Comparison of Next-Day Probabilistic Severe Weather Forecasts from Coarse- and Fine-Resolution CAMs and a Convection-Allowing Ensemble

ERIC D. LOKEN

Cooperative Institute for Mesoscale Meteorological Studies and School of Meteorology, University of Oklahoma, and NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

ADAM J. CLARK

NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

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 17 November 2016, in final form 9 May 2017)

ABSTRACT

Given increasing computing power, an important question is whether additional computational resources would be better spent reducing the horizontal grid spacing of a convection-allowing model (CAM) or adding members to form CAM ensembles. The present study investigates this question as it applies to CAM-derived next-day probabilistic severe weather forecasts created by using forecast updraft helicity as a severe weather proxy for 63 days of the 2010 and 2011 NOAA Hazardous Weather Testbed Spring Forecasting Experiments. Forecasts derived from three sets of Weather Research and Forecasting Model configurations are tested: a 1-km deterministic model, a 4-km deterministic model, and an 11-member, 4-km ensemble. Forecast quality is evaluated using relative operating characteristic (ROC) curves, attributes diagrams, and performance diagrams, and forecasts from five representative cases are analyzed to investigate their relative quality and value in a variety of situations. While no statistically significant differences exist between the 4- and 1-km deterministic forecasts in terms of area under ROC curves, the 4-km ensemble forecasts offer weakly significant improvements over the 4-km deterministic forecasts over the entire 63-day dataset. Further, the 4-km ensemble forecasts generally provide greater forecast quality relative to either of the deterministic forecasts on an individual day. Collectively, these results suggest that, for purposes of improving next-day CAM-derived probabilistic severe weather forecasts, additional computing resources may be better spent on adding members to form CAM ensembles than on reducing the horizontal grid spacing of a deterministic model below 4 km.

1. Introduction

The prospect of increasing a numerical weather prediction model's forecast skill by decreasing its horizontal grid spacing has interested scientists for some time (e.g., Lilly 1990; Brooks et al. 1992; Weygandt and Seaman 1994; Mass et al. 2002). This interest is evidenced, in part, by the decrease in horizontal grid spacing of the U.S. operational North American Mesoscale (NAM) Forecast System from 80 km in 1993 to 12 km in 2001 (Kain et al. 2008) and the advent of the 3-km grid spacing High Resolution Rapid Refresh (HRRR) model (Benjamin et al. 2016). As increasing computing power has permitted models to operate with finer horizontal resolution, it has become clear that 4 km is about the maximum grid spacing that can still produce the dominant circulations in midlatitude mesoscale convective systems without having to use convective parameterization

Corresponding author: Eric D. Loken, eric.d.loken@noaa.gov

DOI: 10.1175/WAF-D-16-0200.1

(e.g., Weisman et al. 1997; Done et al. 2004). These models run without convective parameterization are typically referred to as convection-allowing models, or CAMs.

Decreasing horizontal grid spacing has generally led to clear improvements in forecast skill at convection-parameterizing resolutions (e.g., Mass et al. 2002). Decreasing from convection-parameterizing to convection-allowing grid spacing has also led to clear improvements, especially for fields related to convection (e.g., Clark et al. 2009, 2010, 2012a; Weisman et al. 2008; Done et al. 2004). However, further decreasing the grid spacing of a convection-allowing model has provided mixed results. For example, Kain et al. (2008) compared 4- and 2-km forecasts from the Advanced Research version of the Weather Research and Forecasting Model (WRF-ARW) for simulated lowest-level reflectivity, hourly precipitation, and hourly updraft helicity fields during the 2005 NOAA Hazardous Weather Testbed Spring Forecasting Experiment (HWT SFE), finding that while the finer-resolution forecasts tended to produce more convective detail, they added no significant quality or value relative to the coarser-resolution forecasts. Schwartz et al. (2009) obtained a similar result when comparing 4- and 2-km WRF-ARW-simulated 1 km above ground level reflectivity and 1-h accumulated precipitation forecasts during the 2007 NOAA HWT SFE. Likewise, Clark et al. (2012a) noted that participants in the 2010 HWT SFE gave similar subjective ratings to 1- and 4-km CAM forecasts of deep convection.

Meanwhile, Johnson et al. (2013), who studied 4- and 1-km 30-h forecasts of 1-h accumulated precipitation over 91 days during the 2009–11 NOAA HWT SFEs, found that the 1-km forecasts had a significantly greater median of maximum interest relative to the 4-km forecasts but noted that the two sets of forecasts had similar object-based threat scores. When the 1-km forecasts were mapped onto a 4-km grid, the two forecast configurations had similar verification scores, suggesting that the 1-km forecasts were superior predominantly on scales not fully resolvable with 4-km grid spacing (Johnson et al. 2013). Interestingly, Roberts and Lean (2008), who used the Met Office Unified Model to compare precipitation forecasts from runs with 12-, 4-, and 1-km horizontal grid spacing, found that the 1-km forecasts outperformed the 4- and 12-km forecasts for all scales greater than 15 km. With this said, Roberts and Lean (2008) used a modified form of convective parameterization for their 4-km run, and they focused on forecast time periods of 7 h or less, when differences between the 1- and 4-km forecasts may not yet have been dominated by large-scale errors (e.g., Schwartz et al. 2009; Potvin and Flora 2015). VandenBerg et al. (2014) compared storm motion forecasts from models with 1- and 4-km horizontal grid spacing and concluded that the 1-km storm motion forecasts may offer some improvements over the 4-km forecasts, noting that—at least when viewed relative to environmental flow and for short-lived storms—mean storm speeds produced by the 1-km model were significantly closer to the observed mean storm speeds. Potvin and Flora (2015) studied the impact of varying horizontal model resolution on idealized supercells, concluding that—at least within an idealized framework and at short (i.e., on the order of 1 h) time scales—4-km horizontal grid spacing was too coarse to reliably resolve key supercell processes, since storms tended to decay prematurely and have large track errors. However, Potvin and Flora (2015) noted that their 3-km grid spacing simulations typically resolved important operational features, such as low-level rotation tracks, while their 1-km simulations had the ability to resolve rapid changes in low-level rotation. For a single case of convection over the central United States on 26 May 2008, Xue et al. (2013) found that 1-km grid-spacing forecasts subjectively outperformed the corresponding 4-km forecasts in terms of storm structure and intensity.

Given mixed findings on the benefits of decreasing horizontal grid spacing beyond about 4 km, an open question is whether additional computing power should be spent on increasing horizontal resolution or on adding ensemble members to improve forecasting skill. Indeed, an ensemble may provide an advantage over a similarly configured deterministic model by accounting for forecast uncertainties related to errors in the initial conditions and model parameterizations (e.g., Wandishin et al. 2001). However, it is currently unknown how the skill of an ensemble at coarser—but still convection-allowing—resolution would compare to that of a similarly configured deterministic model with finer grid spacing. The present study seeks to address this question as it applies to next-day probabilistic severe weather forecasts derived from forecast updraft helicity (UH).

UH has been identified as an important severe weather forecasting parameter. For example, Kain et al. (2008) used large values of hourly UH to successfully identify mesocyclones during the 2005 SFE, and Kain et al. (2010) developed a strategy for calculating temporal maximum UH by tracking the largest values occurring at any time step between model output times, thus accounting for the rapid evolution in convective storms. Hereafter, UH refers to the hourly maximum quantity (i.e., the maximum UH value at any time step between hourly output times). Sobash et al. (2011), inspired by the perceived correspondence between large values of UH and severe

TABLE 1. Dates from the 2010–11 NOAA HWT SFEs included in the dataset (63 total dates).

Month	2010	2011
Apr	28, 29, 30	27–29
May	3–7, 10–14, 17–21, 24–28, 31	4, 9, 12–13, 18–20, 22–28, 31
Jun	1–4, 7–11, 14–18	1, 3, 6–10

weather reports during the 2008 SFE, treated "extreme" values of simulated UH as "surrogate" severe weather reports. Applying a spatial smoother to these surrogate reports, Sobash et al. (2011) created a field of surrogate severe probabilistic forecasts (SSPFs) that provided skillful and useful guidance for severe weather forecasters. Clark et al. (2012c) and Clark et al. (2013) investigated whether simulated UH track lengths corresponded with observed tornado track lengths, finding that simulated UH track lengths showed some skill as proxies for tornado track lengths, particularly during the spring months and particularly when the storm environment was used to filter the UH tracks associated with elevated and/or high-based storms. While most previous research has focused on deterministic UH forecasts, Sobash et al. (2016b) investigated the effect of using a 30-member CAM ensemble to create day-1 and day-2 SSPFs over a 32-day period coinciding with the Mesoscale Predictability Experiment (Weisman et al. 2015). Sobash et al. (2016b) found that the ensemble SSPFs were more skillful and reliable relative to the deterministic SSPFs on the mesoscale but not necessarily for larger scales.

This paper builds on the work of Sobash et al. (2011, 2016b) by investigating how a reduction in grid spacing from 4 to 1 km and the creation of a 4-km ensemble influences the quality and value (e.g., Murphy 1993) of next-day SSPFs derived from forecast UH fields. The study is organized as follows. Section 2 details the model specifications and the methodology used for this study, section 3 provides the results and examines five case study days, section 4 summarizes and discusses the results, and section 5 outlines potential future work.

2. Methods

a. Model specifications

During the 2010 and 2011 HWT SFEs, the Center for Analysis and Prediction of Storms (CAPS) ran a 26-member storm-scale ensemble forecast (SSEF) system with 4-km grid spacing (Clark et al. 2012b). The present study analyzes model output from the control member of this SSEF system, an equivalent 1-km version of this control member, and a subset of 11 ensemble members

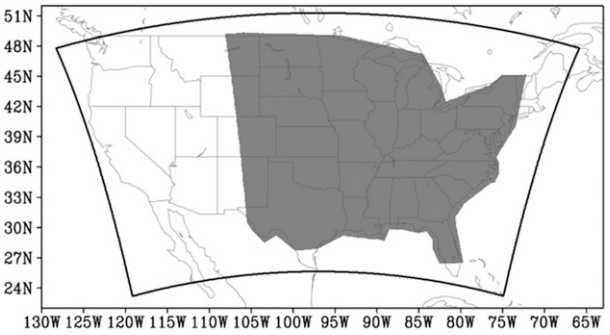

FIG. 1. Model domain (black contour) and analysis domain (gray shading).

for a total of 63 days (38 days from 2010 and 25 days from 2011) over the 2010 and 2011 SFEs (Table 1).[1] Both the 4- and 1-km deterministic models and all 11 members of the ensemble subset use the ARW-WRF Model (Skamarock et al. 2008) dynamic core and have 51 vertical levels. The domain of all models covers the contiguous United States, although the analysis domain is restricted to the eastern two-thirds of the United States (Fig. 1). Analyses from the 0000 UTC 12-km NAM are used as the background for both deterministic models. Then, WSR-88D data are assimilated along with surface and upper-air observations using the Advanced Regional Prediction System three-dimensional variational data assimilation and cloud analysis system (ARPS 3DVAR; Xue et al. 2003; Gao et al. 2004). The subset of 11 SSEF members was chosen for the ensemble because hourly maximum updraft helicity was available from these members (7 of the 26 members that did not use the

[1] Three of the days in the 63-day dataset—26 May 2010, 27 April 2011, and 29 April 2011—contain missing data from at least one ensemble member. Each of the three days is assessed on a case-by-case basis to determine how to appropriately handle the analysis for each day. On 26 May 2010 one forecast hour is missing from two different ensemble members. Data from the 28th forecast hour are missing from the arw_m5 member, and data from the 26th forecast hour are missing from the arw_m6 member. Given that these forecast hours are late in the forecast period (and therefore likely do not contain the maximum UH over the entire period), and given that only one forecast hour is missing from only 2 of 11 ensemble members, the missing data are neglected. In the case of 27 April 2011, data are missing from all forecast hours for the arw_m13 member. Given that only 1 of 11 ensemble members is missing, the decision is made to include 27 April 2011 in the dataset but to evaluate the data as having come from a 10-member ensemble instead of an 11-member ensemble. In the case of 29 April 2011, two ensemble members, the arw_m5 and the arw_m12 members, contain missing data for the final six forecast hours. Given that only two SPC storm reports occurred in the contiguous United States on this day and that both occurred between 1900 and 2000 UTC (i.e., forecast hours 7 and 8), the missing data are neglected.

TABLE 2. Deterministic model and ensemble member specifications. NAMf refers to the 12-km NAM forecast, and ARPSa refers to the Advanced Regional Prediction System three-dimensional variational data assimilation (Xue et al. 2003; Gao et al. 2004). Elements in the ICs column followed by a plus sign (+) or minus sign (−) denote SREF perturbations added or subtracted from the ICs of the arw_cn member. Ensemble member boundary layer schemes included Mellor–Yamada–Janjić (MYJ; Mellor and Yamada 1982; Janjić 2002), Yonsei University (YSU; Noh et al. 2003), Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi 2000, 2001; Nakanishi and Niino 2004, 2006), and quasi-normal scale elimination (QNSE; Sukoriansky et al. 2006). Ensemble member microphysics schemes included Thompson et al. (2004), WRF single-moment 6-class (WSM6; Hong and Lim 2006), WRF double-moment 6-class (WDM6; Lim and Hong 2010), Ferrier et al. (2002), and Morrison et al. (2005). All ensemble members used the Rapid Radiative Transfer Model longwave radiation scheme (RRTM; Mlawer et al. 1997) and the Goddard shortwave radiation scheme (Chou and Suarez 1994). Land surface models included the Noah (Chen and Dudhia 2001) and RUC (Smirnova et al. 1997, 2000) models.

Ensemble member/model	ICs	LBCs	Microphysics	Land surface model	Boundary layer
arw_cn (4-km)[a,b]	0000 UTC ARPSa	0000 UTC NAMf	Thompson	Noah	MYJ
arw_cn (1-km)[a]	0000 UTC ARPSa	0000 UTC NAMf	Thompson	Noah	MYJ
arw_m5[c]	arw_cn + em-p1 + recur pert	em-p1	Morrison	RUC	YSU
arw_m6[b]	arw_cn + em-p1_pert	em-p1	Morrison	RUC	YSU
arw_m7[b]	arw_cn + em-p2_pert	em-p2	Thompson	Noah	QNSE
arw_m8[b]	arw_cn − nmm-p1_pert	nmm-p1	WSM6	RUC	QNSE
arw_m9[b]	arw_cn + nmm-p2_pert	nmm-p2	WDM6	Noah	MYNN
arw_m10[b]	arw_cn + rsmSAS-n1_pert	rsmSAS-n1	Ferrier	RUC	YSU
arw_m1[b]	arw_cn − etaKF-n1_pert	etaKF-n1	Ferrier	Noah	YSU
arw_m12[d]	arw_cn + etaKF-p1_pert	etaKF-p1	WDM6	RUC (2010)/ Noah (2011)	QNSE
arw_m13 (2010)[d]	arw_cn − etaBMJ-n1_pert	etaBMJ-n1	WSM6	Noah (2010)/ RUC (2011)	MYNN
arw_m14[c]	arw_cn + etaBMJ-p1_pert	etaBMJ-p1	Thompson	RUC	MYNN
arw_m13 (2011)[e]	arw_cn + rsm-p1_pert	rsm-p1	MYJ	Noah	MYJ

[a] Deterministic models that were used for both 2010 and 2011.
[b] Ensemble members that were part of both the 2010 and 2011 ensembles.
[c] Ensemble members that were part of the 2010 ensemble only.
[d] Ensemble members that had different land surface models for 2010 and 2011 but were otherwise the same for both years.
[e] Ensemble members that were part of the 2011 ensemble only.

WRF-ARW dynamic core did not produce hourly maximum UH), and these were the only members that accounted for both model and analysis error with mixed physics and perturbed initial conditions and lateral boundary conditions (ICs/LBCs), respectively. The other members shared the same set of ICs/LBCs or LBCs to study various IC perturbation methods, the impact of radar data assimilation, and physics sensitivities. Thus, this set of members was less diverse and tended to cluster around the arw_cn member solution. The ensemble IC/LBC perturbations are derived from evolved (through 3 h) perturbations of 2100 UTC NCEP operational Short-Range Ensemble Forecast (SREF; Du et al. 2006) system members and added to the ARPS 3DVAR analyses. Corresponding SREF forecasts are used for the LBCs. Full model specifications are provided in Table 2.

One notable difference between the 2010 and 2011 forecasts is forecast length: the models produced 30-h forecasts in 2010 but 36-h forecasts in 2011. Hence, for 2010 the 18-h period from 1200 to 0600 UTC on the next day is investigated (12–30-h forecast times), while for 2011 the 24-h period from 1200 to 1200 UTC on the next day is examined (12–36-h forecast times). Because the primary goal is to assess next-day severe weather forecast guidance—and because the Storm Prediction Center's (SPC's) Day 1 Convective Outlook forecasts span from 1200 to 1200 UTC—output from the first 12 forecast hours after model initialization is ignored for both forecasts.

b. Producing SSPFs from UH

As in Sobash et al. (2011, 2016b), extreme values of 2–5-km UH are treated as surrogate severe weather reports (SSRs). The 2–5-km UH is computed using the following formula, as in Kain et al. (2008) and Sobash et al. (2011):

$$\mathrm{UH} = \sum_{z=2000\,\mathrm{m}}^{z=5000\,\mathrm{m}} \overline{w\zeta}\Delta z = (\overline{w\zeta_{2,3}} + \overline{w\zeta_{3,4}} + \overline{w\zeta_{4,5}}) \times 1000, \quad (1)$$

where w is vertical velocity ($\mathrm{m\,s^{-1}}$), ζ is vertical vorticity ($\mathrm{s^{-1}}$), and Δz is the vertical distance between computation levels (here, 1000 m). The subscripts indicate the bottom and top computational levels (km), and the overbars

denote an average over the layer between the two given computational levels.

SSPFs are derived from SSRs using the following methodologies for deterministic and ensemble simulations. First, the maximum UH value that occurred at each grid box over the entire 18- (for 2010) or 24-h (for 2011) period of interest is found for the 4- and 1-km models each day. These maximum daily UH values are remapped onto an 80-km grid using the maximum UH from all of the finer-resolution grid points falling within the 80-km grid boxes. Remapping to an 80-km grid is done to match the verification scales used by the SPC and to reduce the computational expense of creating the SSPFs. It should be noted that the SSPFs produced by remapping to an 80-km grid are very similar to those produced on the native 4- or 1-km grids using a 40-km radius of influence when the same 4- and 1-km thresholds are used on both the native and 80-km grids. Remapping to the coarser grid is therefore done to save computation time. After remapping, a UH threshold is applied to produce a binary field, with ones assigned to points equal to or exceeding the UH threshold and zeros assigned to all other points. UH thresholds from 25 to 125 $m^2 s^{-2}$ in increments of 25 $m^2 s^{-2}$ are used for the 4-km data since these represent typical extreme values based on subjective experience and previous research (e.g., Kain et al. 2008; Sobash et al. 2011, 2016b). Because UH is grid-spacing dependent (i.e., increasing resolution results in higher UH), the same thresholds could not be similarly applied to the 1-km UH data. Instead, UH values at equivalent percentiles for each of the thresholds used for the 4-km data are found for the 1-km data. The percentiles are computed using the distributions of UH from all cases after remapping the maximum values to the 80-km grid. This procedure ensures that the number of SSRs in the 1- and 4-km forecasts are similar, thus minimizing the impact of differences in biases. The percentiles are computed separately for 2010 and 2011 data because it was thought that the forecast length difference between the two years (30 h in 2010 versus 36 h in 2011) might cause the characteristics of the distributions to be slightly different. Indeed, the 1-km UH values from 2011 are usually higher than 2010. Table 3 lists the UH thresholds used for the 4-km data, their percentiles, and the UH value from the 1-km data at each of these percentiles for 2010 and 2011. Finally, as in Sobash et al. (2011), a Gaussian kernel is applied to the binary field to produce forecast probabilities using the following formula:

$$f = \sum_{n=1}^{N} \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2}\left(\frac{d_n}{\sigma}\right)^2\right], \quad (2)$$

where f is the probability value at a given grid point, N is the total number of grid points containing an SSR, d_n is the distance from the grid point to the point of the nth SSR, and σ is the standard deviation of the Gaussian kernel (hereafter referred to as the spatial smoothing parameter). Sobash et al. (2011) found that $\sigma = 160$ and 200 km produced the best reliability for the smallest UH thresholds tested. Herein, $\sigma = 120$ km is used because it produces reliable forecasts for some of the larger UH thresholds examined and because resolution is not sacrificed as much as with larger σ values (i.e., more frequent larger probabilities can occur). To produce SSPFs from the 11-member 4-km ensemble, a similar procedure is used. However, after UH from each member is remapped to the 80-km grid and a specified threshold is applied, the ratio of members that exceed the threshold is calculated for each point. Then, the Gaussian smoother is applied to produce the SSPF field. Note that creating the ensemble SSPF field using this procedure is identical to creating the field by averaging the individual SSPFs from each ensemble member (Sobash et al. 2016b). For the ensemble SSPFs, σ is varied from 60 to 120 km in 30-km increments to identify the optimal value of σ within the ensemble framework.

TABLE 3. The 4- and 1-km equivalent UH threshold values. The 2010 percentile and 1-km UH values are located above the corresponding 2011 percentile and 1-km UH values.

4-km UH ($m^2 s^{-2}$)	Percentile	1-km UH ($m^2 s^{-2}$)
25	0.9821775 (2010)	248.8906 (2010)
	0.9811282 (2011)	280.6609 (2011)
50	0.9930652 (2010)	467.4218 (2010)
	0.9927213 (2011)	505.1666 (2011)
75	0.9967307 (2010)	670.5211 (2010)
	0.9967047 (2011)	718.7805 (2011)
100	0.9985068 (2010)	893.3596 (2010)
	0.9983788 (2011)	935.0260 (2011)
125	0.9993252 (2010)	1158.1796 (2010)
	0.9990868 (2011)	1098.2116 (2011)

c. *Verification*

To verify the SSPFs, archived observed storm reports (OSRs) are obtained from the SPC website. These OSRs include reports of wind 58 mi h^{-1} or greater, hail measuring 1 in. or greater in diameter, and tornadoes. The OSRs are filtered to include only those incidents that fall within the designated forecast time periods. Hence, OSRs from 1200 to 0600 UTC on the following day are considered for 2010, and OSRs from 1200 to 1200 UTC on the following day are considered for 2011. As with the SSRs, a binary 80-km grid of OSRs is constructed. Grid boxes with at least one OSR are assigned a value of 1, while all other grid boxes are assigned a value of 0. Relative operating characteristic (ROC) curves (Mason 1982), attributes diagrams

(Hsu and Murphy 1986), and performance diagrams (Roebber 2009) are constructed to help evaluate the quality of the SSPFs.

ROC curves plot the probability of detection (POD), defined as

$$\text{POD} = \frac{\text{hits}}{\text{hits} + \text{misses}}, \tag{3}$$

against the probability of false detection (POFD), defined as

$$\text{POFD} = \frac{\text{false alarms}}{\text{false alarms} + \text{correct negatives}}. \tag{4}$$

Herein, POD and POFD are computed at specified levels of probability: 1%, 2%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, and 95% to create the ROC curves. Each probability level is used to convert the probabilistic forecasts into binary (e.g., yes/no) forecasts; grid boxes meeting or exceeding the given probability level are considered to be yes forecasts at that probability level. One method of determining forecast quality from a ROC curve is by assessing the area under the ROC curve (AUC; e.g., Marzban 2004), a single-number metric that measures a forecast's ability to discriminate between the occurrence and nonoccurrence of an observed event (e.g., Mason and Graham 2002). In the present case, an observed event is defined as the occurrence of an OSR within a given 80-km grid box. An AUC value of 1.0 indicates a perfect forecast, while an AUC value of 0.5 or less represents a random forecast. An AUC value of 0.70 is typically considered to represent the lower limit of skill for probabilistic forecast systems (Buizza et al. 1999; Sobash et al. 2011). In the present study, AUC values are computed over the entire 63-day dataset as a means of evaluating the overall performance of each of the three forecasts. AUC values are also computed over individual days to evaluate the performance of each individual daily forecast. In both cases, a trapezoidal approximation is used to compute AUC (Wandishin et al. 2001).

While ROC curves and AUC values assess a forecast's ability to discriminate between the occurrence and nonoccurrence of events, these metrics do not give information about a forecast's bias (e.g., Wilks 2001). For this reason, attributes diagrams, which contain information about forecast bias, make good complements to ROC curves. Attributes diagrams plot the observed relative frequency against the forecast probability; herein, the attributes diagrams are made using the same levels of probability used for the ROC diagrams. Reliable forecasts are those in which the forecast probabilities correspond to the observed relative frequencies; therefore, points that fall along a diagonal line of slope 1 from the bottom left to the top right of the diagram (called the perfect reliability line) are said to have perfect reliability. Points that fall above (below) the perfect reliability line represent under- (over) forecasts. In addition to the perfect reliability line, attributes diagrams display horizontal and vertical lines at the sample climatological frequency (abbreviated herein as "sample climatology"), which is found by taking the total number of yes observations (i.e., occurrences of severe weather) divided by the total number of forecasts in all forecast bins. The horizontal sample climatology line is also referred to as the no-resolution line, since points along this line have no resolution. Attributes diagrams also contain a no-skill line, located halfway between the perfect reliability and no-resolution lines. Points along the no-skill line do not contribute to the Brier skill score for a reference forecast of climatology. Meanwhile, points falling between the vertical sample climatology line and the no-skill line contribute positively to the Brier skill score, since these points are closer to the perfect reliability line than they are to the no-resolution line (Wilks 1995).

Performance diagrams plot POD against success ratio (SR), which is defined as

$$\text{SR} = 1 - \frac{\text{false alarms}}{\text{hits} + \text{false alarms}}. \tag{5}$$

In addition, performance diagrams give information about a forecast's bias, defined as

$$\text{bias} = \frac{\text{hits} + \text{false alarms}}{\text{hits} + \text{misses}}, \tag{6}$$

and critical success index (CSI), defined as

$$\text{CSI} = \frac{\text{hits}}{\text{hits} + \text{misses} + \text{false alarms}}, \tag{7}$$

since POD and SR both depend on bias and CSI (Roebber 2009). Herein, POD, SR, bias, and CSI are computed at the same 21 probability levels mentioned previously to produce the performance diagrams. Indeed, each of the 21 probability levels from a given forecast is explicitly plotted onto a performance diagram; therefore, the diagrams are useful for users who wish to determine the probability level that yields a certain value of POD, SR, bias, or CSI for a given forecast. Performance diagrams are also useful for comparing multiple forecasts. Since POD, SR, bias, and CSI are all optimized at 1.0, points that lie closer to the top-right-hand corner of a performance diagram represent more skillful forecasts (Roebber 2009).

A resampling technique outlined by Hamill (1999) is used to test for significant differences in aggregate AUC between the three forecast sets over the entire dataset. A resampling significance test is chosen because the AUC depends on contingency table elements, and small changes in contingency table elements may produce large changes in the AUC (Hamill 1999). Therefore, more common significance tests, such as the paired *t* test, may be inappropriate to use in this case. Conceptually, the resampling technique builds a null distribution of the differences in the aggregate AUC between two forecast sets (e.g., the 1-km deterministic forecast set and the 4-km deterministic forecast set) by repeated random sampling of the contingency table elements of those forecast sets (Hamill 1999). The actual difference in aggregate AUC (computed by subtracting the aggregate AUC of the second forecast from the aggregate AUC of the first forecast) is compared to the null distribution to determine whether or not the difference in aggregate AUC between the two sets of forecasts is significant.

To build a null distribution of aggregate AUC differences between two forecast sets, two separate lists of contingency table elements are created. To start, contingency table elements are obtained from each of the two forecast sets for each of the 63 days in the dataset. The elements of each of the 63 days from the first forecast set (forecast set 1) are assigned to list 1, while the elements of each of the 63 days from the second forecast set (forecast set 2) are assigned to list 2. Next, for each of the 63 days in the dataset, it is randomly determined whether the two lists will exchange contingency table elements for a given day. After the procedure is completed for all 63 days, the aggregate AUC is computed for lists 1 and 2, respectively. Finally, the difference between the aggregate AUC from list 1 and the aggregate AUC from list 2 is computed. This entire procedure is repeated 1000 times in order to form a null distribution of aggregate AUC differences. Finally, the actual aggregate AUC difference is compared with the null distribution to determine significance. If the actual aggregate AUC difference exceeds the 97.5th percentile or falls beneath the 2.5nd percentile of the null distribution, the AUC difference between the two forecast sets is deemed to be significant at the 95% level.

3. Results

a. Comparing 1- and 4-km deterministic forecasts

ROC curves for the 1- and 4-km deterministic forecasts at each UH threshold (i.e., UH = 25, 50, 75, 100, and 125 $m^2 s^{-2}$ for the 4-km forecasts and the corresponding 1-km values for the 1-km forecasts) suggest

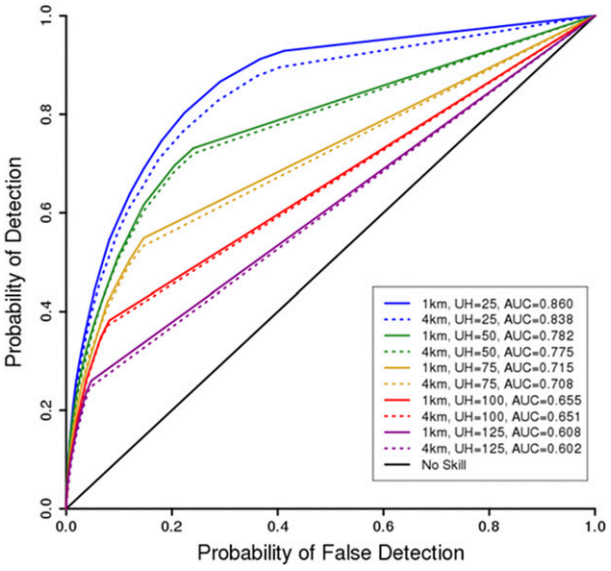

FIG. 2. ROC curves for the 1- (solid) and 4-km (dashed) deterministic models for UH threshold values corresponding to 25 (blue), 50 (green), 75 (goldenrod), 100 (red), and 125 $m^2 s^{-2}$ (violet) on the 4-km grid. Here, as in the text, the 4-km UH threshold values are used to additionally refer to the corresponding 1-km UH values to simplify the analysis. The solid black line indicates the no-skill line.

that the lower UH threshold forecasts have greater forecast skill than the higher threshold forecasts. The 25 $m^2 s^{-2}$ 4-km forecasts and corresponding 249 $m^2 s^{-2}$ (or 281 $m^2 s^{-2}$ for 2011) 1-km forecasts have the greatest AUC values, while the 125 $m^2 s^{-2}$ 4-km forecasts and corresponding 1158 $m^2 s^{-2}$ (or 1098 $m^2 s^{-2}$ for 2011) 1-km forecasts have the lowest AUC values (Fig. 2). Hereafter, to simplify the analysis, the 4-km UH threshold values are used to additionally refer to the corresponding 1-km UH thresholds in the text and figures (refer to Table 3 for the equivalence). For a given UH threshold, the 1-km deterministic forecasts have greater AUC values than the 4-km deterministic forecasts; this pattern holds for all five thresholds. The greatest difference between the 1- and 4-km deterministic AUCs occurs for the UH threshold of 25 $m^2 s^{-2}$. However, for all five UH thresholds examined, the AUC differences are *not* significant at the 95% level (Table 4).

Varying the UH threshold directly influences the reliability of the forecasts. The 25 $m^2 s^{-2}$ forecasts represent overforecasts at nearly all forecast probabilities; moreover, both the 1- and 4-km 25 $m^2 s^{-2}$ forecasts fall slightly below the no-skill line for many of the probabilities, indicating that the forecasts contribute negatively to the Brier skill score at these probabilities. The 50 $m^2 s^{-2}$ forecasts, by contrast, fall

TABLE 4. Results from the two-sided resampling hypothesis test between the 1- and 4-km deterministic forecasts. None of the 1-km AUC − 4-km AUC differences fall outside of the range given in the final column, indicating that none of the differences are significant at the 95% level.

UH threshold ($m^2 s^{-2}$)	1-km AUC	4-km AUC	1-km AUC − 4-km AUC difference	1-km AUC − 4-km AUC 2.5nd percentile (97.5th percentile)
25	0.860	0.838	0.022	−0.0326 (0.0342)
50	0.782	0.775	0.007	−0.0487 (0.0516)
75	0.715	0.708	0.007	−0.0630 (0.0552)
100	0.655	0.651	0.004	−0.0585(0.0559)
125	0.608	0.602	0.006	−0.0537 (0.0494)

near the line of perfect reliability, although these forecasts slightly underforecast at low probabilities and slightly overforecast at higher probabilities. The $75\,m^2\,s^{-2}$ forecasts slightly underforecast at most probabilities, while the 100 and $125\,m^2\,s^{-2}$ forecasts display a greater degree of underforecasting.

For a given UH threshold, the reliability of the 1-km deterministic forecast is generally quite similar to the reliability of the 4-km deterministic forecast (Fig. 3a). However, the 1- and 4-km reliabilities diverge slightly for the two higher UH threshold forecasts (i.e., UH = 100 and $125\,m^2\,s^{-2}$) at higher forecast probability bins. These bins contain relatively few forecasts and therefore must be interpreted cautiously, since a single data point can exert undue influence on the reliability values (Fig. 3b).

In the performance diagrams, the lower threshold forecasts fall closer to the top-right corner of the diagram than the higher threshold forecasts, indicating that—for a given probability—the lower threshold forecasts have greater values of CSI relative to the higher threshold forecasts (Fig. 4). For a given UH threshold, the 1-km deterministic forecast demonstrates slightly greater skill than the 4-km deterministic forecast, as evidenced by the 1-km forecast's higher POD, SR, and CSI compared with the corresponding 4-km forecast. Nonetheless, the differences are generally slight, consistent with the lack of significance in AUC between the 1- and 4-km forecasts. The performance diagrams also show that, for a given model forecast, bias and CSI values are optimized at lower probability levels as the UH threshold is increased.

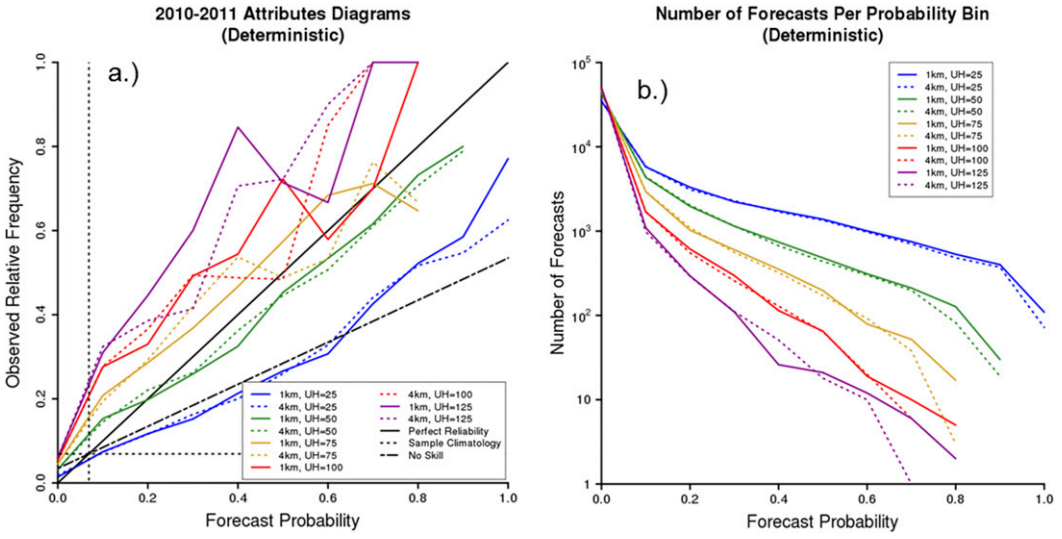

FIG. 3. (a) Attributes diagrams for the 1- (solid) and 4-km (dashed) deterministic models for UH threshold values corresponding to 25 (blue), 50 (green), 75 (goldenrod), 100 (red), and $125\,m^2\,s^{-2}$ (violet) on the 4-km grid. Here, as in the text, the 4-km UH threshold values are used to additionally refer to the corresponding 1-km UH values to simplify the analysis. The solid black line indicates the line of perfect reliability, the long dashed line indicates the no-skill line, and the short dashed lines represent a sample climatological frequency (abbreviated as sample climatology). (b) Number of forecasts per forecast probability bin for the 1- (solid) and 4-km (dashed) deterministic models. The colors represent the same UH thresholds as in (a). Note the logarithmic *y* axis.

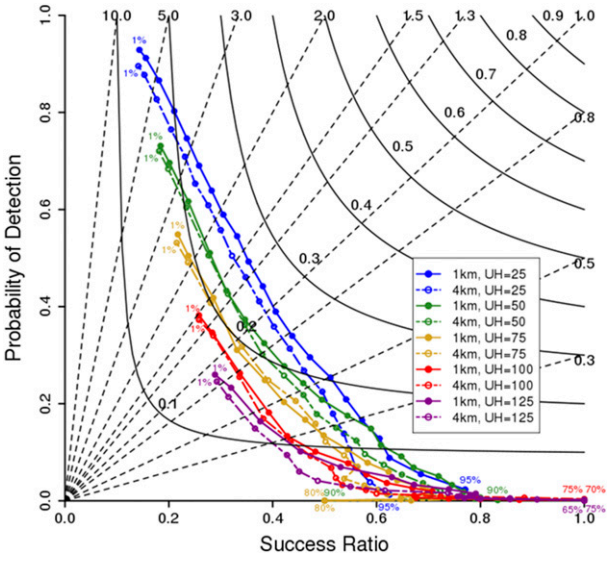

FIG. 4. Performance diagrams for 1- (solid lines with filled points) and 4-km (dashed lines with open points) deterministic models for UH threshold values corresponding to 25 (blue), 50 (green), 75 (goldenrod), 100 (red), and 125 $m^2 s^{-2}$ (violet) on the 4-km grid. Here, as in the text, the 4-km UH threshold values are used to additionally refer to the corresponding 1-km UH values. For the 1- and 4-km UH = 25 $m^2 s^{-2}$ forecasts, the following 21 probability levels are plotted: 1%, 2%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, and 95%. A subset of these probability levels is plotted for the remaining eight forecasts, since these forecasts never produce 95% severe probabilities. The first and last probability levels are labeled for each forecast. Solid black lines indicate lines of constant CSI, while dashed black lines represent lines of constant bias.

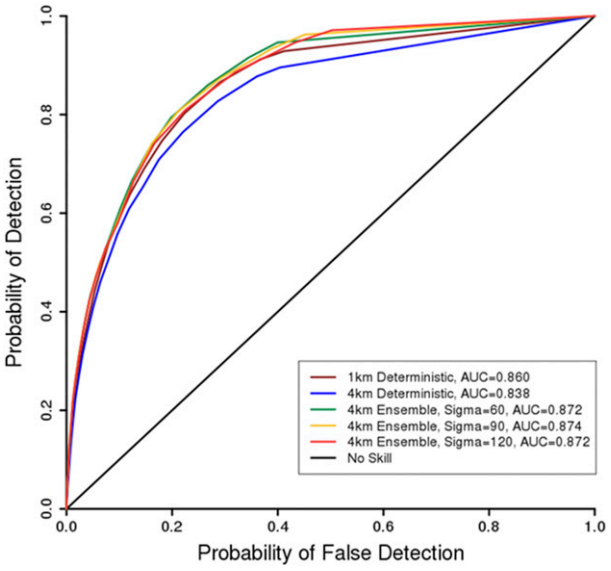

FIG. 5. ROC curves for 1-km deterministic (dark red), 4-km deterministic (blue), σ = 60 km 4-km ensemble (green), σ = 90 km 4-km ensemble (goldenrod), and σ = 120 km 4-km ensemble (red) forecasts. All forecasts use the UH threshold value corresponding to 25 $m^2 s^{-2}$ on the 4-km grid. The solid black line indicates the no-skill line.

b. *Comparing the deterministic and 4-km ensemble forecasts*

The forecast quality of the 1- and 4-km deterministic models is compared with that of an 11-member 4-km ensemble to determine the effect of adding ensemble members on forecast skill. Because of the finding in the previous subsection that the greatest differences in AUC between the forecasts came from the 25 $m^2 s^{-2}$ UH threshold forecasts, only the 25 $m^2 s^{-2}$ forecasts are analyzed for the 4-km deterministic and ensemble comparison.

All deterministic and ensemble forecasts with a UH threshold of 25 $m^2 s^{-2}$ have similar ROC curves (Fig. 5). The AUC for each curve exceeds 0.80, indicating that all forecasts show considerable skill. The 4-km deterministic forecast has the lowest AUC (0.838), while the 4-km ensemble forecast with σ = 90 km has the greatest AUC (0.874). At the 95% level, a (weakly) significant difference in AUC exists between the 4-km deterministic forecast and the ensemble forecasts with σ = 90 and 120 km (Table 5). No statistically significant differences are found between the 1-km deterministic forecast and any of the ensemble forecasts.

While the three ensemble forecasts (i.e., σ = 60, 90, and 120 km) have similar ROC curves and AUC values, the ensemble forecasts do have notably different reliability. The ensemble forecast with σ = 120 km has the best reliability, as its curve in the attributes diagram lies closest to the perfect reliability line (Fig. 6a). This result makes sense, given that the σ = 120 km ensemble forecast benefits from a high degree of spatial smoothing as well as ensemble smoothing. As a result, the σ = 120 km ensemble forecast has fewer high (i.e., ≥0.60) probabilities compared with the other ensemble and deterministic forecasts (Fig. 6b), which helps to reduce the overforecasting bias. However, even the σ = 120 km ensemble forecast has a tendency to overforecast at nearly all probabilities. All three ensemble forecasts' curves on the attributes diagram mostly reside above the no-skill line, representing at least a slight improvement over either of the deterministic forecasts.

The performance diagrams indicate that the three ensemble forecasts have similar skill levels, as the ensemble forecast points are clustered very close to each other and are located about the same distance from the top-right-hand corner of the plot (Fig. 7). The 1-km deterministic forecast has less skill than any of the ensemble forecasts but greater skill than the 4-km deterministic forecast. These results corroborate the

TABLE 5. Results from the two-sided resampling hypothesis test between the 4-km ensemble and the 4-km deterministic forecasts for a UH threshold of 25 $m^2 s^{-2}$.

Model/ensemble	AUC	4-km deterministic AUC − model/ensemble AUC difference	4-km deterministic AUC − model/ensemble AUC 2.5nd percentile (97.5th percentile)
4-km ensemble, σ = 60 km	0.872	−0.0340	−0.0363 (0.0345)
4-km ensemble, σ = 90 km	0.874	−0.0360[a]	−0.0359 (0.0340)
4-km ensemble, σ = 120 km	0.872	−0.0340[a]	−0.0335 (0.0357)
4-km deterministic	0.838	—	—

[a] Significant at the 95% level.

implications of the ROC curves and the AUC analysis; namely, that over the entire 63-day dataset, the ensemble forecasts have greater skill than the 1-km deterministic forecast, which in turn has greater skill than the 4-km deterministic forecast. For the five sets of forecasts examined, the performance diagrams indicate that a probability level of 40%–50% optimizes bias, while a probability level of 35%–40% optimizes CSI.

c. *Comparing the AUC for individual days*

For 61 of the 63 days in the dataset, an individual-day AUC is computed from the 1-km deterministic, 4-km deterministic, and the 4-km ensemble forecasts (σ = 90 km) using the 25 $m^2 s^{-2}$ UH threshold. No AUC is computed for 29 April or 4 May 2011 because no OSRs occurred inside of the analysis domain on those days, resulting in an indeterminate POD [i.e., hits/(hits + misses) = 0/0]. For each of the 61 days analyzed, the 4-km deterministic AUC is subtracted from the 1-km deterministic AUC and the 4-km ensemble AUC, respectively, to obtain two distributions, which describe how the 1-km deterministic and 4-km ensemble forecasts perform relative to the 4-km deterministic forecasts. When the daily 4-km deterministic AUC is subtracted from the corresponding daily 1-km deterministic AUC, the distribution peaks just to the right of the zero line, indicating that, for most days, the 1-km deterministic model gives a slightly better forecast (in terms of AUC) than the 4-km deterministic model (Fig. 8a). The distribution has a small left tail, suggesting that the 4-km deterministic model performs notably better than the 1-km deterministic model for only a handful of days; the vast majority of the data points are located to the right of the zero line. When the daily 4-km deterministic AUC is subtracted from the corresponding daily 4-km ensemble AUC, the distribution peaks to

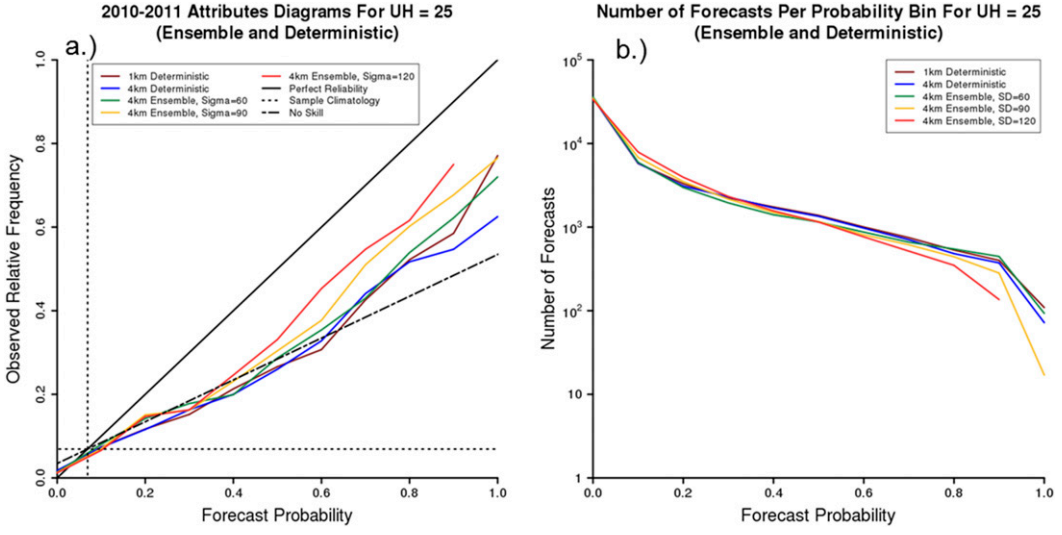

FIG. 6. (a) Attributes diagrams for 1-km deterministic (dark red), 4-km deterministic (blue), σ = 60 km 4-km ensemble (green), σ = 90 km 4-km ensemble (goldenrod), and σ = 120 km 4-km ensemble (red) forecasts. All forecasts use the UH threshold value corresponding to 25 $m^2 s^{-2}$ on the 4-km grid. The solid black line indicates the line of perfect reliability, the long dashed line indicates the no-skill line, and the short dashed lines represent sample climatological frequency (abbreviated as sample climatology). (b) As in (a), but for the number of forecasts per forecast probability bin. Note the logarithmic *y* axis.

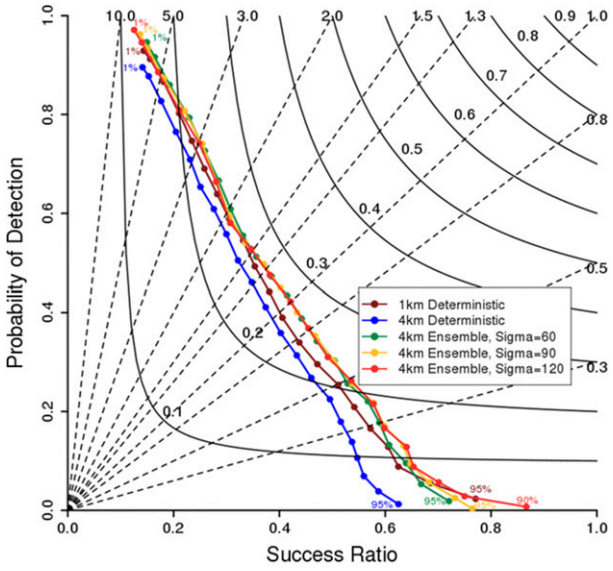

FIG. 7. Performance diagrams for 1-km deterministic (dark red), 4-km deterministic (blue), σ = 60 km 4-km ensemble (green), σ = 90 km 4-km ensemble (goldenrod), and σ = 120 km 4-km ensemble (red) forecasts. All forecasts use the UH threshold value corresponding to 25 $m^2 s^{-2}$ on the 4-km grid. Except for the σ = 120 km 4-km ensemble forecasts, which produced no 95% or greater probabilities, each of the five forecasts have the following 21 probability levels plotted: 1%, 2%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, and 95%. The first and last probability levels are labeled for each forecast. Solid black lines indicate lines of constant CSI, while dashed black lines represent lines of constant bias.

the right of the zero line and has a long right tail, indicating that the ensemble performs better—and sometimes substantially better— than the 4-km deterministic forecast on the vast majority of the days (Fig. 8b). Interestingly, when the daily 1-km deterministic AUC is subtracted from the corresponding daily 4-km ensemble AUC, the distribution looks similar to that in Fig. 8b: it peaks to the right of the zero line and has a right tail (Fig. 8c), suggesting that the ensemble performs objectively better than the 1-km deterministic forecast as well as the 4-km deterministic forecast on the majority of days in the analysis period.

Five days that span the distributions given in Figs. 8a and 8b are chosen for individual analysis. These days are 11 May 2010, 15 June 2010, 7 June 2011, 18 May 2011, and 27 April 2011. The analysis of these individual days offers insight into what (if any) additional forecast quality and/or value can be gained by either reducing the horizontal grid spacing from 4 to 1 km or by adding members to form 4-km convection-allowing ensembles on a given day. The five case study examples are presented below.

1) 11 MAY 2010

On 11 May 2010—the day for which the 4-km ensemble forecast (AUC = 0.805) performed best relative to the 4-km deterministic forecast (AUC 0.564) in terms of AUC—the threat of severe weather existed across multiple regions. A weakening surface cyclone, located over southern Iowa at 1200 UTC, tracked east-northeastward and brought storms to the Ohio valley around 0000 UTC 12 May 2010. Meanwhile, low-level southerly flow from the Gulf of Mexico helped to destabilize a broad region of the central plains. Despite weak large-scale forcing, several severe-hail-producing storms formed in western Oklahoma before 0200 UTC 12 May 2010. A cluster of severe storms also formed in eastern Missouri and western Kansas around 0300 UTC along a warm front. Additionally, several storms formed in northeastern Colorado ahead of an eastward-moving upper-level low during the evening hours; however, no observed severe weather was associated with these storms.

Interestingly, the three model configurations produced drastically different forecasts on this day: the 1-km deterministic forecast (AUC = 0.561) highlighted regions near Missouri, Oklahoma, and Colorado for severe weather; the 4-km deterministic forecast highlighted Colorado but focused its secondary threat area on Ohio and surrounding states; while the 4-km ensemble showed lower severe probabilities in Colorado and gave nonzero severe probabilities over a region extending from western Kansas to eastern Ohio (Figs. 9a–c). OSRs were located in southwestern Kansas, northwestern Oklahoma, central Kansas, central Ohio, and western Pennsylvania, while no OSRs occurred in Colorado. Of the three forecasts, the 4-km ensemble forecast produced the greatest AUC on this day, as it had the lowest probabilities in northeastern Colorado (which reduced its POFD) and had nonzero severe probabilities over the three main regions where reports did occur (which increased its POD). While it is important to realize that probabilistic forecasts should be evaluated over multiple cases, this one case does suggest that ensembles can offer enhanced forecast quality not only by identifying regions of potential severe weather missed by a deterministic model, but also by reducing the magnitude of overdone deterministic severe probabilities.

2) 15 JUNE 2010

On 15 June 2010—the day for which the 1-km deterministic forecast (AUC = 0.763) performed objectively best relative to the 4-km deterministic forecast (AUC = 0.669)—a midlevel trough propagated

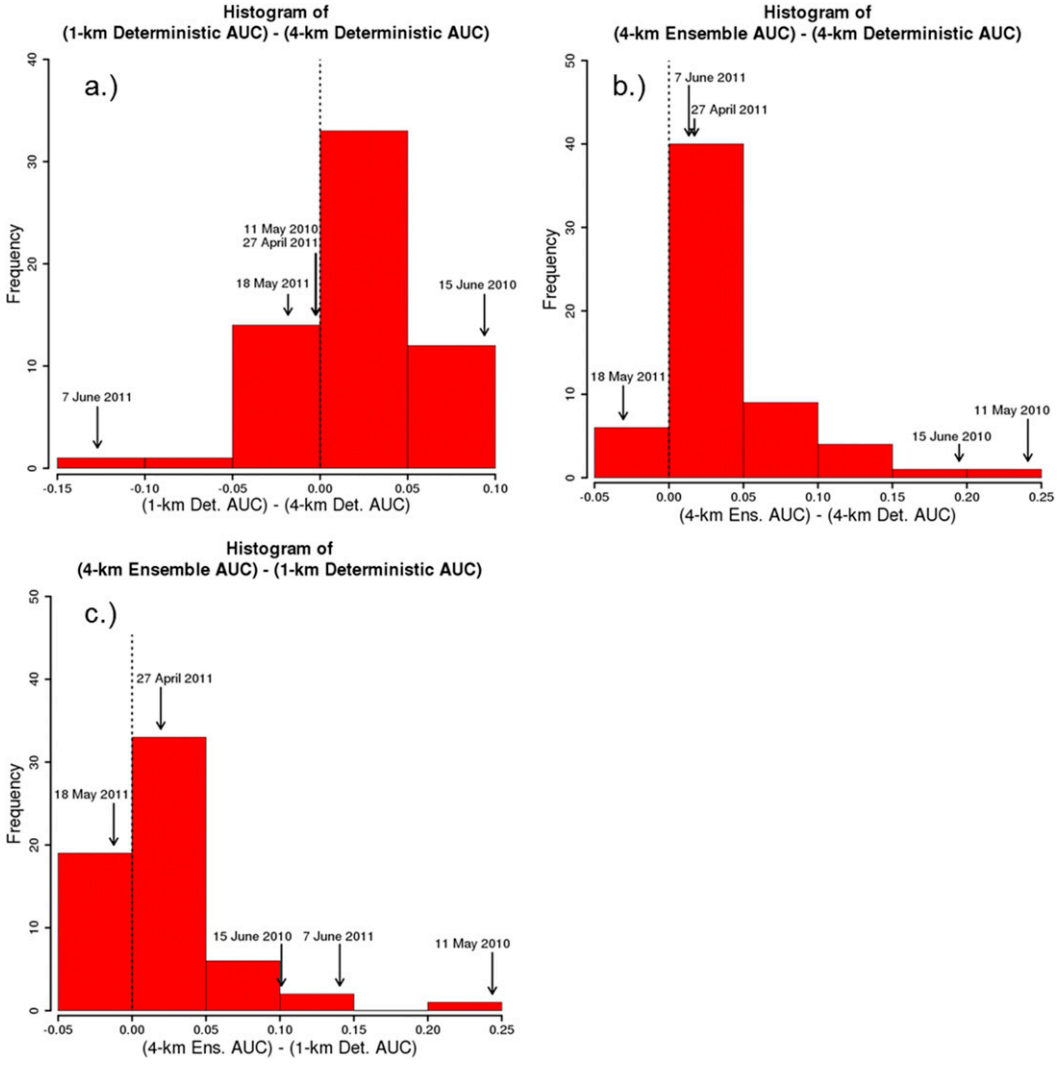

FIG. 8. (a) Histogram showing the distribution of 1-km deterministic AUC − 4-km deterministic AUC for individual days. Positive (negative) values along the x axis indicate days on which the 1-km deterministic forecast had a greater (lower) AUC than the 4-km deterministic forecast. (b) Histogram showing the distribution of 4-km ensemble (σ = 90 km) AUC − 4-km deterministic AUC for individual days. Positive (negative) values along the x axis indicate days on which the 4-km ensemble forecast had a greater (lower) AUC than the 4-km deterministic forecast. (c) Histogram showing the distribution of 4-km ensemble (σ = 90 km) AUC − 1-km deterministic AUC for individual days. Positive (negative) values along the x axis indicate days on which the 4-km ensemble forecast had a greater (lower) AUC than the 1-km deterministic forecast. In each panel, the labeled arrows indicate where the five case study days fall in the distribution, and all forecasts use the 25 $m^2 s^{-2}$ UH threshold.

northeastward into central Illinois, where a warm, moist air mass coincided with an environment containing 25–40 kt (where 1 kt = 0.51 $m s^{-1}$) of 1000–500-hPa wind shear. The short-wave trough initiated convection in eastern Missouri around 1730 UTC, and this convection subsequently moved northeastward, producing a multitude of severe wind and hail reports in Illinois, Indiana, and western Ohio. Meanwhile, broad southerly flow resulted in moist and unstable conditions throughout much of the southeastern United States. Although large-scale forcing for ascent and vertical wind shear were both weak in this region, numerous pulse storms formed, resulting in many severe wind reports.

Two main differences existed between the three forecasts on this day: the distribution (and magnitude) of the higher-end severe probabilities in central Illinois and the spatial coverage of the nonzero severe probabilities in the southeastern United States (Figs. 9d–f). Relative to the 4-km deterministic forecast, the 1-km deterministic forecast had greater severe probabilities in Illinois and gave a more continuous threat area in central Illinois. Additionally, the 1-km deterministic

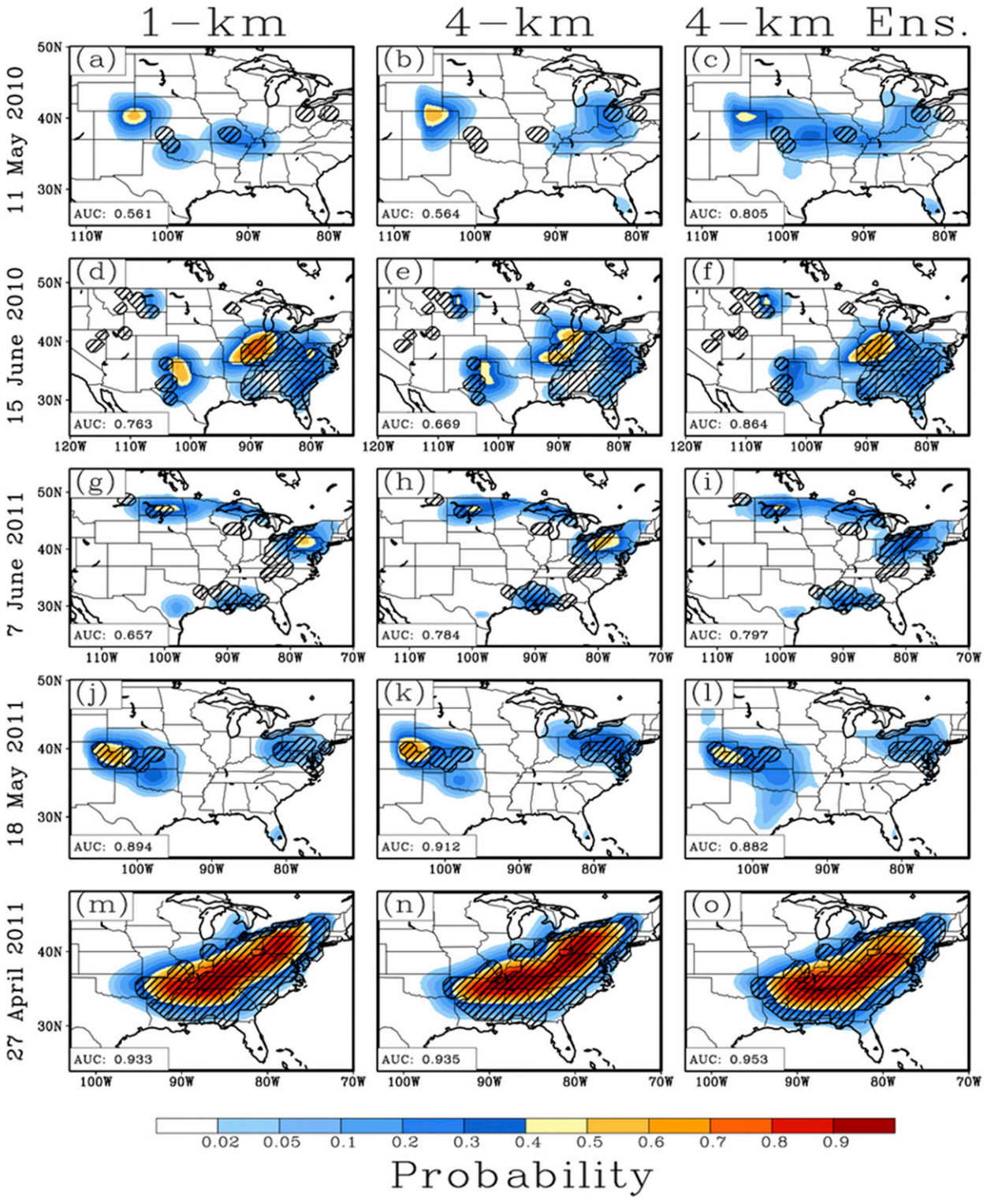

FIG. 9. Probabilistic severe weather forecasts (shaded) for the (a) 1-km deterministic, (b) 4-km deterministic, and (c) 4-km ensemble (σ = 90 km) forecasts for 11 May 2010. Black hatching denotes 80-km grid boxes that contain at least one observed storm report. All forecasts use the UH threshold value corresponding to 25 $m^2 s^{-2}$ on the 4-km grid. As in (a)–(c), results are shown for (d)–(f) 15 Jun 2010, (g)–(i) 7 Jun 2011, (j)–(l) 18 May 2011, and (m)–(o) 27 Apr 2011.

forecast introduced more nonzero severe probabilities into portions of the southeastern United States relative to the 4-km deterministic forecast. The 4-km ensemble forecast (AUC = 0.864), meanwhile, maintained >0.50 severe probabilities over central Illinois but completely filled in the southeastern United States with nonzero severe probabilities, perhaps as a result of ensemble smoothing. The ensemble forecast was therefore rewarded with a greater POD and AUC relative to either deterministic forecast (although the ensemble's lower-magnitude severe probabilities in the Texas Panhandle, where no OSRs were located, may have also helped to elevate the ensemble's AUC over that of the two deterministic models). This case suggests that the 4-km ensemble forecast can potentially offer improvements in forecast quality relative to the 1-km deterministic forecast even on days when the 1-km deterministic forecast performs well relative to the 4-km deterministic forecast.

3) 7 JUNE 2011

On 7 June 2011 the 4-km deterministic forecast (AUC = 0.784) performed best relative to the 1-km forecast (AUC = 0.657). In the northern plains, a 500-hPa short-wave trough and the left-exit region of a 300-hPa jet tracked northeastward through the Dakotas, producing severe weather in south-central North Dakota during the early evening. During the overnight hours, storms fired in southern Wisconsin ahead of a cold front associated with a surface low tracking northeastward through northern Minnesota. Farther eastward, strong θ_e advection and an unstable environment helped sustain a preexisting mesoscale convective system as it tracked southward through Ohio, West Virginia, Virginia, and North Carolina. Along the Gulf Coast, numerous pulse storms formed in an environment of abundant moisture and instability; these storms produced a number of severe wind and hail reports.

Two main differences existed between the forecasts on this day: the magnitude and orientation of the higher-end severe probabilities near Pennsylvania and the spatial extent of the severe probabilities in southern Texas (Figs. 9g–i).

Relative to the 4-km deterministic and 4-km ensemble (AUC = 0.797) forecasts, the 1-km deterministic forecast focused its higher-end severe probabilities in Pennsylvania farther east, where fewer storm reports occurred. The 4-km ensemble had lower-magnitude severe probabilities in Pennsylvania relative to the two deterministic models, but its orientation of 0.3 severe probabilities more closely matched the orientation of the observations. In southern Texas, where no storm reports were observed, the 1-km deterministic forecast produced a much larger area of nonzero severe probabilities relative to the 4-km deterministic and 4-km ensemble forecasts. The three forecasts generally had similar forecast probabilities over the southeastern United States and the upper Midwest.

Given that the three forecasts all highlighted similar regions for severe weather on this day, the forecasts likely all had similar value. It is notable, however, that the 4-km ensemble forecast had a slightly greater AUC than the 4-km deterministic forecast on the day when the 4-km deterministic forecast performed best (in terms of AUC) relative to the 1-km deterministic forecast.

4) 18 MAY 2011

On 18 May 2011 the 4-km ensemble forecast (AUC = 0.882) performed objectively *worst* relative to the 4-km deterministic forecast (AUC = 0.912). Two upper-level troughs dominated the flow pattern on this day: one in the western United States and one in the eastern United States. During the late afternoon, storms formed in eastern Colorado, which was located downstream of an upper trough and in the left-exit region of a 300-hPa jet. Later in the evening, storms also fired in west-central Kansas as a weak local vorticity maximum pivoted northeastward through the state and strong southerly winds helped increase low-level moisture. Perhaps because of a lack of large-scale forcing for ascent, no storms ultimately formed along the dryline in the Texas Panhandle region. Downstream of the trough over the eastern United States, storms formed in the early afternoon, producing severe weather in Pennsylvania, West Virginia, Maryland, and Virginia.

The biggest difference between the three forecasts on this day was the ensemble forecast's large region of ≥0.02 probabilities in Texas, Missouri, and Arkansas (Figs. 9j–l), which perhaps resulted from ensemble smoothing. Since OSRs never occurred in these states, the ensemble forecast had a large POFD relative to the deterministic forecasts. The ensemble forecast also produced slightly lower magnitudes of severe probability in the northeastern United States, where OSRs were located, which decreased its POD relative to either of the deterministic forecasts.

Given the ensemble forecast's inferior POFD and POD, this case shows that the quality of an ensemble forecast does not always exceed that of a deterministic forecast. Nevertheless, objectively, the ensemble's AUC is only about 0.030 lower than the 4-km deterministic forecast's AUC on the "worst day" for the ensemble relative to the 4-km deterministic forecast. This difference is small compared with the 0.241 difference in AUC between the 4-km ensemble and 4-km deterministic forecasts on 11 May 2010.

5) 27 APRIL 2011

We chose 27 April 2011 for individual analysis because it represents a ''high impact'' day. In fact, this was one of the longest and deadliest tornado outbreaks in U.S. history. According to the SPC, some 937 severe reports—including 292 tornado reports—occurred over the contiguous United States. Storms formed downstream of an upper trough and ahead of a cold front in the southeastern United States, where strong low-level southerly flow coincided with abundant vertical wind shear.

All three forecasts demonstrated considerable skill, as all three forecasts had AUC values greater than 0.93. The 1- and 4-km deterministic forecasts were nearly identical and matched the observations quite well (Figs. 9m–o). The ensemble forecast differed slightly from the deterministic forecasts, most notably by extending the region of ≥0.40 severe probabilities farther southward into south-central Mississippi, Alabama, and Georgia. The ensemble's southward shift of the higher severe probabilities likely contributed to its higher AUC relative to the two deterministic forecasts' AUC.

Because the shapes of the three forecasts' higher-end (i.e., ≥0.40) severe probabilities were very similar, the forecasts had similar quality—and likely similar value—on this day. This case is important because it illustrates that all three model configurations—including the ensemble—can produce high-sharpness forecasts.

4. Summary and discussion

Maximum hourly 2–5-km updraft helicity (UH) forecasts from a 4-km grid spacing model, an equivalently configured 1-km grid spacing model, and an 11-member 4-km grid spacing ensemble are remapped to an 80-km grid and used to produce next-day probabilistic severe weather forecasts for 63 days of the 2010 and 2011 NOAA HWT SFEs. As in Sobash et al. (2011), extreme values of UH are treated as surrogate severe weather reports (SSRs). SSRs are smoothed spatially using a two-dimensional isotropic Gaussian smoother to create probabilistic severe weather forecasts.

After testing a variety of 4-km UH values (i.e., UH = 25, 50, 75, 100, and 125 $m^2 s^{-2}$) and their corresponding 1-km UH values as thresholds for SSRs, it is found that the 25 $m^2 s^{-2}$ threshold not only gives the largest AUC for each of the three forecast configurations but also produces the greatest difference in AUC between the three forecast configurations. Meanwhile, the 50 $m^2 s^{-2}$ threshold yields the most reliable forecasts, with thresholds greater than (less than) 50 $m^2 s^{-2}$ producing underforecasts (overforecasts). These results are consistent with Sobash et al. (2011), who found that a 4-km UH threshold of approximately 34 $m^2 s^{-2}$ (the smallest tested) gave the largest AUC values while a threshold near 41 $m^2 s^{-2}$ produced the most reliable forecasts.

Ensemble surrogate severe weather probabilistic forecasts (SSPFs) are created by calculating, at each grid point, the fraction of ensemble members exceeding the specified UH threshold (always 25 $m^2 s^{-2}$ for the ensemble forecasts) and then smoothing these values spatially using a Gaussian smoother. Three values of σ are tested for the ensemble SSPFs (σ = 60, 90, and 120 km). The σ = 90 km forecast produces the best AUC, although varying the spatial smoothing parameter has only a slight impact on forecasts' AUC values for the three values analyzed. Varying the spatial smoothing parameter has a larger impact on the forecasts' reliability values: the σ = 120 km forecast gives the best reliabilities but still overforecasts. These results agree with those of Sobash et al. (2011, 2016b), who found that increasing the spatial smoothing parameter had little effect on a forecast's AUC but resulted in a general progression from overforecasting, to near-perfect reliability, to underforecasting. Such a finding suggests that a value of σ could be found to optimize the ensemble forecast's reliability. However, increasing the spatial smoothing parameter beyond 120 km is not attempted or analyzed here since the ensemble is presumed to account for some of the spatial uncertainty in the SSRs (e.g., Sobash et al. 2016b), eliminating the need to test spatial smoothing parameter values beyond those tested for the deterministic forecasts (i.e., σ = 120 km).

A two-sided resampling hypothesis test is conducted to test for significance between the 4- and 1-km deterministic forecasts and between the 4-km deterministic and 4-km ensemble forecasts, using a UH threshold of 25 $m^2 s^{-2}$. Results suggest that while no significant difference in AUC exists between the 4- and 1-km deterministic forecasts, a weakly significant difference in AUC exists between the 4-km deterministic and 4-km forecasts (with the 4-km ensemble forecasts having the greater AUC). No significant difference is found between any of the three model configurations at the four higher UH thresholds examined (i.e., UH = 50, 75, 100, and 125 $m^2 s^{-2}$), since, as the UH threshold is increased beyond 25 $m^2 s^{-2}$, the three sets of forecasts produce increasingly similar AUC values.

The lack of any significant difference between the 4- and 1-km deterministic forecasts agrees with Kain et al. (2008), who, qualitatively, observed a lack of dramatic differences in the forecast UH fields between the 4- and 2-km horizontal grid-spacing models in the 2005 NOAA HWT SFE. The results of the present study also agree with Schwartz et al. (2009), who found that the

models with 4- and 2-km grid spacing showed similar skill in forecasting heavy rainfall and convective evolution, but who noted that the 2-km forecasts generally contained more realistic-looking convective features than the 4-km forecasts.

The nonsignificant difference between the 4- and 1-km deterministic forecasts appears to contradict the findings of Roberts and Lean (2008). However, as suggested in Schwartz et al. (2009), Roberts and Lean's (2008) use of a modified convective parameterization scheme with their 4-km grid spacing model and their focus on time periods of 7 h or less may help explain the differences in findings between that study and this work. At greater than 7-h time scales, for example, it is possible that large-scale errors may become more important, thus rendering the 4- and 1-km deterministic forecasts similar. Therefore, it is possible that a significant difference in AUC between the 4- and 1-km deterministic forecasts would exist at shorter forecast lead times while no significant difference in AUC exists for next-day forecasts.

The present study's finding of a weakly significant difference between the 4-km ensemble and 4-km deterministic forecasts generally agrees with the results of Sobash et al. (2016b), who found that SSPFs produced from a 30-member ensemble had significantly greater fractions skill scores (FSSs) compared with the SSPFs produced from either of two deterministic forecasts for smaller (i.e., mesoscale) spatial scales. While the present study does not analyze FSS at a variety of spatial scales, subjective inspection of the present study's individual case studies suggests that—despite some noteworthy exceptions—many of the differences between the 4-km deterministic, 1-km deterministic, and 4-km ensemble forecasts occur on the mesoscale. One potential reason for the ensemble forecasts' superior performance relative to the 4-km deterministic forecasts could be the two types of smoothing used to create the ensemble forecasts (i.e., spatial and ensemble smoothing) compared to just the spatial smoothing used to create the deterministic forecasts. Indeed, when no spatial smoothing is applied to the ensemble probabilities, the ensemble AUC is reduced to, approximately, 0.832 over the entire dataset (not shown), compared with an AUC of 0.874 produced by the $\sigma = 90$ km ensemble. Sensitivity tests (not shown) reveal that the AUC is maximized for the $\sigma = 90$ km ensemble forecasts.

To demonstrate the range of day-to-day variation in skill between the three sets of forecasts, five days were selected for individual case study analysis: 11 May 2010, 15 June 2010, 7 June 2011, 18 May 2011, and 27 April 2011. In general, it is found that the 4- and 1-km deterministic forecasts exhibit similar levels of quality (as measured by individual-day AUC) to each other, while the 4-km ensemble forecasts routinely provide enhanced quality relative to either deterministic forecast. Notably, even on days when the ensemble forecast is inferior, the quality of the ensemble forecast tends to remain close to that of the deterministic forecasts. Interestingly, it is found that neither the number of daily SPC storm reports nor the SPC 1200 UTC day-1 convective outlook categories serve as good predictors for determining whether the 1-km deterministic or 4-km ensemble forecast will outperform the 4-km deterministic forecast on a given day (not shown).

Herein, only a single ensemble configuration is used to create the ensemble forecasts. While an ensemble with more members could have been used, previous research has found that increasing the number of members in an ensemble provides diminishing returns to the ensemble's skill (e.g., Clark et al. 2011; Sobash et al. 2016b; Schwartz et al. 2014), suggesting that adding members to the 11-member ensemble may not significantly improve the forecast skill. It is more difficult to predict how the presence of multiple dynamic cores and/or multiple physics parameterizations influences the skill of an ensemble. Previous research has suggested that multicore and multiphysics ensembles generally have enhanced spread and forecast skill relative to single-core and single-physics ensembles, respectively (e.g., Eckel and Mass 2005; Berner et al. 2011). Therefore, it is possible that a multicore, multiphysics ensemble could perform better than the ensemble used herein relative to the two deterministic models, while a single-core, single-physics ensemble could perform worse. However, these results are certainly not guaranteed; increasing an ensemble's spread by introducing members with multiple cores or physics parameterizations does not necessarily improve the ensemble's forecast skill or reliability, especially if the ensemble is not properly calibrated for bias (Eckel and Mass 2005). Moreover, the benefits of a mixed- versus single-physics ensemble may depend on situational factors, such as the amount of large-scale forcing for ascent (Stensrud et al. 2000), or potentially even on the time and spatial scales of the forecast. More work is needed to determine the optimal configuration of CAM ensembles.

5. Future work

Given the abundance of remaining questions, many potential avenues exist for future work. One such avenue, as mentioned above, is to investigate how the configuration of the ensemble (e.g., multi- versus single core and mixed versus single physics) influences the ensemble's skill relative to either deterministic forecast.

Another path is to determine how specific regimes, mesoscale scenarios (e.g., dominant convective mode, convective trigger, etc.), and/or seasons influence the relative skill of the three forecast configurations. Such knowledge would be invaluable, as it could focus forecasters' attention on the forecast model(s) most likely to deliver the greatest quality for a given situation. Future work may additionally wish to examine the three forecast sets' relative skill at varying lead times and/or for other types of forecasts, such as those involving precipitation or specific severe weather hazards. Finally, future work may investigate more complex metrics for diagnosing the probability of next-day severe weather events. For instance, UH may be combined with other parameters to create a more skillful next-day forecast (e.g., Gallo et al. 2016), or a form of UH other than 2–5-km UH may be tested (e.g., Sobash et al. 2016a). It is possible that these other metrics may produce higher quality forecasts, which in turn could alter the relative quality and value of the 1-km deterministic, 4-km deterministic, and 4-km ensemble forecasts.

Acknowledgments. This work was made possible by a Presidential Early Career Award for Scientists and Engineers (PECASE). Additional support was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. In addition to NOAA CSTAR support, CAPS simulations received supplementary support from NSF Grant ATM-0802888. All 2010 and 2011 CAPS forecasts were produced at the National Institute of Computational Sciences (NICS) at the University of Tennessee, and Oklahoma Supercomputing Center for Research and Education resources were used for ensemble postprocessing. The authors would also like to thank the three anonymous reviewers, whose insightful comments and feedback helped dramatically improve the manuscript.

REFERENCES

Benjamin, S., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, doi:10.1175/MWR-D-15-0242.1.

Berner, J., S. Ha, J. P. Hacker, A. Fournier, and C. Snyder, 2011: Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multiphysics representations. *Mon. Wea. Rev.*, **139**, 1972–1995, doi:10.1175/2010MWR3595.1.

Brooks, H. E., C. A. Doswell III, and R. A. Maddox, 1992: On the use of mesoscale and cloud-scale models in operational forecasting. *Wea. Forecasting*, **7**, 120–132, doi:10.1175/1520-0434(1992)007<0120:OTUOMA>2.0.CO;2.

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, doi:10.1175/1520-0434(1999)014<0168:PPOPUT>2.0.CO;2.

Chen, F., and J. Dudhia, 2001: Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: Model description and implementation. *Mon. Wea. Rev.*, **129**, 569–585, doi:10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2.

Chou, M.-D., and M. J. Suarez, 1994: An efficient thermal infrared radiation parameterization for use in general circulation models. NASA Tech. Memo. 104606, Tech. Rep. Series on Global Modeling and Data Assimilation, Vol. 3, 85 pp. [Available online at https://ntrs.nasa.gov/search.jsp?R=19950009331.]

Clark, A. J., W. A. Gallus Jr., M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, doi:10.1175/2009WAF2222222.1.

——, ——, ——, and ——, 2010: Convection-allowing and convection-parameterizing ensemble forecasts of a mesoscale convective vortex and associated severe weather environment. *Wea. Forecasting*, **25**, 1052–1081, doi:10.1175/2010WAF2222390.1.

——, and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, doi:10.1175/2010MWR3624.1.

——, and Coauthors, 2012a: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, doi:10.1175/BAMS-D-11-00040.1.

——, and Coauthors, 2012b: An overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Amer. Meteor. Soc.*, **93** (Suppl.), doi:10.1175/BAMS-D-11-00040.2.

——, J. S. Kain, P. T. Marsh, J. Correia, M. Xue, and F. Kong, 2012c: Forecasting tornado pathlengths using a three-dimensional object identification algorithm applied to convection-allowing forecasts. *Wea. Forecasting*, **27**, 1090–1113, doi:10.1175/WAF-D-11-00147.1.

——, J. Gao, P. Marsh, T. Smith, J. Kain, J. Correia, M. Xue, and F. Kong, 2013: Tornado pathlength forecasts from 2010 to 2011 using ensemble updraft helicity. *Wea. Forecasting*, **28**, 387–407, doi:10.1175/WAF-D-12-00038.1.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) Model. *Atmos. Sci. Lett.*, **5**, 110–117, doi:10.1002/asl.72.

Du, J., J. McQueen, G. DiMego, Z. Toth, D. Jovic, B. Zhou, and H.-Y. Chuang, 2006: New dimension of NCEP Short-Range Ensemble Forecasting (SREF) system: Inclusion of WRF members. *WMO Expert Team Meeting on Ensemble Prediction Systems*, Exeter, United Kingdom, WMO. [Available online at http://www.emc.ncep.noaa.gov/mmb/SREF/WMO06_full.pdf.]

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350, doi:10.1175/WAF843.1.

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, *19th Conf. on Weather Analysis and Forecasting/15th Conf. on Numerical Weather Prediction*, San Antonio, TX, Amer. Meteor. Soc., 10.1. [Available online at http://ams.confex.com/ams/SLS_WAF_NWP/techprogram/paper_47241.htm.]

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting*, **31**, 273–295, doi:10.1175/WAF-D-15-0134.1.

Gao, J., M. Xue, K. Brewster, and K. K. Droegemeier, 2004: A three-dimensional variational data analysis method with recursive filter for Doppler radars. *J. Atmos. Oceanic Technol.*, **21**, 457–469, doi:10.1175/1520-0426(2004)021<0457:ATVDAM>2.0.CO;2.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, doi:10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

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

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285–293, doi:10.1016/0169-2070(86)90048-8.

Janjić, Z. I., 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.]

Johnson, A., X. Wang, F. Kong, and M. Xue, 2013: Object-based evaluation of the impact of horizontal grid spacing on convection-allowing forecasts. *Mon. Wea. Rev.*, **141**, 3413–3425, doi:10.1175/MWR-D-13-00027.1.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, doi:10.1175/WAF2007106.1.

——, S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting unique information from high-resolution forecast models: Monitoring selected fields and phenomena every time step. *Wea. Forecasting*, **25**, 1536–1542, doi:10.1175/2010WAF2222430.1.

Lilly, D. K., 1990: Numerical prediction of thunderstorms—Has its time come? *Quart. J. Roy. Meteor. Soc.*, **116**, 779–798, doi:10.1002/qj.49711649402.

Lim, K.-S. S., and S.-Y. Hong, 2010: Development of an effective double-moment cloud microphysics scheme with prognostic cloud condensation nuclei (CCN) for weather and climate models. *Mon. Wea. Rev.*, **138**, 1587–1612, doi:10.1175/2009MWR2968.1.

Marzban, C., 2004: The ROC curve and the area under it as performance measures. *Wea. Forecasting*, **19**, 1106–1114, doi:10.1175/825.1.

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

——, 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, doi:10.1256/003590002320603584.

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, doi:10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Mellor, G. L., and T. Yamada, 1982: Development of a turbulence closure model for geophysical fluid problems. *Rev. Geophys.*, **20**, 851–875, doi:10.1029/RG020i004p00851.

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**, 16 663–16 682, doi:10.1029/97JD00237.

Morrison, H., J. A. Curry, and V. I. Khvorostyanov, 2005: A new double-moment microphysics parameterization for application in cloud and climate models. Part I: Description. *J. Atmos. Sci.*, **62**, 1665–1677, doi:10.1175/JAS3446.1.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293, doi:10.1175/1520-0434(1993)008<0281:WIAGFA>2.0.CO;2.

Nakanishi, M., 2000: Large-eddy simulation of radiation fog. *Bound.-Layer Meteor.*, **94**, 461–493, doi:10.1023/A:1002490423389.

——, 2001: Improvement of the Mellor–Yamada turbulence closure model based on large-eddy simulation data. *Bound.-Layer Meteor.*, **99**, 349–378, doi:10.1023/A:1018915827400.

——, and H. Niino, 2004: An improved Mellor–Yamada level-3 model with condensation physics: Its design and verification. *Bound.-Layer Meteor.*, **112**, 1–31, doi:10.1023/B:BOUN.0000020164.04146.98.

——, and ——, 2006: An improved Mellor–Yamada level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*, **119**, 397–407, doi:10.1007/s10546-005-9030-8.

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, doi:10.1023/A:1022146015946.

Potvin, C., and M. Flora, 2015: Sensitivity of idealized supercell simulations to horizontal grid spacing: Implications for warn-on-forecast. *Mon. Wea. Rev.*, **143**, 2998–3024, doi:10.1175/MWR-D-14-00416.1.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, doi:10.1175/2007MWR2123.1.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, doi:10.1175/2008WAF2222159.1.

Schwartz, C. S., and Coauthors, 2009: Next-day convection-allowing WRF model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, doi:10.1175/2009MWR2924.1.

——, Z. Liu, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, doi:10.1175/WAF-D-13-00145.1.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., doi:10.5065/D68S4MVH.

Smirnova, T. G., J. M. Brown, and S. G. Benjamin, 1997: Performance of different soil model configurations in simulating ground surface temperature and surface fluxes. *Mon. Wea. Rev.*, **125**, 1870–1884, doi:10.1175/1520-0493(1997)125<1870:PODSMC>2.0.CO;2.

——, ——, ——, and D. Kim, 2000: Parameterization of cold-season processes in the MAPS land-surface scheme. *J. Geophys. Res.*, **105**, 4077–4086, doi:10.1029/1999JD901047.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, doi:10.1175/WAF-D-10-05046.1.

——, G. S. Romine, C. S. Schwartz, D. J. Gagne, and M. L. Weisman, 2016a: Explicit forecasts of low-level rotation from convection-allowing models for next-day tornado prediction. *Wea. Forecasting*, **31**, 1591–1614, doi:10.1175/WAF-D-16-0073.1.

——, C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016b: Severe weather prediction using storm

surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, doi:10.1175/WAF-D-15-0138.1.

Stensrud, D. J., J.-W. Bao, and T. T. Warner, 2000: Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.*, **128**, 2077–2107, doi:10.1175/1520-0493(2000)128<2077:UICAMP>2.0.CO;2.

Sukoriansky, S., B. Galperin, and V. Perov, 2006: A quasinormal scale elimination model of turbulence and its application to stably stratified flows. *Nonlinear Processes Geophys.*, **13**, 9–22, doi:10.5194/npg-13-9-2006.

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, doi:10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2.

VandenBerg, M. A., M. C. Coniglio, and A. J. Clark, 2014: Comparison of next-day convection-allowing forecasts of storm motion on 1- and 4-km grids. *Wea. Forecasting*, **29**, 878–893, doi:10.1175/WAF-D-14-00011.1.

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, doi:10.1175/1520-0493(2001)129<0729:EOASRM>2.0.CO;2.

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, doi:10.1175/1520-0493(1997)125<0527:TRDOEM>2.0.CO;2.

——, 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, doi:10.1175/2007WAF2007005.1.

——, and Coauthors, 2015: The Mesoscale Predictability Experiment (MPEX). *Bull. Amer. Meteor. Soc.*, **96**, 2127–2149, doi:10.1175/BAMS-D-13-00281.1.

Weygandt, S. S., and N. L. Seaman, 1994: Quantification of predictive skill for mesoscale and synoptic-scale meteorological features as a function of horizontal grid resolution. *Mon. Wea. Rev.*, **122**, 57–71, doi:10.1175/1520-0493(1994)122<0057:QOPSFM>2.0.CO;2.

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

——, 2001: A skill score based on economic value for probability forecasts. *Meteor. Appl.*, **8**, 209–219, doi:10.1017/S1350482701002092.

Xue, M., D. Wang, J. Gao, K. Brewster, and K. K. Droegemeier, 2003: The Advanced Regional Prediction System (ARPS), storm-scale numerical weather prediction and data assimilation. *Meteor. Atmos. Phys.*, **82**, 139–170, doi:10.1007/s00703-001-0595-6.

——, F. Kong, K. A. Brewster, K. W. Thomas, J. Gao, Y. Wang, and K. K. Droegemeier, 2013: Prediction of convective storms at convection-resolving 1 km resolution over continental United States with radar data assimilation: An example case of 26 May 2008 and precipitation forecasts from spring 2009. *Adv. Meteor.*, **259052**, doi:10.1155/2013/259052.