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*

January 2009

Submitted to *Weather and Forecasting*

**Corresponding author address:*

Craig Schwartz
University of Oklahoma
120 David L. Boren Blvd. Suite 5642
Norman, Oklahoma 73072
Email: craig.schwartz@ou.edu

Abstract

During the 2007 NOAA Hazardous Weather Testbed Spring Experiment, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma produced a daily 10-member 4 km horizontal resolution ensemble forecast covering approximately three-fourths of the continental United States. Each member used the WRF-ARW core, was initialized at 2100 UTC, ran for 33 hours, and resolved convection explicitly. Different initial condition (IC), lateral boundary condition (LBC), and physics perturbations were introduced in four of the ten ensemble members, while the remaining six members used identical ICs and LBCs, differing only in terms of microphysics (MP) and planetary boundary layer (PBL) schemes. This study focuses on precipitation forecasts from the ensemble.

The ensemble forecasts revealed WRF-ARW sensitivity to MP and PBL schemes. For example, over the 7-week Experiment, the Mellor-Yamada-Janjic PBL and Ferrier MP parameterizations were associated with relatively high precipitation totals, while members configured with the Thompson MP or Yonsei University PBL scheme produced comparatively less precipitation. Additionally, different approaches for generating probabilistic ensemble guidance were explored. Specifically, a "neighborhood" approach is described and shown to considerably enhance probabilistic forecasts for precipitation when combined with traditional techniques of producing ensemble probability fields.

These results have important implications for convection-allowing guidance in both deterministic and ensemble frameworks and are relevant to both operational forecasters and modelers.

1. Introduction

Throughout the history of numerical weather prediction (NWP), computer resources have advanced to enable 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 weather forecast guidance tools. Additionally, these experiments have revealed that running the WRF model at 4 km without CP does not result in grossly unrealistic forecasts, even though a 4 km grid is too coarse to fully capture convective scale 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 the Environmental Modeling Center (EMC) of NCEP, and future plans call for an expansion of the suite of convection-allowing forecasts (G. DiMego, NCEP/EMC, personal communication, 2008).

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 deterministic model solutions. But when convection-allowing models are used to predict intense localized features such as thunderstorms, even small displacement errors can produce large errors in amplitude at individual grid points. In recognition of this problem, post-processing and verification methods have been developed that relax the

requirement that deterministic 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). These "neighborhood" approaches have also been used to generate probabilistic information from deterministic grids. Theis et al. (2005) suggested that a neighborhood approach could be combined with traditional methods of producing probabilistic forecasts, a strategy that is explored herein.

Probabilistic predictions are, by nature, superior to deterministic forecasts at providing guidance for rare events, such as severe thunderstorms or heavy precipitation (Murphy 1991). The probabilistic format allows forecasters to quantify uncertainty such that their forecasts can reflect their best judgments and, perhaps more importantly, allows users to make better decisions as compared to those made with yes-no forecasts (Murphy 1993). Numerical guidance for probabilistic forecasts is commonly derived from an ensemble forecasting system, where an ensemble is comprised of a suite of individual forecasts, each generated from a unique combination 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 (MP) and turbulence.

Ideally, all ensemble members are assumed to be equally likely of representing the "true" condition of the atmosphere at initialization, and thus, have an equal chance of producing the best forecast at a later time. Usually, initial fields differ only slightly, and

forecasts from the members are quite similar at early time steps. However, owing to the chaotic nature of the atmosphere, these differences may amplify with time, such that by the end of the model integration, different ensemble members can arrive at wildly different solutions. The spread of the ensemble members (in terms of standard deviation) 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, information from all the members can be averaged into a mean deterministic field. As errors of different members tend to cancel in the averaging process, this ensemble mean consistently performs better than any of the individual members. Furthermore, numerous studies (e.g., Stensrud et al. 1999; Wandishin et al. 2001; Hou et al. 2001; Bright and Mullen 2002) have shown that an ensemble system, in terms of its ensemble mean, 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 somewhat. Following the recommendation of participants at 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 32-45 km grid spacing (Du et al. 2006) and is run four times daily, starting at 0300, 0900, 1500, and 2100 UTC. Variations in physical parameterizations, dynamic cores, ICs, and LBCs are used to create forecast diversity (Du et al. 2006).

Given the benefits of ensemble forecasting and previous successes of convection-allowing 4 km WRF deterministic forecasts, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma, supported by a pilot three-year NOAA (National Oceanic and Atmospheric Administration) Collaborative Science, Technology, and Applied Research (CSTAR) project, contributed large domain, realtime, 10-member, 4 km convection-allowing ensemble forecasts to the 2007 NOAA Hazardous Weather Testbed Spring Experiment[1] (hereafter SE2007). Variations in ICs, LBCs, and physical parameterizations were used to achieve diversity within the ensemble. 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 goal of this study is to examine the output from the CAPS ensemble for two main purposes. First, it examines forecasts from the different ensemble members and identifies Advanced Research WRF (WRF-ARW; Skamarock et al. 2005) model sensitivities to MP and planetary boundary layer (PBL) parameterizations. Second, a new method of extracting probabilistic ensemble guidance is presented. This technique, suggested by Theis et al. (2005), combines a "neighborhood" approach with more traditional methods of processing ensemble output. 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

[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.

probabilistic forecasts are presented in section 4 and these forecasts are verified in section 5 prior to concluding.

2. Experimental design

a. Model configurations

On each of the ~ 35 days of SE2007, CAPS produced a 10-member ensemble forecast 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 2.2 of the WRF-ARW dynamic core (Skamarock et al. 2005), represented convection explicitly (no CP), 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).

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 Mesoscale model (NAM; Black 1994) (J. Du, NCEP/EMC, 2007, personal communication). 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). LBCs for the LBC/IC members were provided by the four WRF perturbed members [two WRF-ARW (Skamarock et al. 2005) and two WRF-NMM (Nonhydrostatic Mesoscale Model; Janjic et al. 2001; Janjic 2003)] of the 2100 UTC NCEP SREF, and the IC perturbations were extracted from these same four WRF members of the 2100 UTC NCEP SREF. 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 MP and PBL parameterizations. Therefore, comparison of their output allows a robust assessment of WRF-ARW sensitivity to PBL and MP parameterizations in a variety of weather regimes. Additional details on the ensemble configurations can be found in Xue et al. (2007) and Kong et al. (2007).

b. 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 Stage II precipitation grids produced hourly at NCEP (Lin and Mitchell 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. Attention was focused on the 1800-0600 UTC (f21-f33) period to examine the utility of the ensemble as next-day forecast 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."

3. Precipitation sensitivity to physical parameterizations

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

a. 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 maximum. The specific cause of this high bias has not been identified. However, more detailed examinations of selected 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 initialization led to the very high bias (Kong et al. 2008). Nonetheless, as all members were subjected to the same constraints and impacted equally, 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 (Fig. 3). This separation suggests that the combination of PBL and MP 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 MP 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 MP 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 MP 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 MP parameterizations.

b. 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 of 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 (Ferrier and MYJ) 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.

c. Precipitation percentiles

A climatology of precipitation accumulations was constructed by compiling the hourly precipitation forecasts in each grid box within the verification domain on the native grids over all days of SE2007 between 1800-0600 UTC (f21-f33). The values were ranked, and accumulation percentiles (y) (e.g., 95th percentile) were chosen to determine absolute hourly precipitation values (q_y) corresponding to the 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 the yth percentile. This procedure was performed separately for each ensemble 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}).

d. Precipitation bias

To quantitatively determine the biases of individual members, the standard 2 x 2 contingency table for dichotomous (yes-no) events was used (Table 2). The frequency bias (B) is simply the ratio of the areal coverage of forecasts of the event to the coverage

of observed events and can be easily computed from the contingency table [$B = (a+b)/(a+c)$]. For a given value of q, $B > 1$ 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 (Ferrier and MYJ) members overpredicting the most for $q \leq 10.0$ mm hr^{-1}. At thresholds > 10.0 mm hr^{-1}, the n1 and ph2 biases interestingly plummet, leaving the ph1 and p1 members with the highest biases (both configured with Thompson MP and MYJ PBL schemes).

e. Case Study

Figure 7 shows hourly precipitation output from the physics-only members on their native grids over the verification domain. The ensemble was initialized at 2100 UTC 04 June, and the forecast was valid 0000 UTC 06 June—a 27 hour forecast. 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 areal coverages of the ph1 (Thompson and MYJ) and ph3 (WSM6 and YSU) members lied between those of the other two pairs.

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 were in fairly good agreement. However, there were disagreements regarding precipitation intensity over Kentucky and Tennessee, with the ph2 and ph5 members producing the heaviest rainfall.

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 at that time.

f. 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 MP parameterizations used within the ensemble system. Members configured with the same physics schemes behaved similarly, on average, regardless of whether LBC and IC perturbations were introduced.

The MYJ PBL and Ferrier MP parameterizations were associated with relatively high precipitation totals. In contrast, the YSU PBL scheme was associated with comparatively lesser amounts, either in combination with the WSM6 MP scheme (p2 and ph3 members) or the Thompson scheme (n2 and ph4 members). These findings indicate that spread in precipitation 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 WRF-ARW developers and users.

4. Extracting forecast probabilities: Traditional and new approaches

A widely used approach for computing forecast probabilities (FPs) from an ensemble is summarized, followed by discussion of a lesser known post-processing method for extracting FPs from single deterministic predictions. Then, a simple strategy for combining these two approaches is presented. Though these methods can be applied to any meteorological field, they are discussed here within the context of precipitation forecasting.

a. Traditional method

In an uncalibrated ensemble system, all members are assumed to have equal skill, when averaged over many forecasts. Under this assumption, members are weighted equally and the ensemble-based probability can be thought of as the average of the binary probabilities (BPs) for individual members, where the BPs are simply 1 or 0 at a given grid point, depending on the occurrence (1) or non-occurrence (0) of an event; an "event" typically means exceedance of a specified threshold. For example, in the context of

precipitation forecasting, an accumulation threshold (q) is chosen to define an event, and the individual grid-point BPs are given by

$$BP_{ki} = \begin{cases} 1 \ if \ F_{ki} \geq q \\ 0 \ if \ F_{ki} < q \end{cases}, \tag{1}$$

where F is the raw accumulation of precipitation at the grid point, the subscript k refers to the kth ensemble member, and the subscript i denotes the ith grid point. Here, i ranges from 1 to N, the total number of grid points in the computational domain. After a binary grid is generated for each ensemble member according to Eq. (1), the traditional ensemble probability (EP) at the ith grid point can be computed as a mean value according to

$$EP_i = \frac{1}{n} \sum_{k=1}^{n} BP_{ki} , \tag{2}$$

where n is the number of members in the ensemble.

b. A "neighborhood" approach

The above method for computing EP_i utilizes raw model output at individual grid points. 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. Although neighborhood methods are used most often for verification purposes

[e.g., Roberts and Lean (2008)], here they are employed to create non-binary FPs from individual deterministic forecasts [e.g., Theis et al. (2005)].

Application of the neighborhood approach to generate FPs begins with a binary grid, created in accordance with Eq. (1), from a deterministic forecast (e.g., one of the ensemble members). Next, following Roberts and Lean (2008), a radius of influence (r) is specified (e.g., r = 25, 50 km) to construct a "neighborhood" around *each* grid box in the binary field[2]. All grid points surrounding a given point that fall within the radius are included in the neighborhood. Whereas Roberts and Lean (2008) constructed a square neighborhood around each grid box, a circular neighborhood is 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 non-binary FP value at each point, the number of grid boxes with accumulated precipitation $\geq q$ (i.e., the number of 1s in the binary field) within the neighborhood is divided by the total number of boxes within the neighborhood. This "neighborhood probability" (NP) at the ith grid point on the kth ensemble member's grid can be expressed as

$$NP_{ki} = \frac{1}{N_b} \sum_{m=1}^{N_b} BP_{km} \,, \tag{3}$$

where N_b is the number of grid points within the neighborhood of grid point i. Although for a given value of r the number of points within the neighborhood (N_b) is the same for

[2] 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.

each of the N grid boxes, "hidden" in Eq. (3) is the fact that the ith grid box specifies a *unique* set of N_b points on the BP grid that comprise the neighborhood. That is, the specific grid boxes on the BP grid that are used to compute NP_i are *different* for each of the N grid boxes.

Figure 9 illustrates the determination of a neighborhood and computation of NP_i for a hypothetical model forecast using a radius of influence of 2.5 times the grid spacing. 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 $\geq q$ are shaded, and these are assigned a value of 1. In this example, the event occurs in 8 out of 21 grid boxes, so $NP_i = 0.38$, or 38%, at the central grid box.

Figure 10 illustrates the impact of this procedure using a forecast from the control member of the ensemble (cn). The forecast was valid at 0600 UTC 23 May—a lead time of 33 hours—and the model output is displayed on the verification grid. The raw precipitation forecast is shown in Fig. 10a and the binary field (the BP_i field) corresponding to $q = 5.0$ mm hr^{-1} is plotted in Fig. 10b. Note that the binary field can also be considered the NP field generated using $r = 0$ km. As r is increased to 25 km (Fig. 10c) and then 75 km (Fig. 10d), the character of the NP field changes substantially. Specifically, as r increases from 25 to 75 km, maximum probabilities decrease 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 expands to include more points in the neighborhood, few of these newly-included points contain precipitation accumulations $\geq q$. In general, whether NP_i values increase or decrease as the radius of

influence changes is highly dependent on the meteorological situation. However, for most situations, increasing *r* reduces the sharpness (Roberts and Lean 2008) and acts as a smoother that reduces gradients and magnitudes in the NP field.

c. Combining traditional and neighborhood approaches

When the neighborhood method is applied to each ensemble member individually, a set of *n* NP_i grids are generated. These grids are directly analogous to the BP_i grids, but instead of being limited to values of 0 or 1, the point values comprise a continuum from 0 to 1. Just as the BP_i values are averaged over all members to produce traditional ensemble probabilities (EP_i), the NP_i values can be combined to produce a new neighborhood ensemble probability (NEP) according to

$$NEP_i = \frac{1}{n}\sum_{k=1}^{n} NP_{ki} \ . \tag{4}$$

To demonstrate the characters of the traditional and neighborhood probabilistic products, an example is given for the ensemble forecast valid 2100 UTC 15 May, focusing on the 1.0 mm hr^{-1} accumulation threshold (Fig. 11). The traditional probability field (i.e., the EP) is very detailed and rather noisy (Fig. 11a). On the other hand, the NEPs become increasingly smooth as *r* increases from 25 to 125 km (Fig. 11b-e).

In general, the NEP field highlights the same areas as the EP. However, the smoother NEP field is more aesthetically pleasing and 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 fields are now objectively verified and compared with corresponding EP fields.

5. Verification of probabilistic fields

The fractions skill score (FSS) (Roberts 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.

a. 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, one can apply the neighborhood approach to the observations in the same way it is applied to model forecasts, changing the dichotomous observational field into an analogous field of observation-based fractions (or probabilities). The two sets of fraction fields (forecasts and observations) then can be compared directly. Whereas Fig. 9 depicts the creation of a fraction grid for just a model forecast, Fig. 12 shows the creation of a fraction grid for this same hypothetical forecast *and* the corresponding observations. Notice that although the model does not forecast precipitation $\geq q$ 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 = 0.38). Therefore, in the context of a radius *r*, this model forecast is considered correct.

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

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

where $NP_{F(i)}$ and $NP_{O(i)}$ are the neighborhood probabilities at the ith grid box in the model forecast and observed fraction 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

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

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):

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

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 frequency bias (Roberts and Lean 2008).

b. 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. As q increased, the FSS worsened at all scales, indicating the models had less skill at predicting heavier precipitation.

The FSS indicates that at all accumulation thresholds, the NEP produced the most skillful forecasts for $r > 25$ km. Moreover, the advantage of the NEP increased with increasing q. This finding indicates that the NEP [Eq. (3)] improves upon the traditional ensemble probability [Eq. (2)], especially 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.

In a sense, the EP was handicapped in the computation of the FSS because this field did not change as a function of r, while the verifying field (and all the other FPs) did. However, the advantage for the NEP is also evident with other performance measures, such as the relative operating characteristic (ROC; Mason 1982). For the

ROC, a family of contingency tables (Table 2) is constructed for the probabilistic forecasts by selecting different probabilities as yes-no thresholds (i.e., for the 30% threshold, all model grid points with probabilities equal to or greater than 30% are considered to forecast the event) in conjunction with a binary observation field. 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). In this study, a trapezoidal approximation was used to find the area under the ROC curve.

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. Additionally, ROC areas improved as the NEP was computed using progressively larger values of r. This finding further indicates that the NEP improves upon the EP and that the improvement increases as the event becomes more extreme.

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 differed solely in terms of PBL and MP parameterizations.

WRF-ARW sensitivity to MP and PBL schemes was demonstrated using hourly precipitation forecasts. The MYJ PBL and Ferrier MP parameterizations were associated with relatively high precipitation totals, while the YSU PBL scheme, in combination with the Thompson and WSM6 MP parameterizations, produced lesser amounts. Documentation of these biases should be useful to users and developers of the WRF-ARW model. However, users of other NWP systems should be cautious in interpreting these results since the parameterizations examined here were subjected to varying levels of calibration in the WRF-ARW.

In addition to the determination of physics sensitivities, a new method of extracting probabilistic guidance from an ensemble was presented. This method applied a "neighborhood" concept to an ensemble and was found to produce more skillful probabilistic guidance, as measured by the FSS and ROC area, than traditional ensemble-derived probabilistic guidance. Moreover, the neighborhood ensemble probability resulted in smoother, more aesthetically pleasing fields that focused on the spatial scales over which the models were more likely to be accurate. 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 effective and 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 computer resources should be devoted to single high-resolution deterministic forecasts or comparatively coarse-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 next-day guidance for severe storm and heavy precipitation forecasting. Given these conclusions, it seems reasonable that convection-allowing ensembles and post-processing options should continue to be tested, refined, and 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 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 and the NCEP SREF output. The CAPS forecasts were primarily supported by the NOAA CSTAR program and were performed at the PSC supported by the 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, FL, Amer. Meteor. Soc., 355-356. [Available online at http://www.emc.ncep.noaa.gov/mmb/SREF/srefupdate_2001.pdf]

Du, J., J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H. Chuang, 2006: New Dimension of NCEP Short-Range Ensemble Forecasting (SREF) System: Inclusion of WRF Members. Preprints, *WMO Expert Team Meeting on Ensemble Prediction System,* Exeter, UK, Feb. 6-10, 2006. [Available online at http://www.emc.ncep.noaa.gov/mmb/SREF/WMO06_full.pdf]

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.

Hou, D., E. Kalnay, and K. K. Droegemeier, 2001: Objective verification of the SAMEX ’98 ensemble forecasts. *Mon. Wea. Rev.*, **129,** 73-91.

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

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.

Janjic, Z. I., 2003: A nonhydrostatic model based on a new approach. *Meteorology and Atmospheric Physics*, **82**, 271–285.

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*, Savannah, GA, Amer. Meteor. Soc., 12.1. [Available online at http://ams.confex.com/ams/pdfpapers/141723.pdf.]

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. Preprints, *22nd Conf. Wea. Anal. Forecasting/18th Conf. Num. Wea. Pred.*, Salt Lake City, UT, Amer. Meteor. Soc., 3B.2. [Available online at http://ams.confex.com/ams/pdfpapers/124667.pdf.]

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, 7.3. [Available online at http://www.mmm.ucar.edu/wrf/users/workshops/WS2008/presentations/7-3.pdf.]

Lin, Y. and K.E. Mitchell, 2005: The NCEP Stage II/IV hourly precipitation analyses: development and applications. Preprints, *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2. [Available online at [http://ams.confex.com/ams/pdfpapers/83847.pdf.]

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

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.

Schwartz, C. S., J. S. Kain, S. J. Weiss, M. Xue, D. R. Bright, F. Kong, K. W. Thomas, J. J. Levit and M. C. Coniglio, 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*, Savannah, GA, Amer. Meteor. Soc., P10.3. [Available online at http://ams.confex.com/ams/pdfpapers/142052.pdf.]

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. Preprints, *22nd Conf. Wea. Anal. Forecasting/18th Conf. Num. Wea. Pred.*, Salt Lake City, UT, Amer. Meteor. Soc., 3B.1. [Available online at http://ams.confex.com/ams/pdfpapers/124587.pdf.]

Table Captions

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 were used. NAMa and NAMf refer to NAM analyses and forecasts, respectively.

Table 2. Standard 2 x 2 contingency table for dichotomous events.

Figure Captions

Fig. 1. Model domain of the CAPS ensemble forecasts.

Fig. 2. Verification domain used for model climatology.

Fig. 3. Total precipitation over the domain aggregated over all days of SE2007, normalized by number of grid boxes, calculated for each ensemble member on its 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.

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).

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 (27-hr forecast, initialized at 2100 UTC 04 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.

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.

Fig. 10. (a) Control member (cn) 1-hr accumulated precipitation forecast (mm hr^{-1}), (b) binary image (i.e., a BP grid) of precipitation accumulations exceeding 5.0 mm

hr^{-1}, and NP grids computed from (b) using radii of influence of (c) 25 km and (d) 75 km. All panels are valid 0600 UTC 23 May and the control member has been projected onto the verification grid.

Fig. 11. Hourly probability forecasts of precipitation meeting or exceeding 1.0 mm using the (a) EP and NEP (see text) with radii of influence of (b) 25 km, (c) 50 km, (d) 75 km, and (e) 125 km. The observed precipitation is shown in (f). Both the model fields and observations are valid 2100 UTC 15 May. The domain is the same as the verification domain (Fig. 2).

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.

Fig. 13. Fractions skill score (FSS) as a function of radius of influence (r), 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 probability as NEP. Probabilities for the individual members of the ensemble were computed as NPs. Note that the EP field does not change as a function of r, while the others do.

Fig. 14. Fractions skill score (FSS) plotted as a function of forecast hour for a fixed accumulation-rate threshold of 5.0 mm hr^{-1} and radii 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.

Fig. 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}.

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

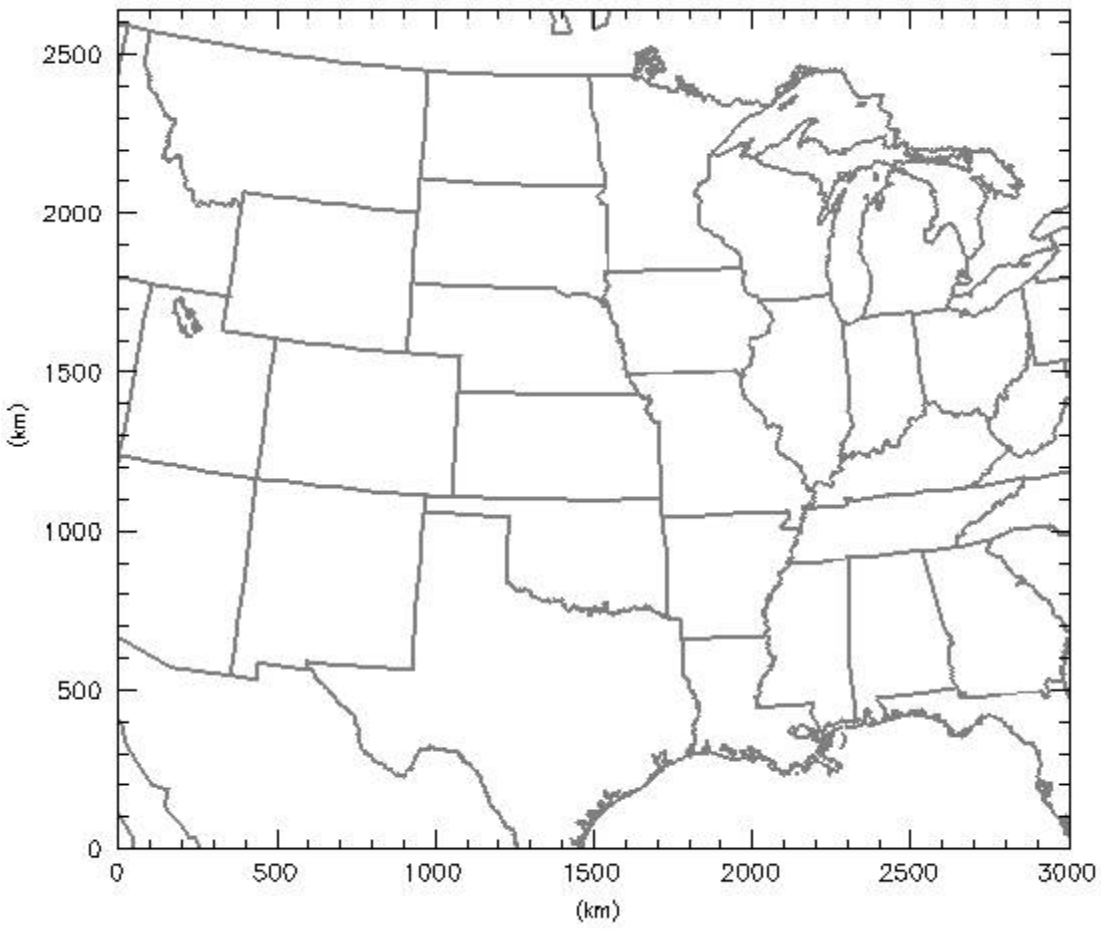

Fig. 1. Model domain of the CAPS ensemble forecasts.

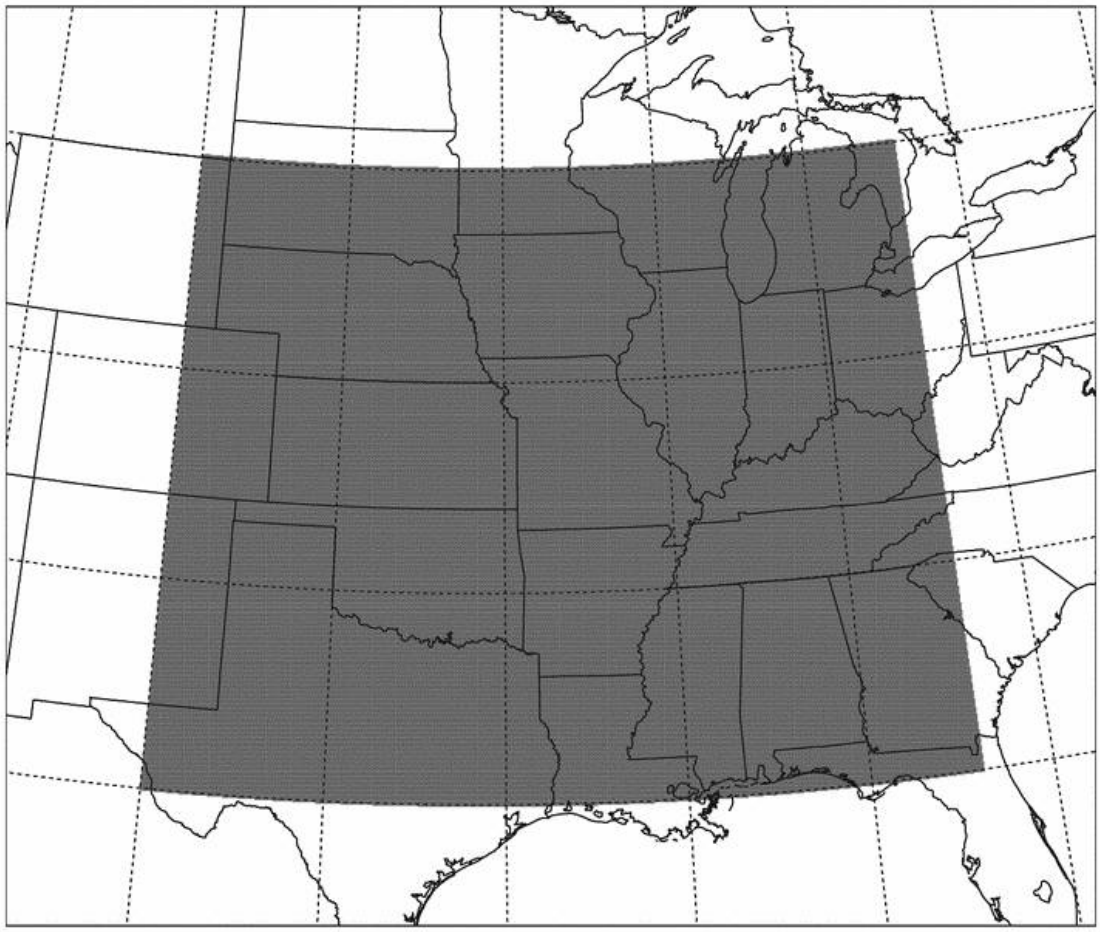

Fig. 2. Verification domain used for model climatology.

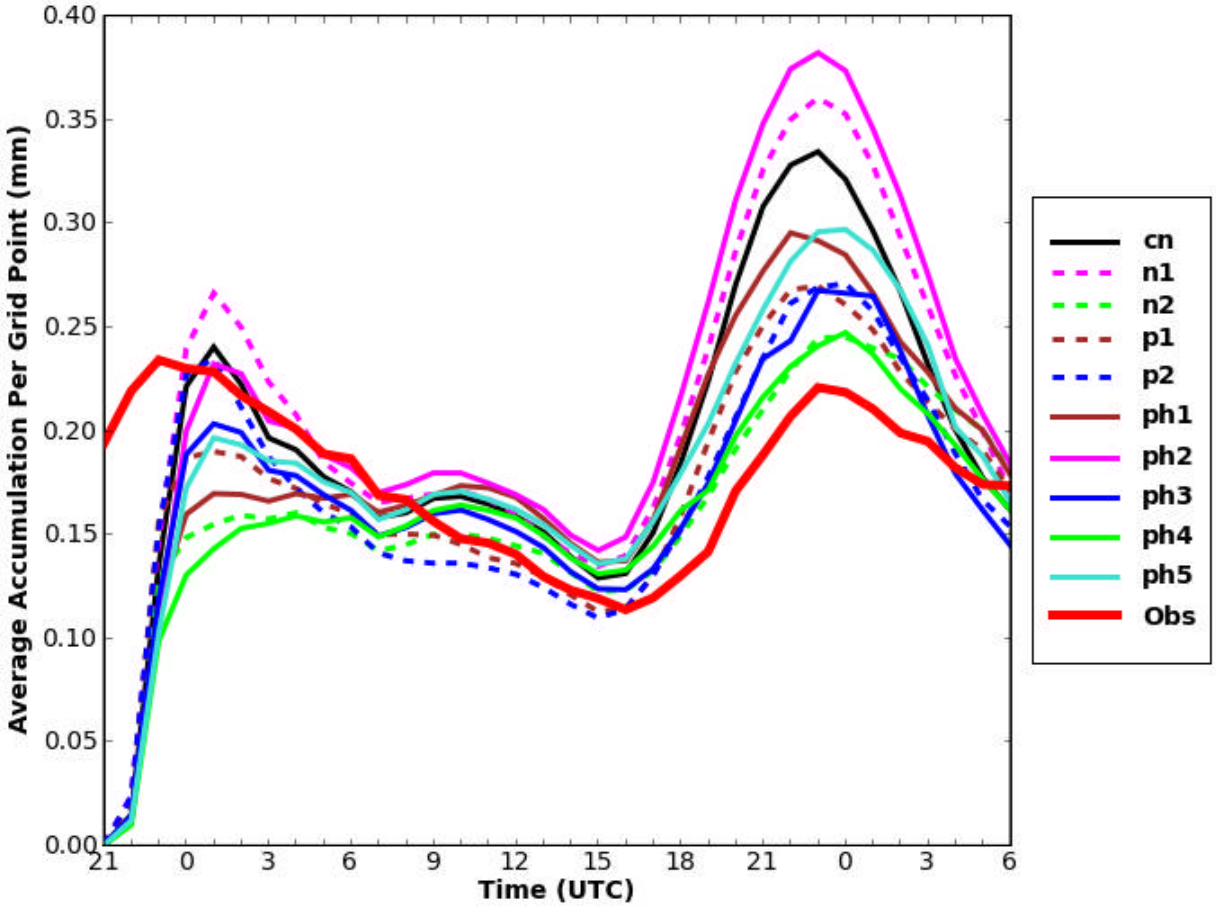

Fig. 3. Total precipitation over the domain aggregated over all days of SE2007, normalized by number of grid boxes, calculated for each ensemble member on its native grid.

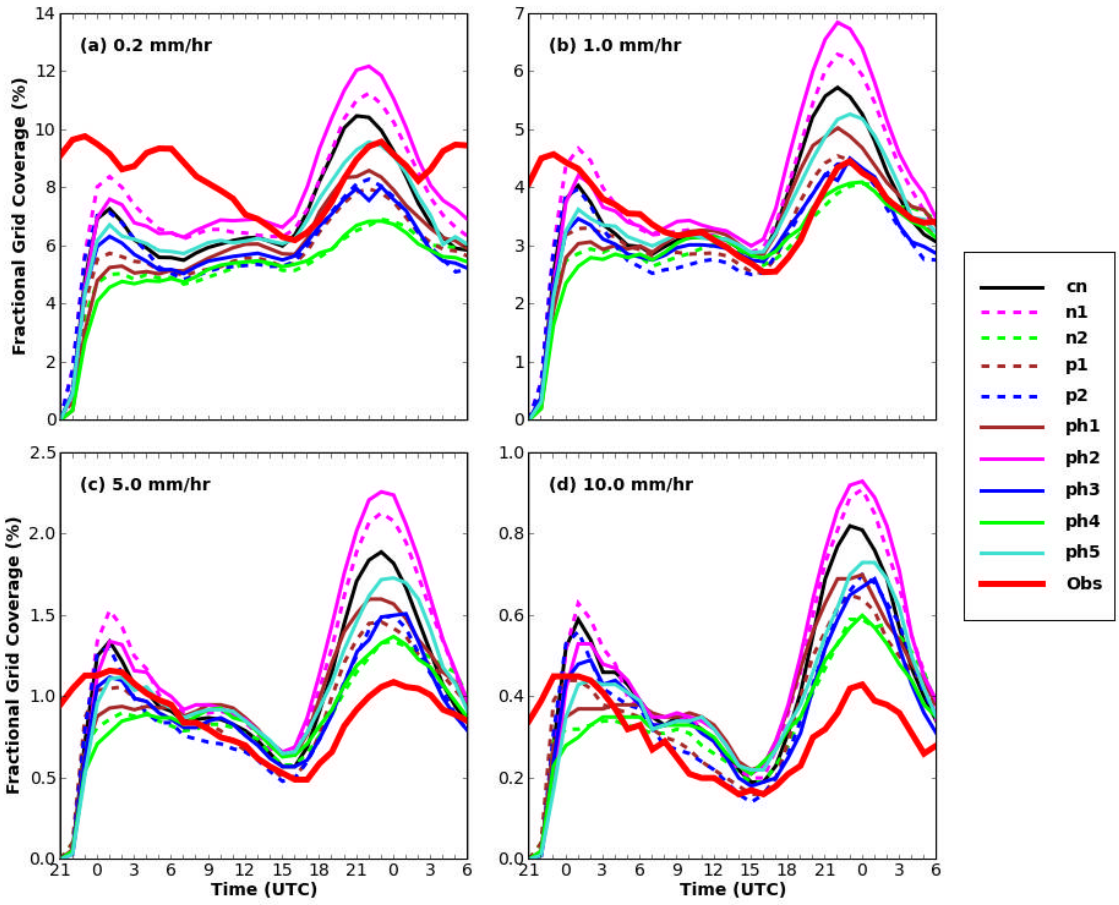

Fig. 4. Fractional grid coverage of hourly precipitation exceeding (a) 0.2 mm hr^{-1}, (b) 1.0 mm hr^{-1}, (c) 5.0 mm hr^{-1}, and (d) 10.0 mm hr^{-1} as a function of time, averaged over all days of SE2007, calculated on each member's native grid.

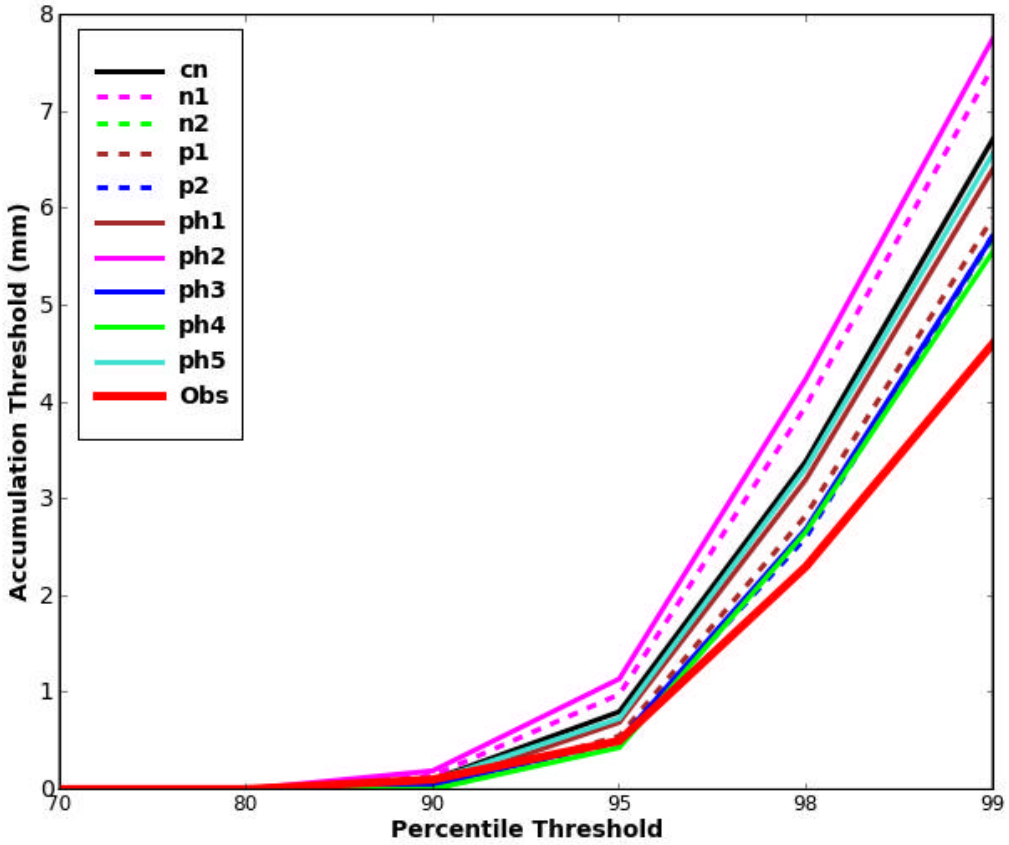

Fig. 5. Precipitation climatology: Percentiles calculated on each member's native grid aggregated between 1800-0600 UTC (f21-f33) over all days of SE2007 (see text).

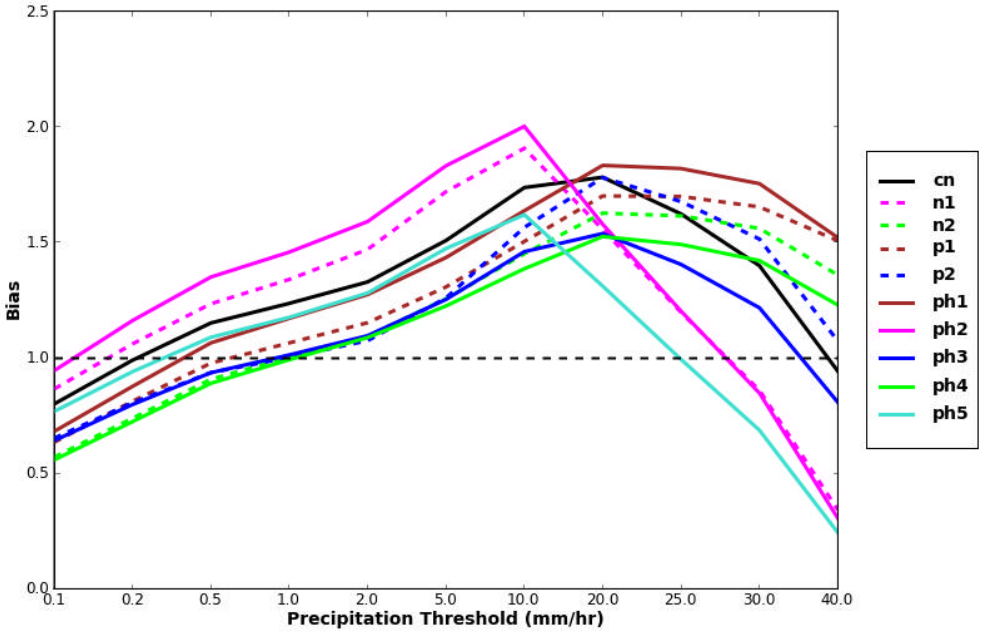

Fig. 6. Bias as a function of accumulation threshold, aggregated during 1800-0600 UTC (f21-f33) over all days of SE2007.

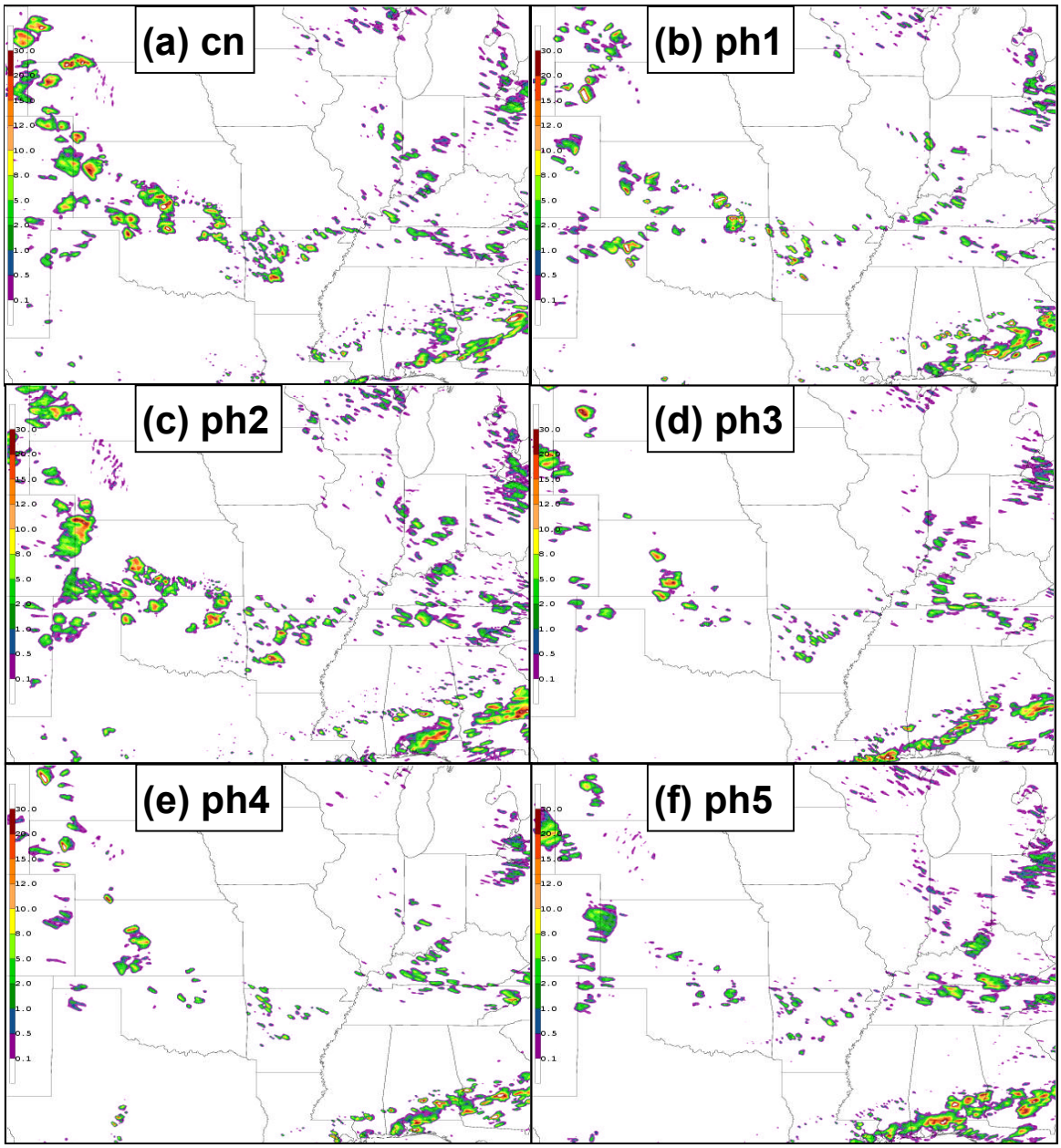

Fig. 7. One-hour (a) cn, (b) ph1, (c) ph2, (d) ph3, (e) ph4, and (f) ph5 forecast accumulated precipitation valid 0000 UTC 06 June (27-hr forecast, initialized at 2100 UTC 04 June). The domain is the same as the verification domain (Fig. 2).

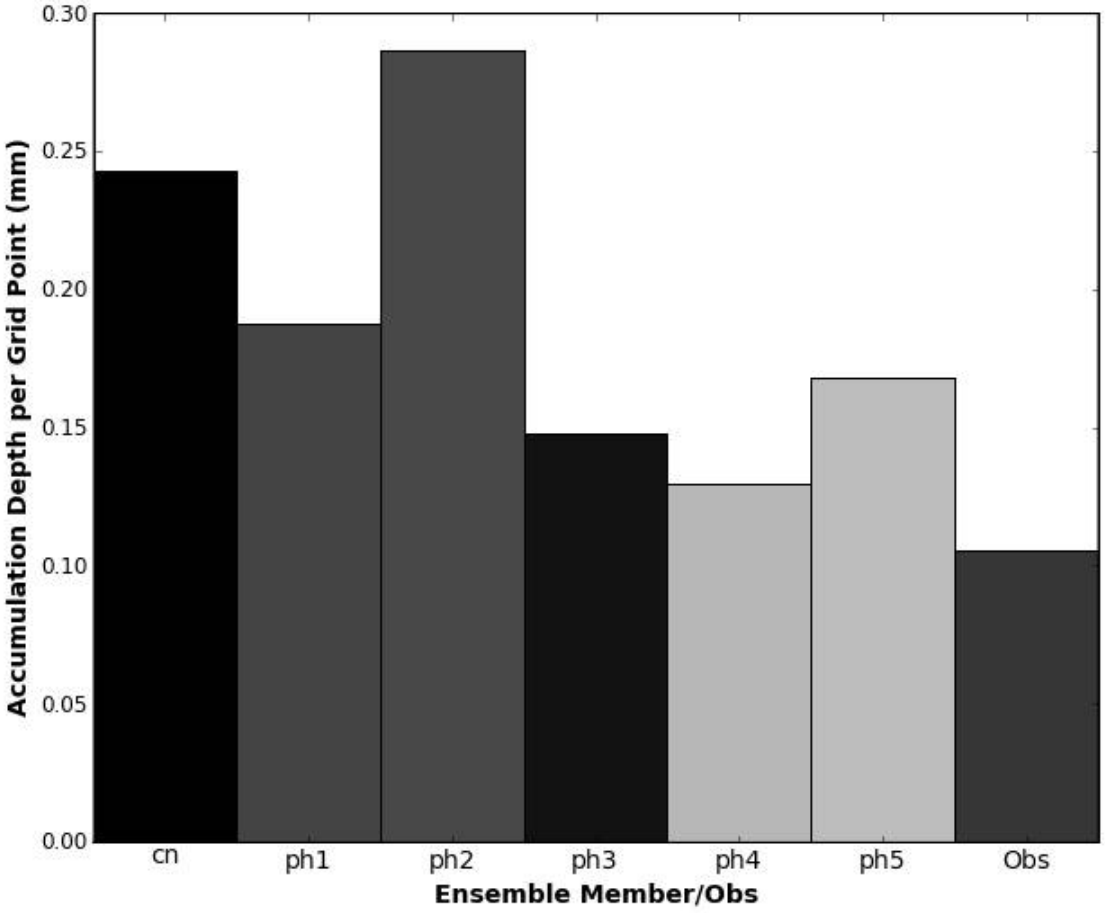

Fig. 8. Total hourly domain-wide precipitation accumulations valid at the same time and calculated over the same domain as Fig. 7.

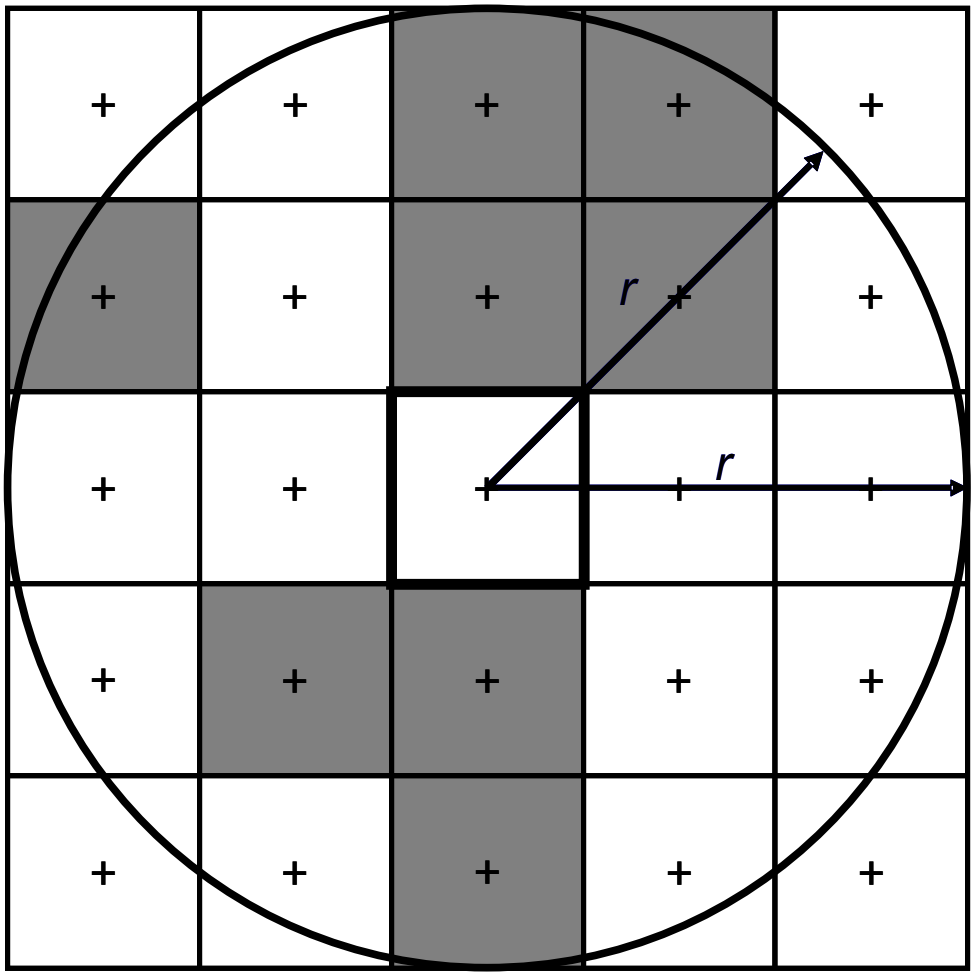

Fig. 9. Schematic example of neighborhood determination and fractional creation for a model forecast. Precipitation exceeds the accumulation threshold in the shaded boxes, and a radius of 2.5 times the grid length is specified.

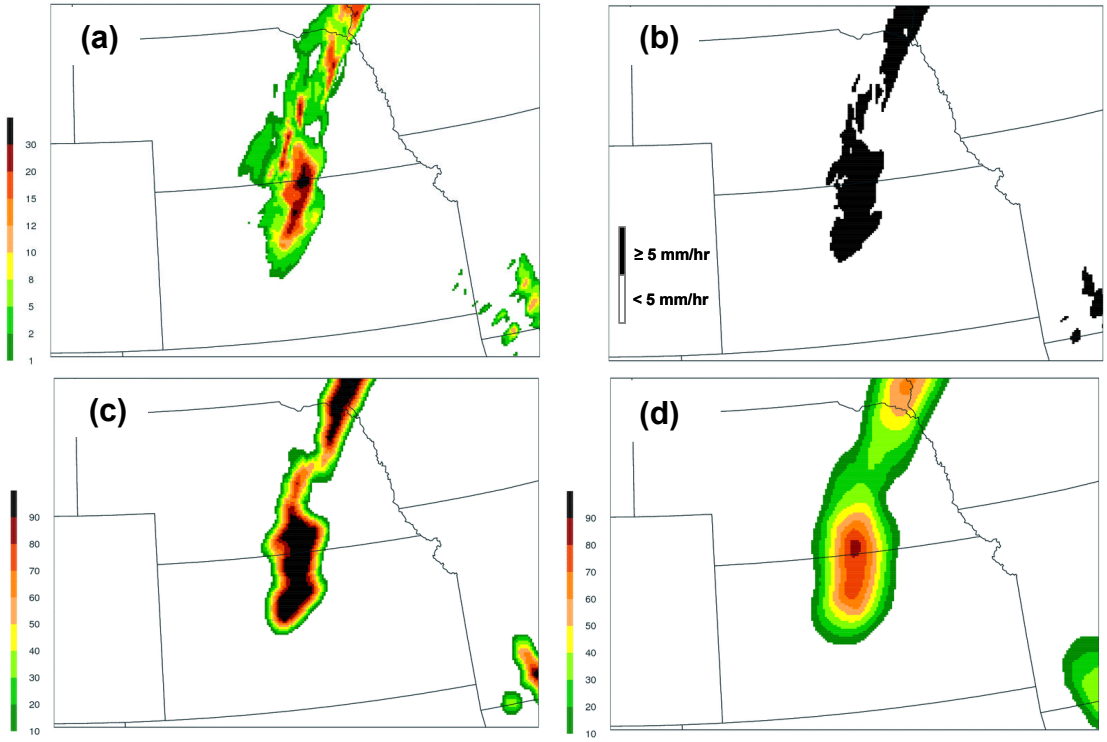

Fig. 10. (a) Control member (cn) 1-hr accumulated precipitation forecast (mm hr^{-1}), (b) binary image (i.e., a BP grid) of precipitation accumulations exceeding 5.0 mm hr^{-1}, and NP grids computed from (b) using radii of influence of (c) 25 km and (d) 75 km. All panels are valid 0600 UTC 23 May and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

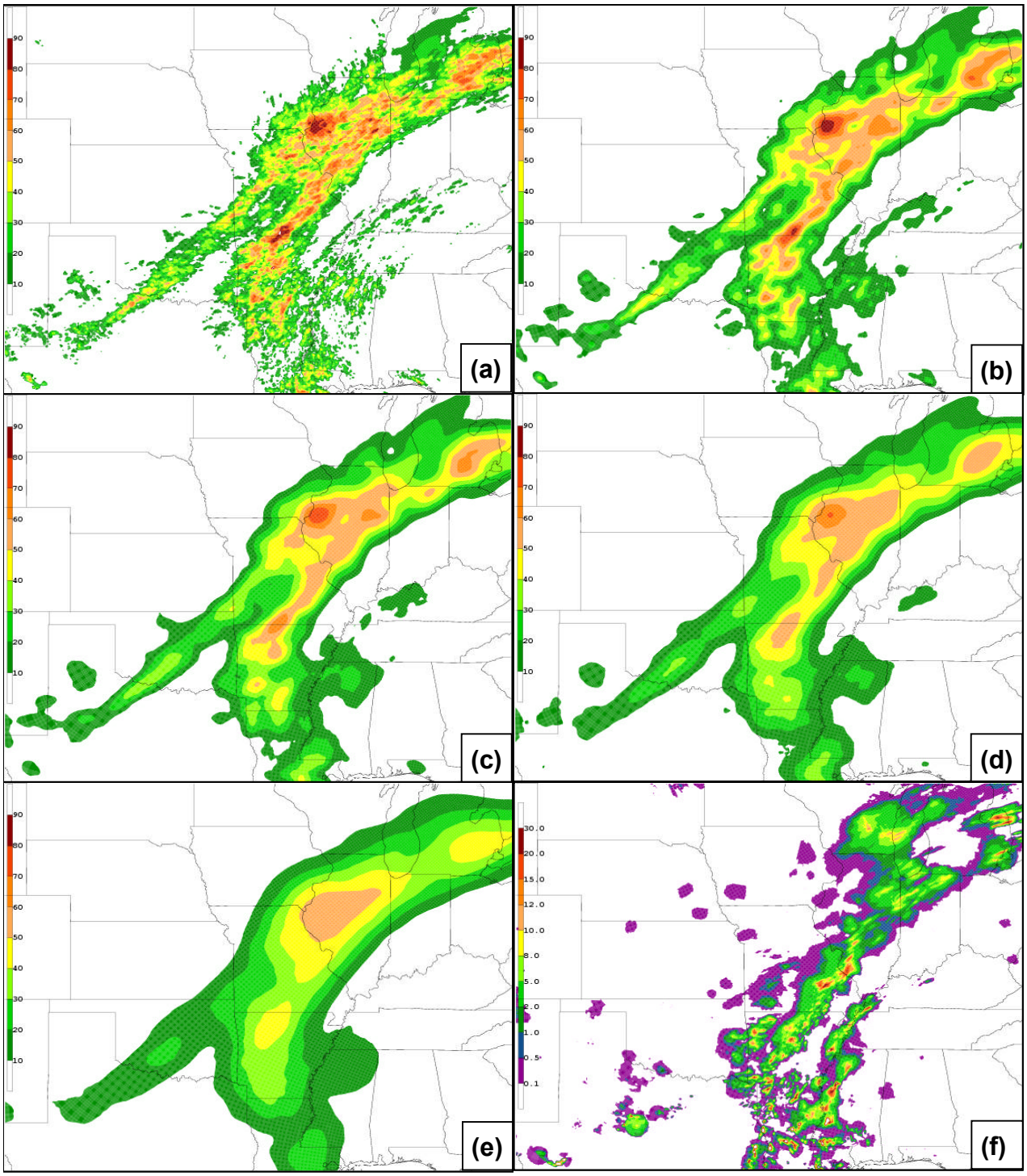

Fig. 11. Hourly probability forecasts of precipitation meeting or exceeding 1.0 mm using the (a) EP and NEP (see text) with radii of influence of (b) 25 km, (c) 50 km, (d) 75 km, and (e) 125 km. The observed precipitation is shown in (f). Both the model fields and observations are valid 2100 UTC 15 May. The domain is the same as the verification domain (Fig. 2).

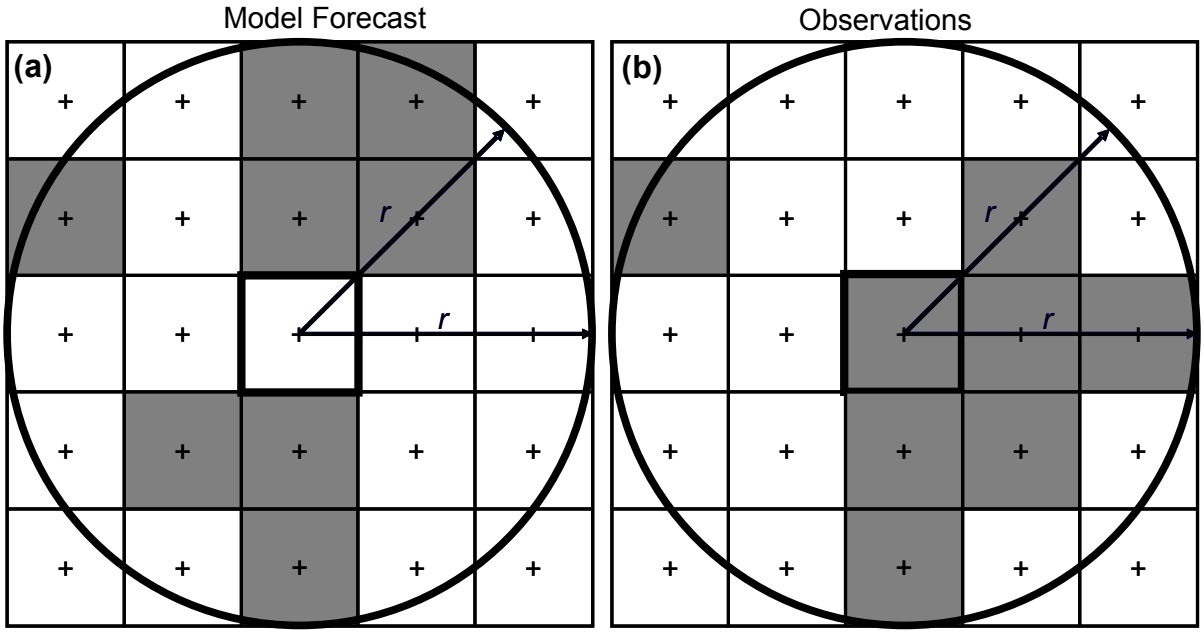

Fig. 12. Schematic example of neighborhood determination and fractional creation for (a) a model forecast and (b) the corresponding observations. Precipitation exceeds the accumulation threshold in the shaded boxes, and a radius of 2.5 times the grid length is specified.

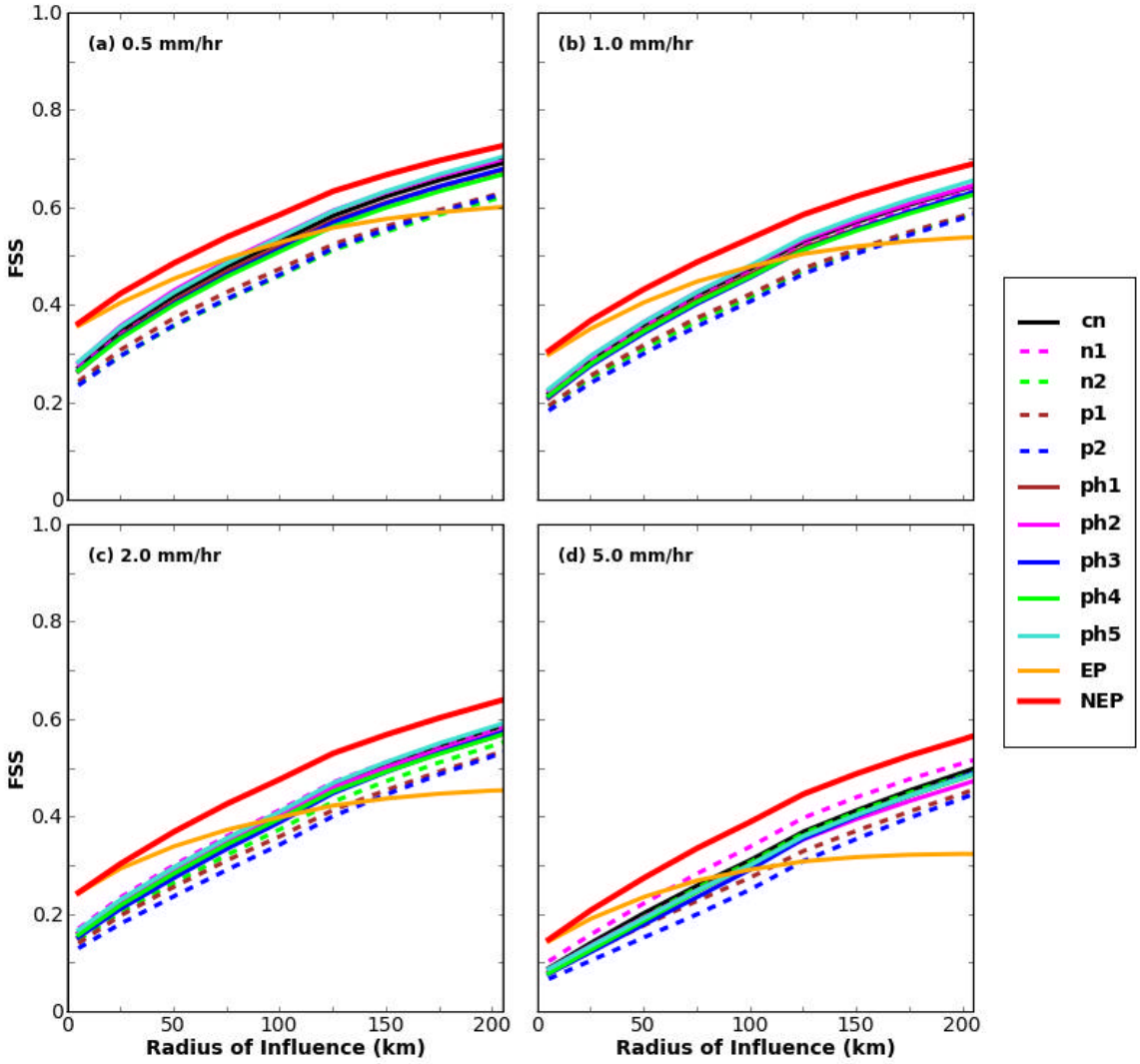

Fig. 13. Fractions skill score (FSS) as a function of radius of influence (r), aggregated during 1800-0600 UTC (f21-f33) over all days of SE2007 using accumulation thresholds of (a) 0.2 mm hr^{-1}, (b) 0.5 mm hr^{-1}, (c) 1.0 mm hr^{-1}, (d) 2.0 mm hr^{-1}, (e) 5.0 mm hr^{-1}, and (f) 10.0 mm hr^{-1}. The traditional ensemble probability is denoted as EP and the neighborhood probability as NEP. Probabilities for the individual members of the ensemble were computed as NPs. Note that the EP field does not change as a function of r, while the others do.

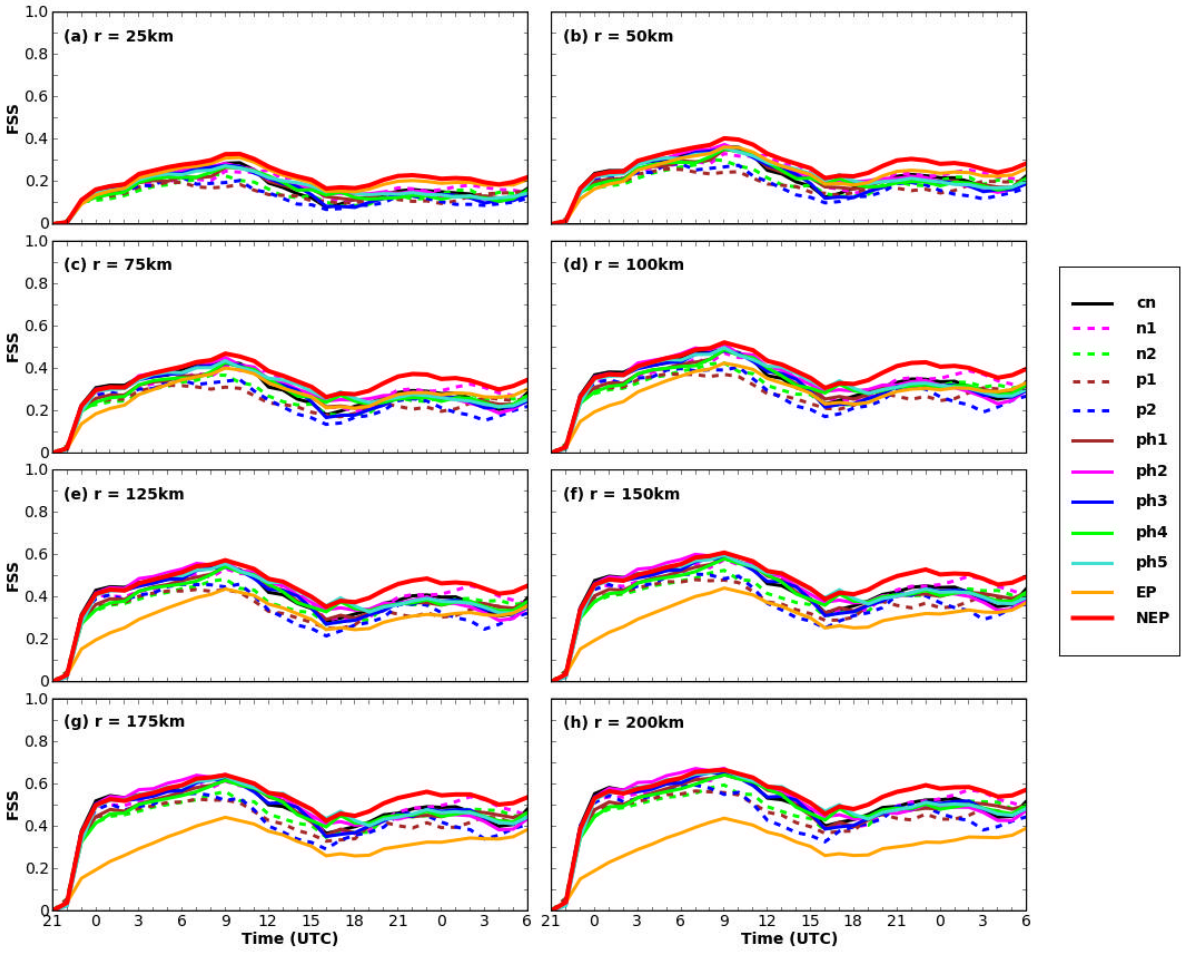

Fig. 14. Fractions skill score (FSS) plotted as a function of forecast hour for a fixed accumulation-rate threshold of 5.0 mm hr^{-1} and radii of influence of (a) 25 km, (b) 50 km, (c) 75 km, (d) 100 km, (e) 125 km, (f) 150 km, (g) 175 km, and (h) 200 km, averaged over all days of SE2007.

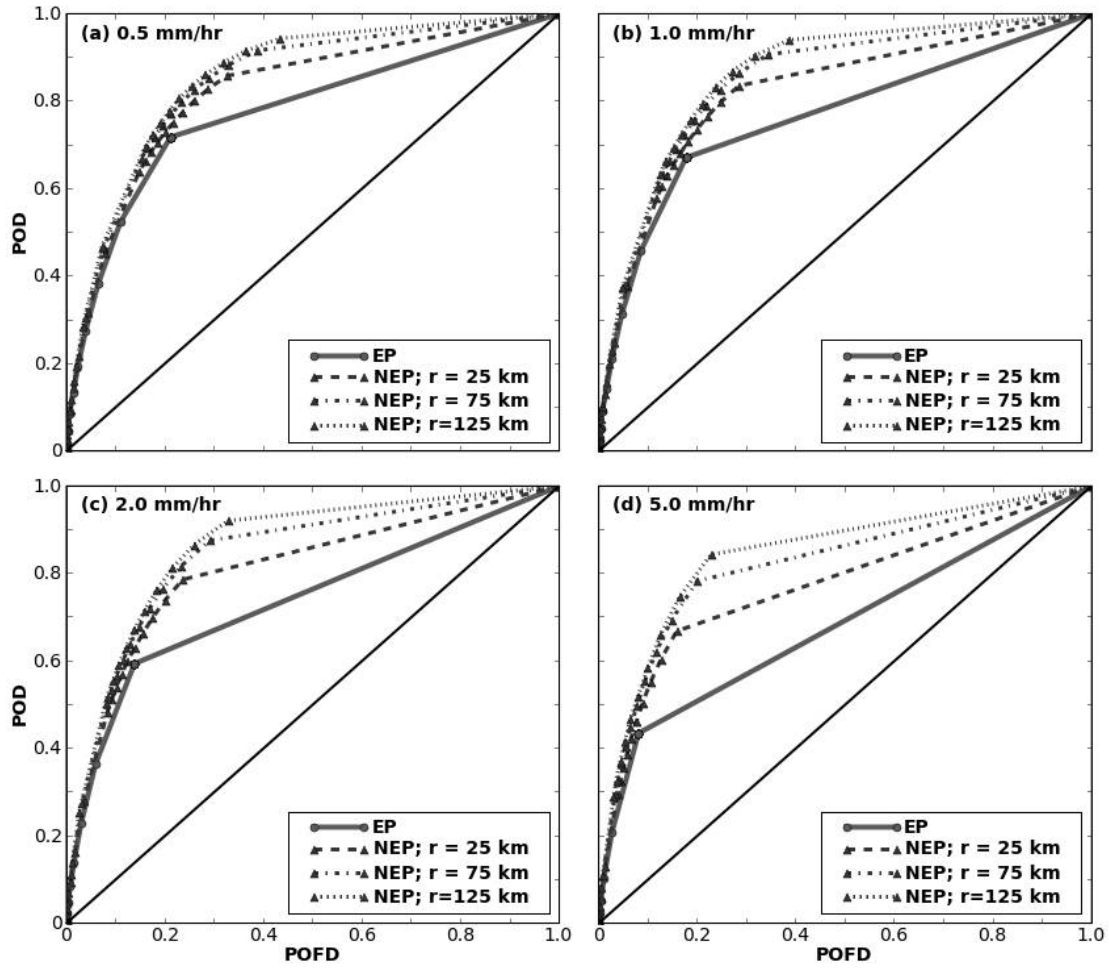

Fig. 15. Relative operating characteristic (ROC) diagrams using data aggregated during 1800-0600 UTC (f21-f33) over all days of SE2007 using accumulation thresholds of (a) 0.5 mm hr^{-1}, (b) 1.0 mm hr^{-1}, (c) 2.0 mm hr^{-1}, and (d) 5.0 mm hr^{-1}.

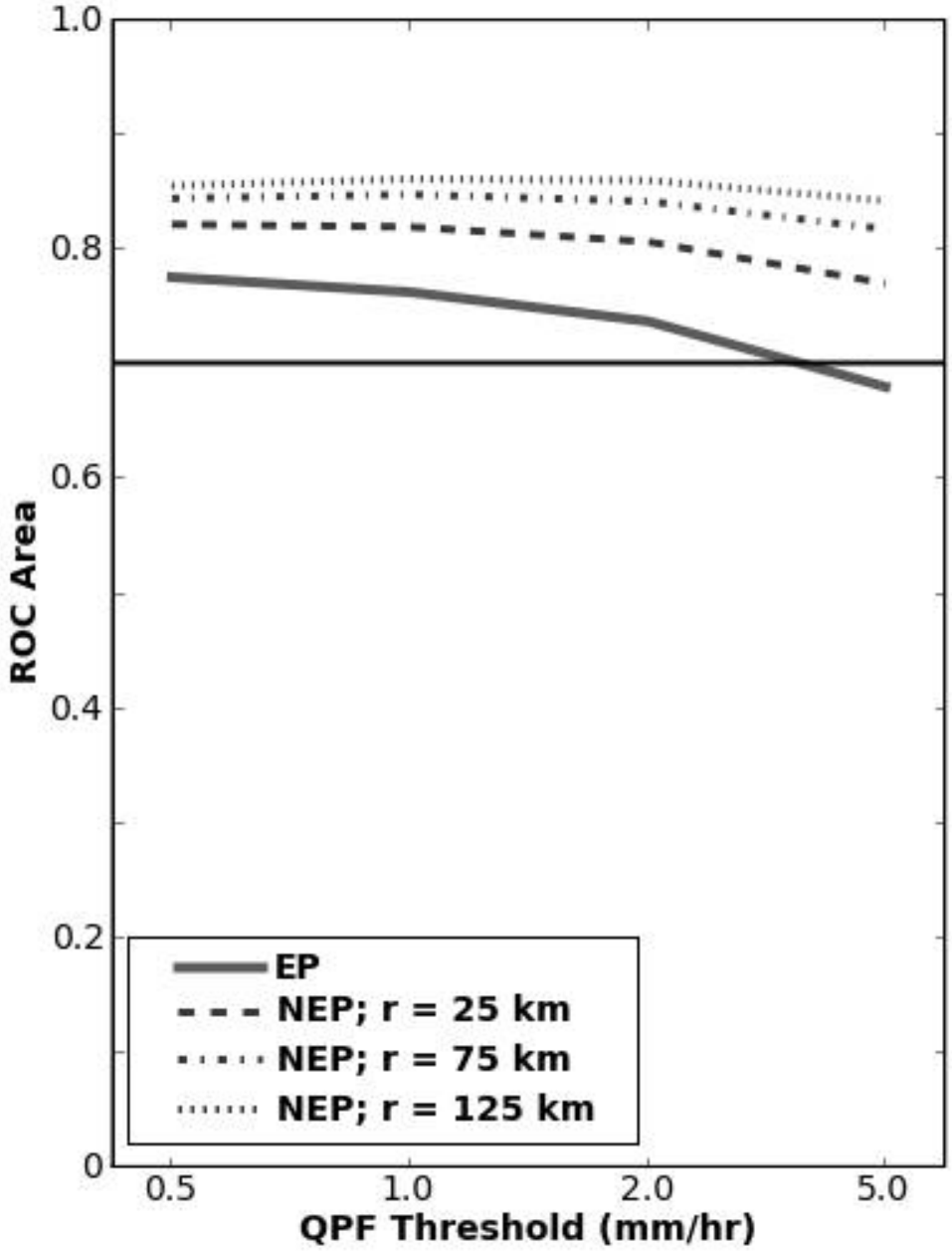

Fig. 16. ROC areas computed from Fig. 15 using a trapezoidal approximation.

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 were used. NAMa and NAMf refer to NAM analyses and forecasts, respectively.

Member	IC	LBC	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
p1	cn + arw_pert	2100 UTC SREF arw_p1	Thompson	MYJ
n2	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 2. Standard 2 x 2 contingency table for dichotomous events.

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