Toward Improved Convection-Allowing Ensembles: Model Physics Sensitivities and Optimizing Probabilistic Guidance with Small Ensemble Membership

CRAIG S. SCHWARTZ,*,** JOHN S. KAIN,+ STEVEN J. WEISS,# MING XUE,*,@ DAVID R. BRIGHT,# FANYOU KONG,@ KEVIN W. THOMAS,@ JASON J. LEVIT,# MICHAEL C. CONIGLIO,+ AND MATTHEW S. WANDISHIN *,&

* *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
+ *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
# *NOAA/NWS/Storm Prediction Center, Norman, Oklahoma*
@ *Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma*
& *Department of Atmospheric Sciences, The University of Arizona, Tucson, Arizona*

(Manuscript received 23 January 2009, in final form 9 June 2009)

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 Advanced Research version of the Weather Research and Forecasting (WRF-ARW) model core, which was initialized at 2100 UTC, ran for 33 h, and resolved convection explicitly. Different initial condition (IC), lateral boundary condition (LBC), and physics perturbations were introduced in 4 of the 10 ensemble members, while the remaining 6 members used identical ICs and LBCs, differing only in terms of microphysics (MP) and planetary boundary layer (PBL) parameterizations. This study focuses on precipitation forecasts from the ensemble.

The ensemble forecasts reveal WRF-ARW sensitivity to MP and PBL schemes. For example, over the 7-week experiment, the Mellor–Yamada–Janjić 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 are explored. Specifically, a ''neighborhood'' approach is described and shown to considerably enhance the skill of probabilistic forecasts for precipitation when combined with a traditional technique of producing ensemble probability fields.

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, 2008; Weisman et al. 2008; Schwartz et al. 2009) 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. For example, Done et al. (2004) demonstrated that convection-allowing 4-km WRF forecasts better predict the frequency and mode of mesoscale convective systems (MCSs) than do 10-km WRF forecasts using CP. Similarly, Kain et al. (2006), Weisman et al. (2008), and Schwartz et al. (2009) all found that 4-km WRF forecasts with explicitly resolved convection yielded better guidance than the CP-using 12-km North American Mesoscale (NAM; Black 1994) model. Additionally, these experiments revealed that running the WRF model at 4 km without CP does not result in egregious failure

** Current affiliation: National Center for Atmospheric Research,++ Boulder, Colorado.

++ The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: Craig Schwartz, NCAR, P.O. Box 3000, Boulder, CO 80307-3000.
E-mail: schwartz@ucar.edu

DOI: 10.1175/2009WAF2222267.1

modes, such as ''numerical point storms'' (Goirgi 1991) that could have significant adverse impacts on meso- and synoptic-scale circulations. Thus, 4-km convection-allowing WRF configurations enable a reasonable evolution of the convective overturning process 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 model configurations have become operational at the National Centers for Environmental Prediction (NCEP) in the form of ''high-resolution window'' deterministic forecasts[1] 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, 2008, personal communication).

When convection-allowing models generate intense localized features such as individual convective cells, even small displacement errors can produce a large disparity in the amplitude of fields such as the rainfall rate at individual grid points. As a result, when the performance of high-resolution models is assessed by traditional verification metrics (such as the equitable threat score), these point errors can lead to poor objective skill scores even though the forecasts may possess substantial value to forecasters and users (Baldwin et al. 2001). In recognition of this inconsistency, postprocessing 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 (Roberts 2005; Theis et al. 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.

Statistically reliable 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). Probabilistic NWP forecasts are commonly derived from an ensemble forecasting system, where an ensemble is composed of a suite of individual forecasts, each generated from a unique combination of initial conditions (ICs), lateral boundary conditions (LBCs), 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 (Houtekamer et al. 1996) recognizes the uncertainties inherent in the parameterizations of small-scale, poorly understood processes, such as cloud microphysics (MP) and turbulence. Stensrud et al. (2000) demonstrated that both IC and physics perturbations are important in creating ensemble spread, but that ensemble spread increases more rapidly when model physics variations are included among the ensemble members.

Ideally, all ensemble members are assumed to be equally likely of representing the ''true'' conditions 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 may be 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 very different solutions (Lorenz 1969). The spread of the ensemble members (in terms of measures such as standard deviation) is typically associated with perceived forecast uncertainty, and point probabilities can be easily 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 (Leith 1974), this ensemble mean consistently performs better than any of the individual members. Furthermore, numerous studies (e.g., Stensrud et al. 1999; Hou et al. 2001; Wandishin 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 day) ensemble forecasts have been produced operationally at NCEP since the early 1990s, primarily for the prediction of large-scale flow features such as 500-hPa geopotential height patterns. However, due to numerous additional challenges, the development and implementation of short-range (0–3 day) ensemble forecasts (SREFs) has lagged somewhat. Among the challenges (see Brooks et al. 1995; Eckel and Mass 2005) is the fact that SREFs are most concerned with forecasting smaller-scale, sensible weather phenomena that are more difficult to predict. Experimental SREF runs began at NCEP in 1995 (Du and Tracton

[1] The high-resolution window forecasts are nested within the 12-km NAM domain. More details on these forecasts are available in Weiss et al. (2008) and online (http://www.emc.ncep.noaa.gov/mmb/mmbpll/nestpage/).

2001) and became operational 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).

Inspired by the documented benefits of both ensemble and convection-allowing forecast systems, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma produced experimental 10-member, convection-allowing ($\delta x = 4$ km) ensemble forecasts for ~35 days during the spring of 2007. These forecasts were evaluated on a daily basis during the 2007 National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT) Spring Experiment[2] and the datasets were archived for later analysis. Clark et al. (2009) used a subset of these data to suggest that a small ensemble (5 members) of high-resolution, convection-allowing configurations appears to be more skillful than a large ensemble (15 members) of coarser-resolution, CP-using members.

This study uses the CAPS dataset for two main purposes. First, it examines precipitation 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. Although previous studies also investigated the impacts of physical parameterizations on WRF rainfall forecasts, they primarily considered coarser-resolution CP-using configurations (e.g., Jankov et al. 2005; Gallus and Bresch 2006), and there is little previous work focusing on convection-allowing WRF model sensitivity to physics [though some discussion is provided in Kain et al. (2005) and Weisman et al. (2008)]. Second, a new method of extracting probabilistic ensemble guidance is presented. This technique, suggested by Theis et al. (2005), combines a ''neighborhood'' approach with a more traditional method of processing ensemble output and dovetails with the work of Clark et al. (2009) by demonstrating another method of enhancing the skill of a small ensemble (i.e., an ensemble with relatively few members). 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 forecasts are presented in section 4 and these forecasts are verified in section 5 prior to concluding.

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

2. Experimental design

a. Model configurations

On each of the ~35 days of the 2007 NOAA HWT Spring Experiment (hereafter SE2007), CAPS produced a 10-member ensemble forecast with 4-km grid spacing (Kong et al. 2007; Xue 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, represented convection explicitly (no CP), had 51 vertical levels, used positive-definite advection, were initialized with a ''cold start'' (no data assimilation) at 2100 UTC, and ran for 33 h 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 NAM. Different IC, LBC, and physics perturbations were introduced into 4 of the 10 ensemble members (n1, n2, p1, and 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 and two from the Nonhydrostatic Mesoscale Model version of the WRF (WRF-NMM; Janjić et al. 2001; Janjić 2003)] of the 2100 UTC NCEP SREF, and the IC perturbations were extracted from these corresponding members. LBCs for the remaining six members (cn, ph1, ph2, ph3, ph4, and 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.[3] 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 considerations

At the conclusion of SE2007, a model climatology was constructed by compiling output from each day of the experiment. Average ensemble performance characteristics were assessed using several statistical measures

[3] Associated with the PBL schemes were corresponding surface layer parameterizations. Those members using the MYJ PBL scheme used the Janjić Eta Model (Janjić 1996) surface layer scheme, while all members using the YSU PBL scheme used the Monin–Obukhov surface layer parameterization. Here, we consider the surface layer scheme to be a component of the PBL parameterization. Thus, when we discuss PBL schemes, we really mean the combination of PBL and surface layer parameterizations.

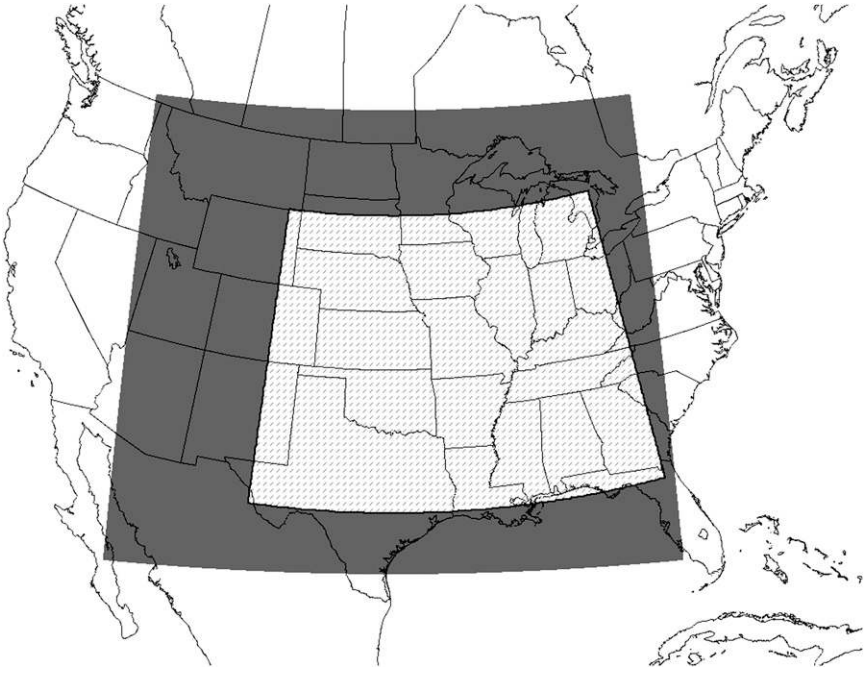

FIG. 1. Model domain of the CAPS ensemble forecasts (shaded) and verification domain (hatched) used for the model climatology.

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 gauge 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. 1). 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 period [corresponding to a 21–33-h forecast (hereafter f21–f33)] 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 in sections 3 and 5), 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 and it appears that these differences can be attributed to the different PBL and MP

TABLE 1. Ensemble member configurations. The WSM6 (Hong et al. 2004), Ferrier (Ferrier 1994), Thompson (Thompson et al. 2004), MYJ (Mellor and Yamada 1982; Janjić 2002), and 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	WSM6	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	WSM6	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	WSM6	YSU
ph4	2100 UTC NAMa	1800 UTC NAMf	Thompson	YSU
ph5	2100 UTC NAMa	1800 UTC NAMf	Ferrier	YSU

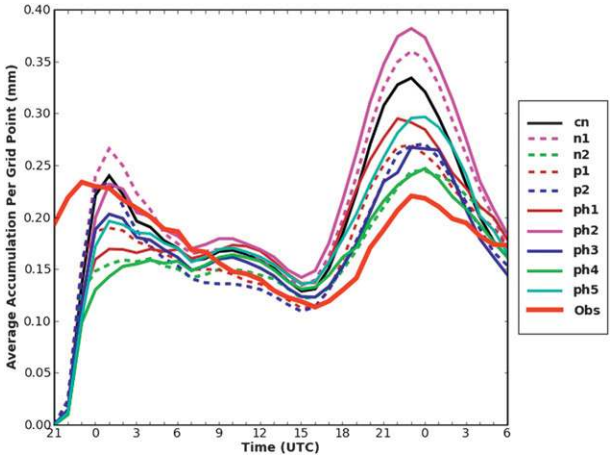

FIG. 2. Total precipitation over the verification domain aggregated over all days of SE2007, normalized by the number of grid boxes, calculated for each ensemble member on its native grid.

schemes. Aggregate statistics over all days of SE2007 are first presented, followed by a brief case study.

a. Domain total precipitation

The total accumulated precipitation throughout the verification domain, calculated on native grids and aggregated over all days of SE2007, is depicted in Fig. 2. 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 during the afternoon diurnal maximum. It should be noted that a high precipitation bias has been observed in other studies using 4-km versions of the WRF-ARW (Kain et al. 2008; Weisman et al. 2008). The specific cause of these high biases has not been explicitly identified but may be related to issues regarding insufficient resolution (Bryan et al. 2003), moisture advection (Skamarock and Weisman 2009), or imperfections in the current parameterizations. However, more detailed examinations of selected events from our study, 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 contributed to the very high bias noted in our study (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 the 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. 2). 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 Mellor–Yamada–Janjić (MYJ) PBL parameterizations (see Table 1 for specific configurations of individual members). 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 Yonsei University (YSU) PBL and Thompson MP schemes. However, the p2 and ph3 members produced the least precipitation during the last 3 h of integration and also during the diurnal minimum. Both members shared the YSU PBL and WRF single-moment six-class scheme (WSM6) MP parameterizations.

b. Areal coverages

Figure 3 depicts fractional coverages of precipitation exceeding various accumulation thresholds (q) (e.g., 1.0 mm h^{-1}), aggregated hourly over all days during SE2007. These statistics were generated from data on each member's native grid and computed over the verification domain. Again, on average, the individual members captured the diurnal cycle fairly well, with the time of peak coverage corresponding well to the observations.

When $q = 0.5$ mm h^{-1} (Fig. 3a), roughly half of the members overpredicted the coverage during the diurnal peak, on average, with the n1 and ph2 members (Ferrier and MYJ) generating the highest coverages. But, as q increased, overall overprediction worsened, such that, by the 5.0 mm h^{-1} threshold (Fig. 3c), all members produced noticeably higher areal coverages than those observed during the afternoon and evening. 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 and 0600 UTC (f21–f33). The values were ranked, and accumulation percentiles (e.g., 95th percentile) were chosen to determine absolute hourly precipitation values (q_y) corresponding to the yth percentile (Fig. 4). For example,

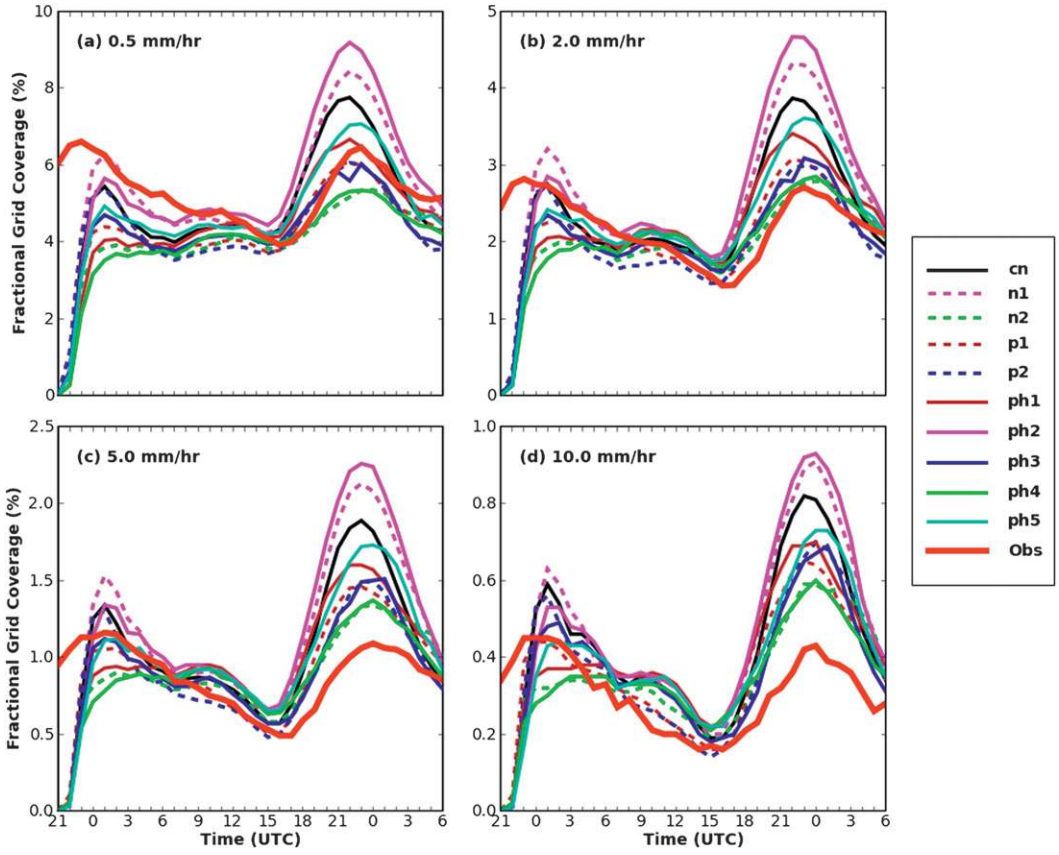

FIG. 3. Fractional grid coverage of hourly precipitation exceeding (a) 0.5, (b) 2.0, (c) 5.0, and (d) 10.0 mm h^{-1} as a function of time, averaged over all days during SE2007, calculated on each member's native grid.

$(100 - y)$ 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 composed the top 1% of all accumulations in the n1 and ph2 (Ferrier and MYJ) hourly precipitation fields (between 1800 and 0600 UTC), while the 99th percentiles in the n2, ph4, p2, and ph3 fields were considerably lower (~5.5 mm h^{-1}).

d. Precipitation bias

To quantitatively determine the biases of individual members, the standard 2 × 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 an 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.

Biases aggregated[4] over all days of SE2007 based on hourly precipitation between 1800 and 0600 UTC (f21–f33) are plotted as a function of precipitation threshold in Fig. 5. A large bias spread is evident, with the n1 and ph2 (Ferrier and MYJ) members overpredicting the most for $q \leq 10.0$ mm h^{-1}. At thresholds >10.0 mm h^{-1}, the n1, ph2, and ph5 biases interestingly plummet, leaving the ph1 and p1 members with the highest biases (both configured with the Thompson MP and MYJ PBL schemes). The n1, ph2, and ph5 members all used the Ferrier MP scheme. Clearly, some aspect of this scheme prohibited these members from consistently generating precipitation rates in excess of 10.0 mm h^{-1}. Although we believe it is important to determine the cause of this behavior, doing so is beyond the scope of this paper.

[4] The 2 × 2 contingency table elements were summed over all days during SE2007 and the bias was computed from these sums.

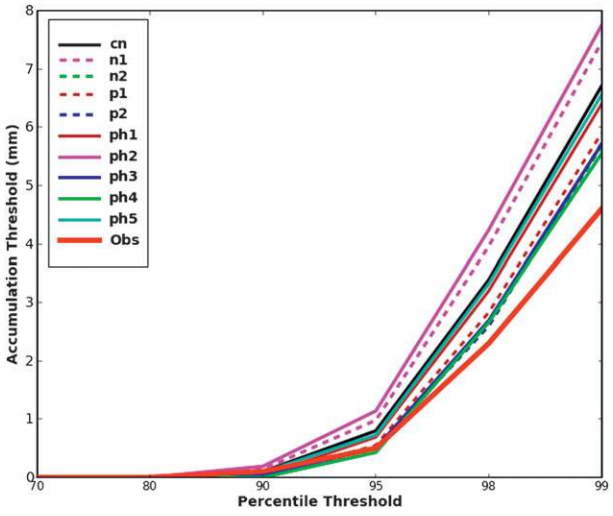

FIG. 4. Precipitation climatology: percentiles calculated on each member's native grid aggregated between 1800 and 0600 UTC (f21–f33) over all days during SE2007 (see text).

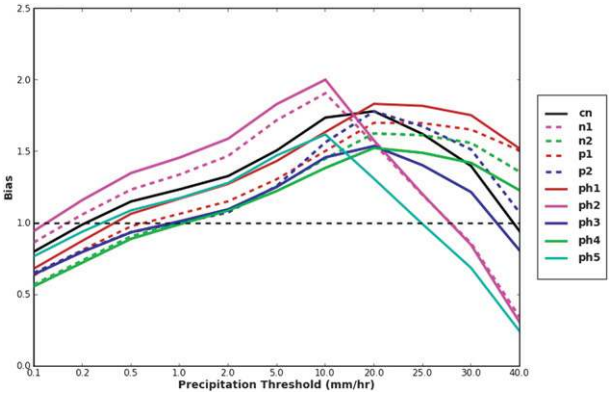

FIG. 5. Bias as a function of the accumulation threshold, aggregated during 1800–0600 UTC (f21–f33) over all days during SE2007.

An example that illustrates typical average characteristics of the ensemble is presented in the next subsection.

e. Case study

Figure 6 shows hourly precipitation output from the physics-only members on their native grids over the verification domain. The ensemble was initialized at 2100 UTC 4 June, and the forecast was valid at 0000 UTC 6 June—a 27-h 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 the 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. 7). The ph2 member produced the most precipitation, while the ph3 and ph4 members 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 members overpredicted the observed hourly precipitation (Fig. 6g) that occurred over the verification domain at that time.

f. Summary

On average, all the ensemble members overpredicted precipitation. However, the high bias was not uniform among the different model configurations. 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 [the Ferrier scheme high bias was also noted in Jankov et al. (2005)]. 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). Similar to Stensrud et al. (2000) and Jankov et al. (2005), 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

TABLE 2. Standard 2 × 2 contingency table for dichotomous events.

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

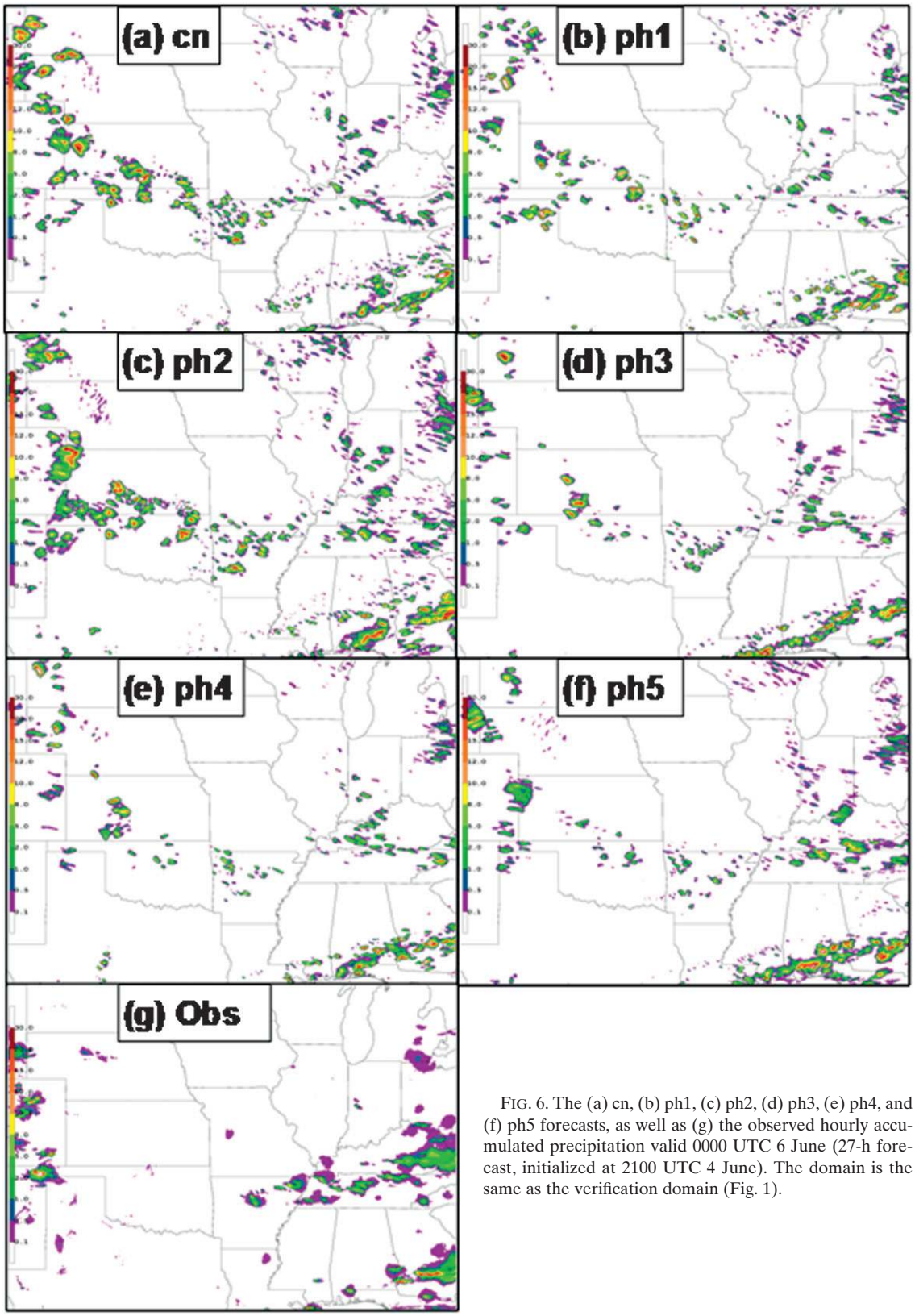

FIG. 6. The (a) cn, (b) ph1, (c) ph2, (d) ph3, (e) ph4, and (f) ph5 forecasts, as well as (g) the observed hourly accumulated precipitation valid 0000 UTC 6 June (27-h forecast, initialized at 2100 UTC 4 June). The domain is the same as the verification domain (Fig. 1).

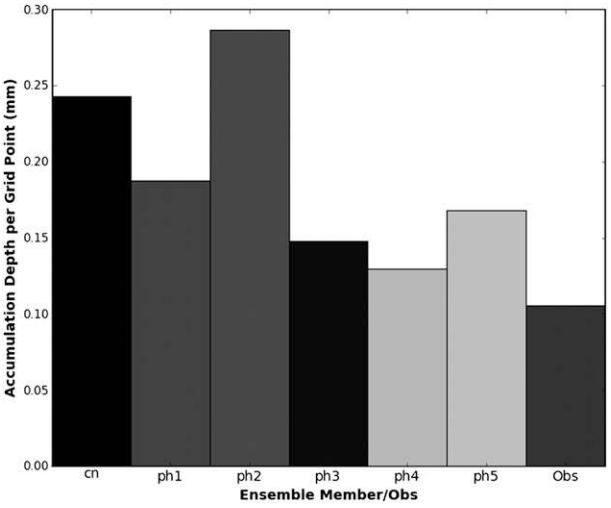

FIG. 7. Total hourly precipitation accumulations valid at the same time and calculated over the same domain as in Fig. 6.

WRF-ARW developers and users and aid bias-correction efforts (e.g., Eckel and Mass 2005; Stensrud and Yussouf 2007) to improve the reliability of future convection-allowing WRF-ARW ensembles.

4. Extracting forecast probabilities: Traditional and new approaches

A widely used approach for computing forecast probabilities (FPs) from an ensemble is summarized, followed by a discussion of a lesser known postprocessing 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 nonoccurrence (0) of an event (an ''event'' typically means exceedance of a specified threshold). For example, within the context of precipitation forecasting, an accumulation threshold (q) is chosen to define an event, and the individual grid-point BPs are given by

$$\mathrm{BP}_{ki} = \begin{cases} 1 & \text{if} \quad F_{ki} \geq q \\ 0 & \text{if} \quad F_{ki} < q \end{cases}, \tag{1}$$

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

$$\mathrm{EP}_i = \frac{1}{n}\sum_{k=1}^{n} \mathrm{BP}_{ki}, \tag{2}$$

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, when an NWP model attempts to place meteorological features that are comparable in scale to its grid length, spatial displacement errors become large. 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 nonbinary 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.[5] All grid points surrounding a given point whose centers 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.

[5] 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.

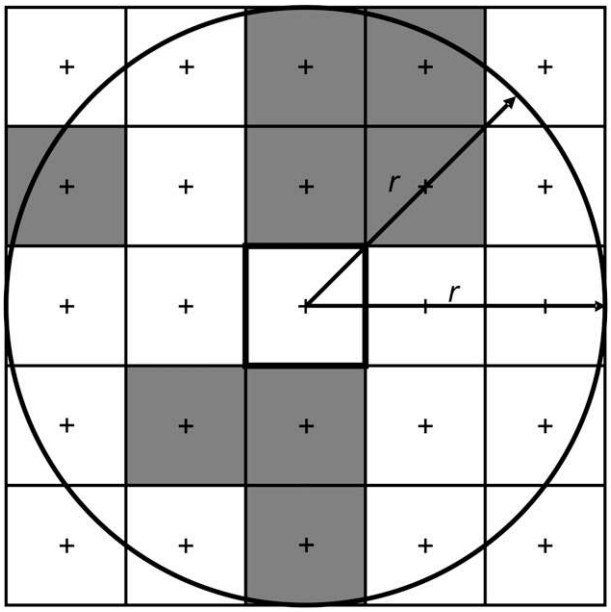

FIG. 8. Schematic example of neighborhood determination and fractional creation for a model forecast. The precipitation exceeds the accumulation threshold in the shaded boxes, and a radius of 2.5 times the grid length is specified.

To generate a nonbinary 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

$$\mathrm{NP}_{ki} = \frac{1}{N_b} \sum_{m=1}^{N_b} \mathrm{BP}_{km}, \tag{3}$$

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 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 comprises 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 8 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 whose centers lie 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 $\mathrm{NP}_i = 0.38$, or 38%, at the central grid box.

Figure 9 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 h—and the model output is displayed on the verification grid. The raw precipitation forecast is shown in Fig. 9a and the binary field (the BP_i field) corresponding to $q = 5.0$ mm h^{-1} is plotted in Fig. 9b. 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. 9c) and then 75 km (Fig. 9d), 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 raw forecast precipitation distribution. However, for most situations, increasing r reduces the sharpness (Roberts and Lean 2008) and acts as a smoother that reduces gradients in and magnitudes of 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 compose 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

$$\mathrm{NEP}_i = \frac{1}{n} \sum_{k=1}^{n} \mathrm{NP}_{ki}. \tag{4}$$

Alternatively, a mathematically equivalent expression for NEP_i can be obtained by averaging the EP_i values themselves over the N_b points within the neighborhood; that is,

$$\mathrm{NEP}_i = \frac{1}{N_b} \sum_{m=1}^{N_b} \mathrm{EP}_m. \tag{5}$$

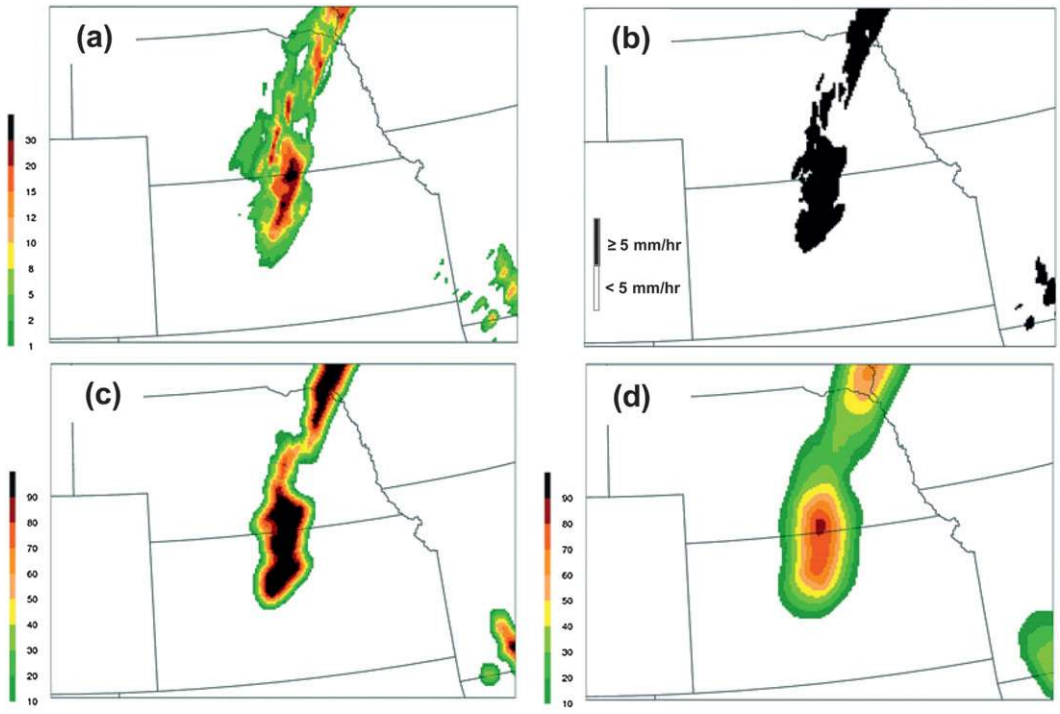

FIG. 9. (a) Control member (cn) 1-h accumulated precipitation forecast (mm h^{-1}), (b) binary image (i.e., a BP grid) of precipitation accumulations exceeding 5.0 mm h^{-1}, and NP grids computed from (b) using radii of influence of (c) 25 and (d) 75 km. All panels are valid 0600 UTC 23 May and the control member has been projected onto the verification grid.

Note that Eq. (5) is analogous to obtaining NPs [Eq. (3)], but instead of averaging binary values, fractions are averaged.

Since an ensemble is naturally designed to provide probabilistic guidance, it may seem counterintuitive or redundant to apply a neighborhood approach to ensemble output. However, an ensemble with small membership can only provide a limited sample of the probability density function (PDF) of the atmospheric state. The neighborhood approach introduces a measure of uncertainty associated with each member and in effect ''fills in'' the PDF, thus increasing the effective size (i.e., number of members) of the ensemble.

To demonstrate the characteristics 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 h^{-1} accumulation threshold (Fig. 10). The traditional probability field (i.e., the EP) is very detailed and rather noisy (Fig. 10a). On the other hand, the NEPs become increasingly smooth as r increases from 25 to 125 km (Figs. 10b–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 attention on spatial scales where there is likely more skill. Additionally, it smoothes out any discontinuities in the EP field. The NEP fields are now objectively verified and compared with the corresponding EP fields.

5. Verification of probabilistic fields

The fractions skill score (Roberts 2005; Roberts and Lean 2008) and relative operating characteristic (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

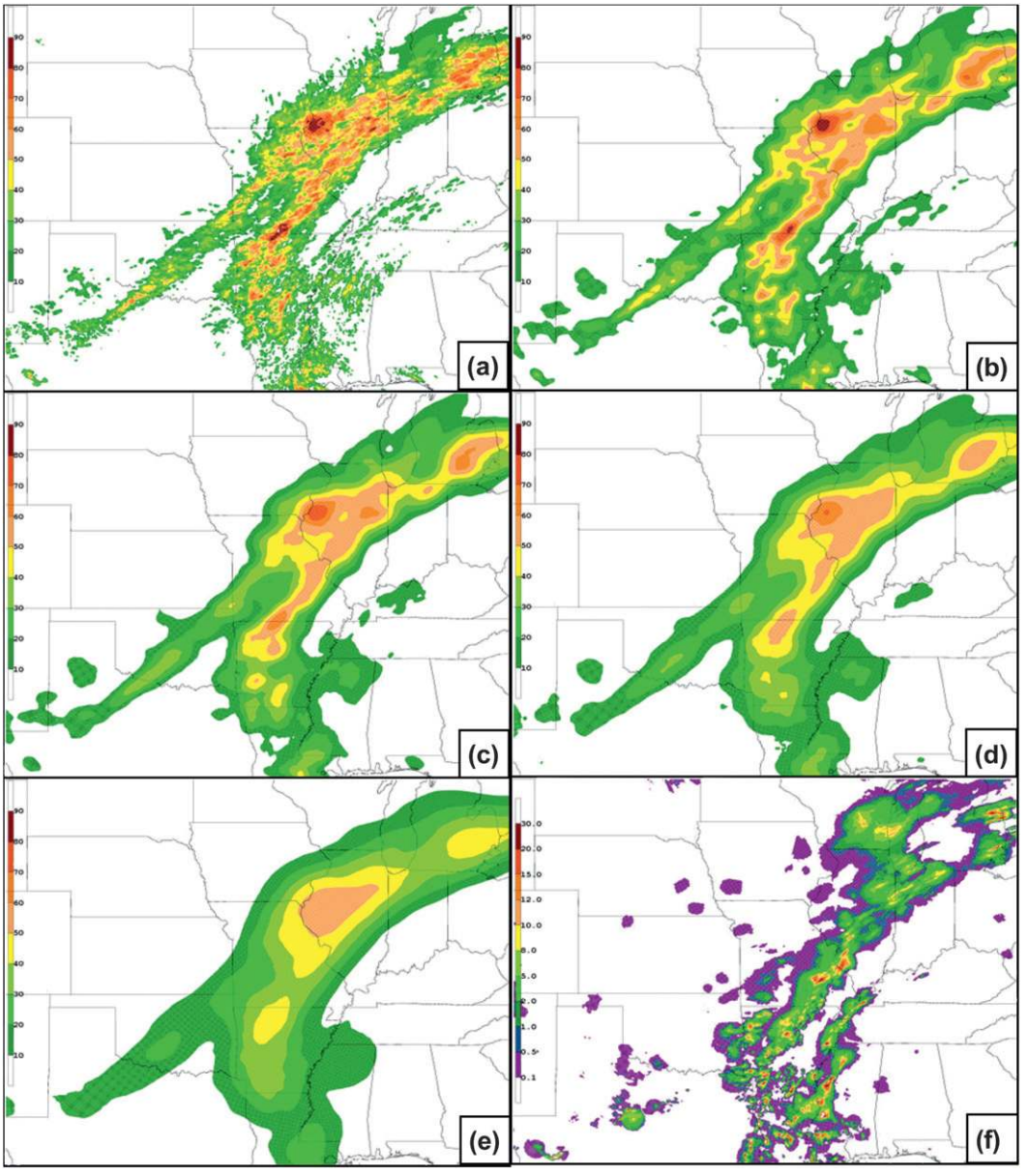

FIG. 10. Hourly probabilistic forecasts of precipitation meeting or exceeding 1.0 mm using the (a) EP and NEP (see text) with radii of influence of (b) 25, (c) 50, (d) 75, and (e) 125 km. The observed precipitation (mm h^{-1}) is shown in (f). Both the model fields and observations are valid 2100 UTC 15 May. The domain is the same as the verification domain (Fig. 1).

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. 8 depicts the creation of a fraction grid for just a model forecast, Fig. 11 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, within the context of a radius *r*, this model forecast is considered to be 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

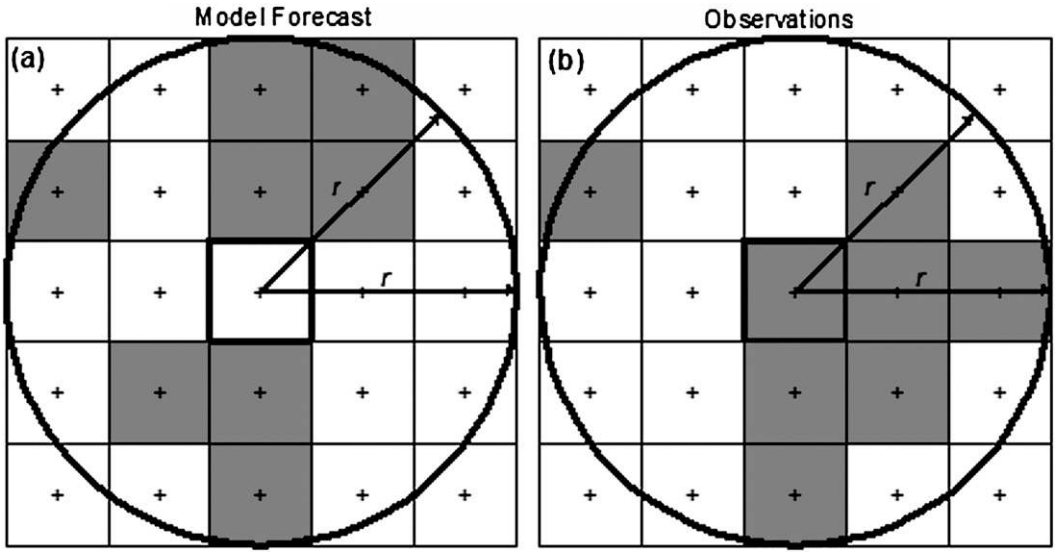

FIG. 11. Schematic example of neighborhood determination and fractional creation for (a) a model forecast and (b) the corresponding observations. The precipitation exceeds the accumulation threshold in the shaded boxes, and a radius of 2.5 times the grid length is specified.

$$\mathrm{FBS} = \frac{1}{N_v}\sum_{i=1}^{N_v}[\mathrm{NP}_{F(i)} - \mathrm{NP}_{O(i)}]^2, \tag{6}$$

where $\mathrm{NP}_{F(i)}$ and $\mathrm{NP}_{O(i)}$ 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. 1), 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 the observations. The worst possible (largest) FBS is achieved when there is no overlap of nonzero fractions and is given by

$$\mathrm{FBS}_{\mathrm{worst}} = \frac{1}{N_v}\left[\sum_{i=1}^{N_v}\mathrm{NP}_{F(i)}^2 + \sum_{i=1}^{N_v}\mathrm{NP}_{O(i)}^2\right]. \tag{7}$$

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-accuracy reference forecast ($\mathrm{FBS}_{\mathrm{worst}}$) and is defined by Roberts (2005) as the fractions skill score (FSS):

$$\mathrm{FSS} = 1 - \frac{\mathrm{FBS}}{\mathrm{FBS}_{\mathrm{worst}}}. \tag{8}$$

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 $2B/(B^2 + 1)$ (Roberts and Lean 2008), where B is the frequency bias (as discussed in section 3d).

Since the FSS can be used to evaluate any probabilistic field, it is an appropriate metric for verification of the EP, NPs, and NEP. However, the EP is handicapped in the computation of the FSS because it does not change as a function of r, while the verifying field, NEP, and NPs do. With this caveat, the FSS values for the EP in Figs. 12 and 13 (section 5b) should be viewed only as references to which the NP and NEP FSS scores can be compared.

b. Verification results

The FSS aggregated[6] over all days of SE2007 during the 1800–0600 UTC (f21–f33) period is shown in Fig. 12 for various hourly absolute precipitation thresholds. As q

[6] FSS was aggregated by summing the FBS and $\mathrm{FBS}_{\mathrm{worst}}$ over all days during SE2007 and computing FSS with these sums.

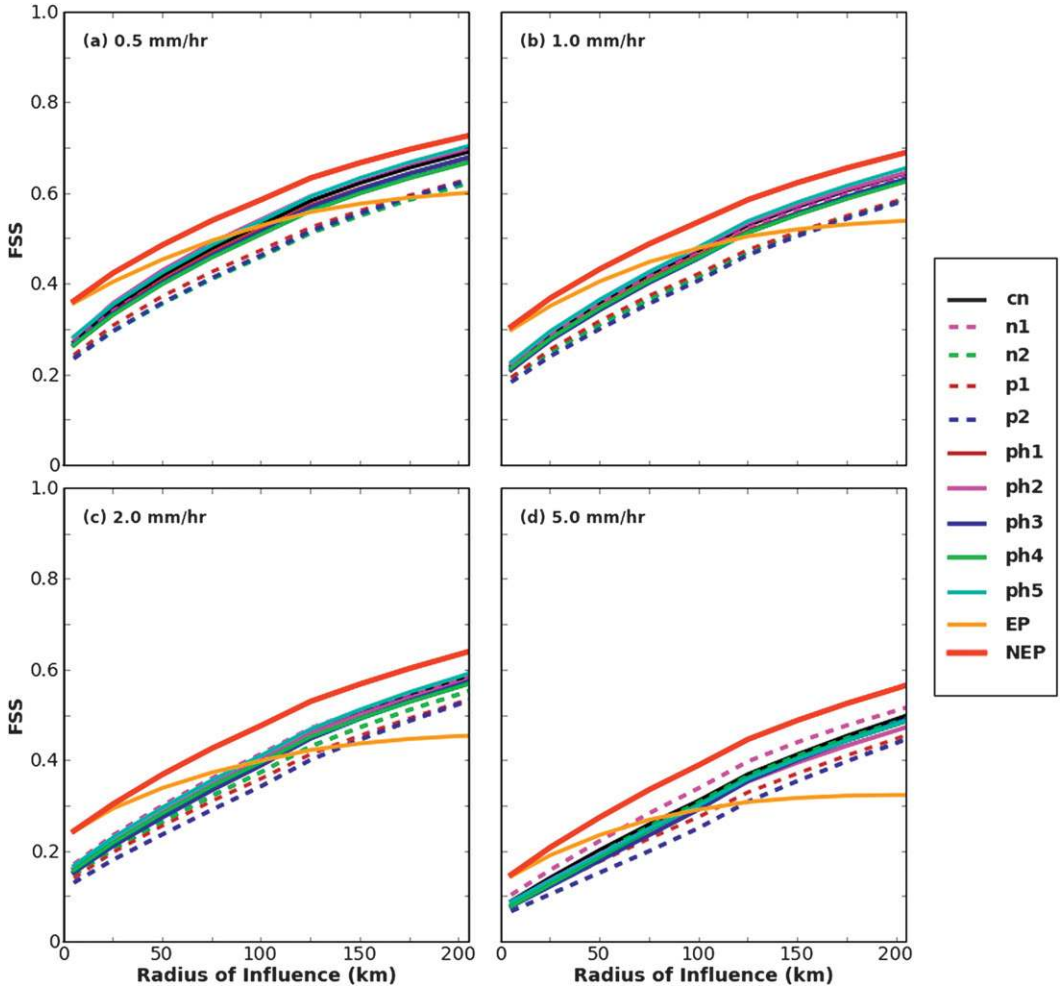

FIG. 12. FSS as a function of radius of influence (r), aggregated during 1800–0600 UTC (f21–f33) over all days during SE2007 using accumulation thresholds of (a) 0.5, (b) 1.0, (c) 2.0, and (d) 5.0 mm h^{-1}. The traditional ensemble probability is denoted as EP and the neighborhood probability as NEP. Probabilities for the individual members of the ensemble were computed as NPs. Note that the EP field is displayed for reference only since it does not change as a function of r, while the NEP and NPs do.

increased, the FSS worsened at all scales, revealing that 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 [Eqs. (4) and (5)] improves upon the traditional point-based 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 h^{-1} (Fig. 13) indicated that NEPs performed the best at nearly all times for all values of r, further demonstrating the benefits of the neighborhood method compared to a point-by-point approach.

The advantage for the NEP is also evident with other performance measures, such as the relative operating characteristic (ROC). 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 greater than or equal to 30% are considered to forecast the event) in conjunction with a binary observation field. Using the elements in 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,[7] and

[7] We used probability thresholds of 0%, 10%, 20%, . . . , 90%, and 100% to compute the ROC.

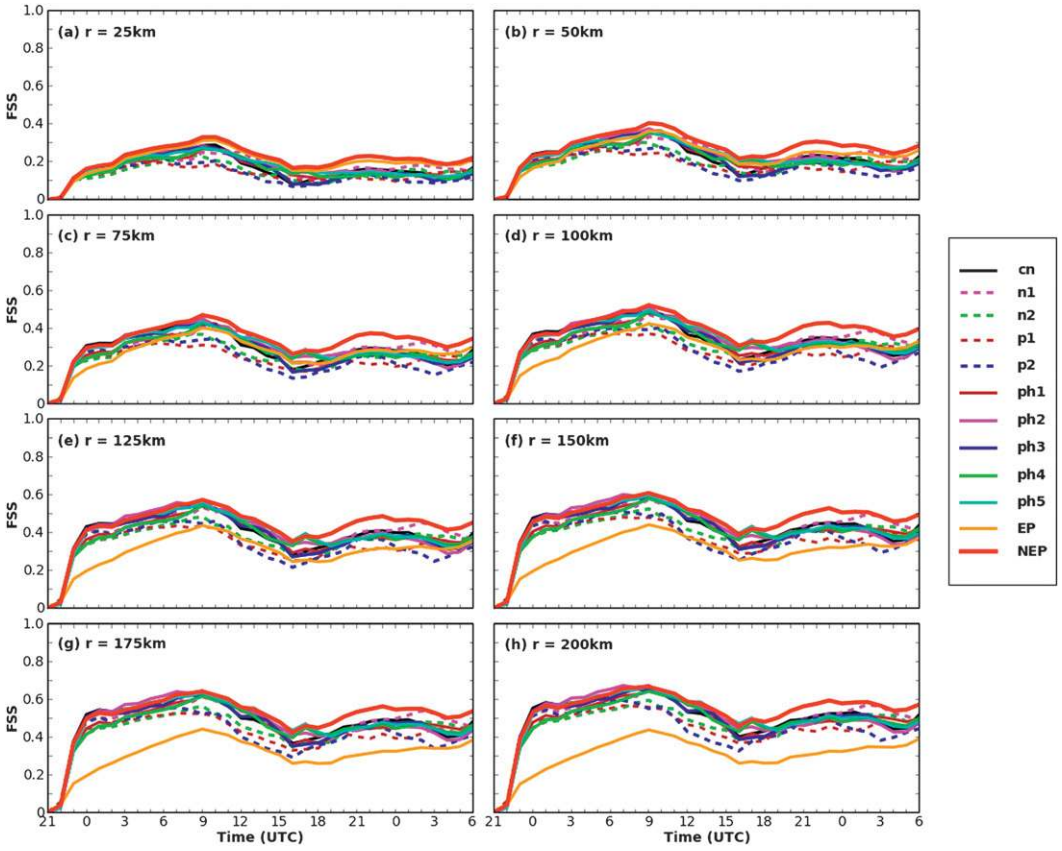

FIG. 13. FSS plotted as a function of forecast hour for a fixed accumulation-rate threshold of 5.0 mm h^{-1} and radii of influence of (a) 25, (b) 50, (c) 75, (d) 100, (e) 125, (f) 150, (g) 175, and (h) 200 km, averaged over all days during SE2007.

the ROC is formed by plotting the POFD against the POD over the range of probabilistic thresholds (Fig. 14). 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 h^{-1} (Fig. 15). 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. These findings further indicates that the NEP improves upon the EP and that the improvement increases as the event becomes more extreme.

6. Summary and conclusions

During SE2007, CAPS produced 33-h convection-allowing 10-member ensemble forecasts using 4-km horizontal grid spacing. LBC, IC, and physics perturbations were introduced into four of the members while the remaining six 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 and precipitation forecasts can be influenced by the dynamic core as well (Gallus and Bresch 2006). It is also important to consider that ICs play an important role in modulating precipitation forecasts (Weisman et al. 2008) and forecasts of variables other than precipitation (such as temperature and dewpoint) may be more heavily influenced

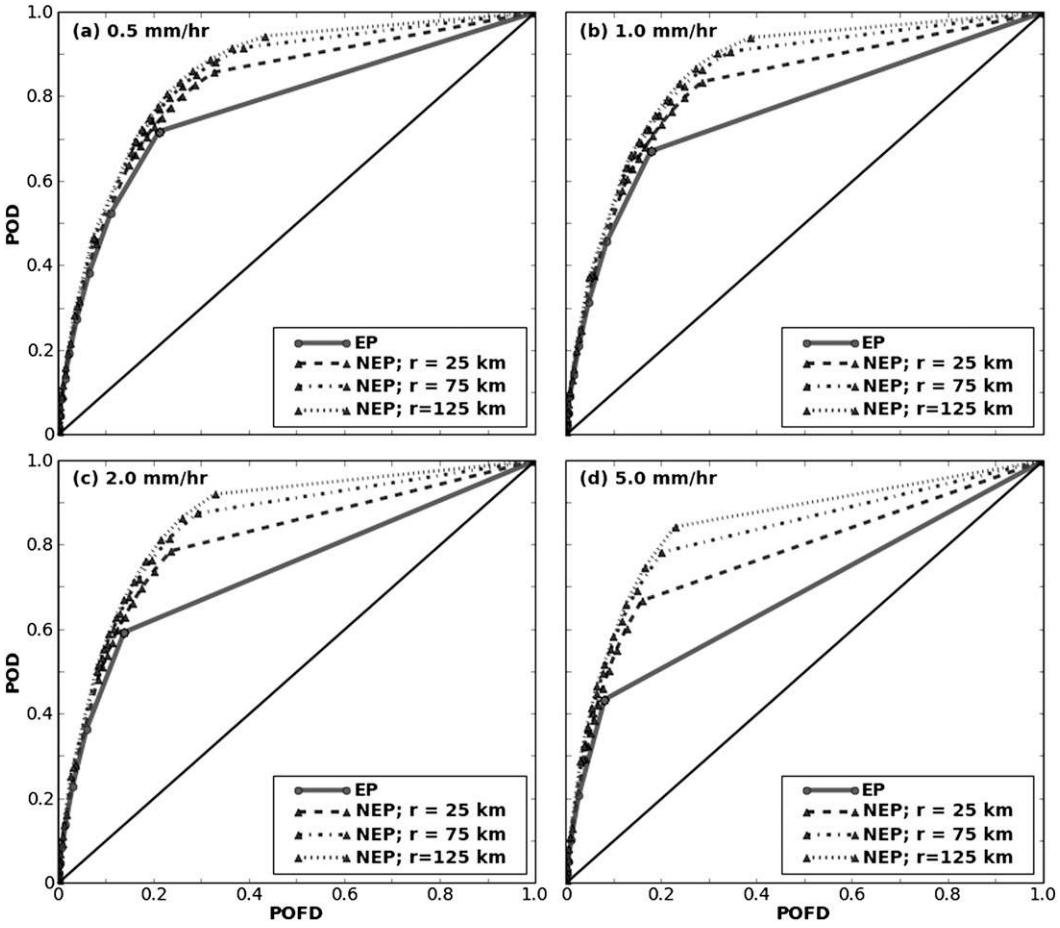

FIG. 14. ROC diagrams using data aggregated during 1800–0600 UTC (f21–f33) over all days of SE2007 using accumulation thresholds of (a) 0.5, (b) 1.0, (c) 2.0, and (d) 5.0 mm h^{-1}.

by IC uncertainty than physics perturbations (Kong et al. 2007; Coniglio et al. 2010).

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 forecasts, as measured by the ROC curve area and FSS, than traditional point-by-point 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 have skill. These findings indicate that simple postprocessing 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.

It appears that numerical forecasts of precipitation benefit considerably from convection-allowing model configurations, whether these configurations are used for deterministic or probabilistic forecasts. Regardless of the specific strategy, there is much to be learned about what type of information can be extracted from convection-allowing model output and how this information might best be used by operational weather forecasters.

Acknowledgments. 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 Collaborative Science, Technology, and Applied Research (CSTAR) program and were performed at the PSC supported by the National Science Foundation (NSF). Supplementary support was provided by NSF ITR project LEAD (ATM-0331594).

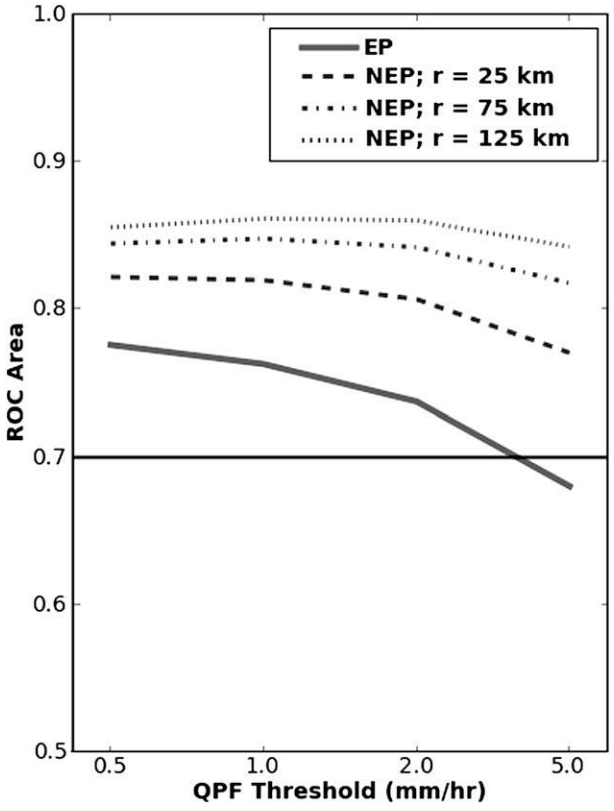

FIG. 15. ROC areas computed from Fig. 14 using a trapezoidal approximation.

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

Baldwin, M. E., S. Lakshmivarahan, and J. S. Kain, 2001: Verification of mesoscale features in NWP models. Preprints, *Ninth Conf. on Mesoscale Processes,* Fort Lauderdale, FL, Amer. Meteor. Soc., 8.3. [Available online at http://ams.confex.com/ams/pdfpapers/23364.pdf.]

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.

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

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting,* **24,** 1121–1140.

Coniglio, M. C., K. L. Elmore, J. S. Kain, S. J. Weiss, M. Xue, and M. L. Weisman, 2010: Evaluation of WRF model output for severe-weather forecasting from the 2008 NOAA Hazardous Weather Testbed Spring Experiment. *Wea. Forecasting,* in press.

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, *Ninth Conf. on Mesoscale Processes,* Fort Lauderdale, FL, Amer. Meteor. Soc., P4.9. [Available online at http://ams.confex.com/ams/pdfpapers/23074.pdf.]

——, 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, *Expert Team Meeting on Ensemble Prediction Systems,* Exeter, United Kingdom, WMO. [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.

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

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

Gallus, W. A., Jr., and J. F. Bresch, 2006: Comparison of impacts of WRF dynamic core, physics package, and initial conditions on warm season rainfall forecasts. *Mon. Wea. Rev.,* **134,** 2632–2641.

Goirgi, F., 1991: Sensitivity of simulated summertime precipitation over the western United States to different physics parameterizations. *Mon. Wea. Rev.,* **119,** 2870–2888.

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.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242.

Janjić, Z. I., 1996: The surface layer in the NCEP Eta Model. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 354–355.

——, 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, 61 pp. [Available online at http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf.]

——, 2003: A nonhydrostatic model based on a new approach. *Meteor. Atmos. Phys.,* **82,** 271–285.

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

Jankov, I., W. A. Gallus, M. Segal, B. Shaw, and S. E. Koch, 2005: The impact of different WRF model physical parameterizations and their interactions on warm season MCS rainfall. *Wea. Forecasting,* **20,** 1048–1060.

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

——, ——, 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.

——, and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting,* **23,** 931–952.

Kong, F., and Coauthors, 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. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction,* Salt Lake City, UT, Amer. Meteor. Soc., 3B.2. [Available online at http://ams.confex.com/ams/pdfpapers/124667.pdf.]

——, M. Xue, K. K. Droegemeier, K. Thomas, and Y. Wang, 2008: 2008 real-time storm-scale ensemble forecast experiment—What's next? Preprints, *Ninth WRF User's Workshop,* Boulder, CO, NCAR, 7.3. [Available online at http://www.mmm.ucar.edu/wrf/users/workshops/WS2008/presentations/7-3.pdf.]

Leith, C., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

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.]

Lorenz, E. N., 1969: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.,* **26,** 636–646.

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

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

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

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. Met Office Tech. Rep. 455, 80 pp.

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

——, 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., and Coauthors, 2009: Next-day convection-allowing WRF model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.,* **137,** 3351–3372.

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., and M. L. Weisman, 2009: The impact of positive-definite moisture transport on NWP precipitation forecasts. *Mon. Wea. Rev.,* **137,** 488–494.

——, J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, W. Wang, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech Note NCAR/TN-468+STR, 88 pp. [Available from UCAR Communications, P.O. Box 3000, Boulder, CO 80307.]

Stensrud, D. J., and N. Yussouf, 2007: Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system. *Wea. Forecasting,* **22,** 3–17.

——, H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.,* **127,** 433–446.

——, J. W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.,* **128,** 2077–2107.

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.

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

Xue, M., and Coauthors, 2007: CAPS real-time storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment. Preprints, *22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction,* Salt Lake City, UT, Amer. Meteor. Soc., 3B. [Available online at http://ams.confex.com/ams/pdfpapers/124587.pdf.]