A Comparison of Precipitation Forecast Skill between Small Convection-Allowing and Large Convection-Parameterizing Ensembles

ADAM J. CLARK AND WILLIAM A. GALLUS JR.

Department of Geological and Atmospheric Sciences, Iowa State University, Ames, Iowa

MING XUE

School of Meteorology and Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

FANYOU KONG

Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

(Manuscript received 22 October 2008, in final form 5 January 2009)

ABSTRACT

An experiment has been designed to evaluate and compare precipitation forecasts from a 5-member, 4-km grid-spacing (ENS4) and a 15-member, 20-km grid-spacing (ENS20) Weather Research and Forecasting (WRF) model ensemble, which cover a similar domain over the central United States. The ensemble forecasts are initialized at 2100 UTC on 23 different dates and cover forecast lead times up to 33 h. Previous work has demonstrated that simulations using convection-allowing resolution (CAR; $dx \sim 4$ km) have a better representation of the spatial and temporal statistical properties of convective precipitation than coarser models using convective parameterizations. In addition, higher resolution should lead to greater ensemble spread as smaller scales of motion are resolved. Thus, CAR ensembles should provide more accurate and reliable probabilistic forecasts than parameterized-convection resolution (PCR) ensembles.

Computation of various precipitation skill metrics for probabilistic and deterministic forecasts reveals that ENS4 generally provides more accurate precipitation forecasts than ENS20, with the differences tending to be statistically significant for precipitation thresholds above 0.25 in. at forecast lead times of 9–21 h (0600–1800 UTC) for all accumulation intervals analyzed (1, 3, and 6 h). In addition, an analysis of rank histograms and statistical consistency reveals that faster error growth in ENS4 eventually leads to more reliable precipitation forecasts in ENS4 than in ENS20. For the cases examined, these results imply that the skill gained by increasing to CAR outweighs the skill lost by decreasing the ensemble size. Thus, when computational capabilities become available, it will be highly desirable to increase the ensemble resolution from PCR to CAR, even if the size of the ensemble has to be reduced.

1. Introduction

Because of inherent errors in modeling and observational systems, perfect deterministic forecasts of atmospheric states are impossible. However, if uncertainty in observations and model processes is properly accounted for by constructing an ensemble of forecasts, the ensemble members can be viewed as sampling a probability density function (PDF) from which reliable (i.e., events occur at frequencies corresponding to their forecast probabilities) probabilistic forecasts can be derived (e.g., Hamill and Colucci 1997; Eckel and Mass 2005; Clark et al. 2008; and many others). In addition, deterministic forecasts computed from the ensemble mean can perform better than individual members because the most uncertain aspects of the forecast are filtered out (e.g., Leith 1974; Holton 2004), which is especially true when considering large-scale features.

In current global ensemble forecast systems [e.g., the National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS; Toth and Kalnay 1993) Ensemble Prediction System; European Centre for Medium-Range Weather Forecasts (ECMWF; Molteni et al. 1996) Ensemble Prediction System], various methods

Corresponding author address: Adam J. Clark, Iowa State University, 3010 Agronomy Hall, Ames, IA 50010.
E-mail: clar0614@iastate.edu

DOI: 10.1175/2009WAF2222222.1

of generating initial condition (IC) perturbations (e.g., Toth and Kalnay 1997; Palmer et al. 1992; Molteni et al. 1996) and model perturbations yield reliable medium-range (2–10 day) forecasts of synoptic-scale parameters like 500-hPa geopotential height and mean sea level pressure. For the purpose of medium-range synoptic forecasting, IC errors make much larger relative contributions than model errors after the synoptic-scale error growth becomes nonlinear [~24 h; Gilmour et al. (2001)]. However, for short-range forecasts of small-scale phenomena like warm-season precipitation, which is the focus of this study, accounting for model error by using different combinations of physical parameterizations (e.g., Houtekamer et al. 1996; Stensrud et al. 2000; Du et al. 2004; Jones et al. 2007) and numerical models (e.g., Wandishin et al. 2001; Du et al. 2004; Eckel and Mass 2005) becomes very important in generating sufficient model dispersion. Unfortunately, even in these ensembles that include IC, model formulation, and physics perturbations, short-range forecasts for sensible weather phenomenon like convective precipitation remain underdispersive (Eckel and Mass 2005). Several factors are probably contributing to this lack of spread including coarsely resolved and temporally interpolated lateral boundary conditions (LBCs; Nutter et al. 2004), inappropriate IC perturbation strategies for short ranges (Eckel and Mass 2005), and an inability to capture small-scale variability because of insufficient resolution (Eckel and Mass 2005).

Because of computational limitations, regional-scale short-range ensemble forecast (SREF) systems like those run at NCEP (Du et al. 2004), the University of Washington (UW; Eckel and Mass 2005), and Stony Brook University (SBU; Jones et al. 2007) have been forced to use relatively coarse grid spacing (32–45 km for NCEP's SREF system, 12 km within a 32-km outer nest in UW's system, and 12 km for SBU's system) and, thus, must use cumulus parameterization (CP). In ensemble systems, using different CPs is an effective way of generating spread in rainfall forecasts (e.g., Jankov et al. 2005), but using CPs introduces systematic errors in rainfall forecasts (e.g., Davis et al. 2003; Liu et al. 2006; Clark et al. 2007), and models using CPs cannot resolve finescale features in rainfall systems. Because of these limitations, significant improvements in rainfall forecasts may be realized by running an ensemble using the explicit representation of convection (i.e., no CP).

Although there is still some debate regarding the grid spacing needed to adequately resolve convection (e.g., Bryan et al. 2003; Petch 2006), results from idealized tests (Weisman et al. 1997) indicate that 4 km is about the coarsest grid spacing at which midlatitude mesoscale convective systems (MCSs) can be resolved. Furthermore, ongoing experiments that began in 2003 in support of the Bow Echo and MCV Experiment (BAMEX; Davis et al. 2004) using various 4-km grid-spacing configurations of the Weather Research and Forecasting (WRF; Skamarock et al. 2005) model to aid convective forecasting have been rather successful (see Kain et al. 2008 for a thorough review). For example, simulations using convection-allowing resolution (CAR) have been found to more accurately depict the diurnal precipitation cycle (Clark et al. 2007; Weisman et al. 2008), as well as MCS frequency and the convective system mode (Done et al. 2004; Weisman et al. 2008) relative to simulations using parameterized-convection resolution (PCR).

Although increasing to CAR may not necessarily increase the forecast skill for deterministic forecasts as measured by traditional ''grid based'' metrics [e.g., equitable threat score (Schaefer 1990) and bias] because of small displacement errors in small-scale features leading to large errors (Baldwin et al. 2001; Davis et al. 2006a), it is possible that significant improvements in probabilistic precipitation forecasts may be obtained from an ensemble using CAR because of superior spatiotemporal representation of statistical properties of convective precipitation in the CAR members (e.g., Fritsch and Carbone 2004; Kong et al. 2006, 2007a). Also, because error growth occurs more rapidly at smaller scales, ensembles using CAR may have a better representation of forecast uncertainty. However, because of current computational limitations, it is difficult to create a CAR ensemble in real time with a domain size and number of members comparable to ensembles that are currently being used operationally. Although, given the potential advantages of CAR, an ensemble composed of a relatively small number of CAR members could potentially outperform an ensemble composed of a large number of PCR members, in which case there will be an incentive for future operational ensemble systems to reduce the numbers of members in order to increase the CAR.

Given these computational considerations, this study aims to compare warm-season precipitation forecast skill between a small (5 member) CAR ensemble using 4-km grid spacing (ENS4) and a relatively large (15-member) PCR ensemble using 20-km grid spacing (ENS20), each covering a similar domain over the central United States (Fig. 1). Because these ensembles have different numbers of members, special care is taken to properly compare probabilistic skill metrics. Although ENS4 has fewer members than ENS20, the computational expense should not be considered equal. In fact, because of the time-step reduction and 3D increase in the number of grid points, a reduction in grid spacing

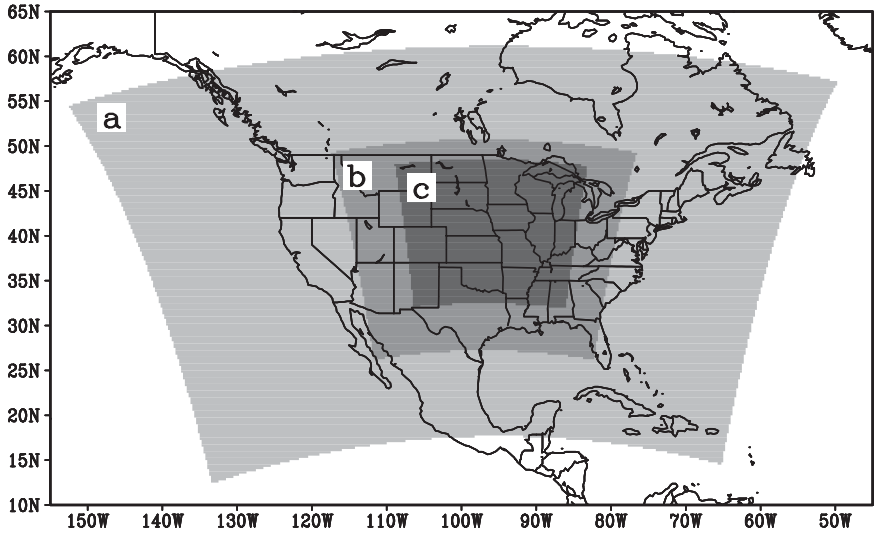

FIG. 1. Domains for (a) SREF ensemble members, (b) ENS4 and ENS20 ensemble members, and (c) the analyses conducted in this study.

from 20 to 4 km increases the computational expense by a factor of ~125. Thus, because ENS4 has 1/3 the members as ENS20, it is still about 42 times more computationally expensive than ENS20 ($125 \times 1/3 = 42$), and to conduct the comparison between ensembles with equal computational expense would require ENS20 to have $125 \times 5 = 625$ members. So, the purpose of this study is not to compare ensembles with similar computational expense, but to determine if at some point when computational capabilities allow, it would be advantageous to reduce ensemble size in order to use CAR. The remainder of this paper is organized as follows: section 2 describes model configurations and cases examined. Section 3 describes the data and methodology, including verification methods. Section 4 provides analyses of deterministic and probabilistic precipitation forecast skill, as well as spread and reliability. Finally, section 5 provides a summary and suggestions for future work.

2. Ensemble descriptions and cases examined

The ENS4 ensemble was obtained from a real-time ensemble forecasting experiment conducted as part of the National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Test Bed (HWT) Spring Experiment (Kain et al. 2008) during April–June 2007 (Xue et al. 2007; Kong et al. 2007b). A 4-km grid-spacing CAR WRF-ARW (version 2.2.0) model ensemble was run by the Center for Analysis and Prediction of Storms (CAPS) of the University of Oklahoma, which was composed of 10 members initialized daily at 2100 UTC and integrated for 33 h over an approximately 3000 km $\times$ 2500 km domain covering much of the central United States (Fig. 1). Four of the members used both perturbed ICs and mixed physical parameterizations (mixed physics), while six members, including the control member, used only mixed physics so that the effects of changing the model physics could be isolated. In this study, only the four members with both mixed physics and perturbed ICs plus the control member—a five-member ensemble (ENS4 ensemble)—are used because the ensemble using mixed physics alone ignores initial condition uncertainty, an important source of forecast uncertainty. For the control member, the 2100 UTC analyses from NCEP's operational North American Mesoscale (NAM; Janjić 2003) model (at 12-km grid spacing) are used for ICs and the 1800 UTC NAM 12-km forecasts are used for LBCs. For the initial perturbed members, perturbations extracted from the 2100 UTC NCEP SREF WRF-ARW and WRF-NMM members are added to the 2100 UTC NAM analyses, and the corresponding SREF forecasts are used for LBCs (3-h updates). Xue et al. (2007) and Kong et al. (2007b) have more details on the configurations.

The ENS20 ensemble was generated at Iowa State University and is also composed of WRF-ARW (version 2.2.0) members with perturbed ICs–LBCs and mixed physics. Different sets of ICs for each ENS20 member are obtained directly from NCEP SREF members (listed in Table 2), rather than adding perturbations to the 2100 UTC NAM analyses, and, similar to ENS4, SREF forecasts are used for LBCs.

All ENS4 and ENS20 members use the rapid radiative transfer model (RRTM) short-wave radiation scheme (Mlawer et al. 1997) and the Goddard longwave radiation scheme (Chou and Suarez 1994), along with

TABLE 1. ENS4 Ensemble member specifications. NAMa and NAMf indicate NAM forecasts and analyses, respectively; em_pert and nmm_pert are perturbations from different SREF members; and em_n1, em_p1, nmm_n1, and nmm_p1 are different SREF members that are used for LBCs. The remaining table elements are described in the text.

Ensemble member	ICs	LBCs	Microphysics scheme	Surface layer scheme	Boundary layer scheme
CN	2100 UTC NAMa	1800 UTC NAMf	WSM-6	Janjić Eta	MYJ
N1	CN – em_pert	2100 UTC SREF em_n1	Ferrier	Janjić Eta	MYJ
P1	CN + em_pert	2100 UTC SREF em_p1	Thompson	Janjić Eta	YSU
N2	CN – nmm_pert	2100 UTC SREF nmm_n1	Thompson	Monin–Obukhov	YSU
P2	CN + nmm_pert	2100 UTC SREF nmm_p1	WSM-6	Monin–Obukhov	YSU

the Noah land surface model physics scheme (Ek et al. 2003). Varied planetary boundary layer parameterizations in both ensembles include the Mellor–Yamada–Janjić (MYJ; Mellor and Yamada 1982; Janjić 2002) and Yonsei University (YSU; Noh et al. 2003) schemes; the microphysics schemes include Thompson et al. (2004), WRF single-moment six-class (WSM-6; Hong and Lim 2006), and Ferrier et al. (2002); surface layer schemes include Monin–Obukhov (Monin and Obukhov 1954; Paulson 1970; Dyer and Hicks 1970; Webb 1970) and the Janjić Eta Model (Janjić 1996, 2002); and CPs in ENS20 include the Betts–Miller–Janjić (BMJ; Betts 1986; Betts and Miller 1986; Janjić 1994), Kain–Fritsch (KF; Kain and Fritsch 1993), and Grell–Devenyi (Grell and Devenyi 2002) schemes. Note that ENS4 does not use a CP. The ENS4 and ENS20 specifications are listed in Tables 1 and 2, respectively.

Forecasts were examined for 23 cases during April–June 2007 (Fig. 2). These cases were chosen based on the availability of the ENS4 real-time forecasts and represent a variety of convective precipitation events [e.g., isolated convection (19 April), heavy rainfall associated with a cutoff upper low (22–25 April), and many nocturnal MCSs (late May–early June)].

3. Data and methodology

This study examines forecasts of 1–3- and 6-hourly accumulated rainfall. Although the 1- and 3-hourly accumulation periods are examined starting at initialization, the 6-h periods begin with forecast hours 3–9, corresponding to the 0000–0600 UTC period. The stage IV (Baldwin and Mitchell 1997) multisensor rainfall estimates available at 1- and 6-hourly accumulation intervals on a 4-km polar stereographic grid are used to verify rainfall forecasts. Both stage IV rainfall estimates and ENS4 rainfall data are remapped to a 20-km grid covering the central United States (Fig. 1), which is just a subdomain of the ENS20 members, using a neighbor-budget interpolation that conserves the total liquid volume in the domain (a procedure typically used at NCEP). The ENS4 forecasts were coarsened to allow direct comparisons to ENS20; additional and potentially useful information on the finer-scale details in the forecast

TABLE 2. ENS20 Ensemble member specifications. The IC–LBC table elements represent various SREF members. The asterisks (*) and plus signs (+) denote the combination of 5 and 10 ensemble members, respectively, with the best statistical consistency. The remaining table elements are described in the text.

Ensemble member	ICs–LBCs	Cumulus scheme	Microphysics scheme	Surface layer scheme	Boundary layer scheme
1^{+}	em_ctl	BMJ	Thompson	Janjić Eta	MYJ
2^{*+}	em_p1	BMJ	WSM-6	Janjić Eta	MYJ
3^{+}	em_n1	BMJ	WSM-6	Monin–Obukhov	YSU
4	nmm_ctl	BMJ	Thompson	Monin–Obukhov	YSU
5^{+}	nmm_p1	BMJ	Ferrier	Monin–Obukhov	YSU
6*	nmm_n1	KF	Thompson	Janjić Eta	MYJ
7	eta_ctl1	KF	WSM-6	Janjić Eta	MYJ
8^{+}	eta_n1	KF	WSM-6	Monin–Obukhov	YSU
9^{+}	eta_n2	KF	Thompson	Monin–Obukhov	YSU
10	eta_n3	KF	Ferrier	Monin–Obukhov	YSU
11^{*+}	eta_n4	Grell	Thompson	Janjić Eta	MYJ
12	eta_p1	Grell	WSM-6	Janjić Eta	MYJ
13^{*+}	eta_p2	Grell	WSM-6	Monin–Obukhov	YSU
14^{+}	eta_p3	Grell	Thompson	Monin–Obukhov	YSU
15^{*+}	eta_p4	Grell	Ferrier	Monin–Obukhov	YSU

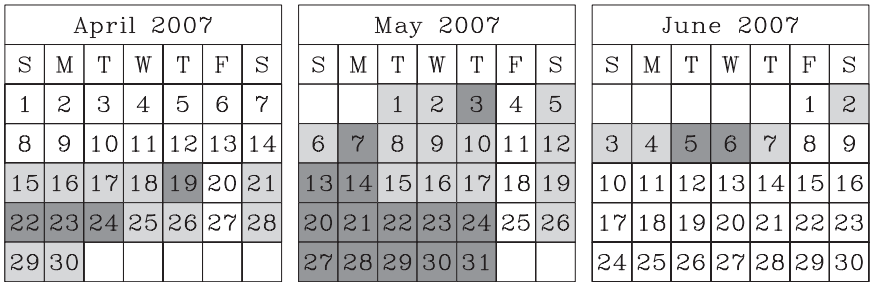

FIG. 2. Dates when SSEF ensemble runs were conducted (light gray) and used in this study (dark gray).

precipitation fields could be gained using the ENS4 data on its original 4-km grid.

Probabilistic and deterministic forecasts derived from each ensemble were verified. Deterministic forecasts were obtained using the probability matching technique (Ebert 2001), which is applied by assuming that the best spatial representation of rainfall is given by the ensemble mean and that the best frequency distribution of rainfall amounts is given by the ensemble member quantitative precipitation forecasts (QPFs). To apply probability matching, the PDF of the ensemble mean is replaced by the PDF of the ensemble member QPFs, which is calculated by pooling the forecast precipitation amounts for all n ensemble members, sorting the amounts from largest to smallest, and keeping every nth value. The ensemble mean precipitation amounts are also sorted and the rank and location of each value are stored. Then, the grid point with the highest precipitation amount in the ensemble mean is replaced by the highest value in the distribution of the ensemble member QPFs, and so on. The ensemble mean obtained from the probability matching procedure (PM) can help correct for large biases in areal rainfall coverage and the underestimation of rainfall amounts that are typically associated with using a standard ensemble mean, and results in a precipitation field with a much more realistic distribution.

Forecast probabilities (FPs) for precipitation were obtained by finding the location of the verification threshold within the distribution of ensemble member forecasts. The reader is referred to Hamill and Colucci (1997, 1998) for a thorough description of the application of this technique for assigning FPs. Similar to Hamill and Colucci (1998), the FPs for thresholds beyond the highest ensemble member forecast are obtained by assuming that the PDF in this region has the shape of a Gumbel distribution (Wilks 1995). Note that this method assumes the rank histograms generated from the ensembles are uniform and forecast uncertainty is accurately represented. These assumptions may not be valid for precipitation forecasts (e.g., Hamill and Colucci 1997, 1998; Jankov et al. 2005; Clark et al. 2008). Thus, in an operational setting, the FPs should be calibrated based on the shape of the rank distribution from past model forecasts (Hamill and Colucci 1998). In addition, Eckel and Mass (2005) recommend adjusting individual ensemble member forecasts based on known biases before FP calibration is performed to obtain maximum ensemble utility. Because the 23 cases examined in this study constitute a relatively small sample from which a useful ''training period'' cannot be obtained, this type of calibration is not performed.

To verify deterministic forecasts, the equitable threat score (ETS; Schaefer 1990) is used, which is computed using contingency table elements representing possible forecast scenarios including hits (observed event correctly predicted), misses (event occurred but not predicted), false alarms (event predicted that did not occur), and correct negatives (event correctly predicted not to occur). A complete description of ETS in terms of contingency table elements can be found in Hamill (1999). ETSs range from $-1/3$ (ETS < 0 has no skill) to 1 (perfect). Average ETSs were calculated by summing contingency table elements from all cases for each forecast hour and rainfall threshold, and computing the scores from the summed elements. This aggregate method gives greater weight to widespread precipitation events than if the ETS for each case was simply averaged.

To verify probabilistic forecasts, the area under the relative operating characteristic curve [ROC score; Mason (1982)] is used, which is closely related to the economic value of a forecast system (e.g., Mylne 1999; Richardson 2000, 2001). The ROC score is computed from members of a contingency table for probabilistic forecasts. To construct the ROC curve, the probability of detection (POD) is plotted against the probability of false detection (POFD) for a set of specified ranges of FPs. The area under this curve is computed using the trapezoidal method (Wandishin et al. 2001). Because the method used to compute FPs in this study allows for

continuous (rather than discrete) values of FPs between 0% and 100%, the same set of FP ranges that make up the points on the ROC curve can be used to verify both ensembles, and problems associated with comparing ROC scores between ensembles of different sizes are avoided. The FPs used are $P < 0.05$, $0.05 \leq P < 0.15$, $0.15 \leq P < 0.25 \ldots 0.85 \leq P < 0.95$, and $0.95 \leq P < 1.00$. The range of values for the ROC score are 0 to 1 with scores above 0.5 showing skill, and a score of 0.7 considered to represent the lower limit of a useful forecast (Buizza et al. 1999).

The resampling methodology described in Hamill (1999) was used to determine whether differences in the ETS and ROC scores were statistically significant ($\alpha = 0.05$; resampling repeated 1000 times). For ETS comparisons, the biases from both ensembles were adjusted to the average bias between them, which minimized the adjustments made to precipitation forecasts to account for bias. Because the ROC score is insensitive to bias (e.g., Harvey et al. 1992; Mason and Graham 2002), no adjustments were made to forecasts prior to its computation.

4. Results

a. *Analysis of diurnally averaged Hovmöller diagrams*

Warm-season precipitation in the central United States tends to form at similar times of day and propagate over similar longitudes so that when diurnally averaged time–longitude (Hovmöller) diagrams of precipitation are constructed, coherent and propagating rainfall axes are observed (Carbone et al. 2002). These coherent axes, which are often composed of long-lived convective ''episodes.'' suggest that an intrinsic predictability is associated with propagating rainfall systems over the central United States, so that predictability limits that have been suggested by past theoretical studies (e.g., Smagorinsky 1969; Lorenz 1969) may be longer than previously thought. However, partly because of shortcomings associated with CPs (e.g., Molinari and Dudek 1992; Kain and Fritsch 1998; Davis et al. 2003; Bukovsky et al. 2006), it is believed that numerical models will not be able to take advantage of this inherent predictability until CAR is utilized. Evidence from some preliminary studies comparing data from CAR and PCR simulations (e.g., Liu et al. 2006; Clark et al. 2007; Weisman et al. 2008) supports this idea. An ensemble of CAR members with a better depiction of the propagating rainfall axis over the central United States than a PCR ensemble should have a considerable advantage because individual CAR members will be more likely to fall within the range of likely solutions if they have an accurate ''model climatology,'' whereas many of the PCR solutions may be very unlikely to verify because of consistent biases in the timing and location for propagating rainfall systems.

To examine whether differences in the diurnal cycle representation exist for the ensemble members in this study, 1-hourly diurnally averaged Hovmöller diagrams for all ensemble member forecasts and stage IV observations are constructed following procedures described in Clark et al. (2007), with latitudinal averages (along constant-longitude slabs) computed between 32° and 48°N latitude. The Hovmöller diagram for the stage IV observations (Fig. 3c) shows that coherent propagating rainfall axes exist even for the relatively small number of cases examined. A primary axis of observed rainfall begins around 2200 UTC (forecast hour 1) at about 102°W and ends around 1500 UTC (forecast hour 18) at about 94°W, while a weaker secondary rainfall axis begins a few hours before model initialization (perhaps 1900 UTC) at 98°W and ends around 0900 UTC (forecast hour 12) at about 90°W. Note that both axes begin to repeat during the second diurnal cycle within the forecast period. Each of the 23 cases examined contributed to 0%–15% of the rainfall within the primary axes, while one case contributed about 25% of the rainfall to the secondary axes and the remaining cases contributed 10% or less rainfall to the secondary axes (not shown). Thus, the primary rainfall axes are more representative of all 23 cases than the secondary axes. The primary rainfall axis is similar to that observed in studies examining longer time periods (e.g., Carbone et al. 2002; Tuttle and Davis 2006), while the secondary axis has not been observed in the studies examining the longer time periods. However, Clark et al. (2007) did observe a secondary rainfall axis during February–April 2007 over the northern portion of the central United States.

The Hovmöller diagrams for the five members of ENS4 (not shown) all reveal coherent propagating rainfall axes resembling both the primary and secondary axes from stage IV observations. The ENS4 ensemble mean (computed using PM; Fig. 3a) also exhibits the propagating axes showing that the averaging process retains the propagating signal and may actually improve its representation relative to individual members. This improvement is suggested by the spatial correlation coefficients computed in Hovmöller space (Fig. 3e), which are higher for the ENS4 ensemble mean than all of its members during forecast hours 4–18, and all but one of its members during forecast hours 19–33.

Some of the Hovmöller diagrams for the members of the ENS20 ensemble (not shown) have propagating

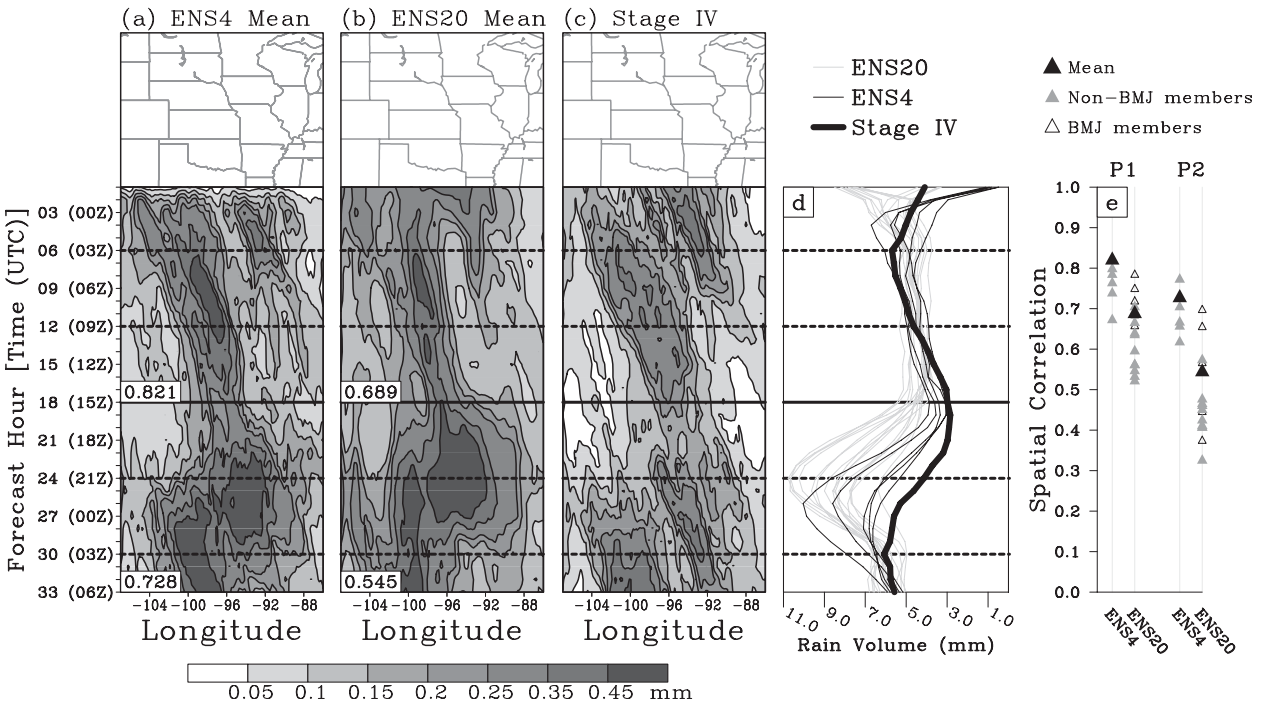

FIG. 3. Diurnally averaged Hovmöller diagrams of ensemble mean (computed using probability matching) 1-hourly precipitation forecasts from (a) ENS4, (b) ENS20, and (c) 1-hourly stage IV observed precipitation. Spatial correlation coefficients computed in Hovmöller space for the ensemble means during forecast hours 4–18 are indicated at the middle left of (a) and (b), and those for forecast hours 19–33 are indicated at the bottom left of (a) and (b). The maps at the tops of (a)–(c) indicate the domains over which the Hovmöller diagrams were computed. (d) Domain-averaged precipitation from the ENS4 members (thin black lines), ENS20 members (thick gray lines), and stage IV observations (thick black line). (e) Spatial correlation coefficients computed in Hovmöller space for the ENS4 and ENS20 members during forecast hours 4–18 (P1) and 19–33 (P2).

signals resembling those in the observations during the first diurnal cycle within the forecast period (before forecast hour 18), in particular the BMJ members. In fact, the spatial correlation coefficients for the ENS20 BMJ members are comparable to the ENS4 members during this first period (Fig. 3e). However, all ENS20 members, except for BMJ members 1 and 2 (Table 2), lack a clear propagating signal during the 19–33-h forecast period. It is worth noting the good relative performance of ENS20 members 1 and 2; perhaps, these two configurations are especially conducive to propagating convective systems despite their relatively coarse grid spacing, which would agree with the results from Bukovsky et al. (2006) in which propagating features were noted and examined in the NCEP operational NAM model containing a similar configuration as ENS20 members 1 and 2. However, note that the propagation mechanism for the features noted by Bukovsky et al. (2006) was not physically realistic, and further examination of these ensemble member forecasts is beyond the scope of this study.

Generally, Hovmöller diagrams and spatial correlation coefficients show that ENS4 has a better diurnal cycle depiction and representation of the propagating rainfall axes than ENS20, especially during forecast hours 19–33. The larger differences during this later forecast period appear to result from the ENS20 members simulating the rainfall maximum that occurs during the second simulated diurnal cycle too early and too intensely, which is reflected in the ENS20 ensemble mean Hovmöller diagram and diurnally averaged time series of the domain-averaged rain volume for ENS4 and ENS20 members (Fig. 3d). These results imply that ENS4 has an inherent advantage over ENS20. The following sections will use various standard verification metrics to determine whether this advantage is enough to compensate for the smaller ensemble size of ENS4 relative to ENS20.

b. Comparison of ensemble equitable threat scores

The skill of deterministic forecasts derived using PM from each ensemble is compared by constructing time series of ETSs for 1-, 3-, and 6-h intervals at the 0.10-, 0.25-, and 0.50-in. rainfall thresholds (Fig. 4). For the ENS20 ensemble, in addition to computations using all 15 members, the ETSs computed using the ensemble

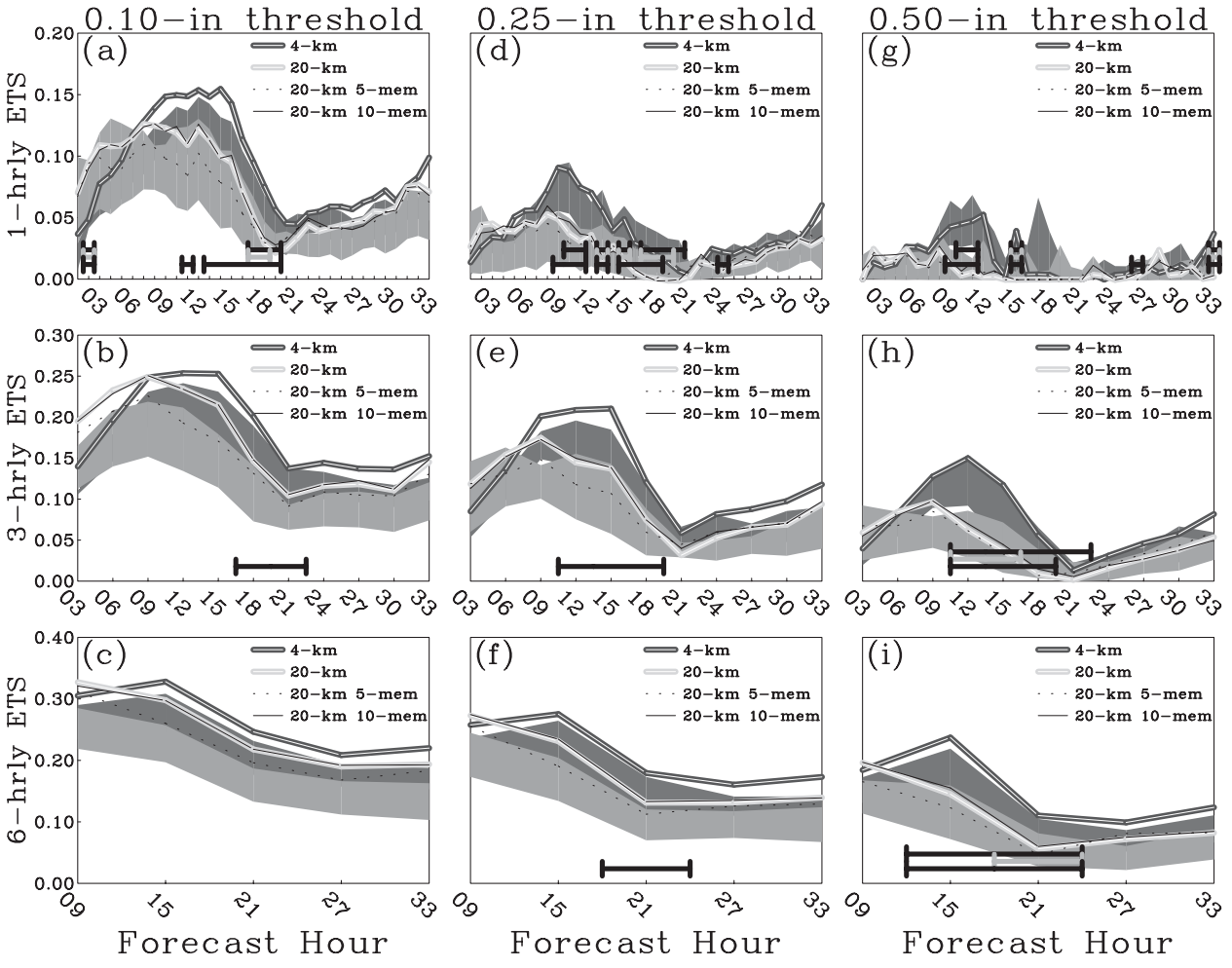

FIG. 4. Time series of average ETSs for the ENS4, ENS20, ENS20(5m), and ENS20(10m) ensemble mean precipitation forecasts at the 0.10-in. precipitation threshold for (a) 1-, (b) 3-, and (c) 6-hourly accumulation intervals; at the 0.25-in. precipitation threshold for (d) 1-, e) 3-, and (f) 6-hourly accumulation intervals; and at the 0.50-in. precipitation threshold for (g) 1-, (h) 3-, and (i) 6-hourly accumulation intervals. ENS20(5m) and ENS20(10m) represent ensembles composed of the combination of 5 and 10 members, respectively, of ENS20 that have the best statistical consistency. The light (dark) shaded areas depict the range of ETSs from the ENS20 (ENS4) ensemble members. Times at which differences between the ETSs from the ENS4 and ENS20 ensemble mean precipitation forecasts were statistically significant are denoted by bars near the bottom of the panels. The highest (middle) (lowest) bars correspond to times at which differences between ENS4 and ENS20 [ENS20(10m)] [ENS20(5m)] were statistically significant.

mean from the 5 and 10 members with the best statistical consistency, as described by Eckel and Mass (2005) for a finite ensemble, are also examined [these members are noted in Table 2 and referred to as ENS20(5m) and ENS20(10m) hereafter]. Thus, comparisons between ensembles with the same number of members can be made, and the impacts of the additional members on ENS20 can be examined. The range of ensemble member ETSs for ENS4 and ENS20 are also shown in Fig. 4.

Generally, both ENS4 and ENS20 tend to have maxima in ETSs between forecast hours 9 and 15 when both models have had sufficient time to "spin up" (e.g., Skamarock 2004) and synoptic-scale error growth should still be relatively small. In addition, these forecast hours correspond to the times at which the propagating rainfall axis in the Midwest is at its maximum amplitude, suggesting some enhanced predictability associated with long-lived MCSs, which occur most frequently at times corresponding to these forecast hours (e.g., Maddox 1983). To verify that the maxima in ETS are not simply artifacts of relatively high biases (Hamill 1999), time series of average 3-hourly biases to which both ensemble mean forecasts were adjusted during computation of ETS are shown in Fig. 5, along with the total number of grid points in which the observed precipitation above the specified threshold occurred. Time series of bias are actually opposite in phase to those of ETS, implying that the maxima in ETS are truly a reflection of skill. In fact,

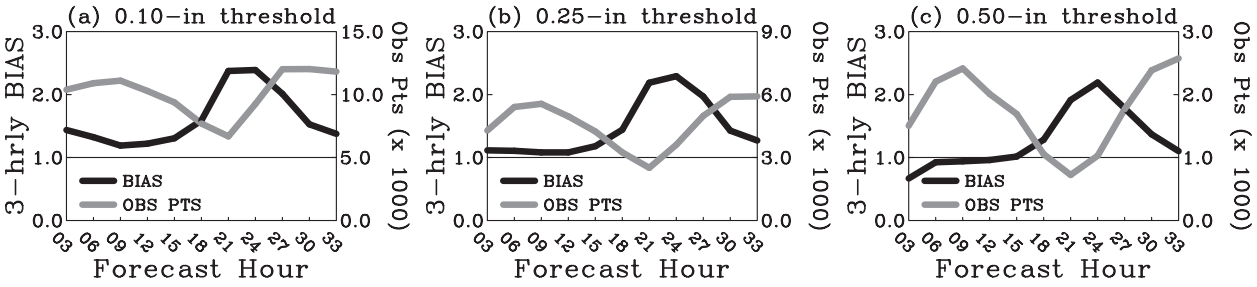

FIG. 5. Time series of 3-hourly bias and total number of grid points in which the observed precipitation occurred above the (a) 0.10-, (b) 0.25-, and (c) 0.50-in. thresholds.

it appears that ETS values are more dependent on observed points above the threshold analyzed, with larger observed rainfall areas corresponding to higher ETSs, than they are bias (Fig. 5). This dependence suggests that widespread precipitation events are more predictable than isolated events, and that the aggregate method for computing average ETS, which gives more weight to widespread precipitation events, should be associated with higher values of ETSs than those computed by simply averaging over all cases.

ENS20(5m) appeared to generally have lower ETSs than ENS20, while ENS20(10m) generally had very similar ETSs to ENS20, indicating that most of the skill realized from increasing ensemble size was obtained with an increase from 5 to 10 members, while very little skill was obtained with the increase from 10 to 15 members. Similar behavior illustrating a ''point of diminishing returns'' has been observed in previous studies (e.g., Du et al. 1997; Ebert 2001), and it is likely that additional model diversity (e.g., addition of members with a different dynamic core) would result in a larger increase in skill as demonstrated by the NCEP SREF system (Du et al. 2006). In addition, note that ensemble mean ETSs from both ensembles are greater than the highest corresponding ensemble member ETSs, illustrating that the ensemble mean forecasts do represent an improvement relative to ensemble member forecasts, which is expected behavior in an ensemble.

After about forecast hour 9, at virtually all forecast lead times, accumulation intervals, and rainfall thresholds examined, ENS4 has higher ETSs than ENS20, with differences that are statistically significant occurring for 1-hourly accumulation intervals at all rainfall thresholds examined, and for 3- and 6-hourly accumulation intervals at the 0.50-in. rainfall threshold. The statistically significant differences generally occur between forecast hours 12 and 21, corresponding to the times near and just after the maxima in ETS. Furthermore, between forecast hours 9 and 12 at the 0.25- and 0.50-in. rainfall thresholds for 1- and 3-hourly accumulation intervals (Figs. 4d, 4g, 4e, and 4h), all of the ENS4 members have higher ETSs than the maximum ETS of the ENS20 members. Also, there appears to be a trend for the differences in ETS between ENS4 and ENS20 to become smaller as increasing accumulation intervals are examined, which implies that timing errors may explain much of the differences, because timing errors decrease as longer accumulation intervals are examined (e.g., Wandishin et al. 2001). The implied influence of timing errors is also supported by the Hovmöller diagrams of diurnally averaged rainfall from each ensemble (Figs. 3a and 3b), and the diurnally averaged time series of the domain-averaged rainfall for all ensemble members (Fig. 4d), which were discussed in the previous section.

To determine which cases contributed the most to the statistically significant differences in ETSs between ENS4 and ENS20, and to determine whether ENS4 consistently performs better than ENS20 over all cases, ETS differences for the 0.50-in. rainfall threshold for 3-hourly accumulation intervals at forecast hours 12–21 (corresponding to Fig. 4h), when differences were statistically significant, are examined (Fig. 6a). Note that these differences [ETS(ENS4) − ETS(ENS20)] were calculated after the bias-correction procedure discussed in section 3 was applied. While the differences from each case do not truly reflect the contributions to the average differences in ETS because rainfall events covering large areas are weighted more than events covering small areas in the averaging procedure used, the larger differences did tend to be associated with relatively widespread rainfall events (not shown). Thus, these larger differences (e.g., highlighted cases in Fig. 6a) are likely making large relative contributions to the average ETS differences.

As revealed in Fig. 6a, ENS4 consistently performs better than ENS20 during forecast hours 12–21 for the 0.50-in. rainfall threshold, and there are only a few times when ENS20 has a higher ETS than ENS4. Two particular cases with relatively widespread precipitation amounts greater than 0.50 in. that were initialized on 29 and 31 May (highlighted in Fig. 6a) were observed to have relatively large differences in ETS. Hovmöller

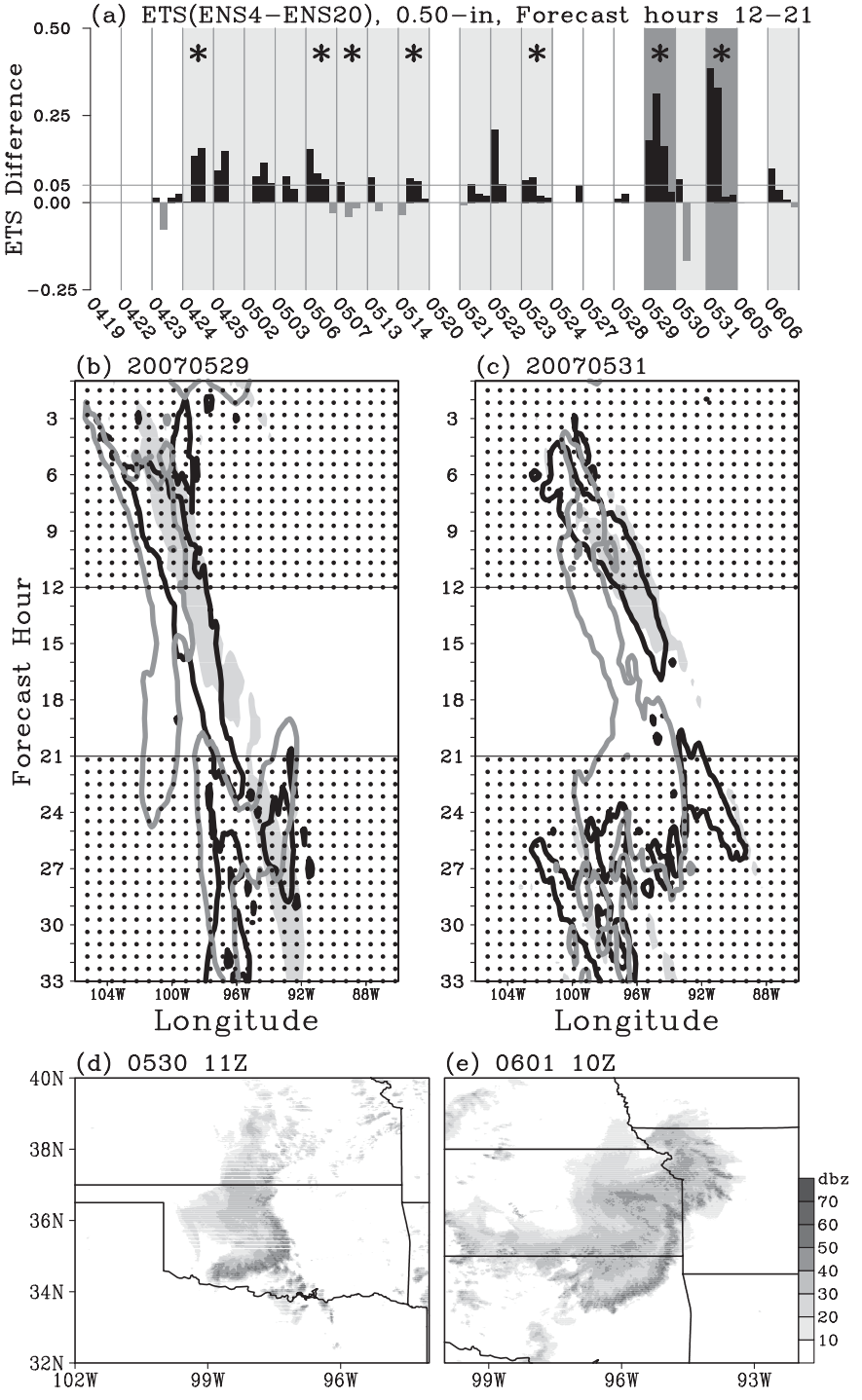

FIG. 6. (a) Difference between ETSs of ENS4 and ENS20 during forecast hours 12–21 at the 0.50-in. rainfall threshold. Hovmöller diagrams of forecast precipitation from ENS4 (black contour), ENS20 (light gray contour), and stage IV observations (light gray shading) are shown for (b) 29 May and (c) 31 May 2007 [cases shaded dark gray in (a)]. The areas outside of the hatching correspond to the forecast hours plotted in (a). Images of the observed composite reflectivity centered on the mesoscale convective systems contributing to the observed precipitation during the cases plotted in (b) and (c) are shown in (d) and (e), respectively.

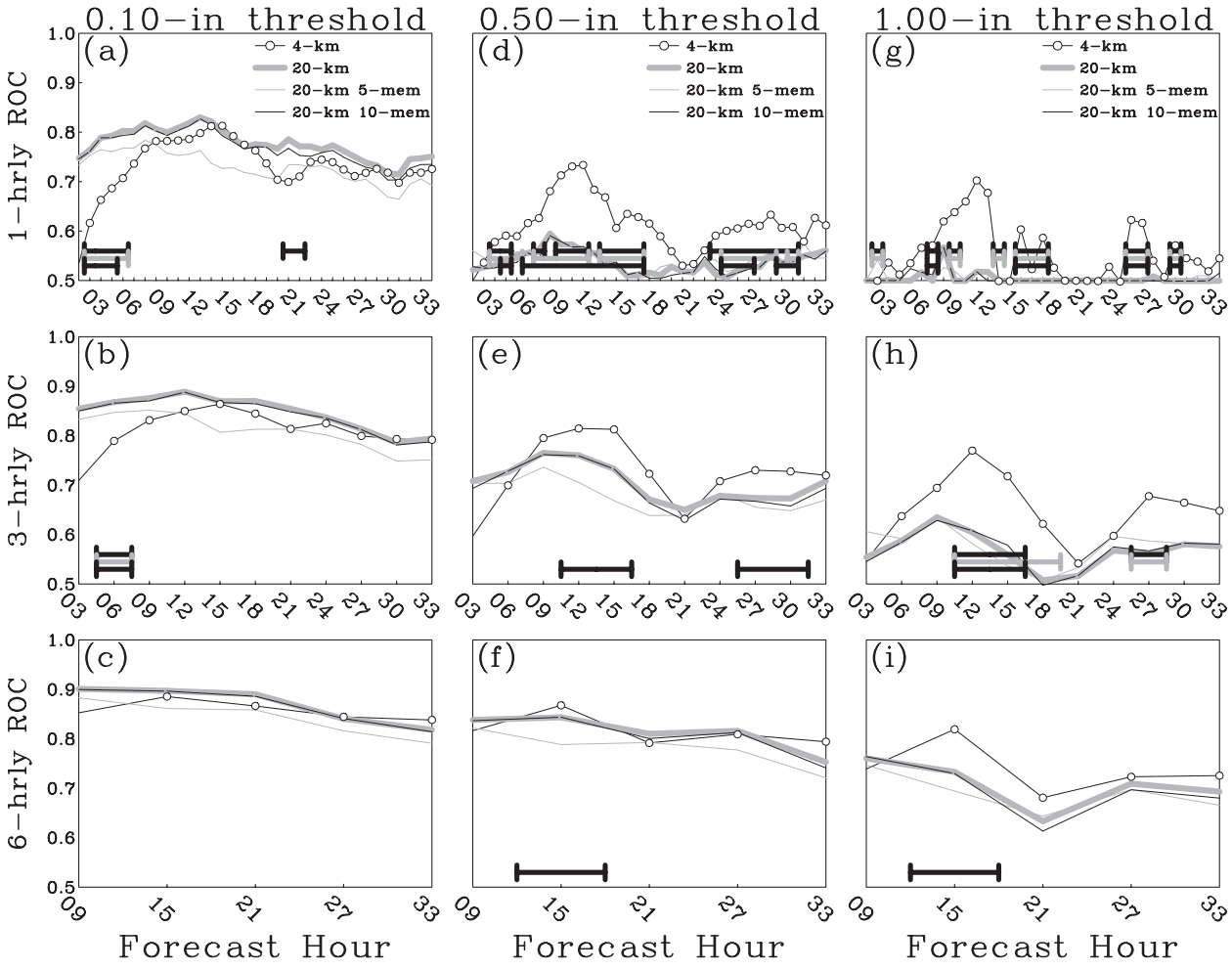

FIG. 7. As in Fig. 4, but for ROC scores at different sets of precipitation thresholds, and ranges among members are not indicated.

diagrams of ENS4 and ENS20 forecast and stage IV observed rainfall for these cases are plotted in Figs. 6b and 6c, revealing that ENS4 was better able to simulate the eastward propagation of rainfall systems from these cases than ENS20. Composite reflectivity images (Figs. 6d and 6e) corresponding to times when large errors were observed in ENS20 show that the errors were associated with mature, nocturnal MCSs. In addition, it was found that out of the 15 cases in which ETS differences were greater than 0.05 during at least one of the periods between forecast hours 12 and 21 (gray shaded cases in Fig. 6a), 7 contained mature MCSs during forecast hours 12–21 (asterisks mark cases in Fig. 6a). These seven MCS cases accounted for 66% of the total number of grid points (over all 23 cases) with observed precipitation greater than 0.50 in. during forecast hours 12–21. These results are a strong indication that the ability of the CAR members in ENS4 to properly simulate propagating MCSs explains the statistically significant differences in ETS between ENS4 and ENS20 observed in Fig. 4.

c. *Comparison of ROC scores*

The skill levels of probabilistic forecasts derived from each ensemble are compared by constructing time series of ROC scores for 1-, 3-, and 6-h intervals at 0.10-, 0.50-, and 1.00-in. rainfall thresholds (Fig. 7). Similar to ETS, ROC scores for the 5 and 10 members of the ENS20 ensemble with the best statistical consistency are also plotted. Because statistically significant differences between ENS4 and ENS20 ROC scores were confined to higher precipitation thresholds than in the ETS analysis, higher thresholds than those shown for ETS are shown for ROC scores in Fig. 7. In general, maxima in ROC scores from both ensembles are observed at forecast hours 9–15. However, the amplitude of the ROC score oscillations is much larger, especially in ENS4, as the rainfall threshold examined increases. The timing of this

ROC score maximum likely is again due to enhanced predictability because of the high relative frequencies of MCSs at these times. There also appears to be a secondary maximum in ENS4 ROC scores at the 0.50- and 1.00-in. rainfall thresholds for all accumulation intervals examined around forecast hour 27 (Figs. 7d–i). This secondary maximum also appears in the ENS20 ROC scores, but only at 6-hourly accumulation intervals (Figs. 7f and 7i). The timing of the secondary ROC score maximum corresponds to the secondary propagating rainfall axis noted in the Hovmöller diagram of the observed precipitation during forecast hours 24–33 (Fig. 3c). Thus, it is also possible that ROC scores are enhanced around forecast hour 27 because of a tendency for propagating MCSs to occur during this time.

Similar to trends seen with ETS, ENS20(5m) generally has lower ROC scores than ENS20, while ENS20(10m) ROC scores are very similar to those of ENS20. Thus, most of the increase in ROC scores realized from increasing the ensemble size is obtained with an increase from 5 to 10 members, with the increase from 10 to 15 members having little impact.

At the 0.10-in. rainfall threshold, at most forecast lead times, ENS20 has similar or slightly higher ROC scores than ENS4. However, the differences are statistically significant only before forecast hour 9, at 1- and 3-hourly accumulation intervals, and at forecast hours 20 and 21 at 1-hourly accumulation intervals (Figs. 7a and 7b). Note that before forecast hour 9, model ''spinup'' processes are still on going and ENS4 takes longer than ENS20 to generate areas of rainfall because grid-column saturation must occur before rainfall is generated in ENS4 members, while grid-column saturation is not required in ENS20 members because a CP is used. At 0.50- and 1.00-in. rainfall thresholds for 1-hourly accumulation intervals, ENS4 ROC scores are higher than ENS20, with differences statistically significant at many forecast lead times (Figs. 7d and 7g). For 3-hourly accumulation intervals, ENS4 ROC scores are higher than ENS20 ROC scores, with differences statistically significant occurring only at the 1.00-in. rainfall threshold (Figs. 7e and 7h), while for 6-hourly accumulation intervals, there are no statistically significant differences (Figs. 7f and 7i). It should be noted that, although there are not statistically significant differences between ENS4 and ENS20 for any of the rainfall thresholds at 6-hourly accumulation intervals, and for the 0.50-in. rainfall threshold for 3-hourly accumulation intervals (Figs. 7e, 7f, and 7i), statistically significant differences between ENS4 and ENS20(5m) ROC scores do occur.

In general, statistically significant differences occurred around forecast hours 9–15 and 24–30, corresponding to the times at which maxima in ROC scores were observed. Also, similar to ETS, there was a trend for the differences between ENS4 and ENS20 to decrease with increasing accumulation intervals, implying the decreasing influence of timing errors with increasing accumulations intervals.

d. Ensemble spread and statistical consistency

1) RANK HISTOGRAM ANALYSIS

Rank histograms are a useful tool for assessing ensemble spread (Hamill 2001) and are constructed by repeatedly tallying the rank of the rainfall observation relative to forecast values from an ensemble sorted from highest to lowest. A reliable ensemble will generally have a flat rank histogram, while too little (much) spread is indicated by a u-shaped (n shaped) rank histogram (Hamill 2001). Furthermore, the skewness of a rank histogram indicates bias, with right skewness (left skewness) indicating a tendency for members to overforecast (underforecast) the variable being examined.

For an ensemble composed of n members, precipitation observations can fall within one of any $n + 1$ bins. The bars that compose a rank histogram represent the fraction of observations that fall within each of these bins. Thus, the ENS4 rank histograms are composed of 6 bars while those of ENS20 are composed of 16 bars. The different numbers of rank histogram bars make it difficult to compare rank histograms from each ensemble. For example, it is obvious that the right skewness of the rank histograms from both ENS4 (gray shaded bars in Fig. 8a) and ENS20 (Fig. 8b) indicates a tendency for members to overpredict precipitation, but it is not clear which rank histogram indicates the greater tendency for overprediction. To allow for a more convenient comparison, the 16 bins composing the ENS20 rank histogram are regrouped into 6 bins, which each contain an equal portion of the original 16 bins (Fig. 8a). Care should be taken when interpreting the regrouped rank histograms. For example, the outer bins in the regrouped ENS20 rank histogram cannot be interpreted as the fraction of observations that fall completely outside the range of all ensemble members, as they are in ENS4, because they contain fractions from 3 of the original 16 bins. Rather, the regrouped rank histograms should be viewed as the rank histogram that would result from ENS20 if it was composed of 5 members, assuming these 5 members had about the same reliability and bias as the 15-member ENS20.

At all forecast lead times, the right skewness of the rank histograms from both ensembles indicates a tendency for members to overpredict precipitation (Figs. 8a and 8b). The right skewness appears to be the most

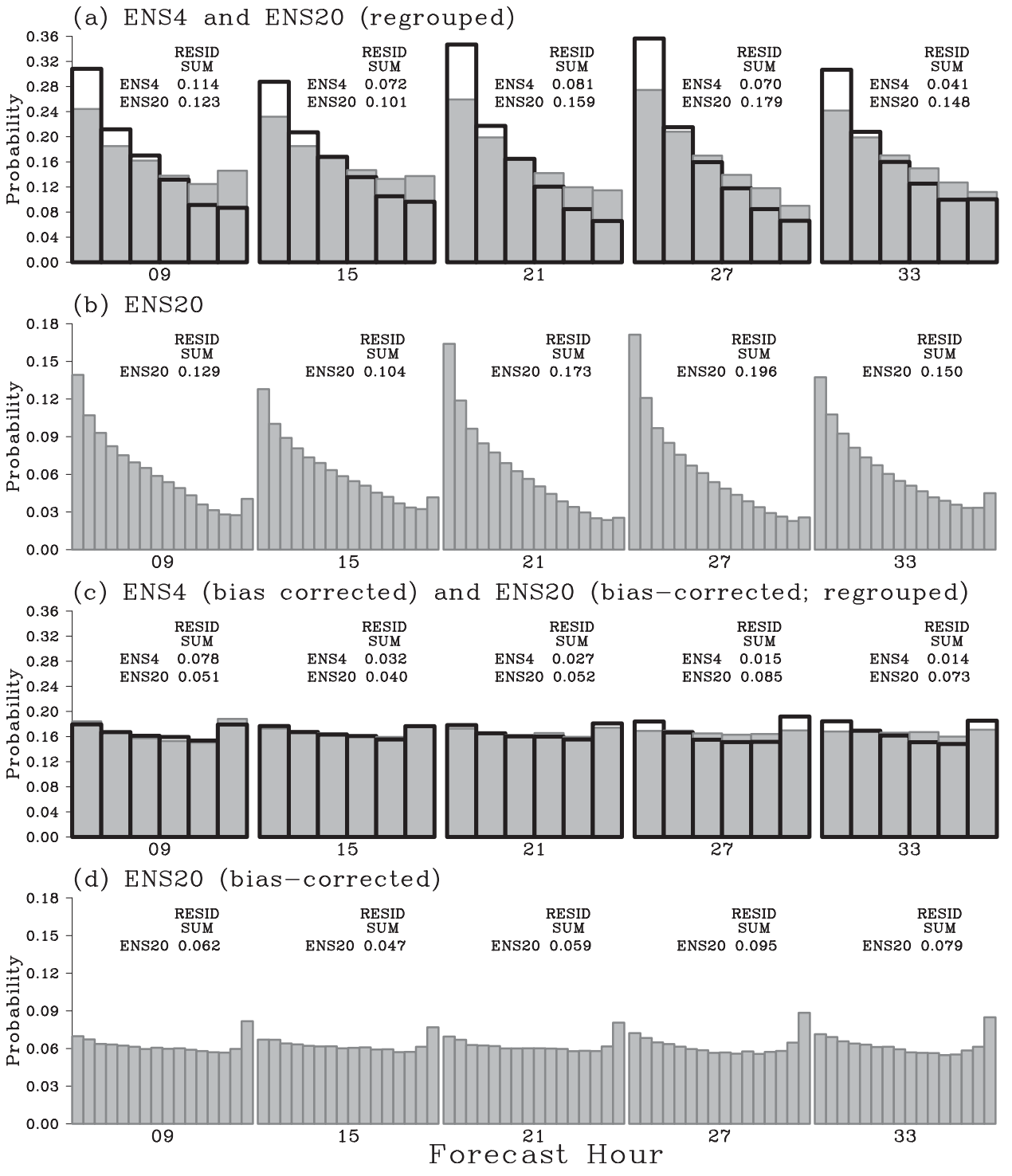

FIG. 8. Rank histograms at various forecast lead times for 6-hourly accumulated precipitation from (a) ENS4 (gray shaded bars) and ENS20 (regrouped; black outlined bars) and (b) ENS20. (c),(d) As in (a),(b), respectively, but with the rank histograms computed using bias-corrected precipitation forecasts from ENS4 and ENS20. The sum of the absolute value of the residuals from fitting a least squares line to the observed frequencies in each rank histogram is indicated above each rank histogram set.

pronounced at forecast hours 21 and 27, which agrees with the time series of observed and forecast domain-averaged rainfall (Fig. 3d) also showing the most pronounced overprediction during these times. A comparison between the ENS4 and ENS20 (regrouped) rank histograms (Fig. 8a) reveals that ENS20 members more severely overpredict precipitation than the ENS4 members. Both ensembles have a slight u shape, indicating a lack of spread (i.e., underprediction of forecast uncertainty), but the right skewness of each ensemble's rank histograms makes it difficult to diagnose which ensemble suffers most severely from this lack of spread. Thus, a procedure is devised to remove the bias from the members of the ensembles. The biases are removed using the PM method applied to each ensemble member forecast, so that forecast precipitation amounts are reassigned using the corresponding distribution of observed precipitation amounts. Thus, the modified forecast precipitation fields have the same patterns and locations as the original forecasts, but the forecast rainfall amounts are adjusted so that their distribution exactly matches that of the observed precipitation. After the modification is applied, a computation of bias at all precipitation thresholds yields a value of 1. An example of a precipitation forecast before and after this procedure is applied is displayed in Fig. 9.

Bias-corrected ENS4 and ENS20 ensemble member forecasts (ENS4* and ENS20* hereafter; see Figs. 8c and 8d) still appear to be slightly right skewed at all forecast lead times. This skewness may result because observations are not drawn from a normal (Gaussian) distribution, while the ensemble FP distributions are assumed to be Gaussian. A more detailed discussion of rank histogram behavior when observations and forecasts are not drawn from the same PDF is found in Hamill (2001). In addition, Wilks (1995) notes that precipitation distributions often fit gamma distributions that are right skewed and do not allow negative values. Thus, the right-skewed rank histograms (Figs. 8c and 8d) may result from observations being drawn from a gamma distribution. Figure 8c reveals that ENS4* and ENS20* have very similar representations of the forecast uncertainty, with both ensembles exhibiting a slight lack of spread, especially up to forecast hour 21. However, there appears to be a trend for the ENS4* rank histograms to become flatter with increasing forecast lead time, while those of ENS20* become slightly more u shaped. By forecast hours 27 and 33, it is clear that ENS4* has a better representation of forecast uncertainty than ENS20*, as indicated by ENS4*'s flatter rank histogram than ENS20*. In addition, in an attempt to quantify rank histogram ''flatness,'' summed absolute values of residuals from least squares fits to the

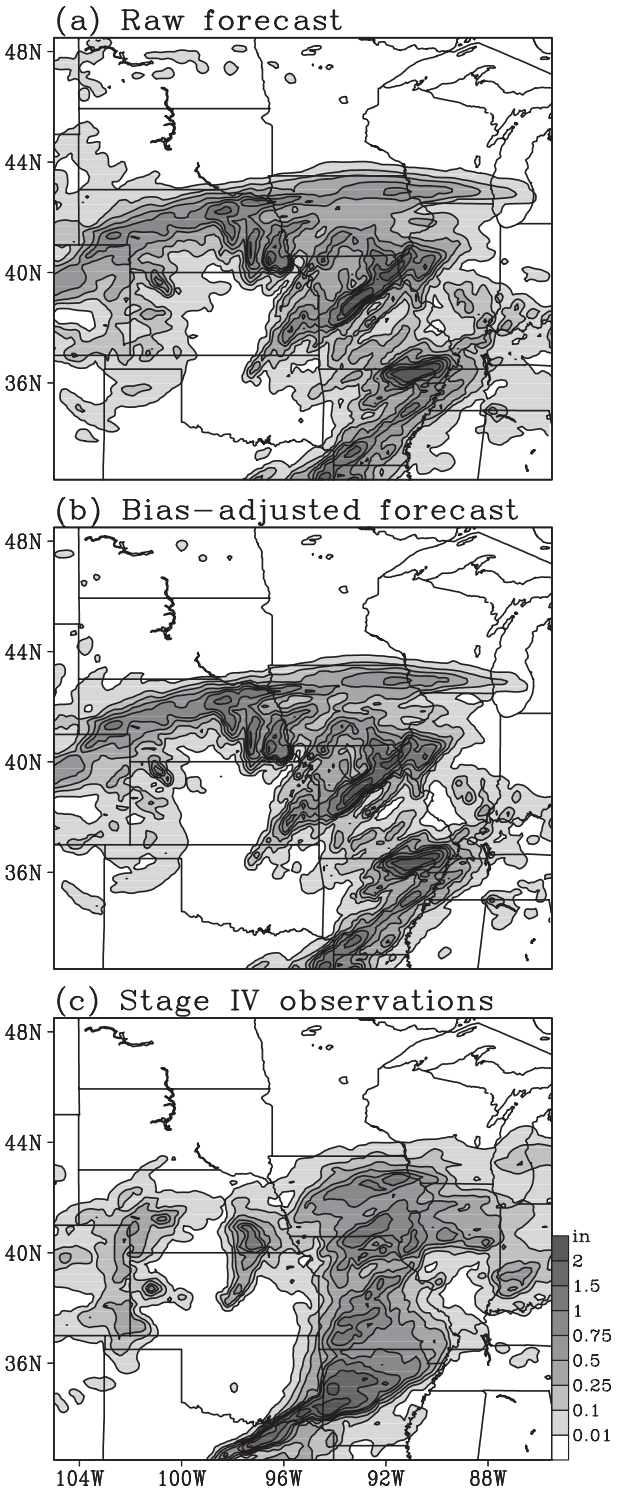

FIG. 9. Example from an ENS4 ensemble member of a (a) raw precipitation forecast and (b) bias-corrected precipitation forecast, along with (c) the stage IV observed precipitation analysis. The forecast was initialized 23 Apr 2007 and valid for forecast hours 27–33.

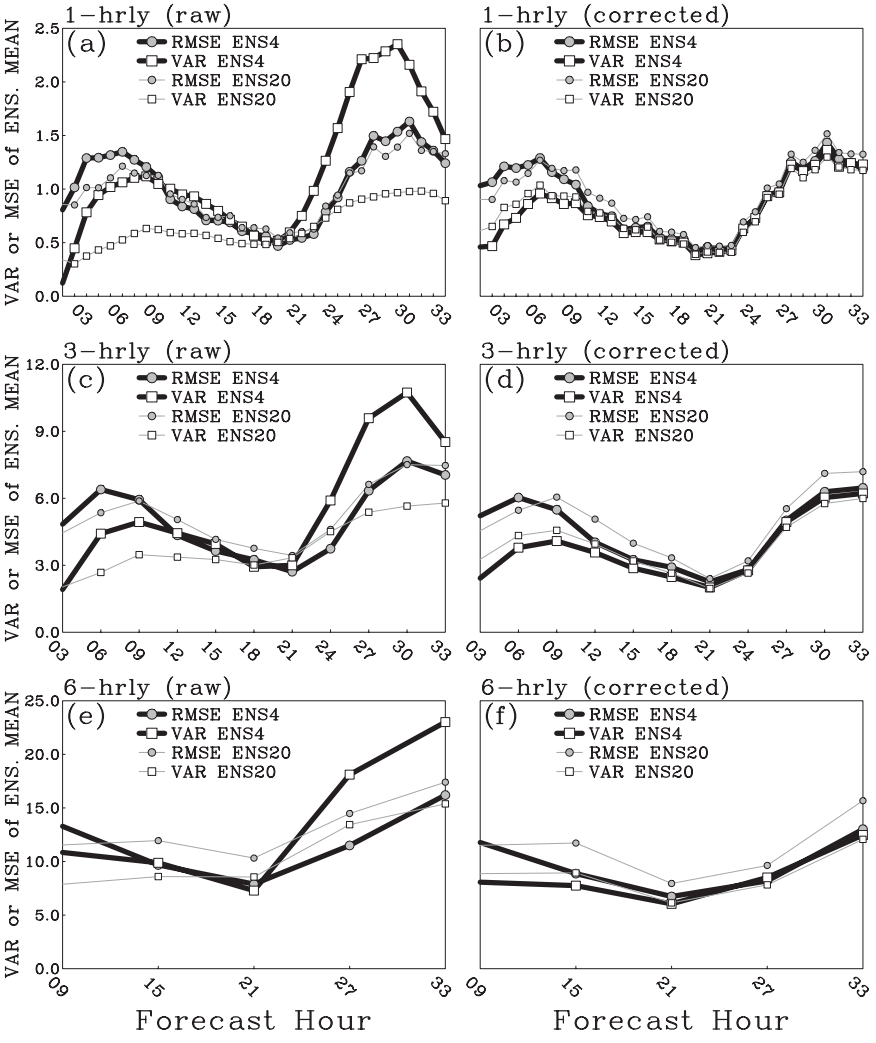

FIG. 10. Time series of average ensemble variance and MSE of the raw ENS4 and ENS20 ensemble mean precipitation forecasts for (a) 1-, (b) 3-, and (c) 6-hourly accumulation intervals, and for bias-corrected ensemble mean precipitation forecasts for (d) 1-, (e) 3-, and (f) 6-hourly accumulation intervals.

observed relative frequencies in each rank histogram set (Fig. 8) verify these visual interpretations (i.e., smaller summed residuals indicate flatter rank histograms).

2) STATISTICAL CONSISTENCY ANALYSIS

Ensembles correctly forecasting uncertainty are considered statistically consistent, and the mean-square error (MSE) of the ensemble mean will match the ensemble variance when averaged over many cases (Talagrand et al. 1999; Eckel and Mass 2005). In this study, the MSE and variance are computed according to Eqs. (B6) and (B7), respectively, in Eckel and Mass (2005), which account for an ensemble with a finite number of members. An analysis of statistical consistency compliments that from rank histograms because the forecast accuracy (i.e., MSE of ensemble mean) and error growth rates (i.e., ensemble variance) between ensembles can be compared, attributes that cannot be inferred from rank histograms. However, note that rank histograms provide information on ensemble bias, while an analysis of statistical consistency does not. The importance of recognizing bias when interpreting statistical consistency is illustrated in this section.

The trends in the MSE of the ensemble mean and ensemble variance of both ensembles follow the diurnal precipitation cycle (Fig. 10). It appears that ENS20 underpredicts forecast uncertainty at most forecast lead times, except around forecast hours 21–24, corresponding to the minimum in the diurnal precipitation cycle. However, the ENS4 ensemble variance increases

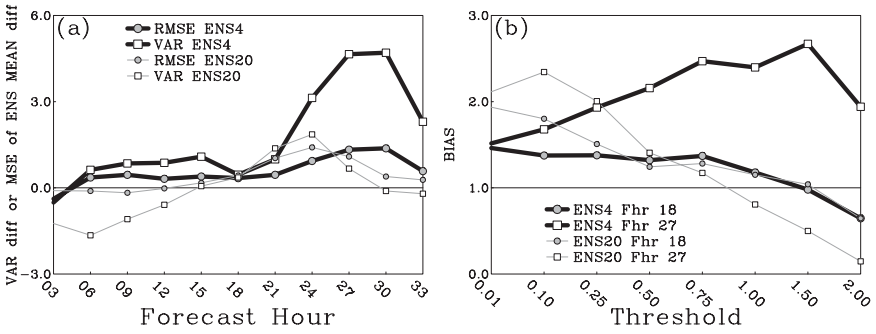

FIG. 11. (a) Time series of differences between ensemble variance and MSE of the ensemble mean before and after bias correction at 3-hourly accumulation intervals for ENS4 and ENS20. (b) Average bias of ENS4 and ENS20 ensemble mean forecasts (generated using probability matching) at increasing precipitation thresholds at forecast hours 18 and 27.

at a much faster rate than that in ENS20, and the ENS4 MSE of the ensemble mean becomes similar to its ensemble variance around forecast hours 9, 12, and 15 for 1-, 3-, and 6-hourly accumulation intervals, respectively (Figs. 10a, 10c, and 10e). After about forecast hour 21, the ENS4 MSE of the ensemble mean becomes smaller than the ensemble variance for all accumulation intervals, implying overprediction of forecast uncertainty, contradicting rank histogram results.

The discrepancy between the rank histogram and statistical consistency results (Figs. 10a, 10c, and 10e) highlights the importance of recognizing the effects of bias when interpreting statistical consistency analyses. When bias is removed using the adjustment process described in the previous section, and MSEs of the ensemble mean and ensemble variance are recomputed (Figs. 10b, 10d, and 10f), the results are consistent with those obtained from rank histogram analyses (i.e., increasing statistical consistency with increasing forecast lead time in ENS4*, with little change in ENS20* statistical consistency as lead time increases). To clearly illustrate the effects of bias on ENS4 and ENS20, differences between the ensemble variance and MSE of the ensemble mean before and after bias correction for 3-hourly accumulation intervals are shown in Fig. 11a. Generally, there are larger-magnitude differences in the ensemble variance than in the MSE of the ensemble mean, which is likely because both quantities used to compute variance (ensemble mean and member forecasts) get adjusted, while only one of the two quantities used to compute the MSE of the ensemble mean is adjusted. Furthermore, the differences between the ENS4 and ENS4* ensemble variances before and after bias correction are usually larger than those of ENS20 (i.e., ensemble variance usually decreases more in ENS4* than in ENS20*). The larger differences between the ENS4 and ENS4* ensemble variances likely result because ENS4 has larger biases than ENS20 at relatively high rainfall thresholds. Because differences between forecasts and the ensemble mean are squared during the computation of the ensemble variance, the biases at relatively high rainfall thresholds have a larger impact on ensemble variance relative to lighter rainfall thresholds. This effect of biases at high rainfall thresholds is supported by Fig. 11b, which shows biases at increasing rainfall thresholds for forecast hours 18 and 27. When the differences in ensemble variance were similar at forecast hour 18 (Fig. 11a), biases at rainfall thresholds above 0.25 in. were also similar (Fig. 11b). However, at forecast hour 27 when the ENS4–ENS4* ensemble variance was much larger than that of ENS20–ENS20* (Fig. 11a), biases above the 0.25-in. precipitation threshold were much larger in ENS4 than in ENS20 (Fig. 11b).

Error growth rates (i.e., rate of increase in spread) can be directly analyzed using the ensemble variance from ENS4* and ENS20* (Figs. 10b, 10d, and 10f). First, note that the faster error growth inferred from the ensemble variance in ENS4 relative to ENS20 up to forecast hour 9 (at 1- and 3-hourly accumulation intervals), and after forecast hour 21 (all accumulation intervals; Figs. 10a, 10c, and 10e), is largely an artifact of bias. After the biases are removed, it becomes clear that the error growth rates of ENS4* and ENS20* are much more similar than was implied by the ensemble variance from ENS4 and ENS20. However, there are still noticeable differences. An approximation of average error growth rates computed by fitting a least squares line to the ensemble variance (displayed in Figs. 10b, 10d, and 10f) for ENS4* (ENS20*) yields slopes of 0.016 (0.010), 0.255 (0.145), and 0.959 (0.527) for 1-, 3-, and 6-hourly accumulation intervals, respectively. So, although the ENS20* ensemble variance begins higher than that of ENS4*, faster error growth likely resulting from

resolving smaller scales in ENS4* than in ENS20* leads to higher ensemble variance in ENS4* after forecast hour 21. Because ENS20* has one more set of varied physics parameterizations (namely, the CP) than ENS4*, it is likely that the larger ensemble variance in ENS20* during the first part of the forecast period results from larger model uncertainty than in ENS4*, which is supported by time series of average ensemble variance from the ENS4* and ENS20* runs with only mixed physics for a set of 20 cases (Fig. 12). The mixed-physics-only ENS4* simulations consisted of five of the ten 4-km grid-spacing ensemble members described in section 2 that were not examined in detail for this study because they neglected IC uncertainty, and the mixed-physics-only ENS20* simulations consisted of fifteen 20-km grid-spacing members with identical configurations as those described in Table 2, but rerun without IC perturbations. Future work is planned to explore the contributions to ensemble spread from different error sources for the CAR and PCR ensembles in more detail.

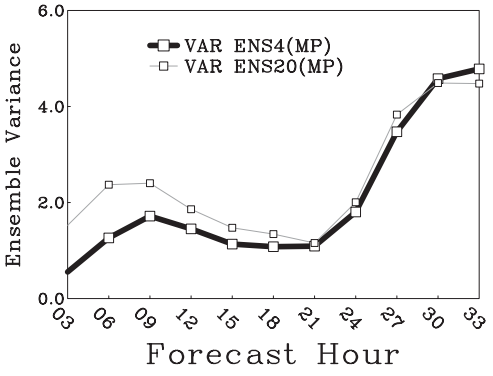

FIG. 12. Time series of average ensemble variance at 3-hourly accumulation intervals for ENS4 and ENS20 using only mixed-physics perturbations.

5. Summary and future work

Precipitation forecast skill from a 5-member, 4-km grid-spacing ensemble (ENS4) was compared to that from a 15-member, 20-km grid-spacing ensemble (ENS20) for 23 warm-season cases during April–June 2007 over the central United States. The goal of this work was to examine, through the use of deterministic and probabilistic skill metrics and Hovmöller diagrams, whether the advantages realized by refining to CAR in ENS4, would outweigh the disadvantages resulting from being forced to use a smaller ensemble size relative to ENS20 due to computational expenses. The main results are summarized below.

Analysis of diurnally averaged Hovmöller diagrams revealed that, as expected, most ENS4 members and the ENS4 ensemble mean had a better diurnal precipitation cycle depiction than ENS20 members and the ENS20 ensemble mean. In addition, the ENS4 ensemble mean diurnal cycle depiction appeared to represent an improvement relative to that of individual members. These results confirmed that ENS4 members should have an inherent advantage over ENS20 with respect to forecasting the timing and location of rainfall systems.

ENS4 ensemble mean precipitation forecasts derived using probability matching generally performed better than those of ENS20, as measured by ETSs, especially at increasing rainfall thresholds. The ENS4 ETSs were higher than those of ENS20 with differences that were statistically significant at all rainfall thresholds examined for 1-hourly accumulation intervals, and at the 0.50-in. rainfall threshold, for 3- and 6-hourly accumulation intervals. The statistically significant differences tended to occur between forecast hours 9 and 15 (0600 and 12000 UTC), corresponding to the times at which MCS occurrence in the central United States is most frequent. In addition, it was found that the biggest differences in ETSs for individual cases occurred when mature MCSs were present.

It was found that ENS4 probabilistic forecasts also generally performed better than those of ENS20, as measured by ROC scores. In addition, similar to ETS, at increasing rainfall thresholds, the differences between ENS4 and ENS20 became larger. However, statistically significant differences tended to be confined to heavier rainfall thresholds than in the ETS analysis.

Bias-corrected rank histograms revealed that both ensembles had an approximately similar representation of forecast uncertainty up to forecast hour 21, but after forecast hour 21, ENS4* rank histograms appeared flatter than those of ENS20*, implying a superior depiction of forecast uncertainty. An analysis of the statistical consistency complemented the rank histograms, also showing a superior depiction of forecast uncertainty in ENS4*. Furthermore, ENS4* had higher error growth rates (i.e., spread increased faster) than ENS20*, but ENS20* had greater spread during the first part of the forecast, probably due to larger model uncertainties associated with the use of multiple CPs in ENS20*. The higher growth rates of the spread in ENS4* likely occurred because smaller scales were being resolved in ENS4* than in ENS20*, and perturbations grow faster on these smaller scales (e.g., Nutter et al. 2004).

Generally, the results from this work are very encouraging for CAR, and the improvements realized from utilizing a CAR ensemble should provide incentive for operational SREF systems to refine their ensemble

resolution to explicitly resolve convection, even if the numbers of members must be reduced due to computational limitations. However, because of the limited time period examined (April–June) and relatively small sample sizes of the cases, it is not clear whether these results are representative of other periods with different flow regimes. For example, the midsummer months (i.e., July–August), characterized by a dominant upper-level ridge over the central United States and ''weakly forced'' convective events, may be even more advantageous to CAR ensembles relative to PCR ensembles because of a stronger diurnal signal during midsummer relative to spring.

Future work should explore CAR ensembles using more members and larger sets of cases for convective and nonconvective precipitation events. In addition, contributions to errors from models and analyses should be quantified at convection-allowing scales because of the implications for ensemble design. Furthermore, it is recommended that entity-based verification techniques (e.g., Ebert and McBride 2000; Davis et al. 2006a) that have been shown to be useful for verifying deterministic CAR simulations (Davis et al. 2006b) be applied to CAR ensembles. In particular, object-based techniques provide an alternative method for examining statistical consistency by comparing average displacement errors between ensemble members to the average displacement error between a mean forecast and observations.

Acknowledgments. The authors thank Huiling Yuan at the Global Systems Division of the Earth System Research Laboratory (ESRL GSD) for assistance in obtaining SREF data in post–real time. In addition, Jon Hobbs at Iowa State University (ISU) and Isidora Jankov at the ESRL GSD assisted with the computational work. This particular research was funded by NSF Grant ATM-0537043. The ENS20 simulations were conducted on the 64-processor computing cluster in the meteorology program at ISU. The CAPS real-time 4-km ensemble forecasts were primarily supported by the NOAA CSTAR program. Supplementary support was provided by NSF ITR project LEAD (ATM-0331594). Drs. Kelvin K. Droegemeier, Keith Brewster, John Kain, Steve Weiss, David Bright, Matt Wandishin, Mike Coniglio, Jun Du, Jimy Dudhia, Morris Weisman, Greg Thompson, and Wei Wang contributed to the ensemble system design and WRF model configuration. Kevin Thomas carried out the real-time runs. The CAPS real-time predictions were performed at the Pittsburgh Supercomputing Center (PSC) supported by NSF. Finally, constructive comments from three reviewers were appreciated.

REFERENCES

Baldwin, M. E., and K. E. Mitchell, 1997: The NCEP hourly multisensor U.S. precipitation analysis for operations and GCIP research. Preprints, *13th Conf. on Hydrology,* Long Beach, CA, Amer. Meteor. Soc., 54–55.

——, 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., 255–258.

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, ATEX and Arctic air-mass data sets. *Quart. J. Roy. Meteor. Soc.,* **112,** 693–709.

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.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Wea. Forecasting,* **14,** 168–189.

Bukovsky, M. S., J. S. Kain, and M. E. Baldwin, 2006: Bowing convective systems in a popular operational model: Are they for real? *Wea. Forecasting,* **21,** 307–324.

Carbone, R. E., J. D. Tuttle, D. A. Ahijevych, and S. B. Trier, 2002: Inferences of predictability associated with warm season precipitation episodes. *J. Atmos. Sci.,* **59,** 2033–2056.

Chou, M.-D., and M. J. Suarez, 1994: An efficient thermal infrared radiation parameterization for use in general circulation models. NASA Tech. Memo. 104606, 3, 85 pp.

Clark, A. J., W. A. Gallus Jr., and T. C. Chen, 2007: Comparison of the diurnal precipitation cycle in convection-resolving and non-convection-resolving mesoscale models. *Mon. Wea. Rev.,* **135,** 3456–3473.

——, ——, and ——, 2008: Contributions of mixed physics versus perturbed initial/lateral boundary conditions to ensemble-based precipitation forecast skill. *Mon. Wea. Rev.,* **136,** 2140–2156.

Davis, C. A., K. W. Manning, R. E. Carbone, S. B. Trier, and J. D. Tuttle, 2003: Coherence of warm season continental rainfall in numerical weather prediction models. *Mon. Wea. Rev.,* **131,** 2667–2679.

——, and Coauthors, 2004: The Bow Echo and MCV Experiment: Observations and opportunities. *Bull. Amer. Meteor. Soc.,* **85,** 1075–1093.

——, B. Brown, and R. Bullock, 2006a: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.,* **134,** 1772–1784.

——, ——, and ——, 2006b: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.,* **134,** 1785–1795.

Done, J., C. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecast (WRF) model. *Atmos. Sci. Lett.,* **5,** 110–117.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.,* **125,** 2427–2459.

——, and Coauthors, 2004: The NOAA/NWS/NCEP Short Range Ensemble Forecast (SREF) system: Evaluation of an initial condition vs multiple model physics ensemble approach.

Preprints, *20th Conf. on Weather Analysis and Forecasting/16th Conf. on Numerical Weather Prediction,* Seattle, WA, Amer. Meteor. Soc., 21.3. [Available online at http://ams.confex.com/ams/pdfpapers/71107.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 System, Exeter, United Kingdom, WMO.

Dyer, A. J., and B. B. Hicks, 1970: Flux-gradient relationships in the constant flux layer. *Quart. J. Roy. Meteor. Soc.,* **96,** 715–721.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.,* **129,** 2461–2480.

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

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

Ek, M. B., K. E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, and J. D. Tarpley, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta Model. *J. Geophys. Res.,* **108,** 8851, doi:10.1029/2002JD003296.

Ferrier, B. S., Y. Jin, Y. Lin, T. Black, E. Rogers, and G. DiMego, 2002: Implementation of a new grid-scale cloud and rainfall scheme in the NCEP Eta Model. Preprints, *15th Conf. on Numerical Weather Prediction,* San Antonio, TX, Amer. Meteor. Soc., 280–283.

Fritsch, J. M., and R. E. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc.,* **85,** 955–965.

Gilmour, I., L. A. Smith, and R. Buizza, 2001: Linear regime duration: Is 24 hours a long time in synoptic weather forecasting? *J. Atmos. Sci.,* **58,** 3525–3539.

Grell, G. A., and D. Devenyi, 2002: A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophys. Res. Lett.,* **29,** 1693, doi:10.1029/2002GL015311.

Hamill, T. M., and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

——, and ——, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

——, 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting,* **14,** 155–167.

——, 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–560.

Harvey, L. O., K. R. Hammond, C. M. Lusk, and E. F. Mross, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.,* **120,** 863–883.

Holton, J. R., 2004: *An Introduction to Dynamic Meteorology*. 4th ed. Academic Press, 535 pp.

Hong, S.-Y., and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Kor. Meteor. Soc.,* **42,** 129–151.

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

——, 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, NOAA/NWS, 61 pp.

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

Jankov, I., W. A. Gallus Jr., 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.

Jones, M. S., B. A. Colle, and J. S. Tongue, 2007: Evaluation of a mesoscale short-range ensemble forecast system over the northeast United States. *Wea. Forecasting,* **22,** 36–55.

Kain, J. S., and J. M. Fritsch, 1993: Convective parameterization for mesoscale models: The Kain–Fritcsh scheme. *The Representation of Cumulus Convection in Numerical Models, Meteor. Monogr.,* No. 46, Amer. Meteor. Soc., 165–170.

——, and ——, 1998: Multiscale convective overturning in mesoscale convective systems: Reconciling observations, simulations, and theory. *Mon. Wea. Rev.,* **126,** 2254–2273.

——, 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., K. K. Droegemeier, and N. L. Hickmon, 2006: Multiresolution ensemble forecasts of an observed tornadic thunderstorm system. Part I: Comparison of coarse- and fine-grid experiments. *Mon. Wea. Rev.,* **134,** 807–833.

——, ——, and ——, 2007a: Multiresolution ensemble forecasts of an observed tornadic thunderstorm system, Part II. Storm-scale experiments. *Mon. Wea. Rev.,* **135,** 759–782.

——, and Coauthors, 2007b: 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,* Park City, UT, Amer. Meteor. Soc., 3B.2. [Available online at http://ams.confex.com/ams/pdfpapers/124667.pdf.]

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

Liu, C., M. W. Moncrieff, J. D. Tuttle, and R. E. Carbone, 2006: Explicit and parameterized episodes of warm-season precipitation over the continental United States. *Adv. Atmos. Sci.,* **23,** 91–105.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus,* **21,** 289–307.

Maddox, R. A., 1983: Large-scale meteorological conditions associated with midlatitude, mesoscale convective complexes. *Mon. Wea. Rev.,* **111,** 1475–1493.

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

Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.,* **128,** 2145–2166.

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

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmosphere: RRTM, a validated correlated-k model for the long-wave. *J. Geophys. Res.,* **102** (D14), 16 663–16 682.

Molinari, J., and M. Dudek, 1992: Parameterization of convective precipitation in mesoscale numerical models: A critical review. *Mon. Wea. Rev.,* **120,** 326–344.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Monin, A. S., and A. M. Obukhov, 1954: Basic laws of turbulent mixing in the surface layer of the atmosphere (in Russian). *Contrib. Geophys. Inst. Acad. Sci. USSR,* **151,** 163–187.

Mylne, K. R., 1999: The use of forecast value calculations for optimal decision making using probability forecasts. Preprints, *17th Conf. on Weather Analysis and Forecasting,* Denver, CO, Amer. Meteor. Soc., 235–239.

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.

Nutter, P., D. Stensrud, and M. Xue, 2004: Effects of coarsely resolved and temporally interpolated lateral boundary conditions on the dispersion of limited-area ensemble forecasts. *Mon. Wea. Rev.,* **132,** 2358–2377.

Palmer, F., R. Molteni, R. Mureau, P. Buizza, P. Chapelet, and J. Tribbia, 1992: Ensemble prediction. ECMWF Research Dept. Tech. Memo. 188, 45 pp.

Paulson, C. A., 1970: The mathematical representation of wind speed and temperature profiles in the unstable atmospheric surface layer. *J. Appl. Meteor.,* **9,** 857–861.

Petch, J. C., 2006: Sensitivity studies of developing convection in a cloud-resolving model. *Quart. J. Roy. Meteor. Soc.,* **132,** 345–358.

Richardson, D. S., 2000: Applications of cost-loss models. *Proc. Seventh Workshop on Meteorological Operational Systems,* Reading, United Kingdom, ECMWF, 209–213.

——, 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.,* **127,** 2473–2489.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting,* **5,** 570–575.

Skamarock, W. C., 2004: Evaluating mesoscale NWP models using kinetic energy spectra. *Mon. Wea. Rev.,* **132,** 3019–3032.

——, 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; and online at http://box.mmm.ucar.edu/wrf/users/docs/arw_v2.pdf.]

Smagorinsky, J., 1969: Problems and promises of deterministic extended range forecasting. *Bull. Amer. Meteor. Soc.,* **50,** 286–311.

Stensrud, D. J., J. 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.

Talagrand, O., R. Vautard, and B. Strauss, 1999: Evaluation of probabilistic prediction systems. *Proc. Workshop on Predictability,* Reading, United Kingdom, ECMWF, 1–25.

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.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

Tuttle, J. D., and C. A. Davis, 2006: Corridors of warm season precipitation in the central United States. *Mon. Wea. Rev.,* **134,** 2297–2317.

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.

Webb, E. K., 1970: Profile relationships: The log-linear range and extension to strong stability. *Quart. J. Roy. Meteor. Soc.,* **96,** 67–90.

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.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.

Xue, M., 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,* Park City, UT, Amer. Meteor. Soc., 3B.1. [Available online at http://ams.confex.com/ams/pdfpapers/142036.pdf.]