13A.6 TOWARD IMPROVED CONVECTION-ALLOWING ENSEMBLES: MODEL PHYSICS SENSITIVITIES AND OPTIMIZING PROBABILISTIC GUIDANCE WITH SMALL ENSEMBLE MEMBERSHIP

Craig S. Schwartz*[1], John S. Kain[2], Steven J. Weiss[3], Ming Xue[1,4], David R. Bright[3], Fanyou Kong[4], Kevin W.Thomas[4], Jason J. Levit[3], Michael C. Coniglio[2], Matthew S. Wandishin[1,5]

[1]School of Meteorology, University of Oklahoma, Norman, Oklahoma
[2]NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma
[3]NOAA/NWS/Storm Prediction Center, Norman, Oklahoma
[4]Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma
[5]Department of Atmospheric Physics, The University of Arizona, Tucson, Arizona

1. Introduction

In the last decade, computer resources have increased to enable numerical weather prediction (NWP) models to run at progressively higher resolutions over increasingly large domains. Several modeling studies (e.g. Done et al. 2004; Kain et al. 2006; Weisman et al. 2008; Kain et al. 2008a; Schwartz et al. 2008) using convection-allowing [no convective parameterization (CP)] configurations of the Weather Research and Forecasting (WRF) model with horizontal grid spacings of ~ 4 km have demonstrated the added value of these high-resolution models as forecast guidance tools. Additionally, these experiments revealed little adverse effects from running the WRF model at 4 km without CP, even though this grid spacing is too coarse to fully capture convective circulations. Given the success of these convection-allowing WRF forecasts, ~ 4 km convection-allowing models have become operational at the United States National Centers for Environmental Prediction (NCEP) in the form of "high-resolution window" deterministic forecasts produced by NCEP's EMC (Environmental Modeling Center).

Thus far, convection-allowing WRF studies (e.g. Done et al. 2004; Kain et al. 2006; Weisman et al. 2008; Kain et al. 2008a; Schwartz et al. 2008) have all focused on single deterministic model output. While the utility of these deterministic forecasts has been shown, deterministic forecasts are, by nature, inferior to probabilistic ones when it comes to providing guidance for rare events, such as severe thunderstorms or heavy precipitation (Murphy 1991). Probabilistic forecasts allow forecasters to quantify uncertainty such that their forecasts can reflect their best judgment and, perhaps more importantly, allow users to make better decisions as compared to yes-no forecasts (Murphy 1993). Moreover, as grid lengths have decreased in NWP models to the size of convective-scale features, grid point values are more likely to incur large errors, thus, hampering deterministic forecasts at the grid scale. In recognition of this problem, post-processing and verification methods have been developed that relax the requirement that model output and corresponding observations match exactly in order for a forecast to be considered correct (Theis et al. 2005; Roberts 2005; Roberts and Lean 2008). Another benefit of this "neighborhood" approach is that it can be used to generate probabilistic information from deterministic grids (Theis et al. 2005). Furthermore, Theis et al. (2005) demonstrated that probabilistic forecasts generated using the neighborhood method are superior to forecasts from direct grid point model output.

A more widely used strategy to generate probabilistic guidance involves an ensemble forecasting system. An ensemble is comprised of a suite of individual deterministic runs, each generated from a unique set of initial conditions (IC), lateral boundary conditions (LBC), physical parameterizations, and/or dynamics formulations. IC and LBC diversity acknowledges the uncertainty of meteorological observations and the data assimilation systems that incorporate observations into the model grids, while differing model physics recognizes the uncertainties inherent in the parameterizations of small-scale, poorly-understood processes, such as cloud microphysics and turbulence. Ideally, all

*Corresponding author address:
Craig Schwartz,
University of Oklahoma, School of Meteorology,
120 David L. Boren Blvd. Suite 5642,
Norman, OK 73072;
E-mail: cschwartz@ou.edu

ensemble members are assumed to be equally likely of representing "truth" at initialization, and thus, have an equal chance of producing the best forecast at a later time. Typically, initial fields vary only slightly at early time-steps, and forecasts from the members are quite similar. However, these differences may amplify with time such that by the end of the model integration, different ensemble members arrive at wildly different solutions. The spread of the members is typically associated with perceived forecast uncertainty, and point probabilities are commonly obtained by considering the total number of members predicting an event at a given grid box. Alternatively, deterministic information from all the members can be averaged into a mean deterministic field. As errors of the different members tend to cancel in the averaging process, this mean almost always performs better than any of the individual members. Furthermore, numerous studies (e.g. Stensrud et al. 1999; Wandishin et al. 2001; Bright and Mullen 2002) have shown that an ensemble system performs comparably to or better than a similarly configured, higher-resolution deterministic forecast, as measured by objective metrics.

Medium-range (3-15 days) ensemble forecasts have been produced operationally at NCEP since the early 1990s, but the development of short-range (0-3 day) ensemble forecasts (SREF) lagged behind. Following the recommendation of participants in a workshop designed to explore future SREF implementation (Brooks et al. 1995), experimental SREF runs were initiated at NCEP in 1995 (Du and Tracton 2001). Given the success of the experimental forecasts, the use of SREFs continued, and they became operational at NCEP in 2001. The current NCEP SREF employs 21 members at 32 km grid spacing and is run 4 times daily. Variations in physical parameterizations, dynamic cores, ICs, and lateral boundary conditions (LBC) are used to create forecast diversity.

Given the benefits of ensemble forecasting and previous successes of convection-allowing 4 km WRF deterministic output, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma contributed large domain, realtime 10-member 4 km convection-allowing ensemble forecasts to the 2007 NOAA (National Oceanic and Atmospheric Administration) Hazardous Weather Testbed Spring Experiment[1] (hereafter SE2007). Variations in ICs, LBCs, and physical parameterizations were used to achieve ensemble diversity. On its own, these ensemble forecasts represented a groundbreaking computational achievement (see Xue et al. 2007) and to our knowledge is the first time a high-resolution, convection-allowing ensemble has been run in a realtime setting.

The purpose of this study is to examine output from the CAPS ensemble for two purposes. First, it examines output from the different ensemble members and identifies Advanced Research WRF (WRF-ARW; Skamarock et al. 2005) model sensitivities to microphysics and PBL (planetary boundary layer) parameterizations. Second, a new method of extracting probabilistic ensemble guidance is presented. This method applies a "neighborhood" approach to the ensemble [as suggested by Theis et al. (2005)]. The ensemble configuration and experimental design are discussed next, followed by a discussion of WRF-ARW sensitivity to physical parameterizations. Traditional and new methods of generating probabilistic ensemble forecasts are presented in section 4 and these forecasts are verified in section 5 prior to concluding.

2. Experimental design

2.1 *Model configurations*

On each of the ~35 days of SE2007, CAPS produced 10-member ensemble forecasts with 4 km grid spacing (Xue et al. 2007; Kong et al. 2007). The ensemble forecasts were generated remotely at the Pittsburgh Supercomputing Center (PSC). All ensemble members used version 2.1 of the WRF-ARW dynamic core (Skamarock et al. 2005), represented convection explicitly (no convective parameterization), resolved 51 vertical levels, were initialized with a "cold-start" (no data assimilation) at 2100 UTC, and ran for 33 hours over a domain encompassing approximately three-fourths of the continental United States (Fig. 1).

[1] This experiment, formerly called the SPC/NSSL (Storm Prediction Center/National Severe Storms Laboratory) Spring Program, has been conducted from mid-April through early June annually since 2000. Details about the experiments can be found at URL http://www.nssl.noaa.gov/hwt.

Ensemble member configurations

Member	IC	BC	Microphysics	PBL physics
cn	2100 UTC NAMa	1800 UTC NAMf	WSM 6-class	MYJ
n1	cn – arw_pert	2100 UTC SREF arw_n1	Ferrier	MYJ
n2	cn + arw_pert	2100 UTC SREF arw_p1	Thompson	MYJ
p1	cn – nmm_pert	2100 UTC SREF nmm_n1	Thompson	YSU
p2	cn + nmm_pert	2100 UTC SREF nmm_p1	WSM 6-class	YSU
ph1	2100 UTC NAMa	1800 UTC NAMf	Thompson	MYJ
ph2	2100 UTC NAMa	1800 UTC NAMf	Ferrier	MYJ
ph3	2100 UTC NAMa	1800 UTC NAMf	WSM 6-class	YSU
ph4	2100 UTC NAMa	1800 UTC NAMf	Thompson	YSU
ph5	2100 UTC NAMa	1800 UTC NAMf	Ferrier	YSU

Table 1. *Ensemble member configurations. The WRF Single-Moment 6-class (WSM6)(Hong et al. 2004), Ferrier (Ferrier 1994); Thompson (Thompson et al. 2004); Mellor-Yamada-Janjic (MYJ)(Mellor and Yamada 1982, Janjic 2002) and Yonsei University (YSU) (Noh et al. 2003) schemes are used. NAMa and NAMf refer to NAM analyses and forecasts, respectively.*

The configurations of the ensemble members are summarized in Table 1. ICs were interpolated to the 4 km grids from a 2100 UTC analysis of the 12 km North American Model (NAM; Black, 1994) (J. Du, NCEP/EMC, personal communication), though the NAM data were coarsened to a 40 km grid before being ingested. Different IC, LBC, and physics perturbations were introduced in four of the ten ensemble members (n1, n2, p1, p2; hereafter collectively referred to as the "LBC/IC" members). The IC and LBC perturbations were based on the NCEP SREF, and the LBCs themselves came from the 2100 UTC SREF for these four members. LBCs for the remaining six members (cn, ph1, ph2, ph3, ph4, ph5; hereafter collectively referred to as the "physics-only" members) were provided by 1800 UTC 12 km NAM forecasts. These six members used identical ICs and LBCs and differed solely in terms of microphysics and PBL (planetary boundary layer) parameterizations. Therefore, comparison of their output allows a robust assessment of WRF-ARW sensitivity to PBL and microphysics parameterizations in a variety of weather regimes.

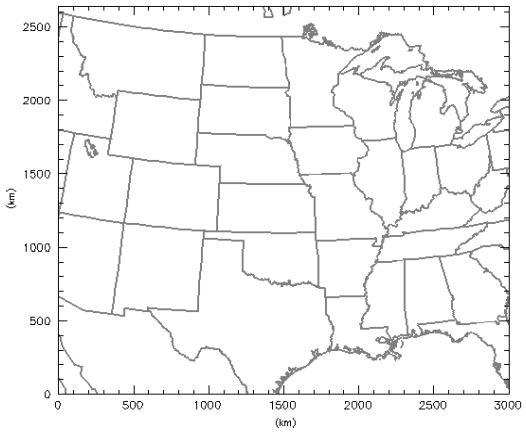

Fig. 1. *Model integration domain.*

2.2 *Verification parameters*

At the conclusion of SE2007, average ensemble performance characteristics were assessed using several statistical measures applied primarily to hourly precipitation fields. Hourly model precipitation forecasts were compared to gridded Stage II precipitation fields produced hourly at NCEP (Lin et. al 2005). Stage II precipitation fields are generated from radar and rain gage data (Seo 1998), and they were regarded as "truth."

Objective verification of the model climatology was performed over a fixed domain comprising most of the central United States (Fig. 2). This domain covered a large area over which Stage II data were robust and springtime weather was active. Additionally, this region was also sufficiently removed from the lateral boundaries so as to minimize contamination.

Attention was focused on the f21-f33 (1800-0600 UTC) period to examine the utility of the ensemble as next-day convective storm guidance. When possible, statistics were computed on native grids. However, in order to calculate certain performance metrics (discussed below), it was often necessary that all data be on a common grid. Therefore, for certain objective verification procedures, model output was interpolated onto the Stage II grid (grid spacing of ~4.7 km), which will be referred to as the "verification grid."

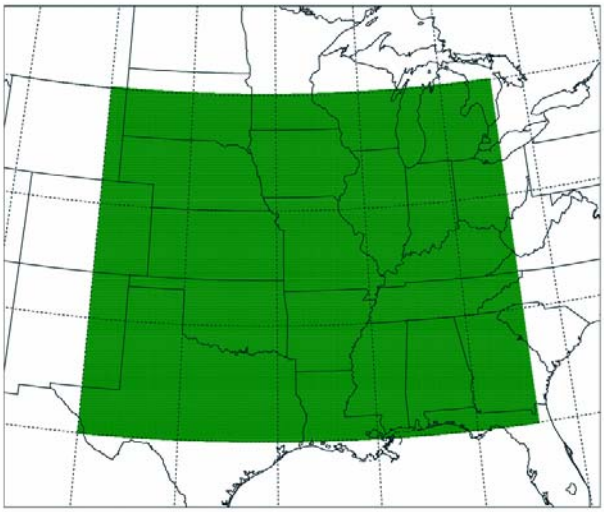

Fig. 2. *Verification domain used for model climatology.*

3. Precipitation sensitivity to physical parameterizations

The individual ensemble members produced varying amounts of precipitation. By consulting model physics configurations (Table 1), it appears these differences can be attributed to the PBL and microphysics schemes. Aggregate statistics over all days of SE2007 are first presented, followed by a brief case study..

3.1 *Domain total precipitation*

Total accumulated precipitation throughout the verification domain, calculated on native grids and aggregated over all days of SE2007, is depicted in Fig. 3. All the members captured the diurnal cycle quite well, with afternoon precipitation maxima within an hour of the observed peak.

All members overpredicted the mean precipitation, especially during the afternoon peak. The specific cause of this high bias has not been identified. However, more detailed examinations of specific events, conducted by CAPS scientists after SE2007, suggested that the bias was significantly reduced when the ensemble was initialized with 0000 UTC ICs and LBCs. Thus, it appears that some aspect of the 2100 UTC ICs, and perhaps the 1800 UTC LBCs, led to the very high bias (Kong et al. 2008). Although this high bias situation was less than optimal, all members were subjected to the same constraints and affected equally. Therefore, differences between the members should still yield a robust assessment of sensitivity to model physics.

Case in point, despite this ubiquitous high bias, there was nonetheless considerable spread between the physics-only members regarding the amplitude of the peak. This separation suggests the combination of PBL and microphysics parameterizations exerts a strong influence on the rainfall fields. This impact is further revealed by examining the amplitudes of the LBC/IC members. In general, members with the same PBL and microphysics parameterizations produced similar amounts of precipitation, regardless of any LBC and IC perturbations. For example, the n1 and ph2 members produced the highest afternoon precipitation totals, and both were configured with the Ferrier microphysics and MYJ PBL parameterizations. On the other hand, the n2 and ph4 members produced the least amount of precipitation during the afternoon maximum, and each was configured with the YSU PBL and Thompson microphysics schemes. However, the p2 and ph3 members produced the least precipitation during the last three hours of integration, and also during the diurnal minimum. Both members shared the YSU PBL and WSM6 microphysics parameterizations.

3.2 *Areal coverages*

Figure 4 depicts fractional coverages of precipitation exceeding various accumulation thresholds (q)(e.g. 1.0 mm hr^{-1}), aggregated hourly over all days SE2007. These statistics were generated from data on each member's native grid. Again, on average, the individual members captured the diurnal cycle fairly well, with the time of peak coverage corresponding well to the observations.

When q = 0.2 mm hr^{-1} (Fig. 4a), all but the n1 and ph2 (Ferrier and MYJ) members generated either a similar or lower fractional coverage than the observations, on average. But, as q increased, overprediction dramatically worsened, such that by the 5.0 mm hr^{-1}threshold (Fig. 4c), all members produced a grossly higher areal coverage than that observed. Again, the areal coverages of members with the same physics schemes were quite similar. During the afternoon hours, the n1 and ph2 members yielded the greatest fractional coverages, while the n2 and ph4 (Thompson and YSU) and p2 and ph3 pairs (WSM6 and YSU) produced the least grid coverage.

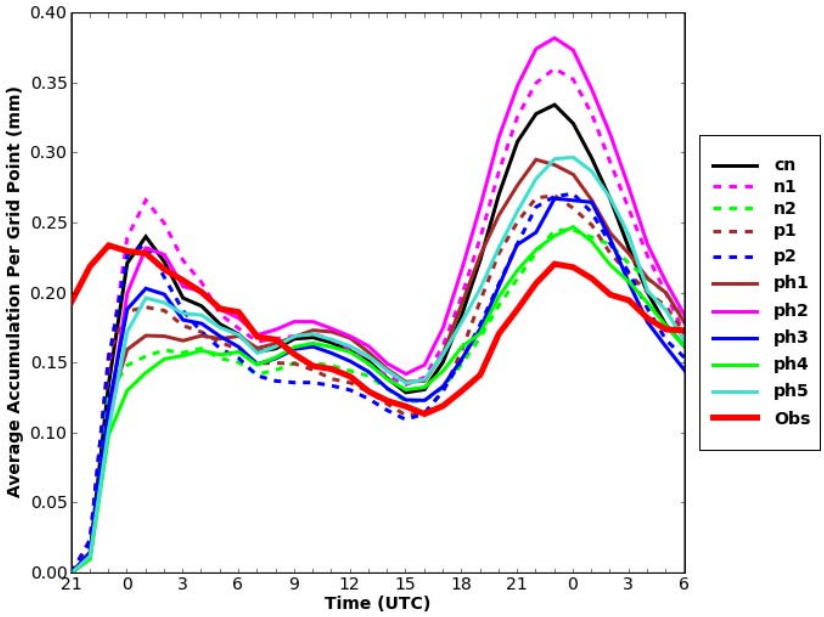

Fig. 3. *Total precipitation over the domain aggregated over all days of SE2007, normalized by number of grid boxes. Calculated on each member's native grid.*

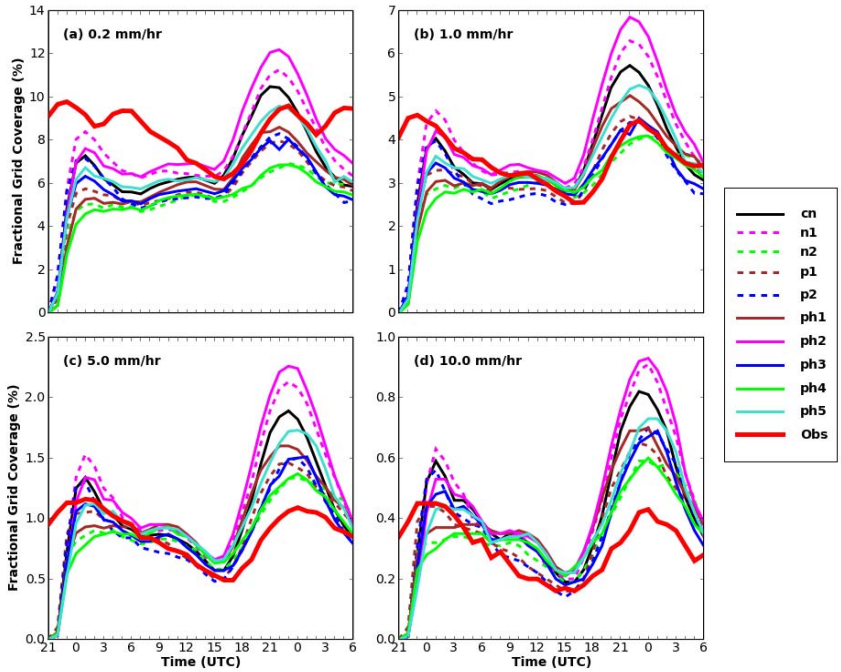

Fig. 4. *Fractional grid coverage of hourly precipitation exceeding (a) 0.2 mm hr^{-1}, (b) 1.0 mm hr^{-1}, (c) 5.0 mm hr^{-1}, and (d) 10.0 mm hr^{-1}as a function of time, averaged over all days of SE2007, calculated on each member's native grid.*

3.3 *Precipitation percentiles*

A climatology of precipitation accumulations was constructed by compiling the hourly accumulated precipitation between 1800-0600 UTC (f21-f33) in each grid box within the verification domain on the native grids over all days of the SE2007. The values were ranked, and accumulation percentiles (y) (e.g. 95th percentile) were chosen to determine absolute hourly precipitation values (q_y) corresponding to yth percentile (Fig. 5). For example, (100-y) percent of all grid points contained accumulations above the value of q_y, which was determined by yth percentile. This procedure was performed separately for each member.

Systematic differences between the members were again evident, as was the tendency for members with common physical parameterizations to behave similarly. For example, hourly accumulations of ~8.0 mm or higher comprised the top 1% of all accumulations in the n1 and ph2 (Ferrier and MYJ) hourly precipitation fields, while the 99th percentile in the n2, ph4, p2, and ph3 fields was considerably lower (~ 5.5 mm hr^{-1}). Additionally, the spread between q_y of the different members increased with the percentile.

3.4 *Precipitation bias*

To quantitatively determine the biases of individual members, the standard 2x2 contingency table for dichotomous (yes-no) events was used (Table 2). The bias (B) is simply the ratio of the coverage of forecasts to the coverage of observations and can be easily computed from the contingency table [$B = (a+b)/(a+c)$]. For a given value of q, a bias > 1 indicates overprediction and B < 1 indicates underprediction at that threshold. Metrics computed from Table 2 require that the models and observations be on the same grid, so the model output was interpolated onto the verification grid.

Bias aggregated over all days of SE2007 between 1800-0600 UTC (f21-f33) are plotted as a function of precipitation threshold in Fig. 6. A large bias spread is evident, with the n1 and ph2 members overpredicting the most for $q \leq 10.0$ mm hr^{-1}. At thresholds > 10.0 mm hr^{-1}, the n1 and ph2 (Ferrier and MYJ) biases interestingly plummet, leaving the ph1 and p1 members with the highest biases (both configured with Thompson microphysics and MYJ PBL schemes).

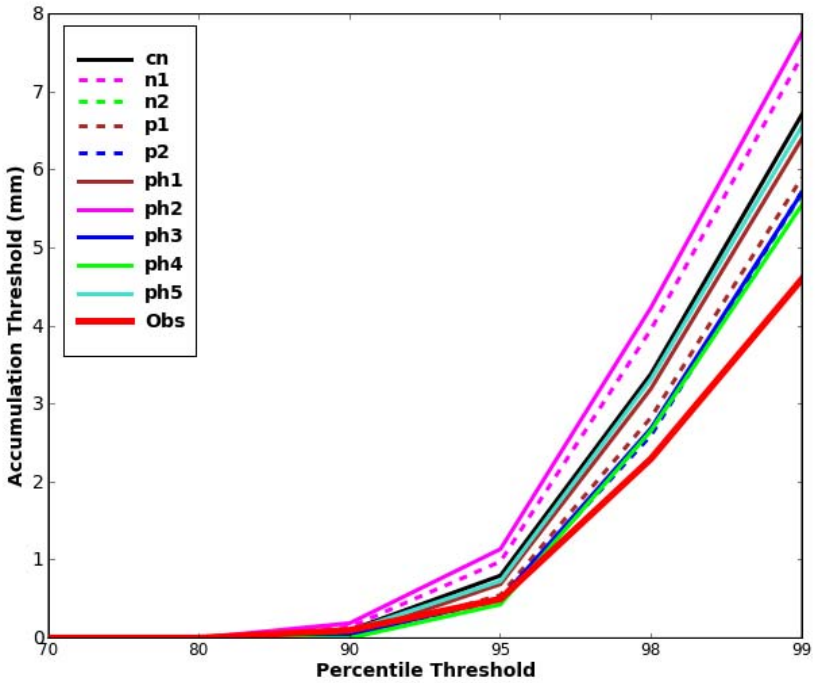

Fig. 5. *Precipitation climatology: Percentiles calculated on each member's native grid aggregated between 1800-0600 UTC (f21-f33) over all days of SE2007 (see text).*

2 x 2 Contingency Table

		Observed		
		Yes	No	
Forecast	Yes	*a*	*b*	*a+b*
	No	*c*	*d*	*c+d*
		a+c	*b+d*	*N*

Table 2. *Standard 2x2 contingency table for dichotomous events.*

3.5 *Case Study*

Fig. 7 shows hourly precipitation output from the physics-only members on their native grids over the verification domain. The ensemble was initialized 2100 UTC 04 June 2007, and the forecast was valid 0000 UTC 06 June. This case illustrates many of the characteristics seen on average throughout SE2007, as previously discussed.

All members produced scattered precipitation from eastern Colorado southeastward into central Arkansas. However, there were differences regarding areal coverage and intensity. The cn (WSM6 and MYJ) and ph2 (Ferrier and MYJ) members were relatively bullish, developing comparatively more and larger areas of precipitation, especially over southern Kansas, northern Oklahoma, and the northern half of Arkansas. On the other hand, the ph4 (Thompson and YSU) and ph5 (Ferrier and YSU) members produced fewer and smaller elements over these same regions. The ph1 (Thompson and MYJ) and ph3 (WSM6 and YSU) members seemed to be in the middle.

Farther east, all the members generated widespread rainfall in southern Alabama and Georgia. While there were some slight differences between the members over this area, they all seemed in fairly good agreement. However, there were disagreements regarding precipitation intensity over Kentucky and Tennessee. The ph2 and ph5 members produced the heaviest rainfall, while the other members predicted lighter showers.

The perceived visual differences are substantiated by a quantitative assessment of the hourly precipitation (Fig. 8). The ph2 member produced the most precipitation, while the ph4 and ph3 generated the least. Although the ph5 member produced less precipitation over the Great Plains, its heavier precipitation over the Ohio Valley and Gulf Coast brought its total precipitation above that of the ph3 member. Note that all the members overpredicted the observed hourly precipitation that occurred over the verification domain.

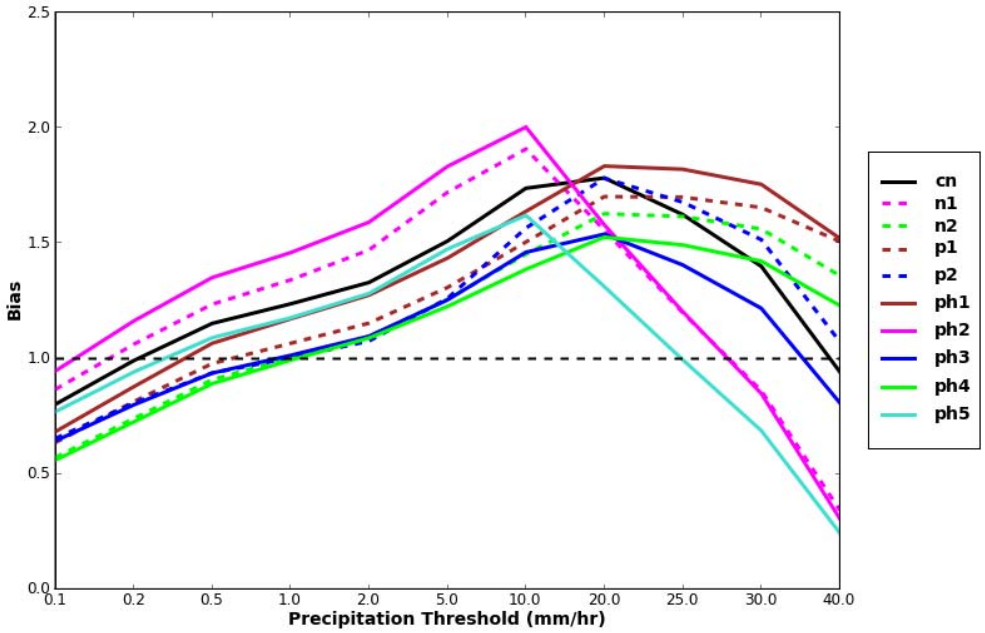

Fig. 6. *Bias as a function of accumulation threshold, aggregated during 1800-0600 UTC (f21-f33) over all days of SE2007.*

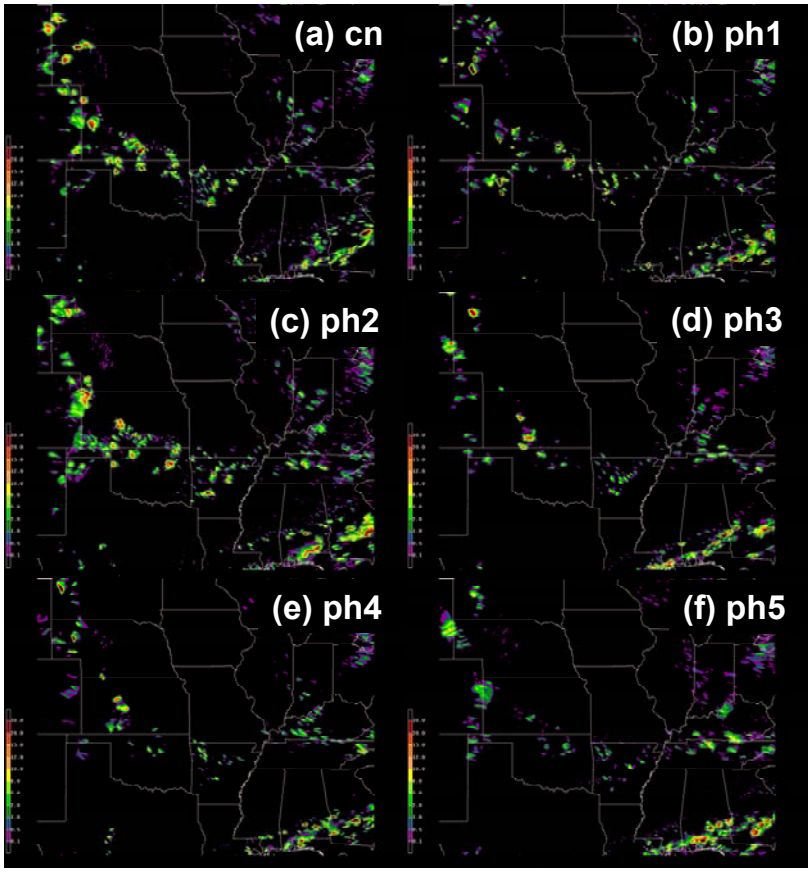

Fig. 7. *One-hour (a) cn (b) ph1, (c) ph2, (d) ph3, (e) ph4, and (f) ph5 forecast accumulated precipitation valid 0000 UTC 06 June. The domain is the same as the verification domain (Fig. 2).*

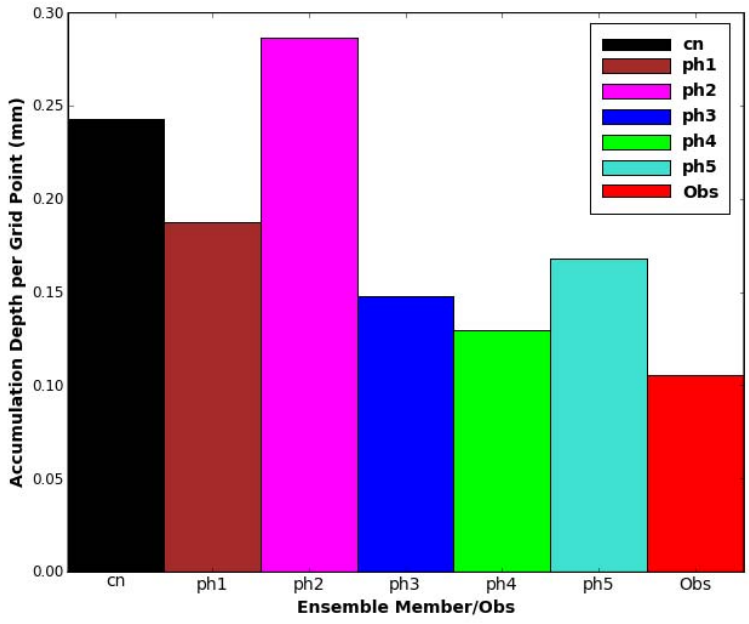

Fig. 8. *Total hourly domain-wide precipitation accumulations valid at the same time and calculated over the same domain as Fig. 7.*

3.6 *Summary*

On average, all the ensemble members produced more precipitation than the observations indicated. However, the bias was not uniform. This spread can be attributed to the different configurations of PBL and microphysics parameterizations used within the ensemble system. Overall, members configured with the same schemes behaved similarly, regardless of whether LBC and IC perturbations were introduced. The MYJ PBL and Ferrier microphysics parameterizations were associated with relatively high precipitation totals. On the other hand, the YSU PBL scheme seemed to be associated with comparatively lesser amounts, either in combination with the WSM 6-class microphysics scheme (p2 and ph3 members) or the Thompson scheme (n2 and ph4 members). These findings indicate that spread can be achieved by varying the physical parameterizations within an ensemble system that uses a single dynamic core. Moreover, documentation of these systematic biases should be valuable to model developers and users.

4. Extracting ensemble guidance: Traditional and new approaches

Two traditional methods of obtaining ensemble guidance are briefly summarized, followed by the presentation of a new post-processing technique to extract ensemble probabilities. Though these methods can be applied to any meteorological field, they are discussed here within the context of precipitation forecasting.

4.1 *Traditional methods*

Traditionally, probabilistic ensemble guidance has been obtained by combining the direct output from each ensemble member. An accumulation threshold (q) is chosen to define an event, and the ensemble probability (EP) of the event occurring at the ith grid box is given by

$$EP_i = \frac{1}{n}\sum_{k=1}^{n} B_{ki} \ , \qquad (1)$$

where n is the number of ensemble members, and

$$B_{ki} = \begin{cases} 1 \ if \ F_{ki} \geq q \\ 0 \ if \ F_{ki} < q \end{cases} \ . \qquad (2)$$

Here, F represents the direct model output field, the first subscript on B and F indicates the contribution from the kth ensemble member, and i ranges from 1 to N, the total number of grid points in the model domain.

Information from each ensemble member can be also considered within the framework of an ensemble mean (M). The value of M at each grid point is given by

$$M_i = \frac{1}{n}\sum_{k=1}^{n} F_{ki} \ . \qquad (3)$$

The ensemble mean is a by-product of the deterministic forecasts of the individual ensemble members and is itself a deterministic forecast. Numerous studies (Stensrud et al. 1999; Wandishin et al. 2001; Bright and Mullen 2002) have, however, shown the ensemble mean performs better than a single deterministic model.

4.2 *A "neighborhood" approach*

The previously described traditional methods utilize direct grid point model output. However, in general, models have little skill at placing features that are comparable in scale to their grid spacing. Thus, as horizontal grid length has decreased in recent years to the sizes of convective-scale features, a variety of methods that incorporate a "neighborhood" around each grid point have been developed to allow for spatial and/or temporal error or uncertainty [reviewed in Ebert (2008)]. As model grid length continues to decrease, these newer methods seem destined to be used more regularly. One of these techniques was developed by Roberts and Lean (2008), and we apply their method to an ensemble (with slight modifications). This approach is outlined below.

4.2.1 *Create binary fields*

As with the formulation of the traditional ensemble mean, precipitation accumulation thresholds were selected to define an event. These thresholds were used to convert the model forecast rainfall fields into binary grids and were applied to each of the n ensemble members individually. Grid boxes with accumulated precipitation ≥ q were assigned a value of 1 and all others a value of 0. For the kth ensemble member, the procedure of binary conversion is given by Equation 2, where i assumes all values between 1 and N.

4.2.2 *Create fractional grids*

After creating binary fields for all the individual members, a radius of influence (r) was specified (e.g., r = 25, 50 km) to construct a "neighborhood" around *each* grid box in the binary fields. All grid points surrounding a given point that fell within the radius were included in the neighborhood. Whereas Roberts and Lean (2008) constructed a square neighborhood around each grid box, a circular neighborhood was used in this study. Essentially, choosing a radius of influence defines a scale over which the model is expected to be accurate, and this scale is applied uniformly in all directions from each grid point.

To generate a fractional value at each point, the number of grid boxes with accumulated precipitation ≥ q within the neighborhood was divided by the total number of boxes within the neighborhood. This fraction can be interpreted as the probability that precipitation will equal or exceed q in the grid box when considering a radius r. In essence, this procedure recognizes the inherent unpredictability at the grid scale and extracts probabilistic information from deterministic grids (Theis et al. 2005).

Fig. 9 illustrates the determination of a neighborhood and computation of a fractional value for a hypothetical model forecast using a radius of influence of 2.5 times the grid spacing.

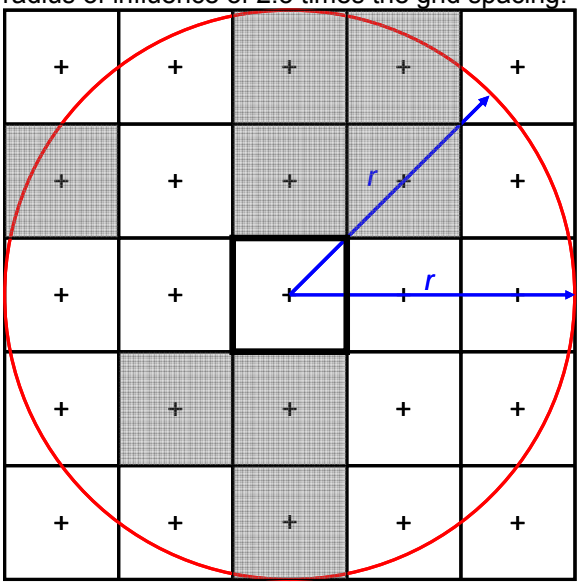

Fig. 9. *Schematic example of neighborhood determination and fractional creation for a model forecast. Precipitation exceeds the accumulation threshold in the shaded boxes, and a radius of 2.5 times the grid length is specified.*

Grid boxes within the radius of influence of the central grid square are included in the neighborhood. Note that by using circular geometry, the corner grid points are excluded, such that the neighborhood consists of 21 boxes. Grid boxes with accumulated precipitation ≥ q are shaded, and these are assigned a value of 1. So, the forecast fraction at the central grid box is 8/21 (eight shaded squares within the neighborhood).

Fig. 10 illustrates fractional grid creation for the control member of the ensemble (cn) using q = 5.0 mm hr^{-1}. The forecast was valid at 0600 UTC 23 May—a lead time of 33 hours—and the model output has been interpolated to the verification grid. The raw, direct model output is shown in Fig. 10a and the binary field in Fig. 10b. Probabilities generated with a radius of influence of 25 km (75 km) are depicted in Fig. 10c (Fig. 10d). Notice that as r increased from 25 to 75 km, probabilities lowered, decreasing from over 90% to 70% (and even lower) over north-central Kansas and extreme southeast South Dakota. Evidently, in this case, as the radius of influence expanded to include more points in the neighborhood, few of these newly-included points contained precipitation accumulations ≥ q. In general, whether probabilities increase or decrease as the radius of influence changes is highly dependent on the meteorological situation. However, for most situations, increasing r typically reduces the sharpness (Roberts and Lean, 2008) and acts as a smoother that reduces probability gradients and usually the probabilities themselves.

When applied to each ensemble member individually, this method generated a set of n probabilistic (fractional) grids. This method can be applied to model output fields other than precipitation, such as radar simulated reflectivity, maximum surface winds, updraft helicity, etc. At this point, the optimal value of r is unknown, and this optimum may vary from model to model. In fact, Roberts (2008) suggests that the optimal radius of influence varies *within* a single model configuration and is a function of lead time.

4.2.3 *Average fractional fields over all ensemble members*

To produce a single product from the n fractional grids, they were all averaged. This step is analogous to producing a traditional ensemble mean (Equation 3), but *fractional* grids were used instead of the direct model output.

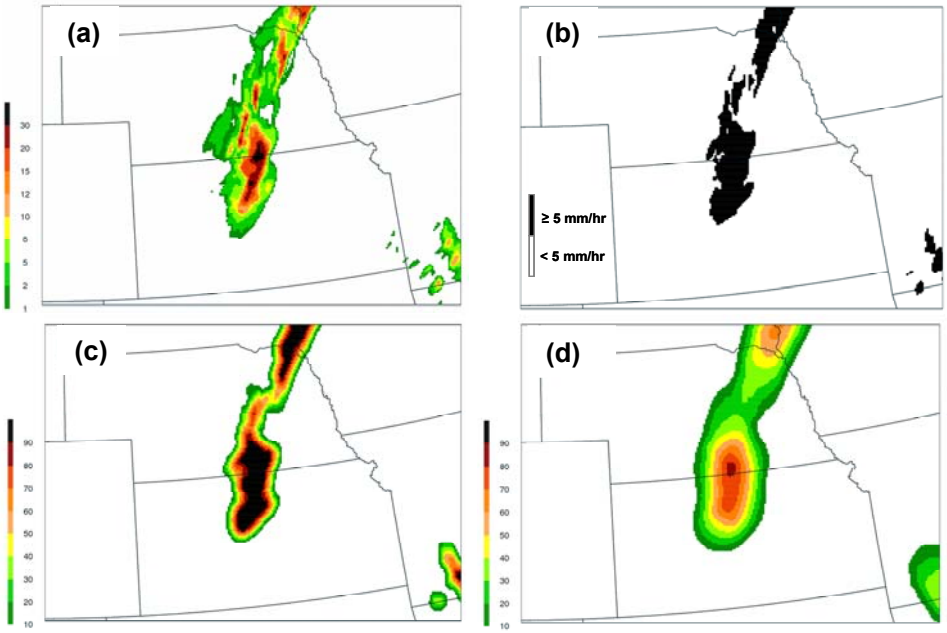

Fig. 10. *(a) Control member (cn) 1-hr accumulated precipitation forecast, (b) binary image of precipitation accumulations exceeding 5.0 mm hr^{-1}, and fractional grids computed from (b) using radii of influence of (c) 25 km and (d) 75 km. All panels are valid 0600 UTC 23 May 2007 and the control member has been projected onto the verification grid.*

That is,

$$NEP_i = \frac{1}{n}\sum_{k=1}^{n} P_{ki} \ , \qquad (4)$$

where P_{ki} is the probability at the *i*th grid point on the *k*th member's probability field, and *NEP* represents the "neighborhood ensemble probability."

4.2.4 *Example*

To demonstrate the character of the traditional and newly described probabilistic products, an example is given for the ensemble forecast valid 2100 UTC 15 May (Fig. 11). The 1.0 mm hr^{-1}accumulation threshold is considered here. As the traditional probability field (Fig. 11d) derived from grid point output, it was very detailed and bordering on noisy. On the other hand, as the NEPs were generated from the neighborhood approach, these fields were relatively smooth, especially as *r* increased to 75 and 125 km (Fig. 11b,c).

In general, the NEP highlighted the same areas as the EP. However, the NEP is more aesthetically pleasing, and it inherently focuses on spatial scales where there is likely to be at least some accuracy. Additionally, it smoothes out any discontinuities in the EP field. The NEP is now objectively verified and compared with the EP.

5. Verification of probabilistic fields

The fractions skill score (FSS) (Roberts et al. 2005; Roberts and Lean 2008) and relative operating characteristic (ROC) (Mason 1982) were adopted to verify the probabilistic guidance considered in this study. To use both of these metrics, it was necessary to project the model forecasts onto the verification grid to directly compare the probability fields with the observations. This interpolation was done before the fractional grids were generated from the individual ensemble members. That is, the direct model output, rather than the fractions, was interpolated to the verification domain.

5.1 *The fractions skill score*

Probabilistic forecasts are commonly evaluated with the Brier Score or Brier Skill Score (Brier 1950) by comparing probabilistic forecasts to a dichotomous observational field. However, the FSS applies the neighborhood approach to the observations in the same way it is applied to model forecast, changing the dichotomous field into an analogous field of

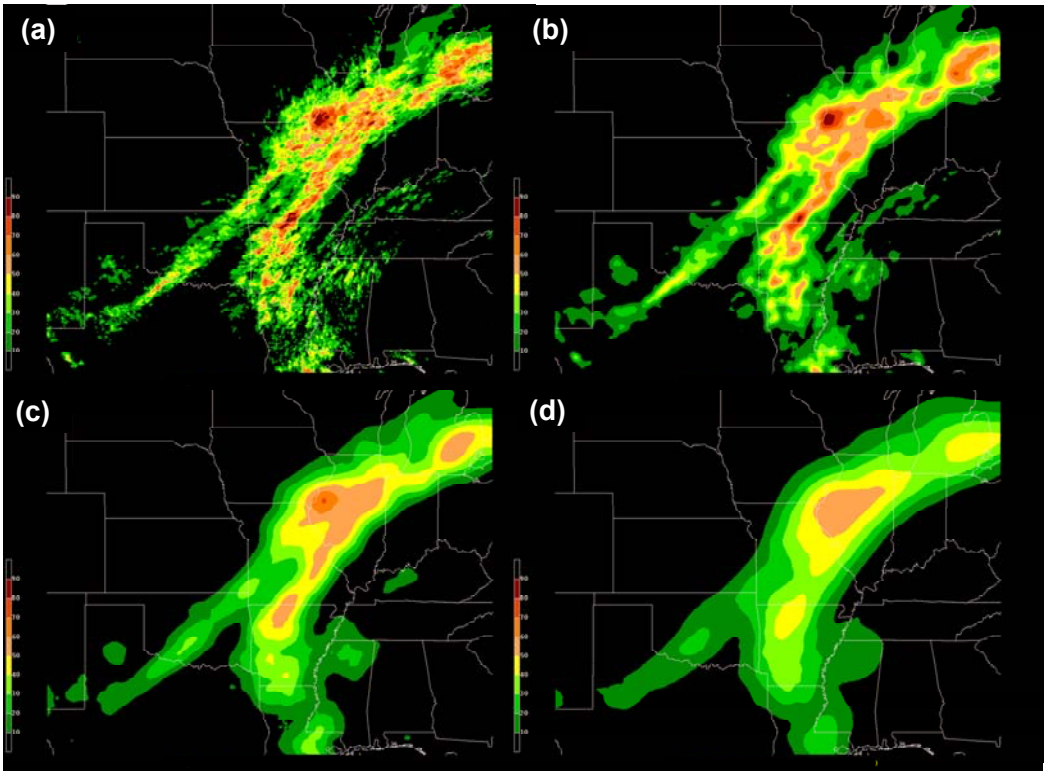

Fig. 11. *Hourly probability forecasts of precipitation meeting or exceeding 1.0 mm from the (a) EP and NEP (see text) using radii of influence of (b) 25 km (c) 75 km (d) 125 km. Forecasts valid 0000 UTC 06 June. The domain is the same as the verification domain (Fig. 2).*

observation-based "probabilities." The two sets of probability fields (forecasts and observations) are then compared directly by the FSS. Whereas Fig. 9 depicts the creation of a fractional grid for just a model forecast, Fig. 12 shows the creation of a fractional grid for this same hypothetical forecast *and* the corresponding observations. Notice that although the model does not forecast precipitation ≥ *q* at the central grid box (quadrant *c* of Table 2, a "miss" using conventional point-by-point verification), when the surrounding neighborhood is considered, the same probability as the observations is achieved (8/21). Therefore, in the context of a radius *r*, this model forecast is considered correct.

As both the model forecast and observational fields have been transformed into fractional grids, the fractional values of the observations and models can be directly compared. A variation on the Brier Score is the Fractions Brier Score (FBS) (Roberts et al. 2005), given by

$$FBS = \frac{1}{N_v}\sum_{i=1}^{N_v}\left(P_{F(i)} - P_{O(i)}\right)^2, \qquad (5)$$

where $P_{F(i)}$ and $P_{O(i)}$ are the fractional (probability) values at the *i*th grid box in the model forecast and observed probability fields, respectively. Here, as objective verification only took place over the verification domain (Fig. 2), *i* ranges from 1 to N_v, the number of points within the verification domain on the verification grid. Note that the FBS compares fractions with fractions and differs from the traditional Brier Score only in that the observational values are allowed to *vary* between 0 and 1.

Like the Brier Score, the FBS is negatively oriented—a score of 0 indicates perfect performance. A larger FBS indicates poor correspondence between the model forecasts and observations. The worst possible (largest) FBS is achieved when there is no overlap of non-zero fractions and is given by

$$FBS_{worst} = \frac{1}{N_v}\left[\sum_{i=1}^{N_v} P_{F(i)}{}^2 + \sum_{i=1}^{N_v} P_{O(i)}{}^2\right]. \qquad (6)$$

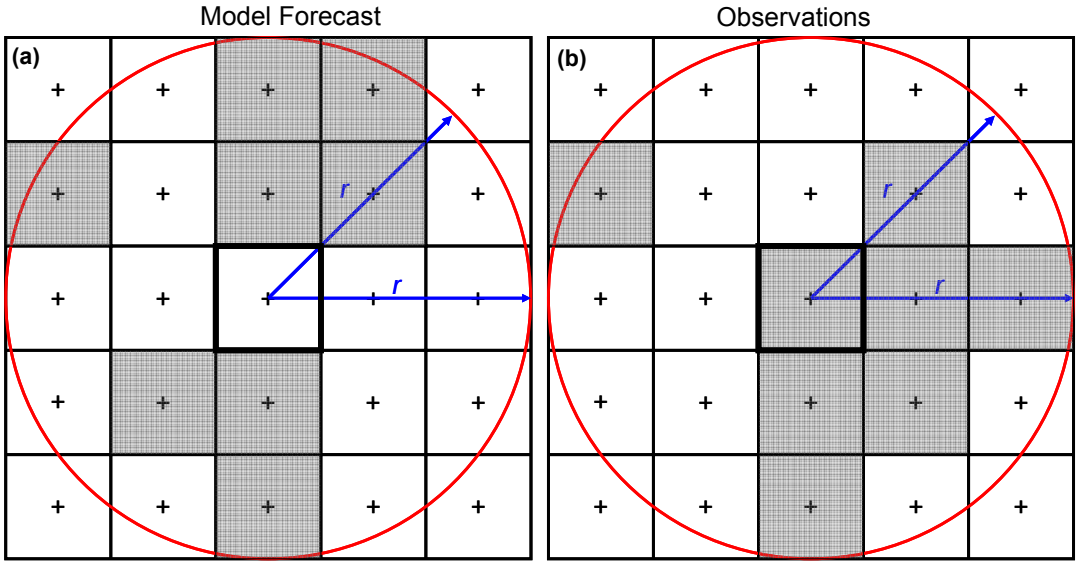

Fig. 12. *Schematic example of neighborhood determination and fractional creation for (a) a model forecast and (b) the corresponding observations. Precipitation exceeds the accumulation threshold in the shaded boxes, and a radius of 2.5 times the grid length is specified.*

On its own, the FBS does not yield much information since it is strongly dependent on the frequency of the event (i.e., grid points with zero precipitation in either the observations or model forecast can dominate the score). However, a skill score (after Murphy and Epstein 1989) can be constructed that compares the FBS to a low-skill reference forecast—FBS_{worst}—and is defined by Roberts (2005) as the fractions skill score (FSS):

$$FSS = 1 - \frac{FBS}{FBS_{worst}}. \qquad (7)$$

The FSS ranges from 0 to 1. A score of 1 is attained for a perfect forecast and a score of 0 indicates no skill. As r expands and the number of grid boxes in the neighborhood increases, the FSS improves as the observed and model probability fields are smoothed and overlap increases, asymptoting to a value of $2B/(B^2 + 1)$, where B is the bias (Roberts and Lean 2008).

5.2 *Verification results*

FSS aggregated over all days of SE2007 during the 1800-0600 UTC (f21-f33) period is shown in Fig. 13 for various hourly absolute precipitation thresholds. The FSS improved as r increased. As q increased, the FSS worsened at all scales, indicating the models had the least skill at predicting heavy precipitation events.

The FSS indicates that at all accumulation thresholds, the NEP produced the most skillful forecasts for r > 25 km. Moreover, the improvement increased with q. This finding indicates that the NEP (Equation 4) improves upon the traditional ensemble probability (Equation 1) for extreme event prediction. Of the individual members, the n2 and p2 members consistently ranked the lowest, while the physics-only members were tightly bunched. FSS as a function of time for q = 5.0 mm hr^{-1} (Fig. 14) indicated NEPs performed the best at nearly all times for all values of r. Also of note is the considerable degradation of the performance of the EP as r increased.

The ROC (Mason 1982) is also used to verify probabilistic forecasting systems. A family of contingency tables (Table 2) can be constructed by selecting different probabilities as yes-no thresholds (i.e., for the 30% threshold, all model grid points with probabilities equal to or

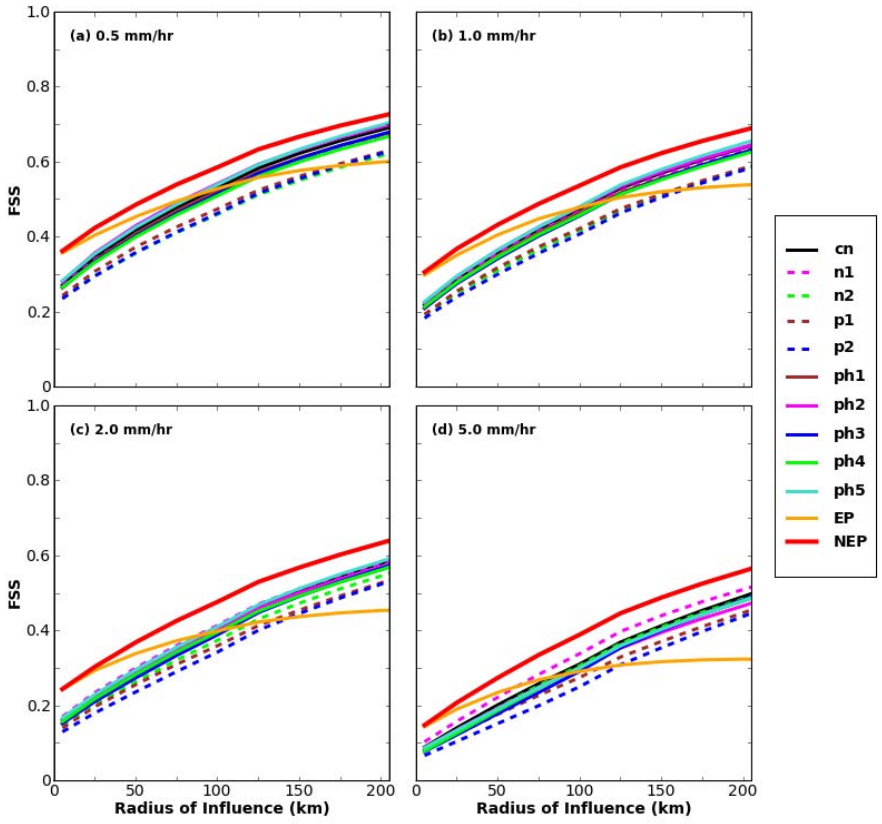

Fig. 13. *Fractions skill score (FSS) as a function of radius of influence, aggregated during 1800-0600 UTC (f21-f33) over all days of SE2007 using accumulation thresholds of (a) 0.2 mm hr^{-1}, (b) 0.5 mm hr^{-1}, (c) 1.0 mm hr^{-1}, (d) 2.0 mm hr^{-1}, (e) 5.0 mm hr^{-1}, and (f) 10.0 mm hr^{-1}. The traditional ensemble probability is denoted as "EP" and the neighborhood probabilities as "NEP."*

greater than 30% are considered to forecast the event). Using the elements of Table 2, the probability of detection [POD = *a/(a+c)*] and probability of false detection [POFD = *b/(b+d)*] can be computed for each probability threshold, and the ROC is formed by plotting POFD against POD over the range of probabilistic thresholds (Fig. 15). The area under this curve is the ROC area, and forecasting systems with a ROC area greater than 0.70 are considered useful (Stensrud and Yussouf 2007).

Using a ROC area of 0.70 as a threshold to determine forecast utility, the EP field was unable to produce useful forecasts when q = 5.0 mm hr^{-1} (Fig 16). However, the NEP field using $r \geq 25$ km provided useful information at all thresholds. As r increased, the ROC area improved. Interestingly, when the neighborhood approach was applied to just the control member using values of $r \geq 75$ km, a greater ROC area equal to or greater than the traditional ensemble probability was achieved. This finding indicates that the neighborhood method applied to an individual ensemble member may provide probabilistic guidance with skill comparable to the EP.

6. Summary and conclusion

During SE2007, CAPS produced convection-allowing 10-member ensemble forecasts. All members used 4 km horizontal grid spacing, ran over the same computational domain, and produced 33 hour forecasts. LBC, IC, and physics perturbations were introduced into 4 of the members while the remaining 6

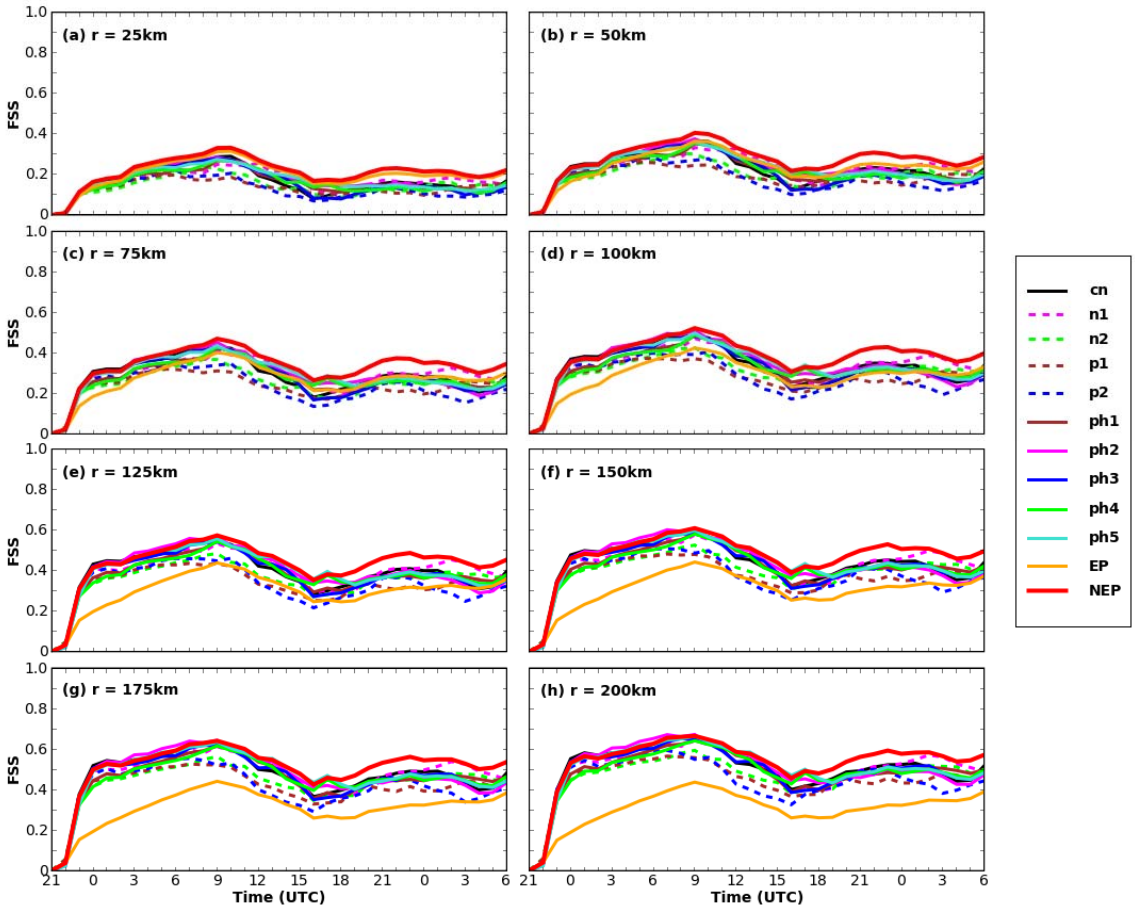

Fig. 14. *Fractions skill score (FSS) using a threshold of 5.0 mm hr^{-1} as a function of time for a radius of influence of (a) 25 km, (b) 50 km, (c) 75 km, (d) 100 km, (e) 125 km, (f) 150 km, (g) 175 km, and (h) 200 km, averaged over all days of SE2007.*

differed solely in terms of PBL and microphysics parameterizations.

WRF-ARW sensitivity to microphysics schemes has been demonstrated using hourly precipitation forecasts. The MYJ PBL and Ferrier microphysics parameterizations were associated with relatively high precipitation totals, while the Thompson microphysics and YSU PBL schemes produced lesser amounts. Documentation of these biases may be useful to model developers and users of NWP systems configured with these parameterization schemes.

Additionally, a new method of extracting probabilistic ensemble guidance by applying a neighborhood approach was presented. This newer approach was found to produce better forecasts, as measured by the FSS, then traditional probabilistic ensemble guidance. Moreover, the NEP resulted in a smoother, more aesthetically pleasing field that unambiguously highlighted those areas most likely to experience extreme events. These findings indicate that simple post-processing can be used to improve high-resolution ensemble forecasts of heavy precipitation and severe weather and provide forecasters with an easy-to-use product. Indeed, it seems that post-processing applied to high-resolution model output offers much promise (see Kain et al. 2008b and references therein).

As high-resolution NWP continues to progress, a central question is whether

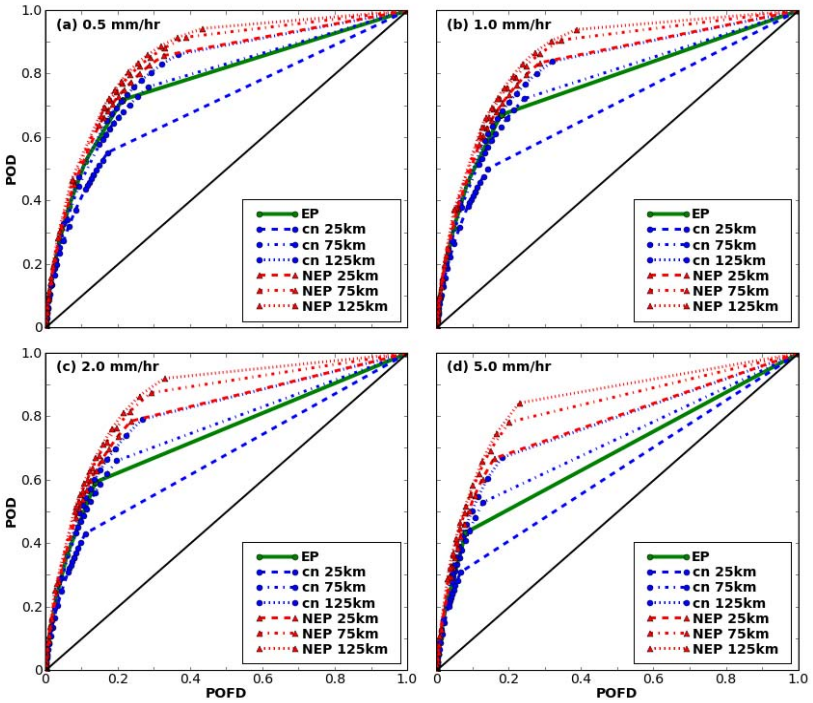

Figure 15. *Relative operating characteristic (ROC) diagrams using data aggregated during 1800-0600 UTC (f21-f33) over all days of SE2007 using accumulation thresholds of (a) 0.5 mm hr^{-1}, (b) 1.0 mm hr^{-1}, (c) 2.0 mm hr^{-1}, and (d) 5.0 mm hr^{-1}*

computer resources should be devoted to single high-resolution forecasts or comparatively coarser-resolution ensemble forecasts. Although there remains debate regarding the current necessity of decreasing grid spacing below 4 km in deterministic models, Kain et al. (2008a) and Schwartz et al. (2008) suggest 4 km WRF-ARW deterministic forecasts provide nearly identical value as 2 km output as guidance for severe storm and heavy precipitation forecasting. Given these conclusions, it seems reasonable that convection-allowing ensembles should continue to be tested and refined, and post-processing options continue to be explored to optimize probabilistic ensemble guidance.

Acknowledgements

Dedicated work by many individuals led to the success of SE2007. At the SPC, HWT operations were made possible by technical

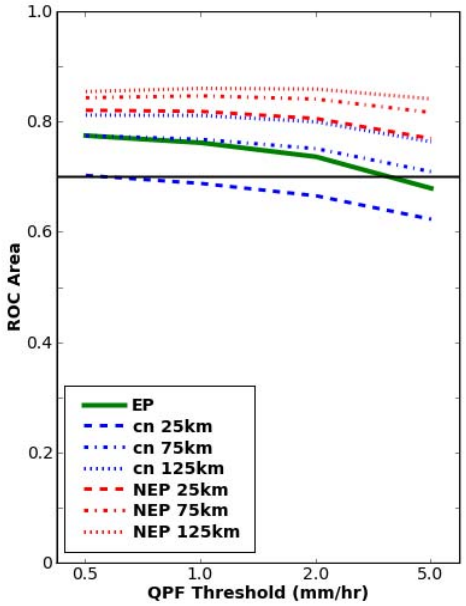

Fig. 16. *ROC areas computed from Fig. 15 using a trapezoidal approximation.*

support from Jay Liang, Gregg Grosshans, Greg Carbin, and Joe Byerly. At the NSSL, Brett Morrow, Steve Fletcher, and Doug Kennedy also provided valuable technical support. We are grateful to Jun Du of NCEP for making available the 2100 UTC NAM analyses. The CAPS forecasts were primarily supported by the NOAA CSTAR program and were performed at the Pittsburgh Supercomputing Center (PSC) supported by NSF. Supplementary support was provided by NSF ITR project LEAD (ATM-0331594). Keith Brewster and Yunheng Wang of CAPS also contributed to the forecast effort. David O'Neal of PSC is thanked for his assistance with the forecasts.

References

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

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

Bright, D.R., and S.L. Mullen, 2002: Short-Range Ensemble Forecasts of Precipitation during the Southwest Monsoon. *Wea. Forecasting*, **17**, 1080–1100.

Brooks, H. E., M. S. Tracton, D. J. Stensrud, G. DiMego, and Z. Toth, 1995: Short-range ensemble forecasting: Report from a workshop, 25–27 July 1994. *Bull. Amer. Meteor. Soc.,* **76,** 1617–1624.

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.

Du, J., and M. S. Tracton, 2001: Implementation of a real-time short-range ensemble forecasting system at NCEP: an update. Preprints, 9th Conference on Mesoscale Processes, Ft. Lauderdale, Florida, Amer. Meteor. Soc., 355-356.

Ebert E. E., 2008: Fuzzy verification of high resolution gridded forecasts: a review and proposed framework. *Meteor. Appl.,* **15**: 53–66.

Ferrier B. S., 1994: A double-moment multiple-phase four-class bulk ice scheme. Part I: Description. *J. Atmos. Sci.*, **51**, 249–280.

Hong, S.-Y., J. Dudhia, and S.-H. Chen, 2004: A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. *Mon. Wea. Rev*., **132**, 103–120.

Janjic, Z. I., 2002: Nonsingular Implementation of the Mellor–Yamada Level 2.5 Scheme in the NCEP Meso model, NCEP Office Note, No. 437, 61 pp.

Kain J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181.

Kain, J. S., S. J. Weiss, D. R. Bright, M. E. Baldwin, J. J. Levit, G. W. Carbin, C. S. Schwartz, M. L. Weisman, K. K. Droegemeier, D. B. Weber, and K. W. Thomas, 2008a: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931-952.

Kain, J.S., S. J. Weiss, S. R. Dembek, J. J. Levit, D. R. Bright, J. L. Case, M. C. Coniglio, A. R. Dean, R. A. Sobash, and C. S. Schwartz, 2008b: Severe-weather forecast guidance from the first generation of large domain convection-allowing models: Challenges and opportunities. *Preprints, 24th Conference on Severe Local Storms*, Amer. Meteor. Soc., Savannah, GA. CD-ROM 12.1.

Kong, F., M. Xue, 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. *22nd Conf. Wea. Anal. Forecasting/18th Conf.*

Num. Wea. Pred., Salt Lake City, Utah, Amer. Meteor. Soc., CDROM 3B.2.

Kong, F., M. Xue, K. K. Droegemeier, K. Thomas, and Y. Wang, 2008: Real-time storm-scale ensemble forecast experiment. Preprints, 9th WRF User's Workshop, NCAR Center Green Campus, 23-27 June 2008, Paper 7.3.

Lin, Y. and K. E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: development and applications. Preprints, 19th Conf. on Hydrology, American Meteorological Society, San Diego, CA, 9-13 January 2005, Paper 1.2.

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

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys. Space Phys*., 20, 851–875.

Murphy A. H., and E. S. Epstein, 1989: Skill scores and correlation coefficients in model verification. *Mon. Wea. Rev.*, **117**, 572–581.

Murphy, A.H., 1991: Probabilities, odds, and forecasts of rare events. *Wea. Forecasting*, **6**, 302–307.

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

Noh, Y., W.G. Cheon, S.-Y. Hong, and S. Raasch, 2003: Improvement of the K-profile model for the planetary boundary layer based on large eddy simulation data. *Bound.-Layer Meteor.*, **107**, 401-427.

Roberts, N. M., 2005: An investigation of the ability of a storm scale configuration of the Met Office NWP model to predict flood-producing rainfall. UK Met Office Technical Report No. 455. (Available from http://www.metoffice.gov.uk/research/nwp/publications/papers/technical_reports/2005/FRTR455/FRTR455.pdf)

Roberts, N.M., and H.W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev*, **136**, 78–97.

Roberts N., 2008: Assessing the spatial and temporal variation in skill of precipitation forecasts from an NWP model. *Meteor. Appl*., **15**, 163–169.

Schwartz, C. S., J. S. Kain, S. J. Weiss, D. R. Bright, M. Xue, F. Kong, K. W. 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 Conference on Severe Local Storms*, Amer. Meteor. Soc., Savannah, GA. CD-ROM 13A.6.

Seo, D. J., 1998: Real-time estimation of rainfall fields using radar rainfall and rain gauge data. *J. Hydrol.*, **208**, 37-52.

Skamarock, W.C., J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, 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].

Stensrud, D.J., H.E. Brooks, J. Du, M.S. Tracton, and E. Rogers, 1999: Using Ensembles for Short-Range Forecasting. *Mon. Wea. Rev.*, **127**, 433–446.

Stensrud, D.J., and N. Yussouf, 2007: Reliable Probabilistic Quantitative Precipitation Forecasts from a Short-Range Ensemble Forecasting System. *Wea. Forecasting*, **22**, 3–17.

Theis S. E., A. Hense, and U. Damrath, 2005: Probabilistic precipitation forecasts from a deterministic model: A pragmatic approach. *Meteor. Appl.*, **12**, 257–268.

Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Mon. Wea. Rev.,* **132**, 519–542.

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.

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.

Xue, M., F. Kong, D. Weber, K. W. Thomas, Y. Wang, K. Brewster, K. K. Droegemeier, J. S. K. S. J. Weiss, D. R. Bright, M. S. Wandishin, M. C. Coniglio, and J. Du, 2007: CAPS realtime storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 spring experiment. *22nd Conf. Wea. Anal. Forecasting/18th Conf. Num. Wea. Pred.*, Salt Lake City, Utah, Amer. Meteor. Soc., CDROM 3B.1.