Next-Day Convection-Allowing WRF Model Guidance: A Second Look at 2-km versus 4-km Grid Spacing

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

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

(Manuscript received 5 January 2009, in final form 20 March 2009)

ABSTRACT

During the 2007 NOAA Hazardous Weather Testbed (HWT) Spring Experiment, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma produced convection-allowing forecasts from a single deterministic 2-km model and a 10-member 4-km-resolution ensemble. In this study, the 2-km deterministic output was compared with forecasts from the 4-km ensemble control member. Other than the difference in horizontal resolution, the two sets of forecasts featured identical Advanced Research Weather Research and Forecasting model (ARW-WRF) configurations, including vertical resolution, forecast domain, initial and lateral boundary conditions, and physical parameterizations. Therefore, forecast disparities were attributed solely to differences in horizontal grid spacing. This study is a follow-up to similar work that was based on results from the 2005 Spring Experiment. Unlike the 2005 experiment, however, model configurations were more rigorously controlled in the present study, providing a more robust dataset and a cleaner isolation of the dependence on horizontal resolution. Additionally, in this study, the 2- and 4-km outputs were compared with 12-km forecasts from the North American Mesoscale (NAM) model. Model forecasts were analyzed using objective verification of mean hourly precipitation and visual comparison of individual events, primarily during the 21- to 33-h forecast period to examine the utility of the models as next-day guidance. On average, both the 2- and 4-km model forecasts showed substantial improvement over the 12-km NAM. However, although the 2-km forecasts produced more-detailed structures on the smallest resolvable scales, the patterns of convective initiation, evolution, and organization were remarkably similar to the 4-km output. Moreover, on average, metrics such as equitable threat score, frequency bias, and fractions skill score revealed no statistical improvement of the 2-km forecasts compared to the 4-km forecasts. These results, based on the 2007 dataset, corroborate previous findings, suggesting that decreasing horizontal grid spacing from 4 to 2 km provides little added value as next-day guidance for severe convective storm and heavy rain forecasters in the United States.

1. Introduction

Convection-allowing numerical weather prediction (NWP) efforts in the United States began in the early to mid-1990s, when experimental model predictions with horizontal grid spacing as fine as 3 km were initialized with real data (Droegemeier et al. 1996a,b; Xue et al. 1996a,b). Although these forecasts focused on relatively small geographic domains and limited time scales, they demonstrated the potential value of convection-allowing model forecasts for the short-range prediction of convective storms. Similar studies in other parts of the world (e.g., Roberts 2003; Speer et al. 2003; Steppeler et al. 2003) yielded equally encouraging results, and convection-allowing forecasts were extended to larger domains and longer time periods when 4-km next-day forecasts were produced in support of the Bow Echo and Mesoscale Convective Vortex (MCV) Experiment (BAMEX)

** Current affiliation: National Center for Atmospheric Research (NCAR), Boulder, Colorado. NCAR 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: 10.1175/2009MWR2924.1

© 2009 American Meteorological Society

field program (Davis et al. 2004). These forecasts were shown to provide more skillful guidance than forecasts from a coarser-resolution model employing convective parameterization (CP; Done et al. 2004). Motivated by this success, detailed evaluations of multiple convection-allowing 4-km model configurations were performed during the 2004 National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT) Spring Experiment[1] and found to corroborate Done et al. (2004) and alleviate concerns that convective initiation might be delayed significantly without CP (Kain et al. 2006).

In light of these additional positive results, scientists at the U.S. National Centers for Environmental Prediction (NCEP) Environmental Modeling Center (EMC) continued to experiment with convection-allowing model configurations in collaboration with forecasters at the NCEP Storm Prediction Center (SPC). This collaborative effort led to the operational implementation of convection-allowing model forecasts at EMC in 2008. Currently, EMC produces ''high-resolution window'' forecasts from two versions of the Weather Research and Forecasting (WRF) model that differ in terms of physical parameterizations and dynamic cores. One configuration uses the Advanced Research WRF model (ARW-WRF; Skamarock et al. 2005) dynamic core, whereas the other uses the WRF Nonhydrostatic Mesoscale Model (WRF-NMM; Janjić et al. 2001; Janjić 2003) core. Both models are run without CP over domains covering three-fourths of the contiguous United States at ~4-km grid spacing. This development represents an important step in the progression of NWP within an operational setting.

Despite this exciting implementation, there remains considerable doubt regarding the appropriateness of 4-km grid spacing for the first generation of convection-allowing forecasts. For example, some operational centers have chosen to continue to parameterize convection at 4 km, though in modified forms, out of concern that abandoning CP altogether will result in unrealistic forecasts [e.g., the Met Office Unified Model (UM; Roberts and Lean 2008, hereinafter RL08); Japanese Meteorological Agency Nonhydrostatic Mesoscale Model (Narita and Ohmori 2007)]. Furthermore, models with grid spacing finer than 4 km almost certainly provide more realistic representations of physical processes in areas of sharp topographic, land use, and land–sea gradients. Additionally, several studies (e.g., Petch et al. 2002; Adlerman and Droegemeier 2002; Bryan et al. 2003; Xue and Martin 2006) have demonstrated that grid spacing on the order of 1 km or less is necessary to begin resolving convective-scale circulations and generate more realism on convective scales. In fact, there seems to be little disagreement among the research community that storm structure becomes more realistic as resolution is increased, perhaps epitomized by Weisman et al. (1997) treating output (albeit cautiously) from their 1-km simulation as ''truth.''

Moreover, the issue of 4-km suitability may have moved even further from clarity when two seemingly contradictory studies were published this past year. The first study (RL08) was based on work conducted at the Met Office using the UM. This work suggested that forecasts of heavy precipitation greatly improve when horizontal grid spacing is reduced from 4 to 1 km. In addition, RL08 found little 4-km improvement over a 12-km forecast. The second study (Kain et al. 2008, hereafter KA08) stemmed from work during the 2005 NOAA HWT Spring Experiment and focused on outputs from 2- and 4-km grid-spacing versions of the ARW-WRF model. KA08 observed that the 2-km configuration produced more realistic storm structures than the 4-km forecasts. Yet, using both objective and subjective verification (Kain et al. 2003a) techniques, KA08 concluded that both models provided virtually identical value in terms of next-day guidance to severe storm forecasters, as both configurations were remarkably similar in their representation of convective initiation, evolution, and mesoscale organizational mode. The two models were also found to be similar in terms of forecast quality. Although systematic differences in experimental designs (see section 5) between RL08 and KA08 may have been largely responsible for these dissimilar findings, the fact that such different results were achieved raises additional questions about 4-km grid spacing nonetheless.

The results of KA08 highlight one of the underlying challenges regarding evaluation of high-resolution models; namely, greater realism does not necessarily translate into greater forecast value or quality (as defined by Murphy 1993). Therefore, it might not be necessary to run NWP models at, for example, 1 km, if the same information can be gleaned from 4-km grid spacing. Additionally, statistical measures of quality do not always corroborate perceptions of value. Numerous studies have highlighted these inconsistencies. For example, Mass et al. (2002) found that 4-km model forecasts produced more realistic and valuable meteorological simulations than 12- and 36-km forecasts over the U.S. Pacific Northwest. However, they noted that objective measures

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

of forecast quality failed to substantiate the perceived 4-km improvement over the 12-km output. Along similar lines, although Done et al. (2004) found 4-km WRF forecasts were more valuable and realistic than 10-km WRF forecasts, the equitable threat score (ETS) applied to precipitation thresholds indicated the two forecasts were of similar quality.

This persistent conflict between realism, quality, and value—coupled with the different findings of KA08 and RL08—significantly increases the difficulty of determining how much resolution to include in future generations of operational NWP models given finite computational resources. The ultimate solution to this problem, though elusive, is very important given practical concerns regarding high-resolution modeling, since finer grid spacing comes at a substantial price. As increased resolution means additional computational demand and storage, doubling horizontal resolution alone requires approximately a 10 time cost increase. Additionally, higher-resolution models take longer to complete their integrations. The challenge is to find an optimal grid spacing that maximizes model forecast quality and value while justifying the cost.

In an attempt to address this important issue, this study provides a second look at 2- and 4-km ARW-WRF outputs and brings us closer to meeting this challenge. As in KA08, the focus is again on the utility of the ARW-WRF as a next-day guidance tool. Similarly, whereas KA08 was based on 2005 Spring Experiment data, this study uses data from the 2007 Spring Experiment (SE2007).

Although both Spring Experiments featured parallel ARW-WRF forecasts generated using 2- and 4-km grid spacing, the 2005 configurations introduced some ambiguity into the assessment of sensitivity to resolution, as the two models used different computational domains, initialization procedures, and vertical resolutions (see KA08). However, in 2007, identical model domains, initialization procedures, and number of vertical levels were used, and the 2- and 4-km model configurations only differed in terms of horizontal grid length. This setup permitted a cleaner isolation of the impact of horizontal grid spacing on ARW-WRF forecasts.

This study again examines 2- and 4-km ARW-WRF forecasts, but it does so from a somewhat different perspective than KA08. Whereas KA08 focused on the representation of severe convection, we shift here to a greater emphasis on heavy rainfall, as in RL08. Additionally, we embrace the verification approach of RL08. Finally, the 2- and 4-km outputs here are compared to operational CP-using 12-km North American Mesoscale (NAM; Black 1994) model forecasts to assess the impact of CP and further investigate grid-spacing sensitivity.

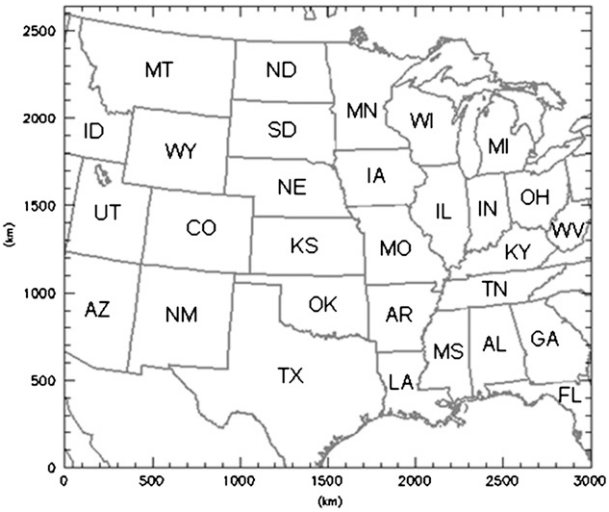

FIG. 1. Model integration domain for the WRF2 and WRF4 forecasts. States are labeled with their standard two-letter abbreviations.

The topics in this paper work in concert to address the question of how much resolution is needed to provide severe weather (including damaging convection and heavy rain) forecasters with reliable guidance at a justifiable cost. Model configurations are described next, followed by an overview of the verification procedures and a presentation of the results. Finally, implications of the results are discussed before concluding.

2. Model configurations

On each of the ~35 days of SE2007, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma produced forecasts from a single deterministic 2-km model and a 10-member ensemble prediction system with 4-km grid spacing (Xue et al. 2007; Kong et al. 2007). The models themselves were run remotely at the Pittsburgh Supercomputing Center (PSC). All ensemble members and the 2-km deterministic model used version 2.2 of the ARW-WRF core (Skamarock et al. 2005) with explicitly represented convection (no CP). The models were initialized at 2100 UTC and ran for 33 h over a domain encompassing approximately three-fourths of the continental United States (Fig. 1).

Other than the difference in horizontal grid spacing, the ensemble control member (WRF4) was configured identically to the 2-km model (WRF2).[2] For example, both used the same physical parameterizations, had

[2] To account for the decrease in grid spacing, the time step in the WRF2 model was approximately half that of the WRF4.

TABLE 1. Model configurations: Mellor–Yamada–Janjić (MYJ; Mellor and Yamada 1982; Janjić 2002); Ferrier (Ferrier 1994); WRF single-moment 6-class microphysics (WSM6; Hong et al. 2004; Hong and Lim 2006); and Betts–Miller–Janjić (BMJ; Betts 1986; Betts and Miller 1986; Janjić 1994).

	NAM	WRF2	WRF4
Dynamic core	WRF-NMM	ARW-WRF	ARW-WRF
Horizontal grid (km)	12	2	4
Initialization (UTC)	0000	2100	2100
Vertical levels	60	51	51
PBL parameterization	MYJ	MYJ	MYJ
Microphysics parameterization	Ferrier	WSM6	WSM6
Cumulus parameterization	BMJ	none	none

51 vertical levels, and employed a ''cold start'' without data assimilation. Initial conditions (ICs) were interpolated to the respective 2- and 4-km grids from the 2100 UTC analysis of the 12-km NAM (J. Du, NCEP EMC, 2007, personal communication) and 1800 UTC NAM forecasts provided the lateral boundary conditions (LBCs).

Additionally, WRF2 and WRF4 (collectively referred to as the high-resolution models) outputs were compared to forecasts from the 12-km operational NAM. However, although the high-resolution models differed just in terms of horizontal grid spacing, there were many differences between the NAM and ARW-WRF configurations (Table 1). Most significantly, the NAM used CP; used a different dynamical core (WRF-NMM; Janjić et al. 2001; Janjić 2003); integrated over a much larger domain; and was initialized at 0000 UTC, three hours later than the high-resolution forecasts.

In light of these many differences, disparities between the NAM and high-resolution forecasts cannot be attributed entirely to resolution and CP. Rather, inclusion of the NAM dataset provides a baseline and operational benchmark to which the high-resolution model performance can be compared. It is important to assess whether the high-resolution models can improve upon coarser-resolution model forecasts to justify the significantly higher computational cost of the high-resolution forecasts.

3. Verification procedures

Meaningful verification of high-resolution model forecasts is challenging. As grid spacing decreases, a model becomes capable of resolving progressively smaller-scale processes and features, such as individual thunderstorms. However, when the scale of these features is comparable to the model grid length, spatial displacement errors become significant and specific point values are likely to incur significant errors. Therefore, when measured by traditional point-by-point metrics, such as ETS, finer grids are the most heavily penalized for timing and displacement errors and often score relatively poorly (Gallus 2002). The information conveyed by the poor objective score, however, may directly contradict perceived value of the forecast.

In an attempt to reconcile the often-seen disparities between realism, value, and quality, a variety of nontraditional objective verification approaches have been developed for mesoscale model verification. One approach is to identify certain features (e.g., bow echoes, supercells) in the model forecast and compare them to corresponding entities seen in the observations. This method is the so-called object-based approach (see Ebert and McBride 2000; Done et al. 2004; Marzban and Sandgathe 2006). Another general approach is to relax the requirement that forecast and observed grid boxes match exactly in order for forecasts to be considered correct (Ebert 2008). This method is referred to as fuzzy verification or a neighborhood approach. In this study, the neighborhood method is used in lieu of the object-based approach. The specific verification methods are now discussed.

a. Subjective verification

A cornerstone of the HWT Spring Experiment is to foster lively discussion between researchers and forecasters. One method to promote this interaction is through the use of systematic subjective verification of experimental human-produced and model forecasts (Kain et al. 2003b, 2006). Experiment participants are encouraged to discuss their observations, and subjective ratings of forecast accuracy are assigned by group consensus. Since an element of personal bias is inherent with subjective evaluations, diversity of viewpoint is essential to minimizing the impact of predispositions (Kain et al. 2006). Such diversity was achieved in SE2007, with participants from both forecasting and research communities, including academic, government, private sector, and military affiliations.

Subjective verification activities occurred each weekday of SE2007. Although many model output fields were examined, subjective verification focused on comparisons

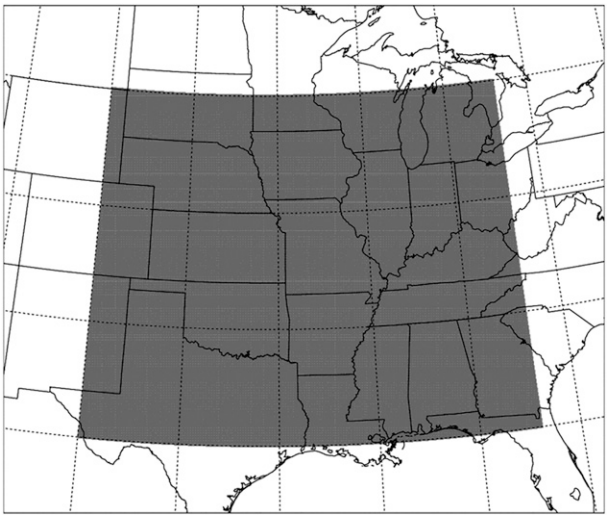

FIG. 2. Verification domain used for the model climatology.

of model-simulated 1 km above ground level (AGL) and observed lowest-elevation-angle radar reflectivity for lead times of 21–33 h (f21–f33). A national mosaic of observed radar reflectivity from Unisys was regarded as truth. These subjective evaluations were conducted over regional spatial domains that were relocated daily to correspond to the region where severe weather was deemed most likely to occur. Although the geographic domain shifted each day, its size remained constant throughout SE2007.

Subjective verification is important because it yields insight about human-perceived forecast value that traditional objective metrics do not measure well (Kain et al. 2003a; Done et al. 2004). Some of these objective measures are now discussed.

b. *Objective verification of model climatology*

At the conclusion of SE2007, average model performance characteristics were assessed by using several statistical measures applied primarily to hourly precipitation fields. Hourly model precipitation forecasts were compared to gridded Stage II precipitation fields produced hourly at NCEP (Lin and Mitchell 2005). Stage II precipitation fields are generated from radar-derived quantitative precipitation estimates 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. 2). This domain covered a large area over which Stage II data were robust and springtime weather was active. Attention was focused on the f21–f33 (1800–0600 UTC) period because this time frame corresponds to the typical diurnal peak of springtime convection and to examine the utility of the high-resolution models as next-day convective storm guidance.

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

		Observed		
		yes	no	
Forecast	yes	a	b	$a + b$
	no	c	d	$c + d$
		$a + c$	$b + d$	N

When possible, statistics were computed on native grids. However, to calculate certain performance metrics (discussed below), it was 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.

Statistical significance was assessed by using a bootstrap technique (Mullen and Buizza 2001). Resamples were randomly drawn from all days of SE2007 for each forecast hour, and verification metrics (discussed in the following subsections) were computed for each resample. This procedure was repeated 10 000 times to estimate the 95% confidence interval of each statistic (as in Figs. 12, 13, 15, 16). When the confidence bounds associated with two different models do not overlap, then the differences are statistically significant at the 97.5% level or higher.

1) POINT-BY-POINT VERIFICATION TECHNIQUES

Dichotomous (yes/no) forecasts are routinely verified against observations by the use of a 2 × 2 contingency table (Table 2). To use this table, the models and observations must be on the same grid, and therefore the model output was interpolated onto the verification grid. By selecting precipitation accumulation thresholds q (e.g., 1.0 mm h^{-1}) to define an event, each of the N grid points on the verification grid within the verification domain (Fig. 2; N = 204 073) were placed into their proper quadrants of Table 2 depending on the correspondence between the forecast F and observations O at that point. The ith grid point fell into category a if the event was correctly predicted ($F_i \geq q$ and $O_i \geq q$); b if the event was forecast but did not occur ($F_i \geq q$ and $O_i < q$); c if an event occurred but was not forecast ($F_i < q$ and $O_i \geq q$); and d if a nonevent was correctly predicted ($F_i < q$ and $O_i < q$). It follows that $a + b + c + d = N$.

A variety of metrics to assess model performance can be computed from the 2 × 2 contingency table. Among the myriad of scores are the bias B, threat score (TS; also known as the critical success index), and ETS. Bias is simply the ratio of the coverage of forecasts to the coverage of observations, given by $B = (a + b)/(a + c)$. For a

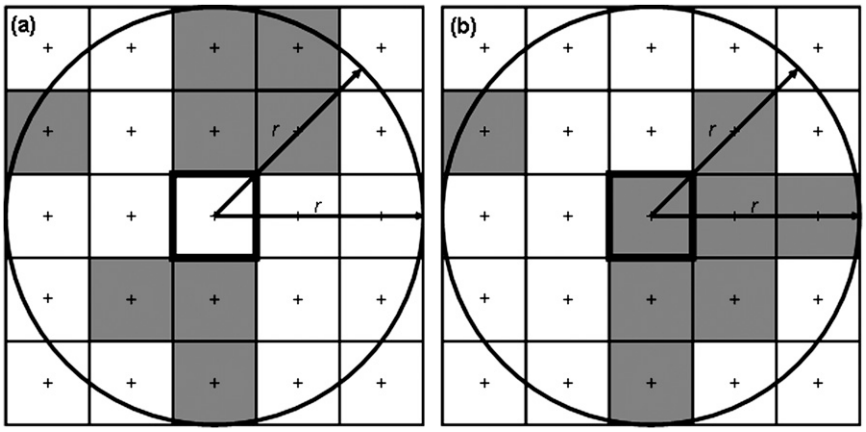

FIG. 3. Schematic example of neighborhood determination and fractional creation for (a) a model forecast and (b) its corresponding observations. Precipitation exceeds the accumulation threshold in the shaded boxes, and a radius of 2.5 times the grid length is specified.

given value of q, a $B > 1$ indicates overprediction and $B < 1$ indicates underprediction at that threshold. The TS is given by TS $= a/(a + b + c)$, ranges from 0 to 1, and is positively oriented. The TS can be made more "equitable" by adjusting the TS to account for "hits" (elements in quadrant a of Table 2) due to random chance. This correction is given by $e = (a + b)(a + c)/N$ and is used in the ETS [ETS $= (a - e)/(a + b + c - e)$]. ETS ranges from $-1/3$ to 1, with a perfect forecast achieving a score of 1.

2) A NEIGHBORHOOD APPROACH TO VERIFICATION

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 spacing has decreased in recent years to the size of convective-scale features, a variety of methods that incorporate a neighborhood around each grid point have been created to allow for spatial and/or temporal error or uncertainty (reviewed in Ebert 2008). One of these neighborhood techniques was developed by Roberts (2005) and RL08 and is adopted (with slight modifications) in this study. This technique is outlined in the following subsections.

(i) Creation of binary fields

As with the contingency table approach, model output was interpolated to the verification grid. Precipitation accumulation thresholds q were selected to define an event and convert both the observed O and model forecast F rainfall fields into binary grids. Grid boxes with accumulated precipitation $\geq q$ were assigned a value of 1 and all others were assigned a value of 0; that is, letting the subscript i denote the accumulated precipitation in the ith grid box,

$$B_{O(i)} = \begin{cases} 1 & \text{if} \quad O_i \geq q \\ 0 & \text{if} \quad O_i < q \end{cases} \quad \text{and} \quad B_{F(i)} = \begin{cases} 1 & \text{if} \quad F_i \geq q \\ 0 & \text{if} \quad F_i < q \end{cases}, \tag{1}$$

where $B_{O(i)}$ and $B_{F(i)}$ denote the newly created binary grids corresponding to the observational field and model output, respectively. Here, i ranges from 1 to N.

In addition to absolute accumulation thresholds, percentile thresholds were also used to create binary fields, as in RL08. For example, the yth percentile threshold (e.g., 95th percentile) selected the top $(100 - y)$ percent of forecast and observed accumulations to determine a new absolute threshold value q_y that corresponded to the yth percentile. We determined these values of q_y from a climatological perspective, where the climatological period included every hour during SE2007. Specifically, all grid points on the verification grid within the verification domain containing nonzero hourly precipitation accumulations were aggregated over all days of SE2007 separately for each model and the observations. The accumulations were ranked and the specific values of q_y were computed for different values of the yth percentile (see Fig. 14). Binary fields were obtained from Eq. (1), where the unique value of q_y corresponding to the particular model or observations was substituted in place of q. Using percentile thresholds removed the effect of bias and allowed for a robust comparison of spatial accuracy among the different models. Note that this approach differed somewhat from that of RL08, who computed q_y based on ranking accumulation values each output time, including points with zero accumulation.

(ii) Creation of fractional grids

After creating binary fields, a radius of influence r was specified (e.g., $r = 25$ or 50 km) to construct a

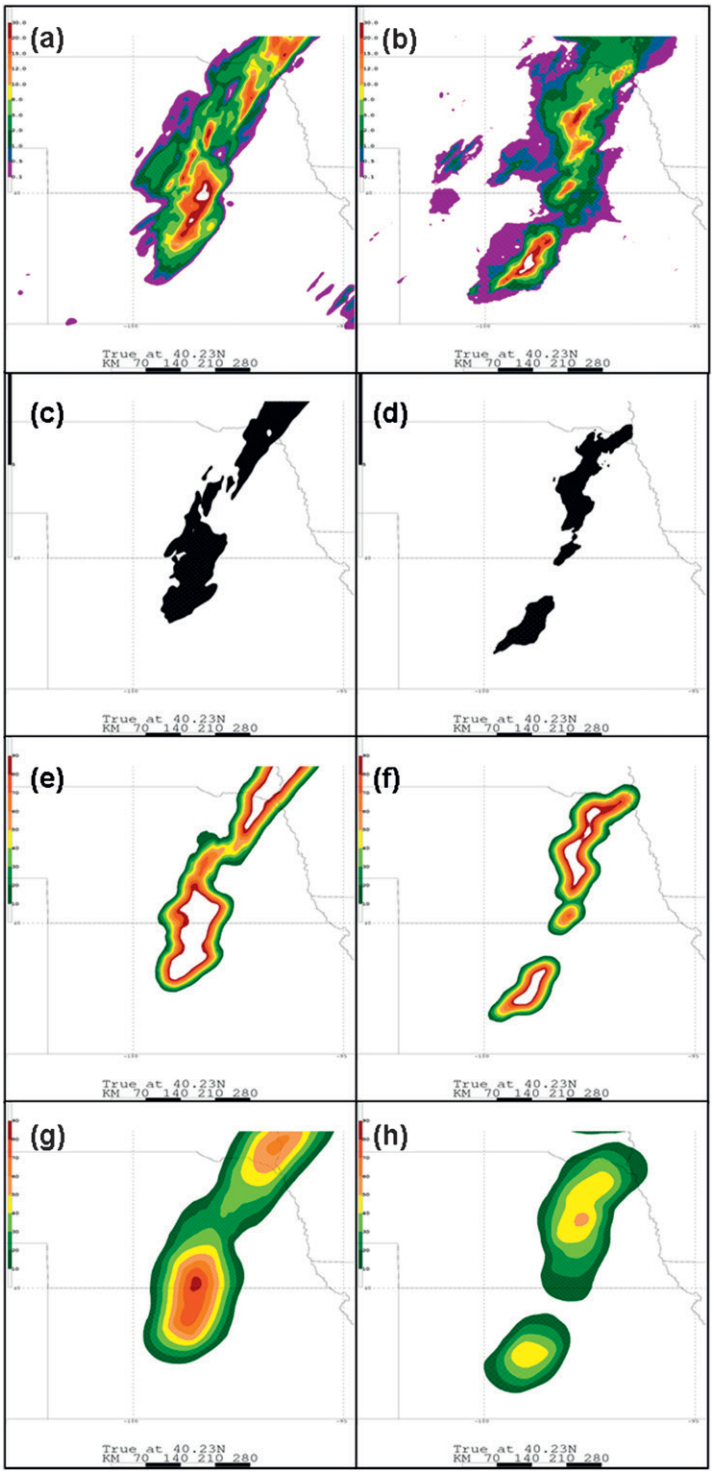

FIG. 4. (a) WRF4 1-h accumulated precipitation forecast; (b) observed 1-h precipitation accumulation. Binary image of precipitation accumulations exceeding 5.0 mm h^{-1} (shaded) using (c) WRF4 output and (d) observations; and fractional grids computed from (b) and (d) using radii of influence of (e),(f) 25 and (g),(h) 75 km. All panels are valid 0600 UTC 23 May 2007, and the WRF4 has been projected onto the verification grid.

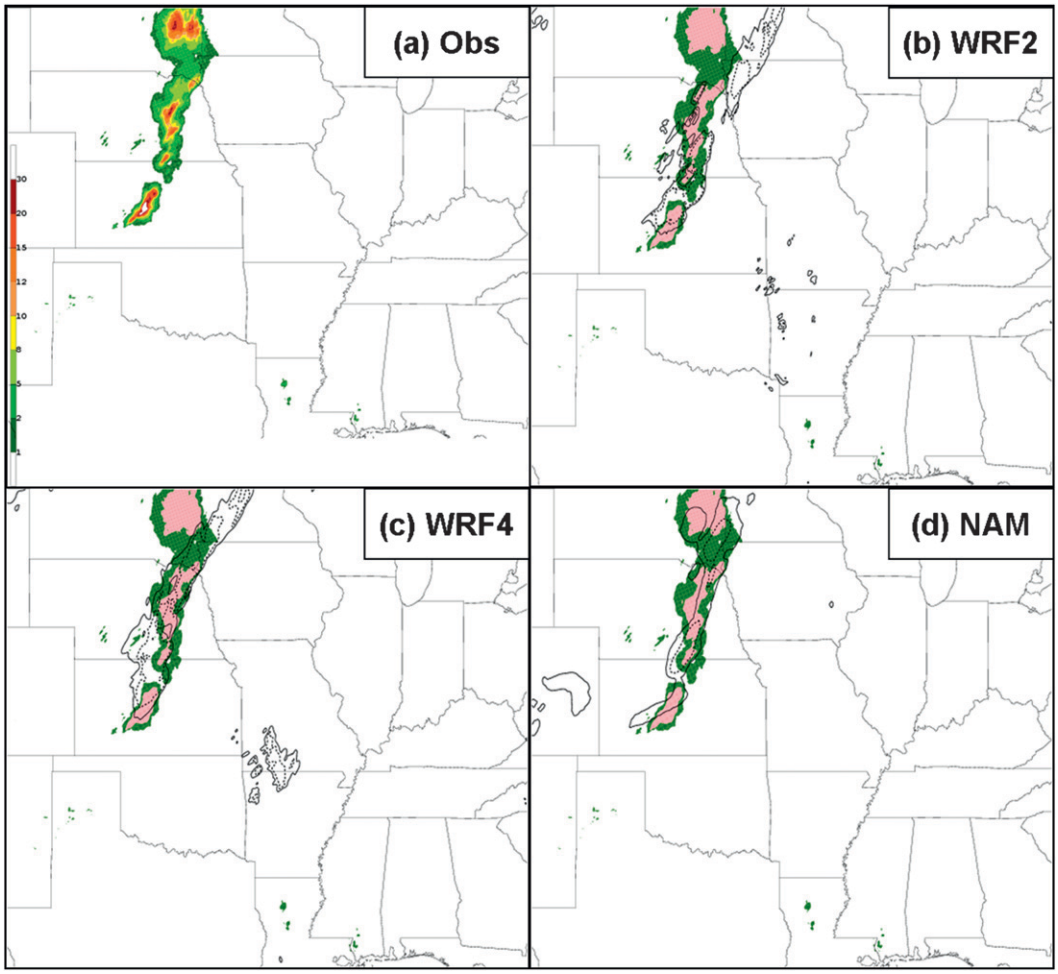

FIG. 5. (a) Observed 1-h precipitation, and 1-h model-forecast precipitation from (b) WRF2, (c) WRF4, and (d) NAM valid 0600 UTC 23 May 2007, with a 33-h forecast for the WRF2 and WRF4 and a 30-h NAM forecast. Shadings in (b),(c), and (d) indicate observed 1-h accumulated precipitation ≥ 1.0 and ≥ 5.0 mm h^{-1}, and dashed lines denote model forecasts at the same thresholds.

neighborhood around each grid box in the observed and forecast binary fields. All grid points surrounding a given point that fell within the radius were included in the neighborhood. Whereas RL08 constructed a square neighborhood around each grid box, a circular neighborhood was used in this study. Essentially, choosing a radius of influence defines a scale over which the model is expected to have accuracy, and this scale is applied uniformly in all directions from each grid point.

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

Figure 3 illustrates the determination of a neighborhood and computation of a fractional value for hypothetical observed and model forecasts valid at the same time, assuming a radius of influence equal to 2.5 times the grid spacing. Grid boxes whose centers lie within the radius 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. So, the forecast and observed fractions at the central grid box are both 8/21 (eight shaded squares within the neighborhood). 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

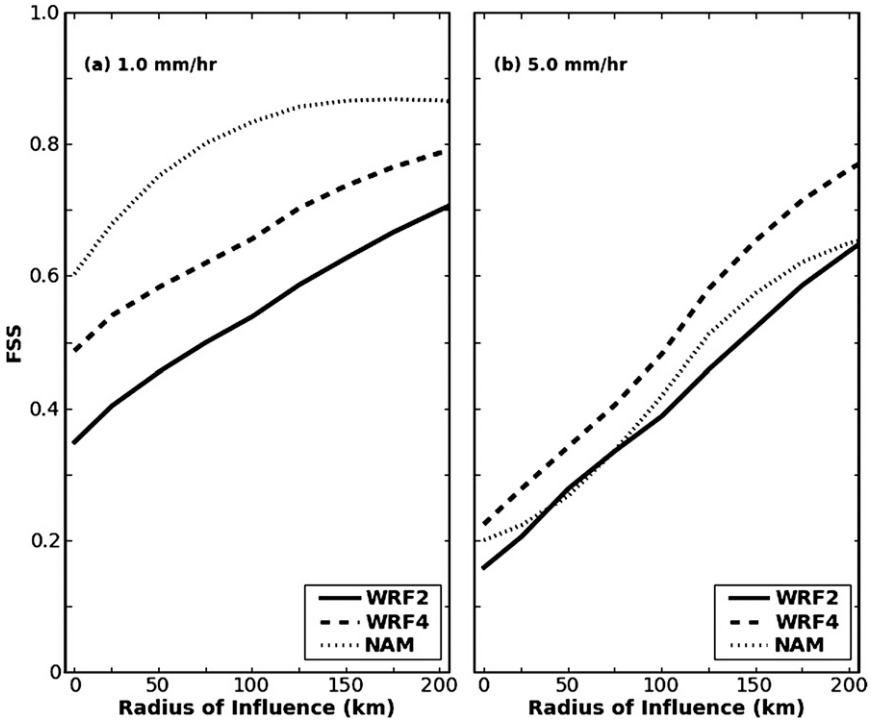

FIG. 6. FSS as a function of radius of influence for 1-h precipitation accumulations valid 0600 UTC 23 May 2007 using accumulation thresholds of (a) 1.0 and (b) 5.0 mm h^{-1}.

point-by-point verification), when the surrounding neighborhood is considered, the same probability as the observations is achieved. Therefore, in the context of a radius r, this forecast is considered correct.

Figure 4 illustrates fractional grid creation for a model forecast and the corresponding observations using q = 5.0 mm h^{-1}. Both the WRF4 forecast (Fig. 4, left column) and the observations (Fig. 4, right column) were valid at 0600 UTC 23 May—a lead time of 33 h. The raw, direct model output is depicted in Fig. 4a and the observations are depicted in Fig. 4b. Binary fields are shown in Figs. 4c,d. Probabilities generated with a radius of influence of 25 km (75 km) are depicted in Figs. 4e,f (Figs. 4g,h). Notice that as r increased from 25 to 75 km, the probabilities lowered, decreasing from over 90% to 70% (and even lower) over north-central Kansas and south-central Nebraska in the WRF4 forecast. The reduction of probabilities in central Kansas was even greater in the observed field. Evidently, in this case, as the radius of influence expanded to include more points in the neighborhood, few of these newly included points contained precipitation accumulations $\geq q$. In general, whether probabilities increase or decrease as the radius of influence changes is highly dependent on the spatial characteristics of the output field. However, for most situations, increasing r reduces the sharpness (RL08) and acts as a smoother that reduces probability gradients and the probability values themselves.

This approach can be applied to model output fields other than precipitation, such as simulated reflectivity, maximum surface winds, and updraft helicity. There exists an optimal value of r that maximizes forecast skill but does not sacrifice excessive mesoscale detail. However, at this time, this optimal value is unknown and this optimum may vary from model to model and parameter to parameter. In fact, Roberts (2008) suggests that the optimal radius of influence varies within a single model configuration and is a function of lead time. Results shown in section 4 include multiple values of r.

(iii) Calculation of fractions skill scores

The forecast and observed fractional grids were then compared to each other by use of a simple skill score. Formulation of this score begins with a variation on the Brier score (Brier 1950) called the fractions Brier score (FBS; Roberts 2005), which is given by

$$\mathrm{FBS} = \frac{1}{N}\sum_{i=1}^{N}[P_{F(i)} - P_{O(i)}]^2, \qquad (2)$$

where $P_{F(i)}$ and $P_{O(i)}$ are the fractional (probability) values at the ith grid box in the model forecast and observed

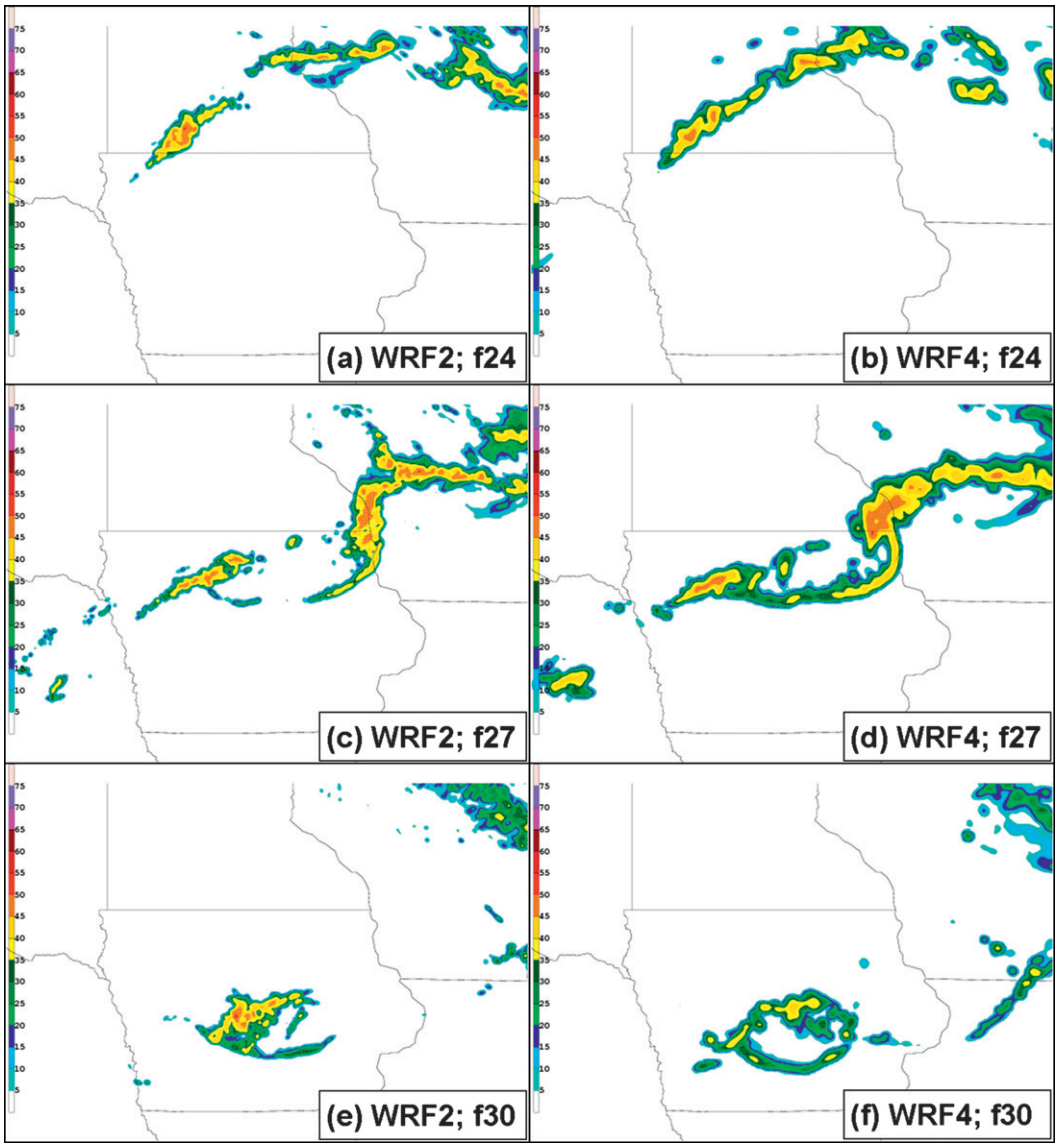

FIG. 7. (left) WRF2 and (right) WRF4 simulated 1-km AGL reflectivity forecasts valid (a),(b) 2100 UTC 30 Apr, (c),(d) 0000 UTC 1 May, and (e),(f) 0300 UTC 1 May.

probability fields, respectively. 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 nonzero fractions and is given by

$$\mathrm{FBS}_{\mathrm{worst}} = \frac{1}{N}\left[\sum_{i=1}^{N} P_{F(i)}^2 + \sum_{i=1}^{N} P_{O(i)}^2\right]. \tag{3}$$

On its own, the FBS does not yield much information because it is strongly dependent on the frequency of the event (i.e., points with no rain in either the observations or forecast can dominate the score). However, a skill score (after Murphy and Epstein 1989) can be constructed that compares the FBS with a low-skill reference forecast—$\mathrm{FBS}_{\mathrm{worst}}$—and is defined by Roberts (2005) as the fractions skill score (FSS):

$$\mathrm{FSS} = 1 - \frac{\mathrm{FBS}}{\mathrm{FBS}_{\mathrm{worst}}}. \tag{4}$$

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

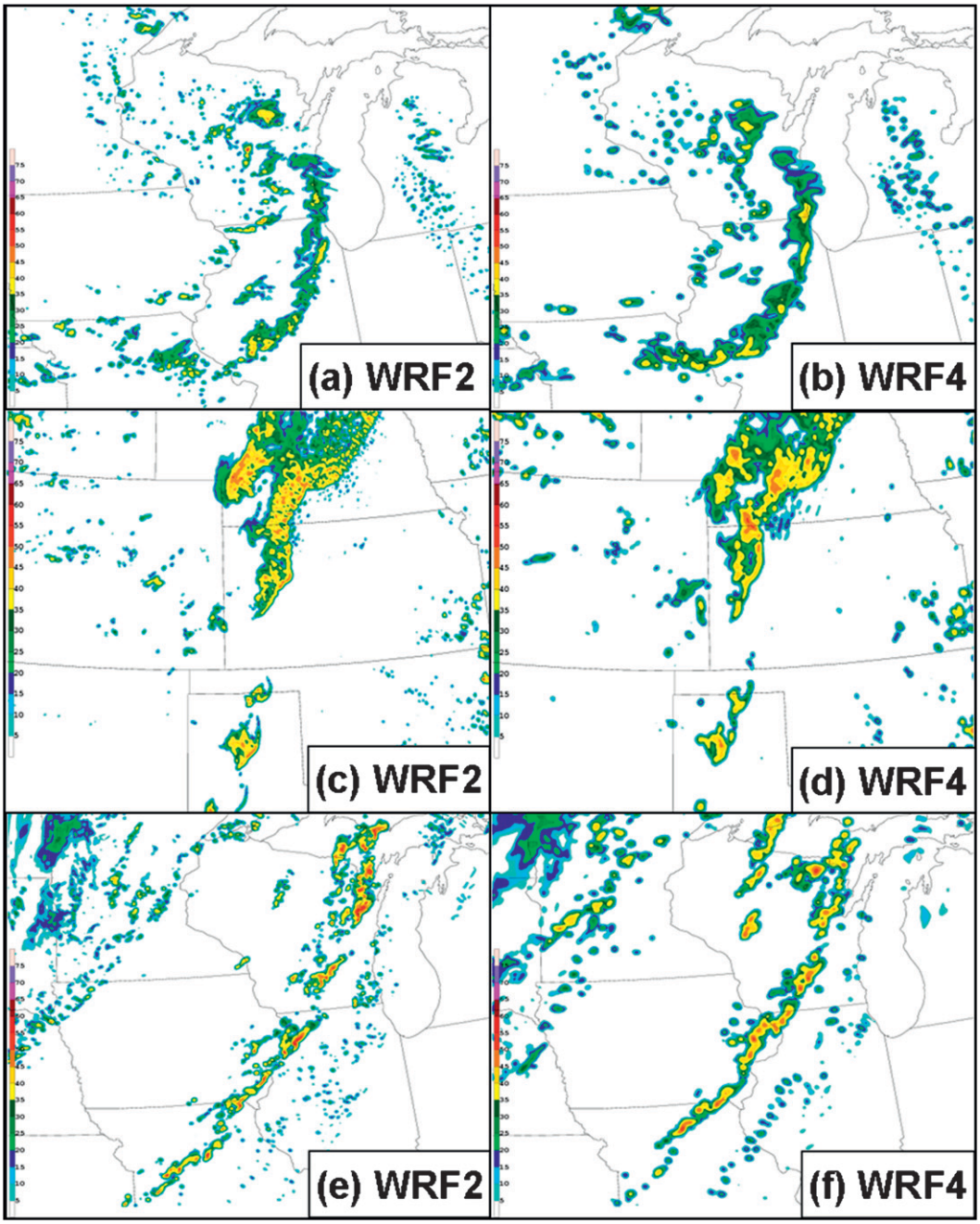

FIG. 8. As in Fig. 7, but valid 0000 UTC (a),(b) 17 May, (c),(d) 30 May, and (e),(f) 8 Jun.

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, asymptotically approaching a value of $2B/(B^2 + 1)$, where here again B is the bias (RL08).

(iv) FSS example

To illustrate how the FSS reflects visual impressions of actual model output, an example from SE2007 is presented. Figure 5 shows observed and model-forecast 1-h accumulated precipitation valid 0600 UTC 23 May 2007—a 33-h lead time for WRF2–WRF4 and a 30-h lead time for the NAM. All three models developed precipitation ahead of a cold front advancing through the Northern Plains region of the United States. However, the forecasts differed with regard to precipitation placement.

The color shadings in Figs. 5b–d outline areas of observed precipitation exceeding 1.0 and 5.0 mm h^{-1}.

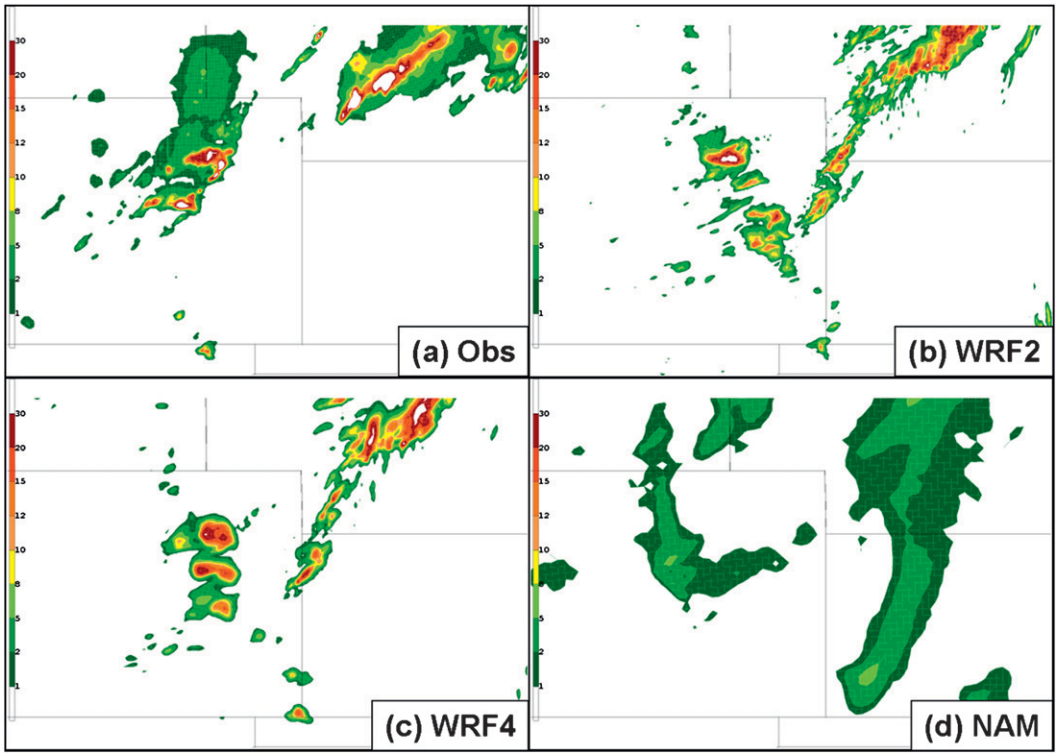

FIG. 9. One-hour (a) observed, (b) WRF2 forecast, (c) WRF4 forecast, and (d) NAM forecast accumulated precipitation valid 2100 UTC 29 May.

Model forecasts are overlaid, with the solid line corresponding to the 1.0 mm h^{-1} threshold and the dashed line enclosing areas forecast to receive at least 5.0 mm h^{-1} of precipitation.

The observations (Fig. 5a) indicate precipitation was oriented primarily north-northeast–south-southwest through southeastern South Dakota, eastern Nebraska, and central Kansas. At the 1.0 mm h^{-1} threshold, the NAM forecast (Fig. 5d) appeared to be in general agreement with the observations. However, the WRF4 (Fig. 5c) predicted more of a northeast–southwest alignment, bringing the precipitation well into Minnesota—the WRF2 (Fig. 5b) even more so. Additionally, both the WRF2 and WRF4 developed spurious convection in northern Arkansas and southwestern Missouri, whereas the NAM produced erroneous precipitation in eastern Colorado. From the figure, subjectively, it appears as if the NAM produced the best forecast at the 1.0 mm h^{-1} threshold and WRF2 produced the worst. This impression is confirmed by the FSS (Fig. 6a), with the NAM (WRF2) receiving the highest (lowest) score at all values of r.

At a threshold of 5.0 mm h^{-1}, however, the NAM did not maintain a precipitation area as large as the observations. Both the WRF4 and WRF2 produced a wider area of precipitation exceeding 5.0 mm h^{-1}, but much of it was displaced to the northeast in the WRF2 forecast. Although a northeastward displacement was also evident in the WRF4 output, an area exceeding 5.0 mm h^{-1} in northeast Nebraska was at least partially collocated with the observations. Given this partial overlap, WRF4 appeared to be best. Distinguishing between the WRF2 and NAM was more difficult. Although the NAM underpredicted the areal coverage, WRF2 generated the precipitation in the wrong area. Again, the corresponding FSS (Fig. 6b) confirms the subjective interpretation, with the highest score assigned to the WRF4 and roughly equal values for the NAM and WRF2.

4. Results

a. Subjective assessment of simulated reflectivity

Subjective ratings of 1-km AGL simulated reflectivity forecasts produced from the WRF2 and WRF4 (on their native grids) were assigned each day of SE2007. Specifically, the model representation of convective evolution, including initiation, coverage, mesoscale configuration, orientation, and movement, was assessed between 1800 and 0600 UTC (f21 and f33). Given their impressions

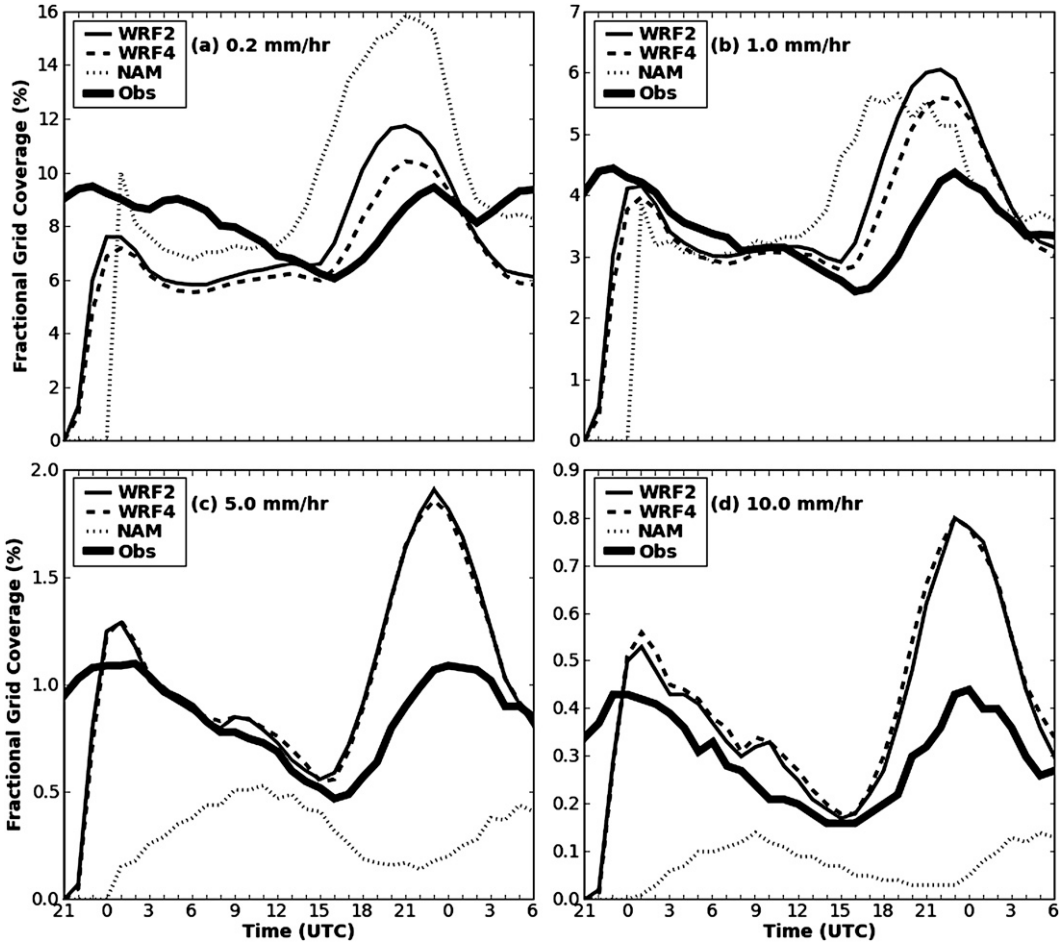

FIG. 10. Fractional grid coverage of hourly precipitation exceeding (a) 0.2, (b) 1.0, (c) 5.0, and (d) 10.0 mm h^{-1} as a function of time averaged over all days of SE2007 and calculated on each model's native grid over the verification domain.

over this 12-h period, participants scored each model's forecast on a scale from 0 to 10, where a score of 10 was reserved for a superior forecast and extremely poor forecasts received a score of 0.

Of the 22 days where subjective ratings for the WRF2 and WRF4 were available, the two models were assigned an equal score 16 times. Of the remaining six days, WRF2 scored a point higher than WRF4 four times, and WRF4 scored a point higher twice. It is noteworthy and revealing that the subjective ratings never differed by more than one point, implying that on the occasions when SE2007 participants perceived differences in convective evolution, they were not extreme.

The similar subjective ratings suggest that WRF2 and WRF4 1-km AGL reflectivity forecasts provided comparable value. These principles are further illustrated in Fig. 7, which depicts simulated reflectivity forecasts from the model runs initialized at 2100 UTC 29 April 2007. By 2100 UTC 30 April (f24), both models developed convection over southern Minnesota that stretched northeastward into northwestern and central Wisconsin (Figs. 7a,b). At this time, the convective mode was similar in both model forecasts with a broken line evident in each. By 0000 UTC 1 May (f27), bowing structures were present in both the WRF2 and WRF4 simulated reflectivity fields over roughly the same location in northeastern Iowa. Additionally, both models also increased in convective coverage and linear structure in central Wisconsin. By 0300 UTC 1 May (f30), both models weakened the convection as it moved through southern Wisconsin but maintained similarly oriented convection over central Iowa.

Upon closer scrutiny, there were subtle differences. For example, the WRF4 developed a more solid line than the WRF2 and exhibited greater curvature at f27. In addition, there was more finescale detail in the WRF2 reflectivity representation (e.g., Figs. 7c,d over central Wisconsin: the WRF2 was not as "blobular" as the WRF4).

This increased detail came as little surprise and was expected from a higher-resolution model. However, this added detail emerged on scales approaching the grid spacing, where there is little predictive skill. Overall, the character of mesoscale convective organization and evolution—including the development, propagation, and eventual decay of a bowing line segment—was remarkably insensitive to the grid spacing in this case.

Additional snapshots of simulated 1-km AGL reflectively valid at 0000 UTC (f27) 17 May (Figs. 8a,b), 30 May (Figs. 8c,d), and 8 June (Figs. 8e,f) are presented to further illustrate the similarity of the WRF2 and WRF4 reflectivity patterns that were observed routinely during SE2007. Though the individual elements were typically smaller on the WRF2 grids, these forecasts conveyed essentially the same mesoscale convective information as forecasts from the WRF4.

b. *Case study of precipitation fields*

To illustrate the differences between the high-resolution and NAM outputs, observed and forecast 1-h precipitation accumulations are shown in Fig. 9, valid at 2100 UTC 29 May 2007. The observations (Fig. 9a) indicated localized areas of intense precipitation in east-central Colorado on the southern end of a plume of lighter precipitation extending northward into Wyoming. These pockets of heavy precipitation corresponded to supercell thunderstorms that produced a few tornadoes. A second area of precipitation was observed over southern Nebraska.

The NAM (Fig. 9d) predicted an area of precipitation over eastern Colorado. However, it developed an extensive area of spurious precipitation in Kansas and overpredicted the areal coverage of precipitation in Nebraska. The WRF2 and WRF4 (Figs. 9b,c) forecasts appeared rather similar to each other. Both developed intense precipitation cores in Colorado slightly too far south and east and produced areas of rain in far northwestern Kansas that were not observed.

Although there were errors in both the NAM and high-resolution forecasts, the high-resolution forecasts revealed far more about the character of the precipitation than the NAM output. The NAM's broad outlines yielded little information about the likely convective mode of the day. On the other hand, the high-resolution (Figs. 9b,c) precipitation fields suggested that discrete cells were likely. This information about storm mode is quite valuable to severe weather forecasters and can increase confidence about the character of severe weather events. However, a forecaster relying solely on the NAM precipitation forecast would be unlikely to gain any such insight.

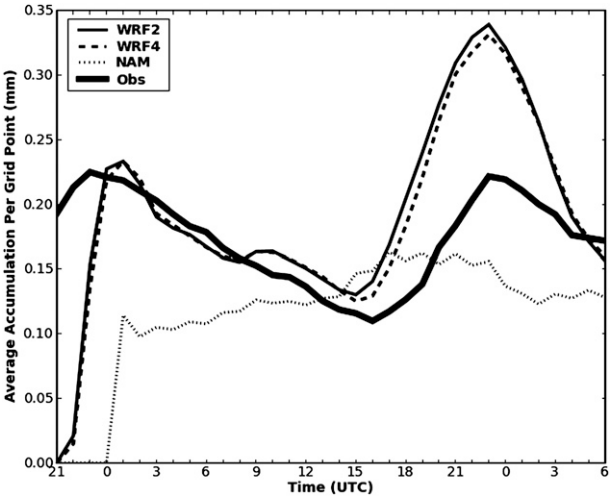

FIG. 11. Total precipitation over the verification domain aggregated over all days of SE2007, normalized by number of grid boxes and calculated on each model's native grid.

c. *Objective assessment of model climatology*

1) AREAL COVERAGES

Figure 10 depicts fractional coverages of precipitation exceeding various accumulation thresholds aggregated hourly over all days of SE2007. These statistics were generated from data on each model's native grid and computed over the verification domain (Fig. 2). The diurnal cycle was well captured in the WRF2 and WRF4 outputs, with an afternoon maximum corresponding well in time to the observations. However, a very high bias was noticed at the time of peak coverage.[3] Areal coverages of the two high-resolution models were similar, though the WRF2 produced, on average, somewhat more coverage of precipitation exceeding relatively low thresholds (<5 mm h^{-1}). Although these differences at lower thresholds were not significant in the statistical sense, this finding suggests that higher resolution enables the WRF2 to initiate more weak precipitation features than the WRF4. However, given the uncertainty regarding bias with the high-resolution models,

[3] The bias values measured in the WRF2 and WRF4 precipitation forecasts were considerably higher than corresponding values from other convection-allowing ARW-WRF forecasts examined during SE2007 that were initialized at 0000 UTC instead of 2100 UTC. Testing by CAPS scientists at the conclusion of SE2007 indicates that the high bias was significantly reduced when the models were initialized with 0000 UTC ICs and LBCs. Thus, it appears that some aspect associated with the 2100 UTC ICs led to the very high bias (Kong et al. 2008). Although this condition was less than optimal, it affected the WRF2 and WRF4 equally and should not detract from a meaningful comparison of the WRF2 and WRF4 forecasts.

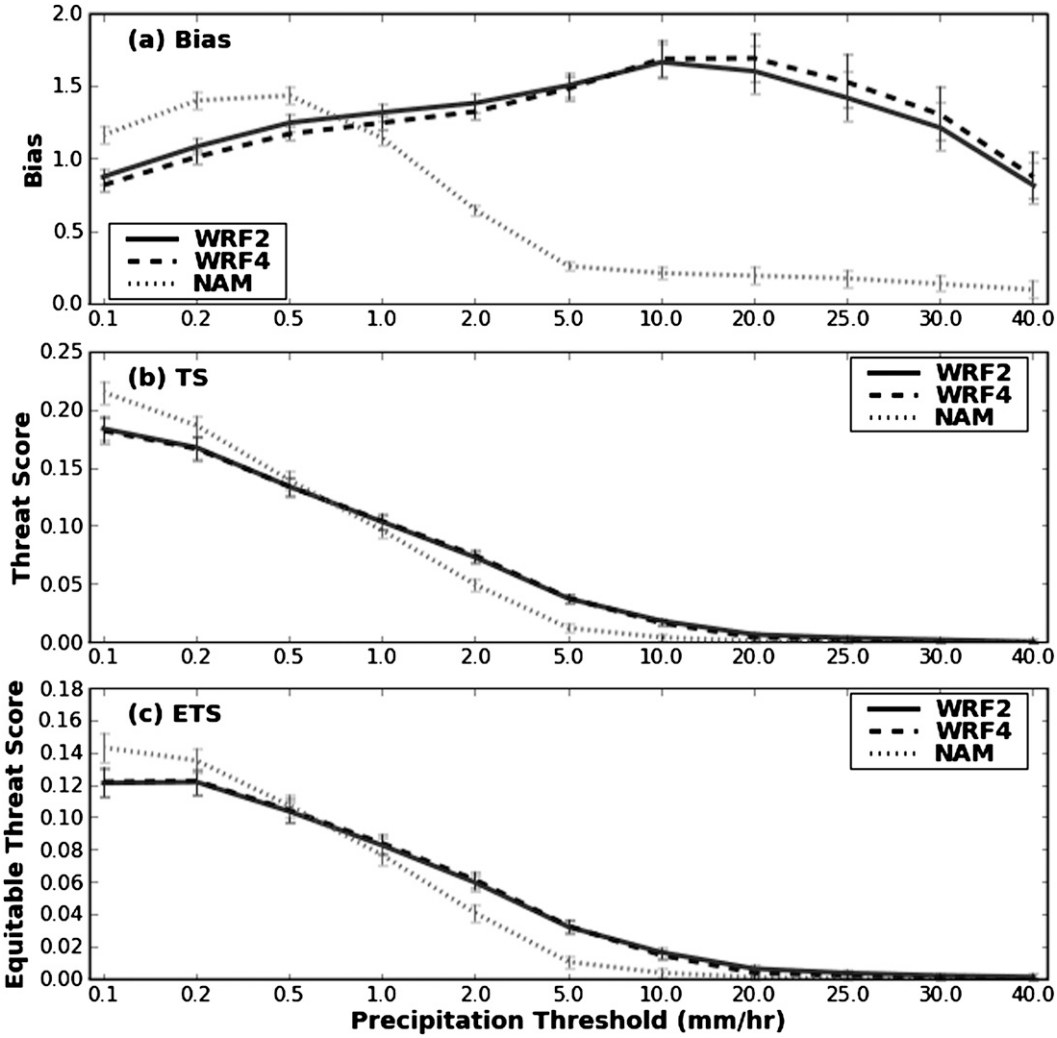

FIG. 12. (a) Bias, (b) TS, and (c) ETS as a function of accumulation threshold, based on hourly precipitation aggregated during 1800–0600 UTC (f21–f33) over all days of SE2007. Error bars denote the bounds of the 95% confidence interval.

it is not clear whether this difference is favorable. Regardless, as q was increased to 5.0 mm h^{-1} and beyond, the WRF2 and WRF4 areal coverages were nearly identical.

In contrast to the WRF2 and WRF4, the diurnal cycle was not well represented by the 12-km NAM at absolute thresholds $\geq$2.0 mm h^{-1}. In fact, at these thresholds, the NAM indicated a maximum coverage when the observations showed a relative minimum, and vice versa. At thresholds $\geq$5.0 mm h^{-1}, the NAM was incapable of resolving areas of heavier precipitation consistently, whereas at lower thresholds (e.g., 0.2 mm h^{-1}) the NAM generated precipitation over too large of an area. The tendency to produce broad areas of light precipitation and to underpredict the occurrence and coverage of heavy precipitation is characteristic of a model configuration that relies on parameterized, rather than explicit, prediction of deep, moist convection.

2) ACCUMULATED PRECIPITATION

Total accumulated precipitation throughout the verification domain, calculated on native grids and aggregated hourly over all days of SE2007, is depicted in Fig. 11. The WRF2 and WRF4 produced nearly the same amount of total precipitation, whereas the NAM generated lesser values. The differences between WRF2 and WRF4 were never statistically different, whereas the separation between the NAM and the high-resolution models was significant throughout the afternoon (f23–f30; 2000–0300 UTC). A high (low) bias was evident in the high-resolution models (NAM) during the afternoon convective period. Again, WRF2 and WRF4 accurately

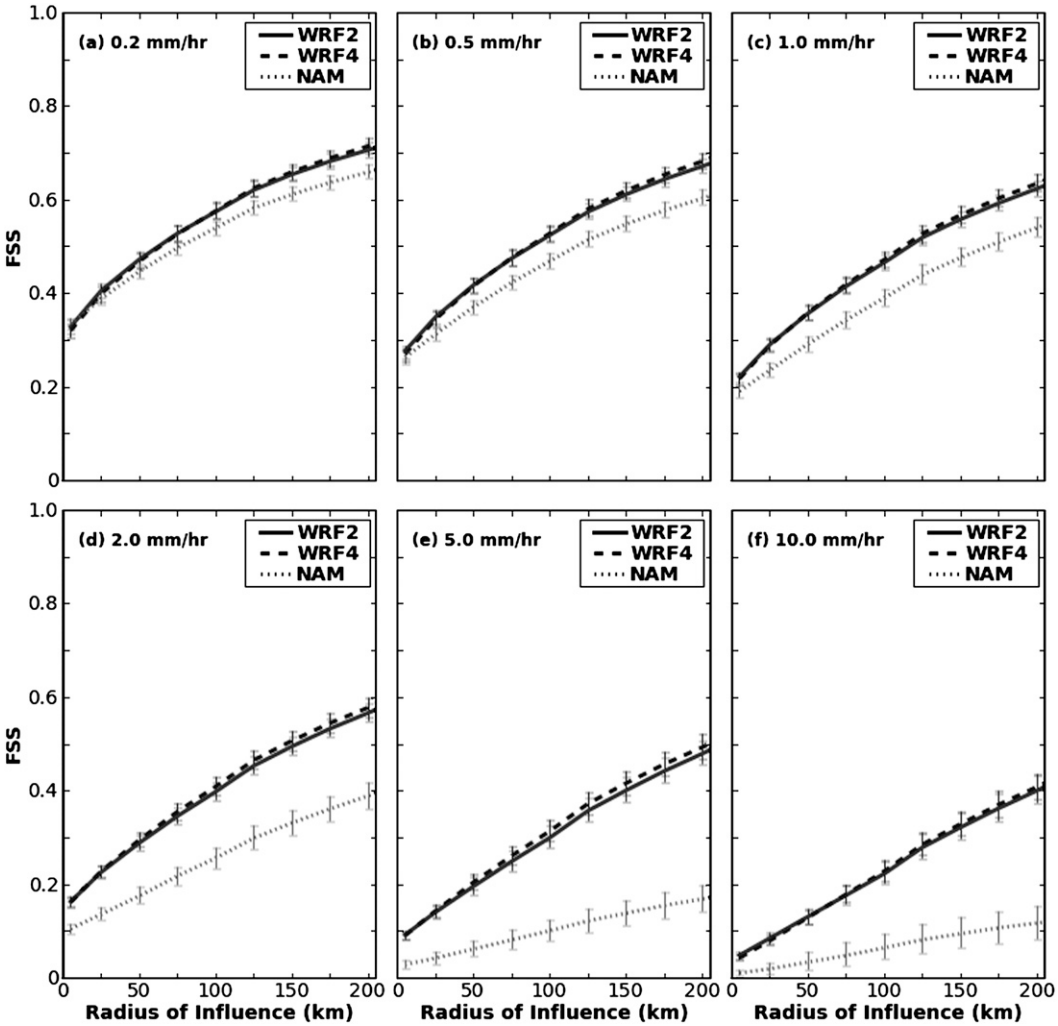

FIG. 13. FSS as a function of radius of influence, based on hourly precipitation aggregated during 1800–0600 UTC (f21–f33) over all days of SE2007 using accumulation thresholds of (a) 0.2, (b) 0.5, (c) 1.0, (d) 2.0, (e) 5.0, and (f) 10.0 mm h^{-1}. Error bars denote the bounds of the 95% confidence interval.

depicted the timing, but not the amplitude, of the diurnal cycle, whereas the NAM struggled with both amplitude and timing.

3) CONTINGENCY TABLE METRICS

Bias, TS, and ETS based on hourly precipitation aggregated over all days of SE2007 between 1800 and 0600 UTC (f21 and f33) are plotted as a function of precipitation threshold in Fig. 12. The WRF2 and WRF4 bias scores (Fig. 12a) indicated overprediction at all but extremely low and high accumulation thresholds. On the other hand, at low exceedance thresholds, the NAM displayed a tendency to overforecast the precipitation area, but at higher thresholds its bias was very low: the NAM was simply unable to generate areas of intense precipitation consistently. As the TS rewards overforecasting (Baldwin and Kain 2006), this skill score was highest for the NAM at $q = 1.0$ and 2.0 mm h^{-1} (Fig. 12b). Differences in TS and ETS for the WRF2 and WRF4 were not statistically significant, indicating little difference in forecast skill using a point-by-point verification approach. However, the high-resolution models outperformed the NAM at thresholds between 2.0 and 10.0 mm h^{-1} (Figs. 12b,c). Above the 10.0 mm h^{-1} threshold, none of the three models showed appreciable skill as measured by the contingency table metrics.

4) FRACTIONS SKILL SCORES

FSS based on hourly precipitation 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 expected, the FSS improved

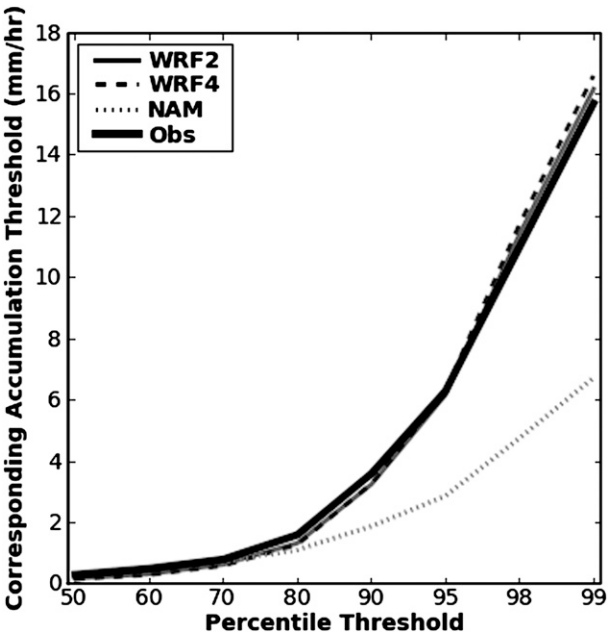

FIG. 14. Precipitation climatology: percentiles calculated from all grid points on the verification grid within the verification domain containing nonzero hourly precipitation accumulations aggregated over all hours and days of SE2007. This procedure was performed separately for each model and the observations (see text).

as r increased. However, as q increased, the FSS worsened at all scales, indicating the models had the least skill at predicting heavy precipitation events. At all precipitation thresholds, there was no improvement of the WRF2 forecasts over those of the WRF4, indicating the two models performed similarly at all thresholds and spatial scales. But, especially for higher values of q and r, the high-resolution models showed a substantial improvement over the NAM output that was statistically significant at the 97.5% level. Thus, even though forecast quality degraded at higher exceedance thresholds, high-resolution improvement over the NAM was maximized at these levels. The large high-resolution improvement at higher absolute thresholds was partially due to the NAM's inability to consistently generate high precipitation totals, as evidenced by its low bias (Fig. 12a). As a result, fractions generated from the NAM output were generally low, leading to a large numerator in Eq. (4), which decreased the FSS.

When climatological percentile thresholds (Fig. 14) were used to define an event, similar results were obtained (Fig. 15). Again, the high-resolution models showed the greatest improvement over the NAM at the highest percentile thresholds and there was no statistical difference between the WRF2 and WRF4 scores. As the use of percentile thresholds in the FSS permits a robust assessment of spatial accuracy, Fig. 15 indicates that the WRF2 and WRF4 were more skillful at placing precipitation than the NAM. These findings suggest that even if the NAM could be downscaled, the high-resolution models would still be superior in terms of spatial accuracy.

FSS aggregated hourly over all days of SE2007 is shown in Fig. 16 for various values of r and an accumulation threshold of 5.0 mm h^{-1}. There was little difference between the high-resolution models, although the WRF4 performed slightly better toward the end of the integration (though this improvement was not statistically significant). Both the WRF2 and WRF4 demonstrated more skill than the NAM through the vast majority of the period, with the gap widening at larger values of r.

5. Discussion

Our results corroborate the findings of KA08. Although the WRF2 in the present study produced more detailed storm structures than the WRF4 output, the differences were on scales approaching the resolution limits of the model configurations, where there is little predictive skill. In terms of overall representation of convective evolution, subjective verification and visual inspection indicated the WRF2 and WRF4 behaved similarly on most days. This general consistency suggests that WRF2 and WRF4 simulations are likely to provide comparable value as guidance for the prediction of next-day disruptive convective events, such as severe thunderstorms and heavy rain/flash floods. Moreover, forecast quality, as measured by the FSS, showed little objective difference between the high-resolution forecasts, on average, over the course of SE2007, indicating similar skill at all spatial scales.

On the other hand, the high-resolution models improved significantly upon the NAM, producing more skillful precipitation forecasts at all spatial scales. It is noteworthy that the high-resolution improvement over the NAM was maximized at higher accumulation thresholds, as these thresholds correspond to relatively extreme events. Additionally, the high-resolution models provided added value in terms of convective-mode guidance, consistent with previous studies (e.g., Kain et al. 2006; Weisman et al. 2008) and an important benefit for severe weather forecasters. As tools to improve heavy precipitation and severe weather forecasting are quite valuable, these findings lend additional evidence to suggest high-resolution, convection-allowing models may have much to offer to the forecasting community.

However, our results seem to contradict those of RL08, who used the FSS to demonstrate 1-km forecast

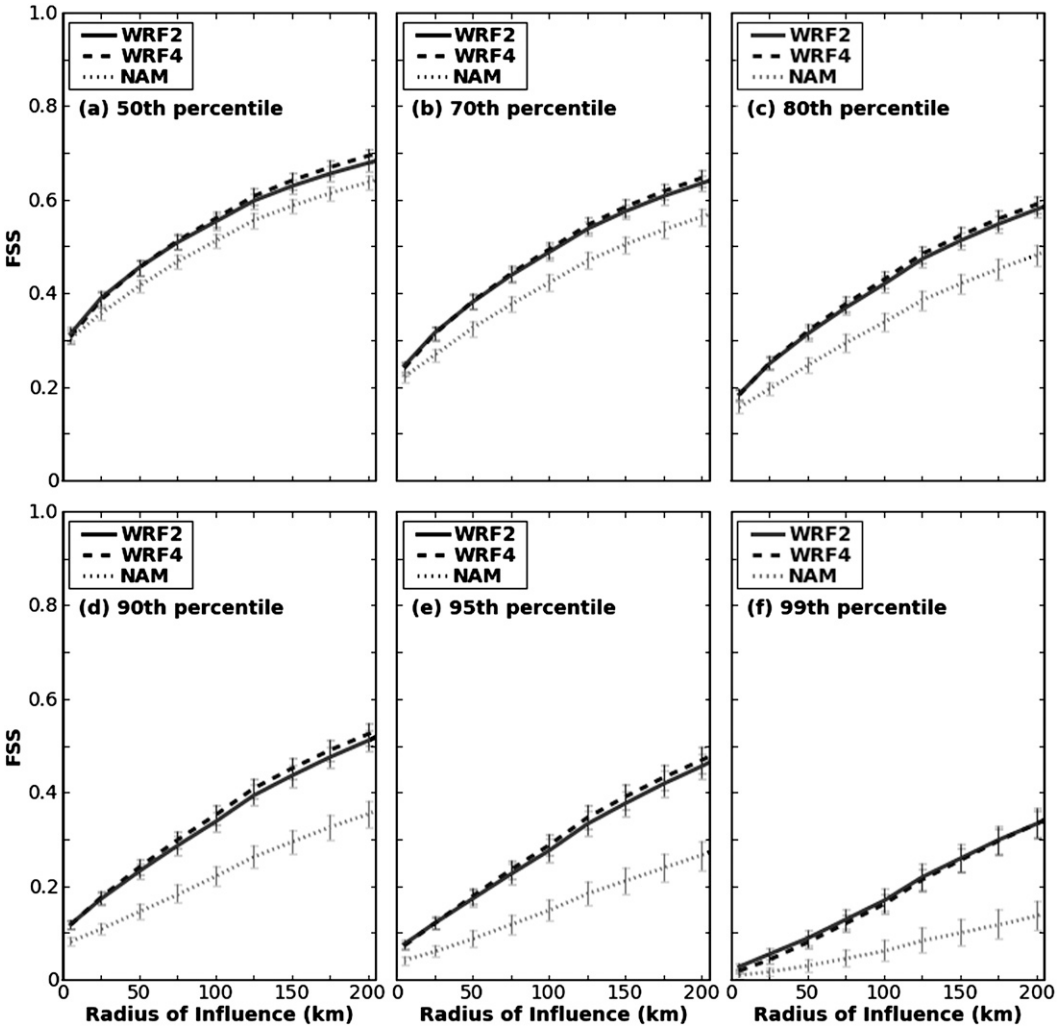

FIG. 15. As in Fig. 13, but by using climatological percentile thresholds (see Fig. 14) of (a) 50%, (b) 70%, (c) 80%, (d) 90%, (e) 95%, and (f) 99%.

superiority to 4-km forecasts in the UM. They also concluded a 4-km configuration of the UM performed little, if any, better than a 12-km version. The dissimilar conclusions of RL08 and the present study can be attributed partly to different experimental designs. For example, in this experiment the 4-km model explicitly resolved convection, whereas RL08 used a modified form of CP at 4 km. Furthermore, the studies examined model forecasts produced by completely different dynamic cores (ARW-WRF versus UM). In addition, all models in this study were run from a cold start without any data assimilation, whereas RL08 employed data assimilation in their 12-km forecasts (which were used to spin up their 4- and 1-km forecasts). Differences in geographical features and typical weather between the two test regions may have played a major role. Perhaps most importantly, RL08 focused on the first 7 h of model integration, whereas the focus here was on model forecast times between 21 and 33 h.

Nonetheless, despite these many differences in experimental design and differing conclusions, the results presented herein can be reconciled somewhat with RL08. Based on the findings of RL08, it seems that higher resolution may be more important early in the model forecast cycle when storm-scale information is present in the initial conditions. However, as storm-scale predictability on both 2- and 4-km grids is lost as forecast time increases (Zhang et al. 2003; Hohenegger and Schär 2007), we hypothesize that higher resolution is less important at later forecast periods, when cumulative small-scale error growth renders the smallest scales less predictable. Clearly, more work is needed to clarify these issues and there remains a substantial challenge of determining how much resolution to include in future NWP models.

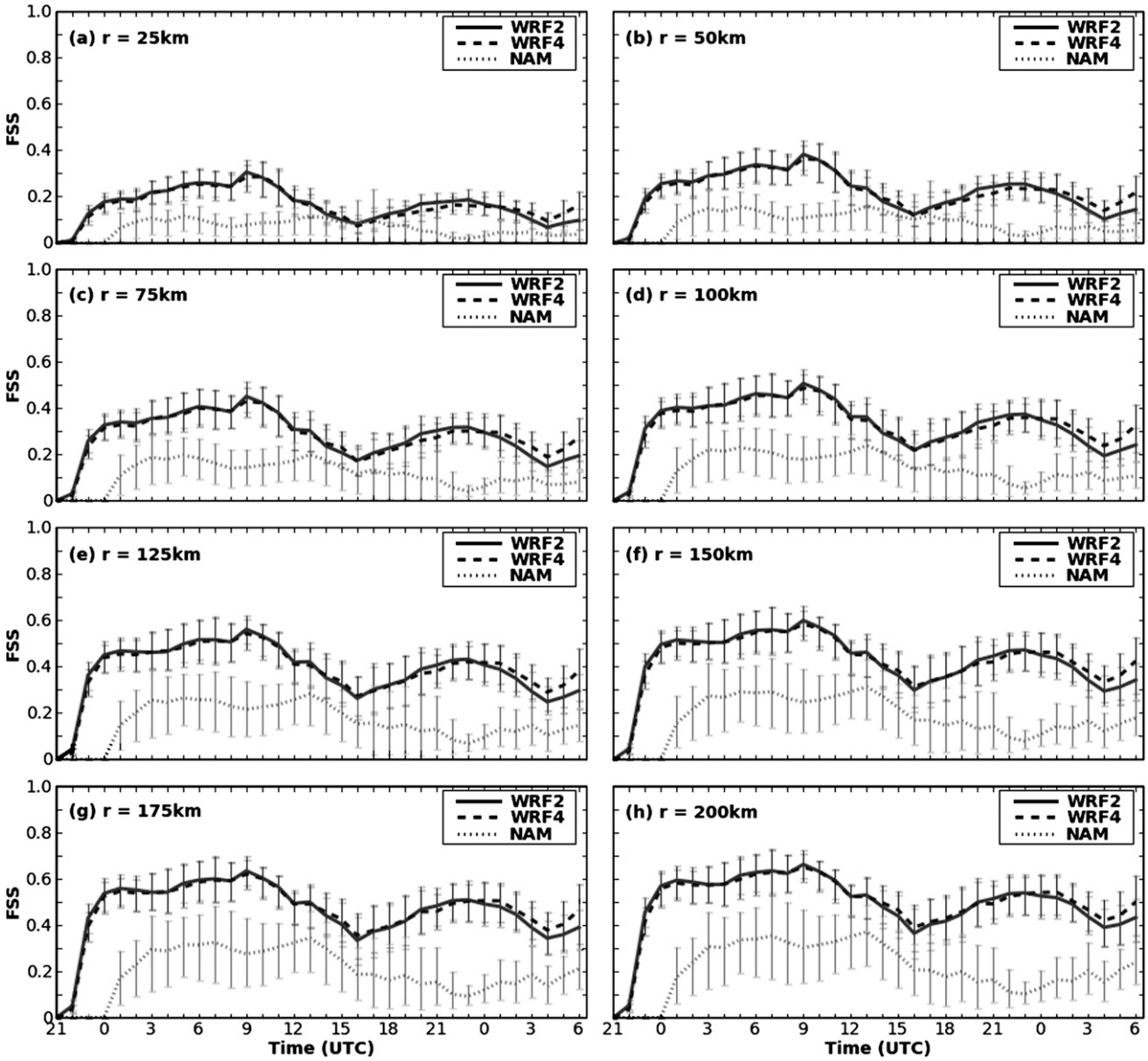

FIG. 16. FSS using a threshold of 5.0 mm h^{-1} as a function of time for a radius of influence of (a) 25, (b) 50, (c) 75, (d) 100, (e) 125, (f) 150, (g) 175, and (h) 200 km averaged over all days of SE2007. Error bars denote the bounds of the 95% confidence interval.

6. Summary and conclusions

During SE2007, convection-allowing 2- and 4-km configurations of the ARW-WRF were run over a large domain encompassing much of the United States. Aside from the difference in horizontal grid spacing, the configurations were otherwise identical, allowing for a clean isolation of the impact of horizontal resolution on next-day ARW-WRF forecasts. Forecasts from the 12-km NAM were also considered in order to provide an operational benchmark for the high-resolution output.

Using subjective verification techniques, the convection-allowing, high-resolution models (horizontal grid spacings of 2 and 4 km) were found to provide significant added value for next-day forecasts compared to the operational NAM. For example, the high-resolution forecasts provided useful information regarding the mesoscale organizational mode of convection that was not available from the NAM. Since the characteristics of severe convection appear to be strongly linked to convective mode, this guidance is particularly valuable. Moreover, these added details did not result in degradation of forecast quality, as indicated by the improved high-resolution FSS relative to the NAM. Additionally, the results suggest that convection-allowing models are substantially more skillful at predicting both the location

and amplitude of heavy rain events than mesoscale models employing CP, thus holding great promise for the hydrometeorological community.

In those areas where the high-resolution models showed distinct improvements over the NAM, 4-km grid length seems to be nearly as advantageous as 2-km spacing. Although the WRF2 produced finer-scale structures that were likely more realistic, the WRF2 and WRF4 appeared to provide comparable value as guidance for the prediction of convective mode and placement and intensity of heavy rainfall. Thus, for severe weather and heavy rainfall forecasting applications, there should be no rush to decrease horizontal grid spacing beyond 4 km, and it seems difficult to justify the added cost to run large-domain forecasts at 2-km grid spacing rather than at 4-km spacing, at least for forecast periods exceeding ~12 h. In essence, it appears that although substantial and obvious benefits are realized when NWP systems are reconfigured from models with ~10-km grid spacing and CP to ~4-km grid length with explicit convection, the added benefits of further incremental decrease in grid spacing are much smaller by comparison.

However, this conclusion should not be seen as pessimistic about the future of operational high-resolution modeling. Rather, instead of immediately increasing the resolution further as computers become evermore powerful, resources can be devoted to ensemble forecasting and postprocessing algorithms (see Schwartz et al. 2009). In fact, some new postprocessing methods have shown promise at outlining areas of severe weather and heavy rainfall potential when applied to output from convection-allowing ARW-WRF models with ~4-km grid spacing (Schwartz et al. 2009; Sobash et al. 2008). However, as data assimilation techniques advance and especially when it becomes conceivable to assimilate and predict the evolution of specific storm-scale features, this conclusion regarding 2-km versus 4-km grid length may change, especially for forecast periods less than ~12 h. However, until these advances are realized and significantly more computational resources become available, 4-km grid spacing seems to provide a robust configuration for the first generation of convection-allowing NWP models.

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 Geoffrey Manikin (NCEP EMC) for providing us with the NAM precipitation data and Jun Du of NCEP for making available the 2100 UTC NAM analyses. Nusrat Yussouf and Dr. Kimberly Elmore are thanked for their assistance with statistical analysis. The CAPS forecasts were primarily supported by the NOAA/CSTAR program and were performed at the PSC supported by the National Science Foundation. 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

Adlerman, E. J., and K. K. Droegemeier, 2002: The sensitivity of numerically simulated cyclic mesocyclogenesis to variations in model physical and computational parameters. *Mon. Wea. Rev.,* **130,** 2671–2691.

Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting,* **21,** 636–648.

Betts, A. K., 1986: A new convective adjustment scheme. Part I: Observational and theoretical basis. *Quart. J. Roy. Meteor. Soc.,* **112,** 677–691.

——, and M. J. Miller, 1986: A new convective adjustment scheme. Part II: Single column tests using GATE wave, BOMEX, and arctic air-mass data sets. *Quart. J. Roy. Meteor. Soc.,* **112,** 693–709.

Black, T. L., 1994: The new NMC mesoscale 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.

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.

Davis, C., and Coauthors, 2004: The Bow Echo and MCV Experiment (BAMEX): Observations and opportunities. *Bull. Amer. Meteor. Soc.,* **85,** 1075–1093.

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.

Droegemeier, K. K., and Coauthors, 1996a: Realtime numerical prediction of storm-scale weather during VORTEX '95. Part I: Goals and methodology. Preprints, *18th Conf. on Severe Local Storms,* San Francisco, CA, Amer. Meteor. Soc., 6–10.

——, and Coauthors, 1996b: The 1996 CAPS spring operational forecasting period: Realtime storm-scale NWP, Part I: Goals and methodology. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 294–296.

Ebert, E. E., 2008: Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.,* **15,** 51–64.

——, and J. L. McBride, 2000: Verification of precipitation in weather systems: Determination of systematic errors. *J. Hydrol.,* **239,** 179–202.

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. J., 2002: Impact of verification grid-box size on warm-season QPF skill measures. *Wea. Forecasting,* **17,** 1296–1302.

Hohenegger, C., and C. Schär, 2007: Predictability and error growth dynamics in cloud-resolving models. *J. Atmos. Sci.,* **64,** 4467–4478.

Hong, S.-Y., and J.-O. J. Lim, 2006: The WRF single-moment microphysics scheme (WSM6). *J. Korean Meteor. Soc.,* **42,** 129–151.

——, 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.

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

——, 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.

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

——, P. R. Janish, S. J. Weiss, M. E. Baldwin, R. S. Schneider, and H. E. Brooks, 2003b: Collaboration between forecasters and research scientists at the NSSL and SPC: The Spring Program. *Bull. Amer. Meteor. Soc.,* **84,** 1797–1806.

——, 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.

——, 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: Real-time storm-scale ensemble forecast experiment—What's new? Preprints, *Ninth WRF Users' Workshop,* Boulder, CO, NCAR, 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.]

Marzban, C., and S. Sandgathe, 2006: Cluster analysis for verification of precipitation fields. *Wea. Forecasting,* **21,** 824–838.

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

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

Mullen, S. L., and R. Buizza, 2001: Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.,* **129,** 638–663.

Murphy, A. H., 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.

Narita, M., and S. Ohmori, 2007: Improving precipitation forecasts by the operational nonhydrostatic mesoscale model with the Kain-Fritsch convective parameterization and cloud microphysics. Preprints, *12th Conf. on Mesoscale Processes,* Waterville Valley, NH, Amer. Meteor. Soc., 3.7. [Available online at http://ams.confex.com/ams/pdfpapers/126017.pdf.]

Petch, J. C., A. R. Brown, and M. E. B. Gray, 2002: The impact of horizontal resolution on the simulations of convective development over land. *Quart. J. Roy. Meteor. Soc.,* **128,** 2031–2044.

Roberts, N. M., 2003: Results from high-resolution simulations of convective events. Met Office Tech. Rep. 402, 47 pp.

——, 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: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting,* in press.

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, and J. G. Powers, 2005: A description of the Advanced Research WRF version 2. NCAR Tech. Note, NCAR/TN-468+STR, 88 pp.

Sobash, R., D. R. Bright, A. R. Dean, J. S. Kain, M. Coniglio, S. J. Weiss, and J. J. Levit, 2008: Severe storm forecast guidance based on explicit identification of convective phenomena in WRF-model forecasts. Preprints, *24th Conf. on Severe Local Storms,* Savannah, GA, Amer. Meteor. Soc., 11.3. [Available online at http://ams.confex.com/ams/pdfpapers/142187.pdf.]

Speer, M. S., L. M. Leslie, and L. Qi, 2003: Numerical prediction of severe convection: Comparison with operational forecasts. *Meteor. Appl.,* **10,** 11–19.

Steppeler, J., G. Doms, U. Schättler, H. W. Bitzer, A. Gassmann, U. Damrath, and G. Gregoric, 2003: Meso-gamma scale forecasts using the nonhydrostatic model LM. *Meteor. Atmos. Phys.,* **82,** 75–96.

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

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

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

Xue, M., and W. J. Martin, 2006: A high-resolution modeling study of the 24 May 2002 case during IHOP. Part I: Numerical simulation and general evolution of the dryline and convection. *Mon. Wea. Rev.,* **134,** 149–171.

——, and Coauthors, 1996a: Real-time numerical prediction of storm-scale weather during VORTEX-95. Part II: Operation summary and examples. Preprints, *18th Conf. on Severe Local Storms,* San Francisco, CA, Amer. Meteor. Soc., 178–182.

——, and Coauthors, 1996b: The 1996 CAPS spring operational forecasting period: Realtime storm-scale NWP, Part II: Operational summary and sample cases. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 297–300.

——, and Coauthors, 2007: CAPS realtime 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.]

Zhang, F., C. Snyder, and R. Rotunno, 2003: Effects of moist convection on mesoscale predictability. *J. Atmos. Sci.,* **60,** 1173–1185.