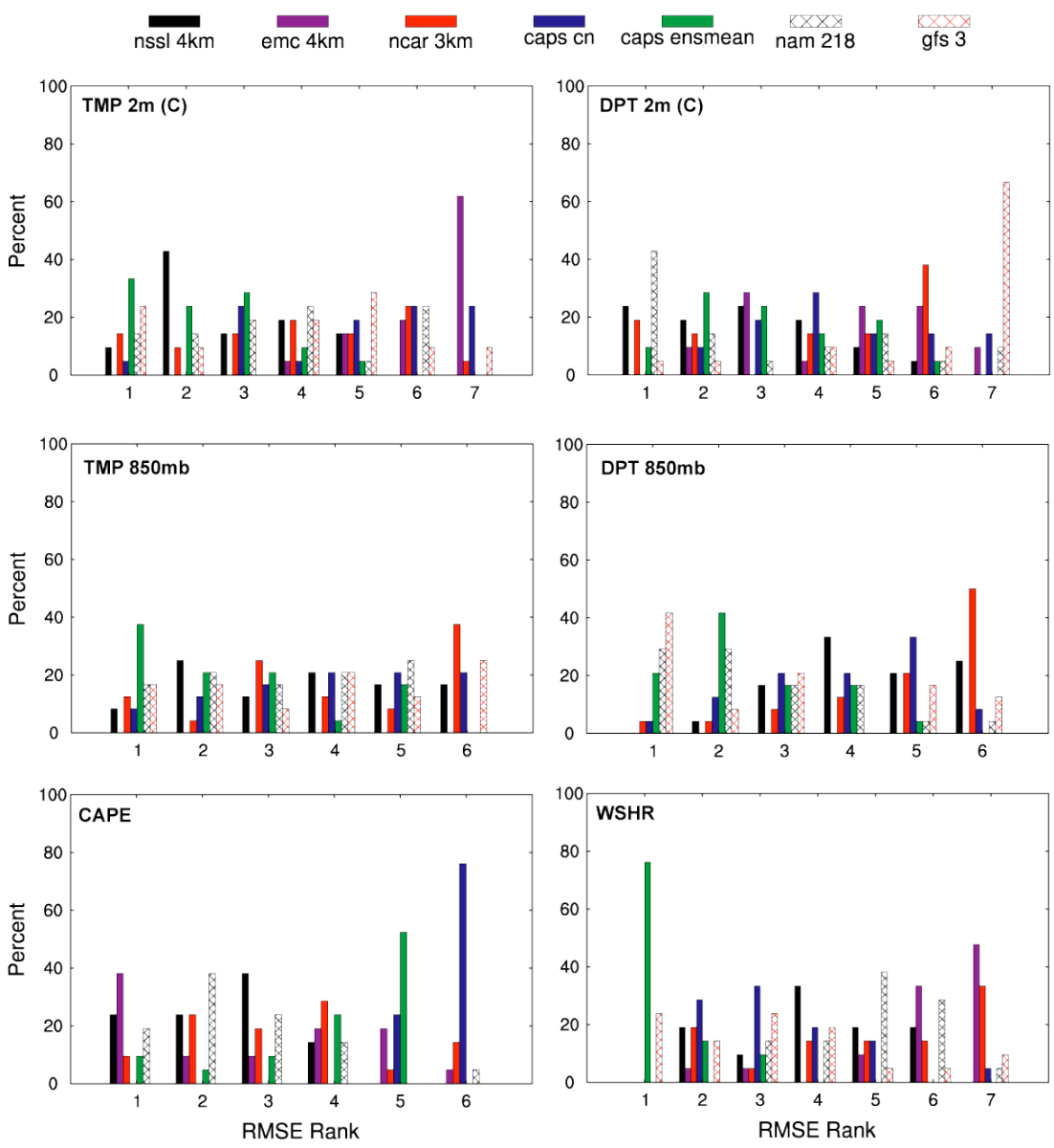

Figure 9. The relative ranks of the RMSE averaged over the 15 h to 21 h forecast period and averaged over the forecast domain for each model and for six model fields (see text for details). For example, the upper-left panel shows that the CAPS ensemble mean 2-m temperature had the lowest mean f15-f21 h RMSE averaged over the forecast domain (ranked the highest) on almost 40% 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.

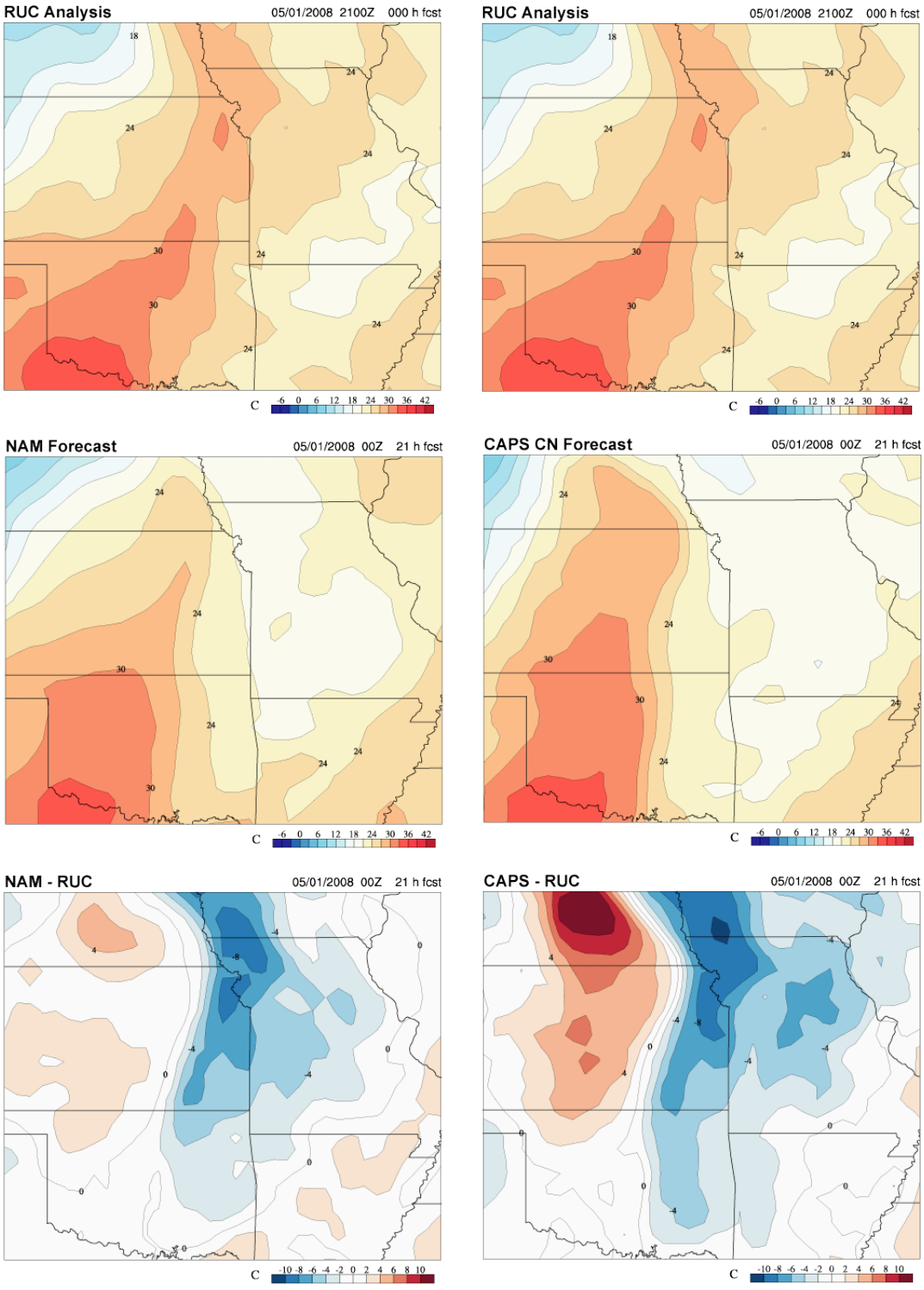

Figure 10. A comparison of the TMP2m RUC analysis at 2100 UTC 1 May 2008 (upper-most panels) and the NAM (middle-left panel) and CAPS CN (middle-right panel) 21 h forecasts valid at the same time within the forecast domain.

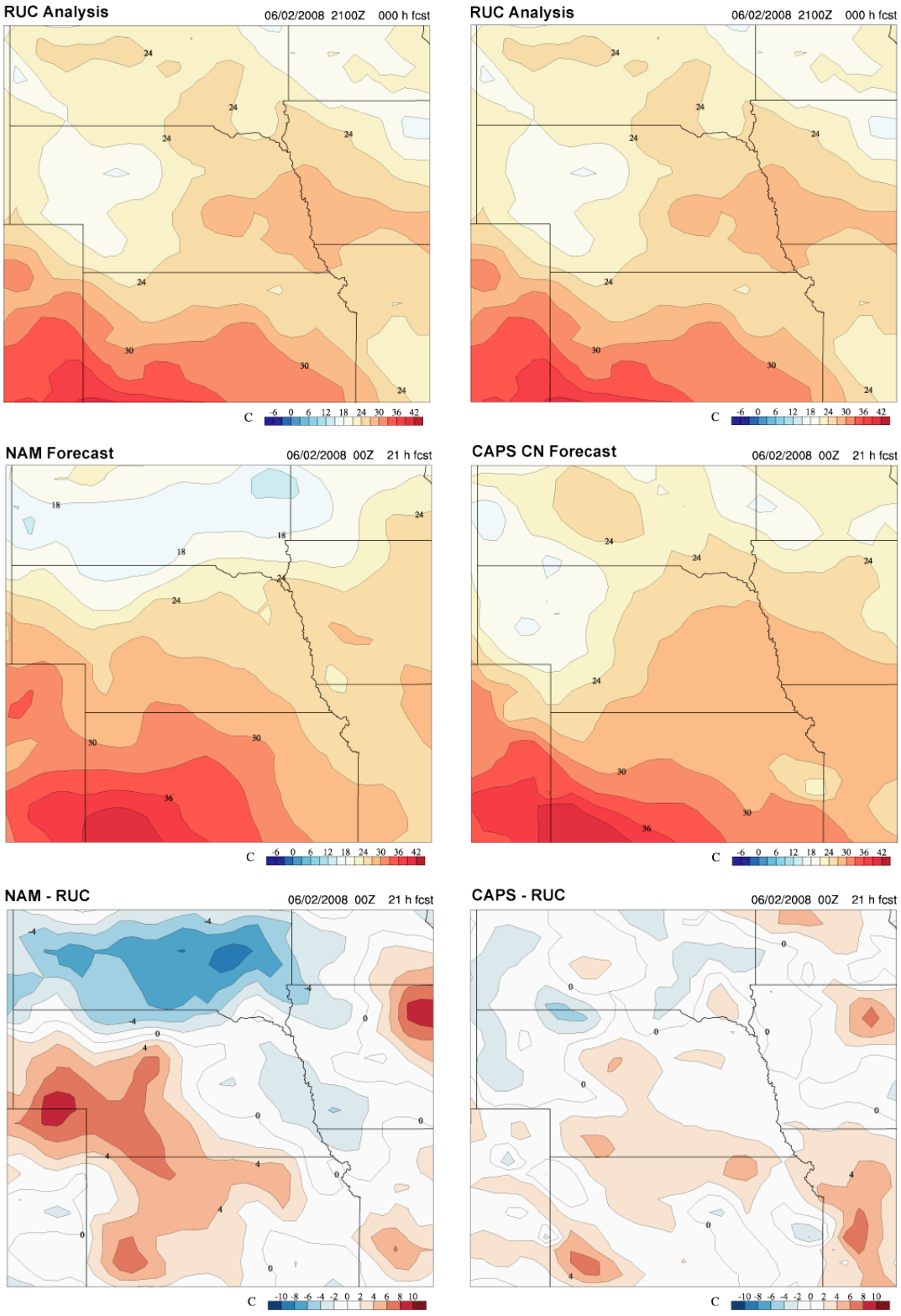

Figure 11. A comparison of the TMP2m RUC analysis at 2100 UTC 2 June 2008 (upper-most panels) and the NAM (middle-left panel) and CAPS CN (middle-right panel) 21 h forecasts valid at the same time within the forecast domain.

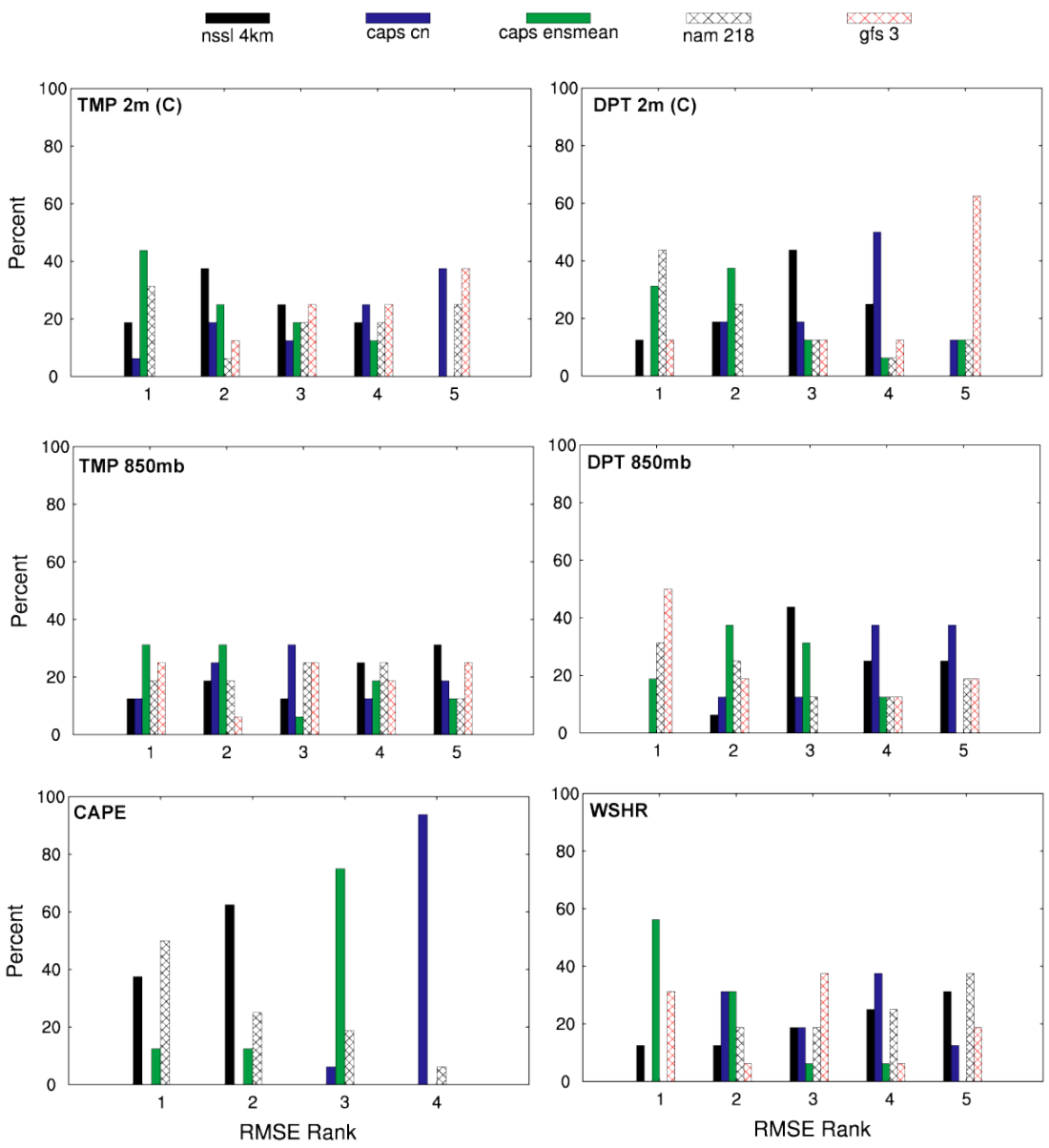

Figure 12. As in Figure 9, except for the 16 "clean-slate" days (see text for details). The EMC and NCAR models are not shown because output was available on only 10 of the 16 clean-slate days for these models.

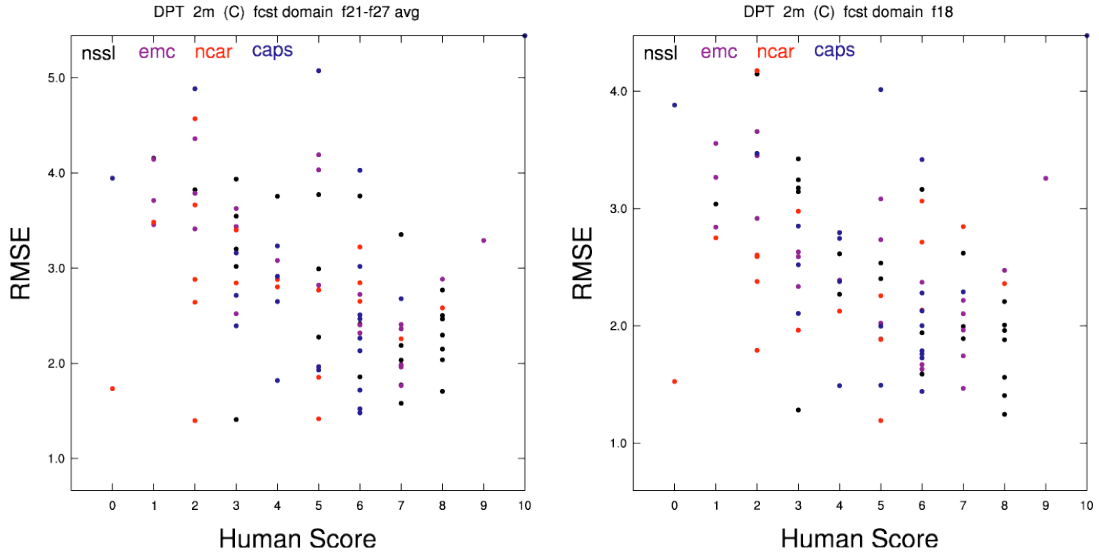

Figure 13. A comparison of the mean RMSE for DPT2m averaged over the forecast domains and the numerical rating given to the models for the assessment of the quality of the reflectivity forecasts by the SE2008 participants (see text for details). The left panel shows the comparison for the mean RMSE over the 21-27 h period (corresponding to the times over which the numerical ratings are based) and at 18 h.