Severe Weather Prediction Using Storm Surrogates from an Ensemble Forecasting System

RYAN A. SOBASH, CRAIG S. SCHWARTZ, GLEN S. ROMINE, KATHRYN R. FOSSELL, AND MORRIS L. WEISMAN

National Center for Atmospheric Research, Boulder, Colorado*

(Manuscript received 9 October 2015, in final form 2 December 2015)

ABSTRACT

Probabilistic severe weather forecasts for days 1 and 2 were produced using 30-member convection-allowing ensemble forecasts initialized by an ensemble Kalman filter data assimilation system during a 32-day period coinciding with the Mesoscale Predictability Experiment. The forecasts were generated by smoothing the locations where model output indicated extreme values of updraft helicity, a surrogate for rotating thunderstorms in model output. The day 1 surrogate severe probability forecasts (SSPFs) produced skillful and reliable predictions of severe weather during this period, after an appropriate calibration of the smoothing kernel. The ensemble SSPFs exceeded the skill of SSPFs derived from two benchmark deterministic forecasts, with the largest differences occurring on the mesoscale, while all SSPFs produced similar forecasts on synoptic scales. While the deterministic SSPFs often overforecasted high probabilities, the ensemble improved the reliability of these probabilities, at the expense of producing fewer high-probability values. For the day 2 period, the SSPFs provided competitive guidance compared to the day 1 forecasts, although additional smoothing was needed to produce the same level of skill, reducing the forecast sharpness. Results were similar using 10 ensemble members, suggesting value exists when running a smaller ensemble if computational resources are limited. Finally, the SSPFs were compared to severe weather risk areas identified in Storm Prediction Center (SPC) convective outlooks. The SSPF skill was comparable to the SPC outlook skill in identifying regions where severe weather would occur, although performance varied on a day-to-day basis.

1. Introduction

The benefits of running convection-allowing models (CAMs) for convective storm prediction have been demonstrated in a variety of contexts over the past decade (e.g., Done et al. 2004; Kain et al. 2006; Lean et al. 2008; Kain et al. 2008; Clark et al. 2009, 2010, 2011; Coniglio et al. 2010; Schwartz et al. 2009; Sobash et al. 2011; Weisman et al. 2013). As these studies show, much of the value of running CAM forecasts comes from their ability to provide explicit information about convective properties such as initiation, mode, motion, longevity, and intensity. To effectively use CAM output in operational forecasting settings, novel forms of guidance are needed to summarize these forecast attributes in ways that can be easily understood by forecasters and other end users.

Toward this end, Sobash et al. (2011, hereafter S11) documented a proof of concept for producing a next-day (from 1200 to 1200 UTC) severe convective weather guidance product using output from deterministic CAM forecasts initialized at 0000 UTC. In S11, areas of intense simulated convective storms were identified using updraft helicity (UH; Kain et al. 2008), a diagnostic designed to detect midlevel mesocyclones in CAM output. A threshold was applied to the UH field to identify ''severe'' simulated convection and produce a field of ''surrogate'' severe weather reports, which were then spatially smoothed to create a surrogate severe probabilistic forecast (SSPF) from the deterministic model output. The SSPF was designed to be a simple post-processing procedure used with output from existing deterministic CAMs to produce guidance that could be

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

Corresponding author address: Dr. Ryan A. Sobash, NCAR/MMM, P.O. Box 3000, Boulder, CO 80307.
E-mail: sobash@ucar.edu

DOI: 10.1175/WAF-D-15-0138.1

directly compared to existing operational probabilistic severe weather forecasts [e.g., Storm Prediction Center (SPC) convective outlooks] and used by forecasters to summarize the day's threat areas as depicted by a particular model solution. The SSPFs were deemed to be subjectively valuable when evaluated during the 2008 NOAA Hazardous Weather Testbed (Coniglio et al. 2010), and S11 showed that the SSPFs possessed objective skill when verified against observed severe weather reports.

While S11 used a deterministic CAM, they suggested that the SSPF technique should be easily and readily extended to convection-allowing ensemble prediction systems (EPSs), with expected benefits over using a deterministic forecast to produce SSPFs. A similar methodology to S11, using neighborhood smoothing, was applied by Schwartz et al. (2010) and Duc et al. (2013) to produce probabilistic forecasts of precipitation from an EPS, resulting in forecast guidance with superior skill when compared to the raw ensemble probabilities. In addition, other recent studies have used EPS output to produce probabilistic forecast guidance for convective weather hazards (e.g., Clark et al. 2012, 2013; Kain et al. 2013; Schwartz et al. 2015b, hereafter S15). Given these successes, extending the SSPF approach to an ensemble system is likely to produce superior forecasts of severe weather.

Much of this work (e.g., Clark et al. 2012, 2013; Kain et al. 2013) used convection-allowing EPSs that were initialized by downscaling analyses from external data assimilation and modeling systems [e.g., NOAA/NCEP Short-Range Ensemble Forecasting (SREF) or Global Forecast System (GFS) analyses]. An alternative approach is to utilize a continuously cycled ensemble Kalman filter (EnKF) analysis system (e.g., Schwartz and Liu 2014; Schwartz et al. 2015a). The EnKF produces ensemble analyses during data assimilation that can be used as initial conditions for ensemble forecasts. Several recent studies have used continuously cycled EnKF systems to produce mesoscale analyses that subsequently initialized convective-scale ensemble forecasts (Romine et al. 2014; Schwartz et al. 2014; S15; Schumacher and Clark 2014). These forecasts were shown to be skillful at predicting general areas of precipitation (Schwartz et al. 2014) and demonstrated promise in predicting regions of severe weather (S15).

This paper extends S11's work by applying the surrogate severe approach to produce SSPFs from output of a convection-allowing EPS that was initialized with mesoscale EnKF analyses. The EnKF-based EPS was run as a part of the Mesoscale Predictability Experiment (MPEX; Weisman et al. 2015) during the spring of 2013 and generated 30-member, 48-h, 0000 and 1200 UTC ensemble forecasts each day for a 32-day period over the central and eastern United States. This large set of ensemble forecasts provides an opportunity to evaluate the SSPF technique, as initially described in S11, and establish its effectiveness when applied to EPS output. While S15 used this approach to verify hourly SSPFs from the same EPS, it was done within the broader context of overall ensemble system performance. This work aims to more thoroughly understand the differences between SSPFs produced from ensemble and deterministic systems and the variations in skill from different initializations.

To achieve these goals, we create SSPFs for both day 1 and day 2 predictions (1200–1200 UTC) of severe weather hazards (e.g., strong wind gusts, large hail, and tornadoes) using output from several deterministic and ensemble model configurations. As in S11, we focus on using UH as a surrogate for all varieties of severe weather and verify the forecasts against observed severe weather reports. First, the ensemble SSPFs are compared to two baseline deterministic SSPFs (generated from forecasts initialized by EnKF analysis ensemble means and NCEP Global Forecast System analyses) to investigate the added benefit of running an ensemble system for SSPFs and to understand the behavior of SSPFs derived from deterministic versus ensemble output. Second, day 1 SSPFs are compared to day 2 SSPFs to quantify forecast skill in these two periods. Most previous studies have focused largely on the day 1 time frame, while less attention has been paid to the day 2 convective period, largely because of forecast lengths that do not extend into this period. Third, the skill of the 0000 and 1200 UTC SSPFs is compared to assess the potential benefits of running convection-allowing EPSs twice per day for predictions of severe weather hazards. Finally, in addition to verifying the SSPFs against severe weather reports, SSPFs are also evaluated against daily operational severe weather forecasts produced by the SPC. This comparison has not been previously attempted; thus, we document one method to do so and demonstrate how the SSPFs can be utilized as a performance baseline for human forecasts.

In section 2, we discuss the methodology for creating SSPFs from deterministic and ensemble forecasts and document the configuration of the forecast sets used in this work. Further, we outline the verification methods used to quantify SSPF skill. We explore the verification results from the ensemble in section 3, as well as compare the ensemble to the deterministic SSPFs and the day 1 SSPFs to the day 2 SSPFs. Section 4 compares the performance of the SSPF approach to that of SPC convective outlooks, while section 5 provides the conclusions and discussion of the findings.

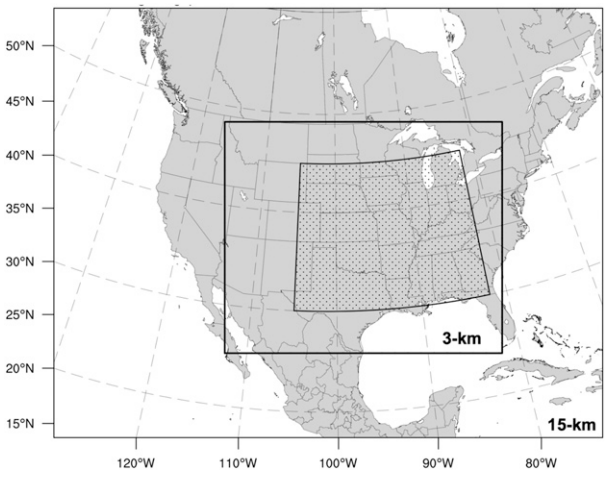

FIG. 1. Computation domain for all forecasts and SSPF verification domain (speckled) (from S15).

2. Methodology

a. CAM forecast dataset description

A continuously cycled EnKF data assimilation system was run during the spring of 2013 in support of MPEX between 15 May and 15 June. This period was particularly active across the continental United States (CONUS) with many multiday severe weather outbreaks and several notable tornadoes [e.g., the Granbury, Texas, tornado, rated 4 on the enhanced Fujita scale (EF4); the Moore, Oklahoma, EF5 tornado; and the El Reno, Oklahoma, EF3 tornadoes]. To produce the ensemble forecasts used here, the EnKF assimilated surface and upper-air observations every 6 h into a 50-member, 15-km horizontal grid spacing ensemble using the Advanced Research version of the Weather Research and Forecasting (WRF) Model (Skamarock et al. 2008) and the Data Assimilation Research Testbed (DART; Anderson et al. 2009). At 0000 and 1200 UTC, 30-member, 48-h, 3-km horizontal grid spacing ensemble forecasts were initialized by downscaling the first 30 analyses from the 15-km ensemble onto the 3-km grid. The 3-km nest domain covered the central and eastern CONUS, while the 15-km parent domain covered the CONUS plus adjacent areas (Fig. 1). For each member, the forecasts were integrated forward for 48 h, with boundary conditions for the 15-km parent domain coming from perturbed GFS forecasts (e.g., Torn et al. 2006). Additional details about the data assimilation and forecast system, including model configuration, assimilation settings, and observation types, can be found in S15.

In addition to the 0000 and 1200 UTC 30-member ensemble forecast datasets (henceforth ENS30-00 UTC and ENS30-12UTC, respectively), 0000 and 1200 UTC deterministic forecasts were produced retrospectively using 0000 and 1200 UTC GFS analyses for the initial conditions (GFS-00UTC and GFS-12UTC, respectively). Additional 0000 UTC deterministic forecasts were produced using ensemble mean analyses from the 50-member, 15-km ensemble (EMEAN-00UTC) as the initial conditions. All forecast model configurations were identical and only varied in the source of their initial conditions. The forecasts are summarized in Table 1.

b. Producing surrogate reports and SSPFs from CAM output

For this work, hourly maximum 2–5-km UH is used as the model surrogate for severe weather, because of its ability to identify rotating convection in CAM output (Kain et al. 2008) as well as other forms of intense convection (e.g., Clark et al. 2013). To produce SSPFs, the procedure outlined in S11 and S15 is used with the deterministic and ensemble CAM output as follows. First, a threshold is applied to the two-dimensional UH output at each forecast hour to produce a 3-km binary grid. These points are mapped onto a coarser grid, with grid spacing ~80 km, the same grid used in S11, by flagging an 80-km grid box if it contains one or more of the 3-km points that exceeded the UH threshold. The flagged 80-km grid boxes serve as the locations of surrogate storm reports (SSRs) and are aggregated over

TABLE 1. Experiment summary. All experiments used an identical 3-km forecast model configuration and domain.

Forecast name	Initial time (UTC)	Initial conditions	Ensemble or deterministic
ENS30-00UTC	00	Downscaled from first 30 members of 0000 UTC 50-member EnKF 15-km analyses	30-member ensemble
ENS30-12UTC	12	Downscaled from first 30 members of 1200 UTC 50-member EnKF 15-km analyses	30-member ensemble
EMEAN-00UTC	00	Downscaled from ensemble mean of 0000 UTC 50-member EnKF 15-km analyses	Deterministic
GFS-00UTC	00	Downscaled from 0000 UTC GFS analyses	Deterministic
GFS-12UTC	12	Downscaled from 1200 UTC GFS analyses	Deterministic

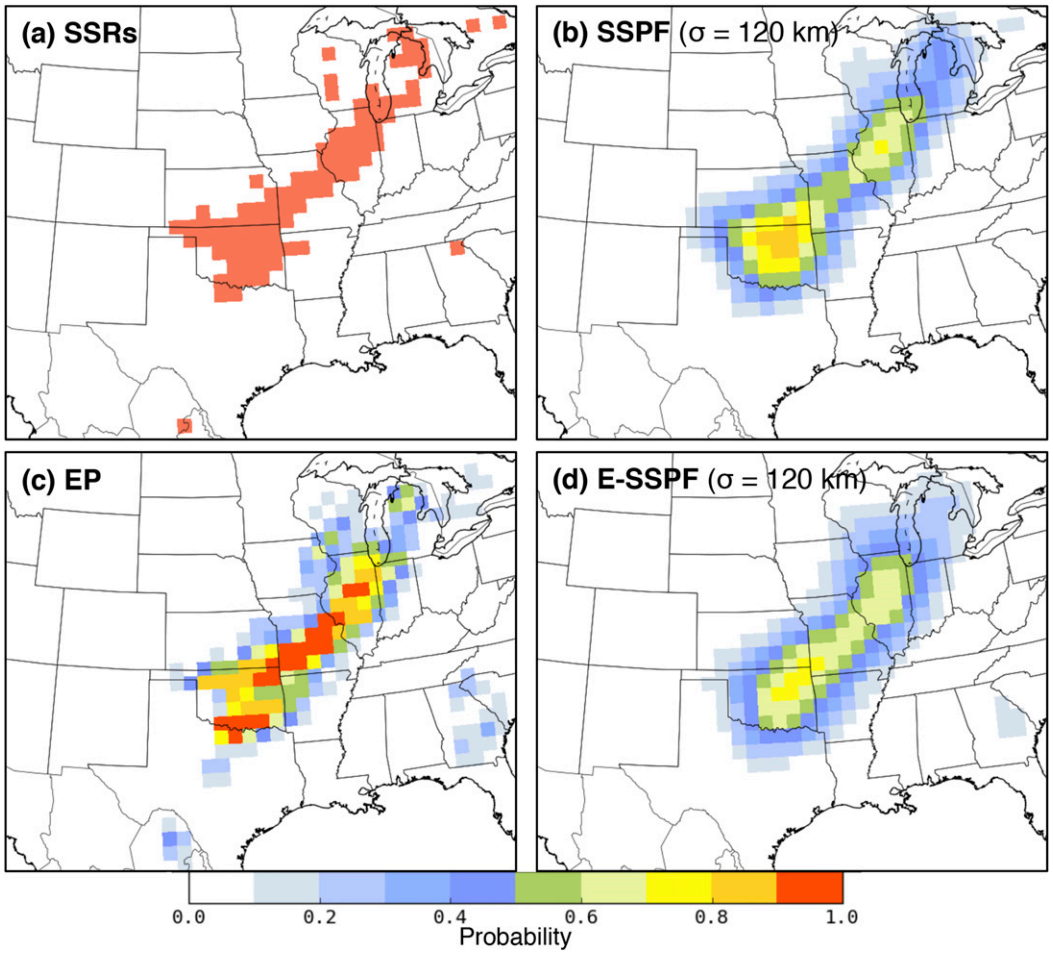

FIG. 2. Comparison of guidance produced from ENS30-00UTC output on 20 May 2013: (a) SSRs from ensemble member 1 using UH = 75 $m^2 s^{-2}$; (b) the SSPF produced from the field of SSRs in (a), with σ = 120 km; (c) the average of the binary SSRs from ENS30-00UTC, i.e., the EP of UH $\geq$ 75 $m^2 s^{-2}$; and (d) the E-SSPF produced by averaging the individual member SSPFs, or by smoothing the EP field in (c), as described in the text.

24-h periods to produce day 1 and day 2 SSR fields valid from 1200 to 1200 UTC (e.g., Fig. 2a). In other words, the final SSR field indicates the 80-km grid boxes where at least one 3-km grid point exceeded the UH threshold at least once within the corresponding 24-h period. The 24-h period was chosen to correspond to the period used in SPC convective outlooks, facilitating a direct comparison between the SSPFs and SPC forecasts in section 4.

Instead of determining the SSR thresholds based on UH frequency thresholds on the native model grid as in S11 (e.g., the 99.99th percentile), we chose UH thresholds from 25 to 200 $m^2 s^{-2}$ in 25 $m^2 s^{-2}$ increments. This range is based on subjective experience with typical UH values in the model output, as well as a comparison of which UH thresholds produced similar numbers of SSRs as observed storm reports (OSRs) during the 32-day period, when the OSRs are placed on the same 80-km grid (Table 2). Two additional UH thresholds, 70 and 80 $m^2 s^{-2}$, were used in an attempt to match SSR and OSR totals even more closely to produce a bias of 1 to ensure a fair comparison between the forecast datasets. Examination of SSR biases from each of the forecasts will be discussed in section 3a.

Each 24-h SSR field was smoothed with a two-dimensional Gaussian filter to produce an SSPF (e.g., Fig. 2b). This process weights each SSR within a window around each 80-km grid point using a Gaussian weighting function (i.e., a convolution of the SSR field with a Gaussian kernel) to produce a probability at that grid point. The Gaussian weights sum to 1, ensuring a probability no larger than 100%. Standard deviations (σ) from 20 to 200 km were used to examine the scale dependence of skill. In practice, the weights were only

TABLE 2. Number of SSRs and SSR forecast bias (in parentheses) during the 32-day forecast period for SSRs within the verification domain from the five sets of forecasts examined in this work for days 1 and 2. The 0000 UTC ensemble and 1200 UTC ensemble biases are for the ensemble mean number of SSRs. The day 1 and day 2 periods contained 1332 and 1385 OSRs, respectively. Boldface numbers indicate the UH threshold for each forecast period that produced an SSR bias closest to 1.

UH ($m^2 s^{-2}$)	ENS30-00UTC day 1	ENS30-12UTC day 1	EMEAN-00UTC day 1	GFS-00UTC day 1	GFS-12UTC day 1	ENS30-12UTC day 2	GFS-12UTC day 2
25	3835 (2.88)	4167 (3.13)	3929 (2.95)	3683 (2.77)	4062 (3.05)	3862 (2.79)	3799 (2.74)
50	2129 (1.60)	2367 (1.78)	2198 (1.65)	2064 (1.55)	2338 (1.76)	2168 (1.57)	2092 (1.51)
70	1411 (1.06)	1587 (1.19)	1456 (1.09)	**1352 (1.02)**	1600 (1.20)	1449 (1.05)	**1395 (1.01)**
75	**1279 (0.96)**	1449 (1.09)	**1326 (1.00)**	1218 (0.91)	1443 (1.08)	**1316 (0.95)**	1286 (0.93)
80	1165 (0.87)	**1325 (1.00)**	1207 (0.91)	1112 (0.83)	**1309 (0.98)**	1200 (0.87)	1166 (0.84)
100	817 (0.61)	936 (0.7)	844 (0.63)	775 (0.58)	929 (0.70)	838 (0.6)	783 (0.57)
125	538 (0.40)	626 (0.47)	546 (0.41)	521 (0.39)	625 (0.47)	549 (0.4)	519 (0.37)
150	367 (0.28)	425 (0.32)	359 (0.27)	334 (0.25)	440 (0.33)	364 (0.26)	355 (0.26)
175	257 (0.19)	291 (0.22)	266 (0.20)	221 (0.17)	310 (0.23)	248 (0.18)	256 (0.18)
200	179 (0.13)	205 (0.15)	197 (0.15)	159 (0.12)	215 (0.16)	172 (0.12)	190 (0.14)

applied to SSRs within a physical distance of 4σ, beyond which the weights become negligible. For example, using $\sigma = 20$ km produces weights such that for each 80-km grid point, an SSR located at the grid point is weighted by 0.91, while SSRs located at the four nearest grid points would be weighted by 0.02. In this case, a single isolated SSR would produce an SSPF value of 0.91 at the point of the SSR, with values of ~0.02 at the four grid points surrounding the SSR point. This approach has its roots in kernel density estimation (Silverman 1986), which estimates the probability density function (PDF) of a random variable without assuming a particular distribution (i.e., is nonparametric).

For the deterministic forecasts, an SSPF is produced from each forecast's SSR field, while for the ensemble forecasts, an SSPF is produced for each member's SSR field, and the individual member SSPFs are averaged to produce an ensemble SSPF (E-SSPF; e.g., Fig. 2d). The E-SSPFs can be compared to the solutions from each individual ensemble member's SSPF, as well as the SSPFs produced from the deterministic forecasts. Averaging each member's SSPF to produce the E-SSPF is mathematically equivalent to producing an ensemble probability (EP) of UH exceeding a given threshold (i.e., the ensemble mean SSR field; Fig. 2c) and then applying the smoothing procedure on this probability field to create an SSPF. In the latter case, the Gaussian smoother is only applied once, on the ensemble mean SSR field, while the individual member SSPFs are not explicitly produced.

SSPFs using $\sigma < 20$ km produced nearly identical verification scores to those of SSPFs produced with $\sigma = 20$ km. Consequently, SSPFs using $\sigma = 20$ km can be considered to be the skill of the 80-km grid-scale SSR forecasts without smoothing (i.e., in the case of a deterministic SSPF, forecasts of only 1 or 0, e.g., Fig. 2a, or in the case of an E-SSPF, the raw EP values, e.g., Fig. 2c). While the kernel choice can impact the SSPF, it has less of an impact than the kernel length scale. The Gaussian kernel used here weights SSRs less in far regions of the neighborhood, unlike a uniform kernel [i.e., a simple average within the neighborhood, as used in Schwartz et al. (2010)] that weights all points equally. SSPFs were also produced using the uniform kernel, and the results were nearly identical to those using a Gaussian kernel, although the specific settings needed to be tuned appropriately for each kernel type (e.g., a Gaussian kernel with $\sigma = 120$ km produced nearly identical verification results to those of a uniform kernel with neighborhood size of 240 km).

c. *Verification methods*

The SSPFs were verified with OSRs retrieved from the National Climatic Data Center's (NCDC) *Storm Data* publication. This dataset is available from the SPC Warning Coordination Meteorologist (WCM) web page (http://www.spc.noaa.gov/wcm) and consists of OSRs (wind gusts $\geq$ 50 kt, where 1 kt = 0.51 m s^{-1}; hail $\geq$ 2.54-cm diameter; and tornadoes) following the application of NCDC quality control procedures, which generally result in a reduction of reports compared to the total number received by the NWS, as duplicates are removed and thinning is applied. No attempt was made to discriminate between report types; all storm reports were used for verification. The OSRs were produced similarly as the SSRs, in that each report was mapped to the 80-km grid and aggregated over a 24-h period. Using this method, the final OSR field indicates the grid boxes where at least one severe report occurred within the 24-h period beginning at 1200 UTC. In addition to the raw field of OSRs, a smoothed observed severe probabilistic field (OSPF) was produced analogously to the SSPF

technique described in the previous subsection. For verification, the OSPFs and SSPFs were masked and verification scores were computed using only the grid points within the subdomain depicted in Fig. 1.

Several probabilistic verification measures were used to assess the SSPF quality, including the fractions skill score (FSS; Roberts and Lean 2008) and measures to assess the reliability, resolution, and sharpness of the probabilistic forecasts (Wilks 2006). The FSS provides a way of assessing the scale dependence of the SSPF skill by comparing SSPFs and corresponding OSPFs across a range of smoothing length scales. Statistical significance between two forecast FSSs (e.g., ENS30-00UTC and GFS-00UTC) was assessed using a pairwise difference bootstrap resampling technique (Wilks 2006). This procedure was identical to that used in S15 to compute the 90% FSS confidence intervals of the difference in FSSs except that a larger number of resamples were performed here (10 000).

To compute reliability and sharpness, attributes diagrams were created by binning the forecasts into 10% probability bins (0%–<10%, 10%–<20%, etc.). The forecast points in each bin were aggregated over the 32 SSPFs, and the corresponding observed relative frequencies at these points were computed using the OSRs. If the observed event of interest (here, a severe weather report) occurs at the same frequency as the forecast probability, the points will align along the diagonal. A no-skill line is also included where the forecast points contribute positively to the Brier skill score (BSS), assuming a reference of climatology. To evaluate forecast resolution, relative operating characteristic (ROC) curves were constructed by using SSPF probability thresholds in 5% increments to create a contingency table for each threshold, following the procedure in Schwartz et al. (2014). Then, the area under the ROC curve (denoted AUC) was computed and used as a summary measure of forecast resolution. In addition, the Brier score (BS; Brier 1950) and its decomposition (Murphy 1973) into reliability, resolution, and uncertainty components were computed to obtain an additional quantitative measure of forecast reliability and resolution.

3. Evaluation of SSRs and SSPFs

a. Bias characteristics of the SSRs

An examination of the bias between the forecast sets is performed to identify differences in the SSR climatology during the 32-day forecast period that can affect interpretation of the SSPF verification results. The total number of SSRs for the day 1 and day 2 periods ranges from ~4000 at the lowest UH threshold (~125 SSRs day^{-1}) to ~200 (~6 SSRs day^{-1}) at the highest threshold (Table 2). The total number of OSRs during this period was 1332 for day 1 and 1385 for day 2 (~40 day^{-1}). Thus, the UH threshold that produced an SSR total closest to the observed totals (i.e., a forecast bias closest to 1) occurred between 70 and 80 $m^2 s^{-2}$ for all forecasts.

As discussed in S15, the 1200 UTC initialized forecasts (ENS30-12UTC and GFS-12UTC) produced a larger number of SSRs during the day 1 period, along with a corresponding high precipitation bias, compared to both the 0000 UTC day 1 forecasts, as well as the 1200 UTC day 2 forecasts (one would expect these numbers to be similar over a large enough sample of cases). When comparing results from different sets of forecasts, we will use the UH threshold that produces biases closest to 1. For example, ENS30-00UTC day 1 SSPFs produced using UH = 75 $m^2 s^{-2}$ will be compared to ENS30-12UTC day 1 SSPFs using UH = 80 $m^2 s^{-2}$. This approach is similar to computing percentile thresholds from forecast data, as performed by Mittermaier and Roberts (2010), which reduces the impact of bias.

b. Skill of ENS30-00UTC E-SSPFs

The skill of the ENS30-00UTC E-SSPFs is examined with the FSS, attributes diagrams, and the BS. First, the FSS is used to examine the overall scale dependence of the skill for the ENS30-00UTC E-SSPFs. As in Mittermaier and Roberts (2010), the length scale where FSS = 0.5 is considered to be the minimum "skillful" scale. For all UH thresholds, E-SSPF skill increases with increased smoothing length scale, as expected, although the range of E-SSPF skill for each UH threshold varies substantially (Fig. 3). For UH thresholds of 70–75 $m^2 s^{-2}$, where the SSR bias is nearest to 1, FSS values approach 0.5 for σ = 20 km and increase above 0.8 for $\sigma \geq 120$ km. At UH thresholds that produce SSR forecast biases far from 1 (e.g., UH $\geq$ 25 $m^2 s^{-2}$ and UH $\geq$ 125 $m^2 s^{-2}$), the corresponding E-SSPFs have reduced FSSs compared to FSSs using UH near 75 $m^2 s^{-2}$. In fact, for E-SSPFs produced with UH $\geq$ 150 $m^2 s^{-2}$, the FSS does not exceed the minimum skillful scale at any smoothing length scale, suggesting there is no skill for $\sigma \leq 200$ km using these large UH thresholds to predict severe weather.

While FSSs improved as σ increased, forecast reliability often improved only up to a point, beyond which the reliability was reduced. This behavior was initially described by S11 and is demonstrated here with UH $\geq$ 75 $m^2 s^{-2}$ E-SSPFs (Fig. 4a). At small smoothing length scales (e.g., $\sigma < 60$ km), the forecasts overpredict large probabilities while underpredicting small probabilities. By increasing σ (e.g., from 60 to 120 km), the frequency

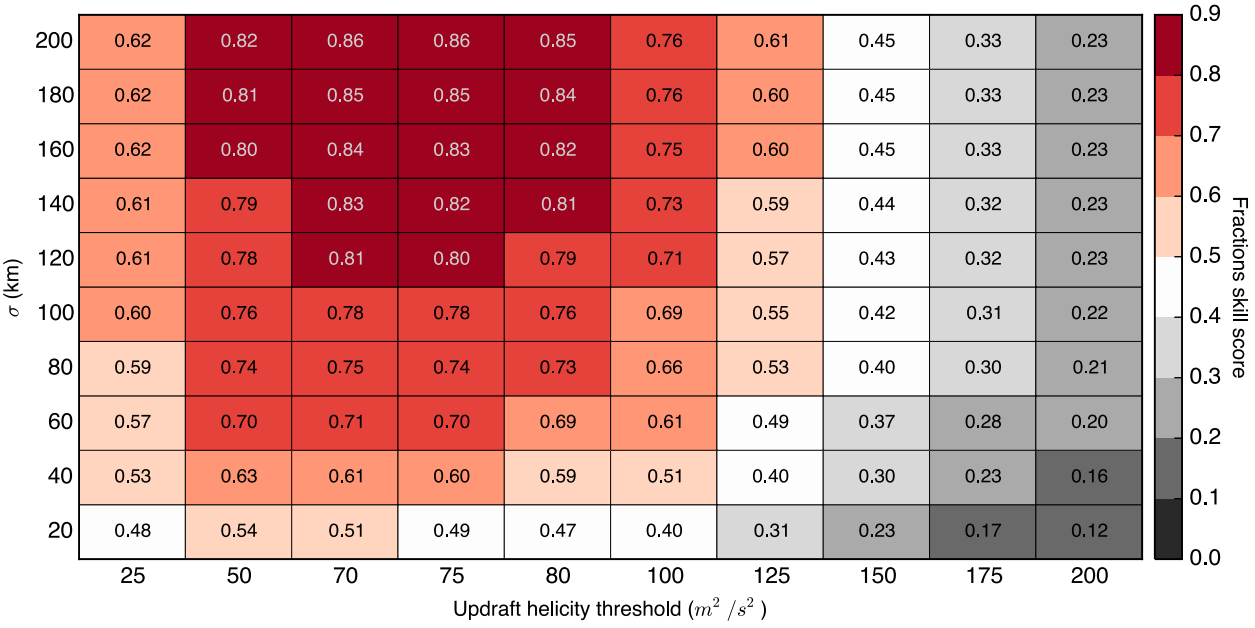

FIG. 3. FSSs of ENS30-00UTC SSPFs for all UH thresholds and Gaussian kernel standard deviations.

of high probabilities is reduced, while the number of low probabilities is increased (Fig. 4b), reducing the amount of under- and overforecasting (i.e., reducing forecast sharpness). This effect is largest for probabilities $> 40\%$, with large changes in reliability occurring as σ increased from 20 to 180 km, while smaller changes occurred for lower probabilities, consistent with the much larger number of forecast points. As σ is increased to 180 km and beyond, the highest forecast probabilities are underforecast. Thus, increasing σ from 60 to 120 km eliminates probabilities $> 80\%$, and increasing σ further to 160 km eliminates all probabilities $> 70\%$.

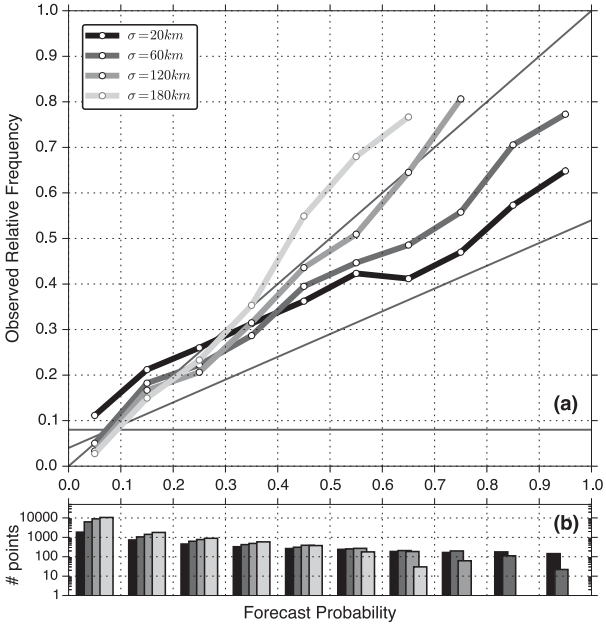

FIG. 4. (a) Attributes diagram for ENS30-00UTC SSPFs for UH $\geq 75\ m^2\ s^{-2}$ and for σ = 20, 60, 120, and 180 km. Forecasts are binned in 10% increments, i.e., [0, 0.1), [0.1, 0.2), etc. (b) Number of forecast points for each set of forecasts within each bin.

To quantify the changes in reliability and resolution of the SSPFs, the BS, its decomposition into reliability and resolution (BS_{rel} and BS_{res}, respectively), the BSS, and the ROC AUC were computed for UH $\geq 75\ m^2\ s^{-2}$ ENS30-00UTC E-SSPFs (Table 3). More reliable forecasts are associated with smaller BS_{rel} magnitudes, while forecasts with better resolution are associated with larger BS_{res} magnitudes. For the ENS30-00UTC E-SSPFs, a minimum in BS_{rel} occurs near σ = 120–140 km, reaffirming the subjective impression from the attributes diagram in Fig. 4a that this value of σ produces the best overall forecast reliability. Both the ROC AUC and

TABLE 3. Verification metrics for ENS30–00UTC SSPFs using UH $\geq 75\ m^2\ s^{-2}$. The uncertainty component of the Brier score is 0.0716. Boldface numbers represent the best values in each category.

σ(km)	BS	BS_{rel}	BS_{res}	BSS	ROC AUC
20	0.05928	0.00388	0.01572	0.165	0.803
40	0.05722	0.00316	0.01645	0.186	0.819
60	0.05567	0.00188	0.01656	0.205	0.823
80	0.05492	0.00126	**0.01677**	0.216	0.827
100	0.05457	0.00092	0.01666	0.220	0.828
120	**0.05450**	0.00078	0.01668	**0.222**	**0.829**
140	0.05465	**0.00070**	0.01624	0.217	0.828
160	0.05495	0.00085	0.01638	0.217	0.828
180	0.05537	0.00102	0.01624	0.213	0.827
200	0.05588	0.00115	0.01581	0.205	0.826

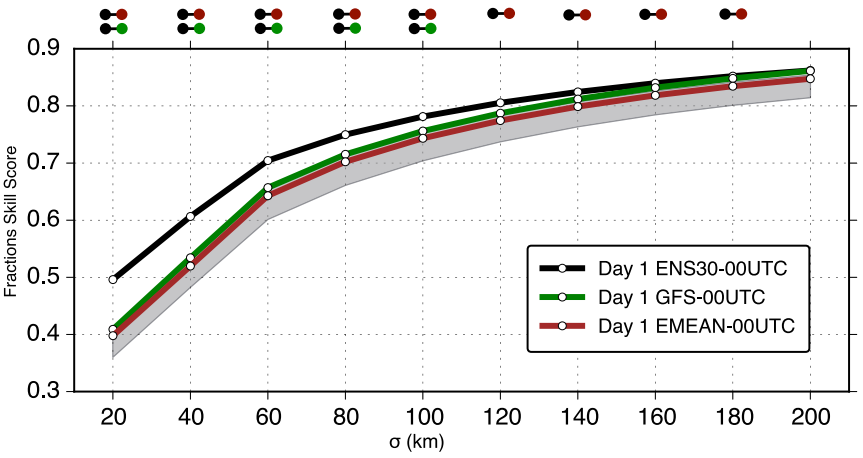

FIG. 5. FSSs for day 1 SSPFs from ENS30-00UTC (black), GFS-00UTC (green), and EMEAN-00UTC (brown). The range of the 30 deterministic SSPF scores is indicated by gray shading. The SSPFs are computed using UH thresholds that produce an SSR bias near 1, as described in the text and in Table 2. Dot pairs above the plot indicate that the 90% confidence interval for the differences in the aggregate FSSs between the two experiments do not include zero.

BS_{res} assess forecast resolution, and both are less sensitive to changes in σ, compared to BS_{rel}, as described in S11. Both the ROC AUC and BS_{res} reach maxima between 80 and 120 km, with slightly less resolution at small (<60 km) and large (>160 km) values of σ. Finally, the BS and BSS, which account for both reliability and resolution, suggest the errors in the E-SSPFs are minimized at $\sigma = 120$ km.

The BS, BSS, and ROC AUC results indicate that an optimal value of σ to produce the best reliability appears to be near 120–140 km for the ENS30-00UTC E-SSPFs using UH $= 75\,m^2\,s^{-2}$ as the surrogate threshold. Combined with the FSS results, these parameters seem to be the most appropriate choice for producing E-SSPFs from this ensemble system. While increasing σ can be used to improve E-SSPF reliability, the reduction in forecast sharpness and resolution is a trade-off that must be considered when choosing an appropriate smoothing length scale to produce SSPFs in both ensemble and deterministic forecasting systems.

c. *Comparison of E-SSPF to deterministic SSPF skill*

The skill of the ensemble forecasts compared to the deterministic forecasts, both initialized at 0000 UTC, will be examined in this section, using FSS and attributes diagrams. The ENS30-00UTC E-SSPFs possess larger FSSs than SSPFs produced from GFS-00UTC and EMEAN-00UTC, primarily when $\sigma \geq 120$ km (Fig. 5). The difference in FSS between the ensemble and two deterministic forecasts is largest at $\sigma = 20$ km, where the ENS30-00UTC SSPFs produce an FSS of ~0.5, just at the skillful-scale cutoff, while the two deterministic forecasts lie below this cutoff near 0.4. These differences shrink with increasing values of σ, with the FSS from the ensemble possessing statistically significant differences from the FSS from the two deterministic forecasts up through $\sigma = 100$ km for the GFS-00UTC SSPFs and through $\sigma = 180$ km for the EMEAN-00UTC SSPFs. In other words, the differences between the E-SSPFs and the deterministic SSPFs occur mostly on the mesoscale, while they converge at synoptic scales.

The lack of appreciable difference in the FSS scores at large scales implies that both the ensemble and deterministic forecasts produce similar forecasts over large-scale regions where severe weather is likely to occur. This is not surprising given the presence of many strongly forced convective outbreaks during the forecast period, as discussed in S15 and Weisman et al. (2015), with small synoptic-scale differences in the two forecasts, but larger differences on the mesoscale. Another factor possibly leading to similar forecasts at large scales is the use of GFS boundary conditions for both the ensemble and deterministic forecasts. However, this likely plays a lesser role given the relatively short lead times (12–36 h) for the day 1 forecasts and the placement of the verification domain away from the boundaries (Fig. 1).

Among the two deterministic forecasts, the GFS-00UTC SSPFs had slightly larger FSSs than the EMEAN-00UTC SSPF FSSs by a consistent amount for all values of σ, yet the differences in the FSSs were not statistically significant at the 95% level (Fig. 5). The EMEAN-00UTC SSPF FSSs generally fall within the range, but near the upper half, of SSPF skill produced by the individual ensemble members, consistent with the precipitation verification in Schwartz et al. (2014). On the other hand, the GFS-00UTC SSPFs lie either at or above the upper range of the FSS values of the individual 30 deterministic SSPFs within the ensemble, suggesting that the

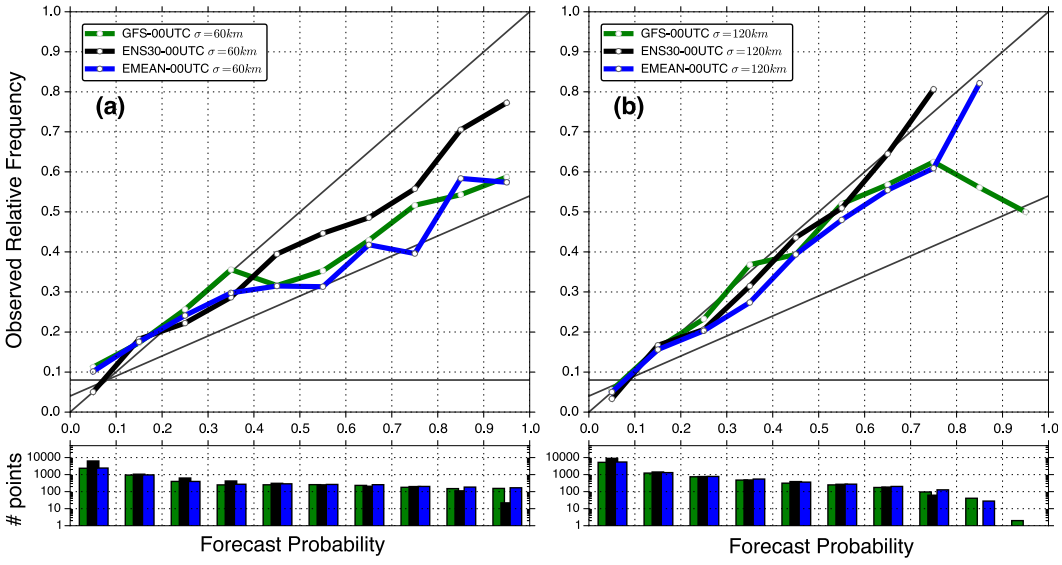

FIG. 6. As in Fig. 4, but for ENS30-00UTC, GFS-00UTC, and EMEAN-00UTC SSPFs using σ = (a) 60 and (b) 120 km. The UH thresholds used for each forecast are indicated in Table 2.

GFS-initialized forecasts provide slightly more accurate forecasts of severe weather than SSPFs produced by any of the individual ensemble members.

The largest differences in forecast reliability between the E-SSPFs and deterministic SSPFs occur at small scales, with ENS30-00UTC producing more reliable forecasts at $\sigma = 60$ km than the two deterministic SSPFs for forecast probabilities exceeding 40% (Fig. 6a). The E-SSPFs produce a larger number of forecast points at low probabilities than do the deterministic SSPFs, while producing a smaller number at higher probabilities. This behavior is due to the averaging that takes place in producing the E-SSPFs compared to the deterministic SSPFs. At $\sigma = 120$ km, ENS30-00UTC produces similar reliability to GFS-00UTC and EMEAN-00UTC for forecast probabilities below 50% (Fig. 6b). Above 50%, the ENS30-00UTC forecasts are more reliable for probabilities between 60% and 80%, while the deterministic SSPFs are overconfident at these probabilities, although for forecasts >80% the number of forecast points is <100, giving less confidence in the specific reliability results.

At both smoothing length scales, the deterministic SSPFs were most overconfident at the highest probabilities. Since SSPF values from the deterministic forecasts are based on the spatial distribution of SSRs, a concentrated region of SSRs needs to exist in the deterministic forecast for large probability values to occur in the SSPF. One way this could occur is if the model predicts a long-lived, large area of convection, such as a mesoscale convective system. Previous work has shown that these types of systems can produce extreme UH values (Clark et al. 2013), potentially producing a large region of SSRs, leading to large probability values. In this situation, the high confidence is due to the size of the system and not related to its likelihood of occurrence or anticipated location, potentially leading to overconfident probabilities. The ensemble forecasts are able to account for this uncertainty and produce more reliable forecasts of high probabilities.

d. Impact of initialization time on SSPF skill

For both the day 1 E-SSPFs and deterministic SSPFs, the 1200 UTC forecasts possess greater FSSs than SSPFs initialized 12 h earlier at 0000 UTC, with the day 1 ENS30-12UTC E-SSPFs possessing the largest FSS at all scales (Fig. 7a). As seen in earlier results, the largest separation in FSS between the 0000 and 1200 UTC SSPFs occurs at small scales, with the 0000 and 1200 UTC SSPFs producing very similar FSS scores at large scales. While the GFS SSPFs appear to show a greater increase in skill from 0000 to 1200 UTC than the E-SSPFs at small scales, the deterministic GFS-12UTC SSPF FSS remains smaller than the ENS30-00UTC E-SSPFs for $\sigma < \sim 100$ km (cf. blue and black lines in Fig. 7a), suggesting that even deterministic SSPFs initialized 12 h later are not as skillful as 0000 UTC initialized ensemble forecasts at these scales. At larger scales, the GFS-12UTC SSPF FSSs converge toward the E-SSPF FSSs, with the GFS-00UTC SSPF skill the smallest of the four forecasts for most scales. These results are similar to those demonstrated for hourly SSPF verification by S15 (see their Fig. 14), although a larger benefit was observed for the 1200 over the 0000 UTC SSPFs when evaluating within shorter time periods.

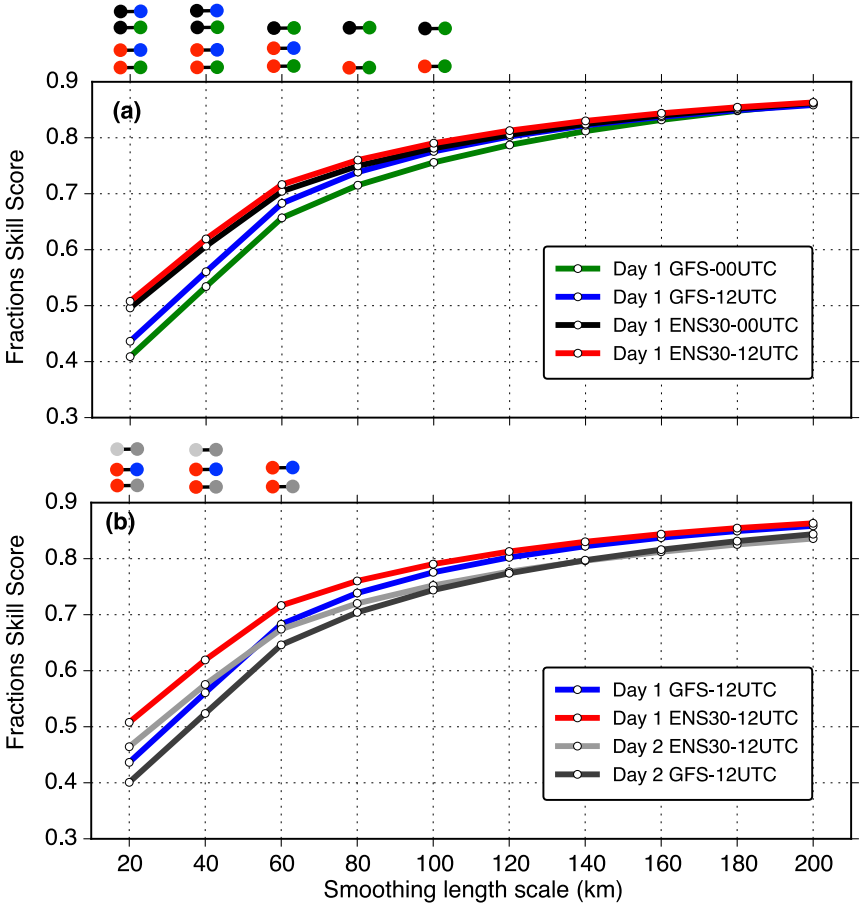

FIG. 7. As in Fig. 5, but for (a) 0000 and 1200 UTC-initialized SSPFs on day 1 and (b) 1200 UTC-initialized SSPFs on days 1 and 2. The UH thresholds used for each forecast are indicated in Table 2.

e. Skill of day 1 versus day 2 SSPFs

In this section, we compare the skill of the day 1 and day 2 forecasts from the 1200 UTC–initialized forecasts (ENS30-12UTC and GFS-12UTC). The 1200 UTC forecasts, rather than 0000 UTC forecasts, are used since they encompass the full day 2 period (1200–1200 UTC, or forecast hours 24–48). As described by S15, ENS30-12UTC and GFS-12UTC produce differing SSR biases during the day 1 and day 2 diurnal periods, with the day 1 period producing a larger number of surrogate reports compared to the day 2 period (Table 2). To produce a fair comparison between these two periods, we use the UH threshold that produces an SSR bias closest to 1.0 for each of the periods. The day 1 ENS30-12UTC and GFS-12UTC results use UH = $80\,m^2\,s^{-2}$, while the day 2 ENS30-00UTC results use UH = $75\,m^2\,s^{-2}$ and the day 2 GFS-12UTC results use UH = $70\,m^2\,s^{-2}$. While this produces an equitable comparison, the results are robust to changes in the UH threshold between 70 and $80\,m^2\,s^{-2}$.

Similar to the results from the previous subsection describing the 0000 UTC day 1 forecasts, the greatest differences in FSS between the ENS30-12UTC and GFS-12UTC day 1 and day 2 forecasts occur at small scales, decreasing as σ increases (Fig. 7b). When comparing the day 1 and day 2 forecast periods from the same model (e.g., day 2 ENS30-12UTC to day 1 ENS30-12UTC), the skill difference is nearly constant across all length scales, with the day 1 forecasts being more skillful. However, the FSS differences between the day 1 and day 2 SSPFs from both ENS30-12UTC and GFS-12UTC are not statistically significant at the 95% level for most σ. This is likely due to the increased predictability of many of the events during this period, with only a small decline in forecast skill during the day 2 period. Interestingly, the day 2 E-SSPFs from ENS30-12UTC possessed slightly larger FSS values on the grid–scale than the day 1 SSPFs from GFS-00UTC, suggesting that the GFS-12UTC grid-scale forecasts of severe weather for the day 1 period were quite poor, possibly because of

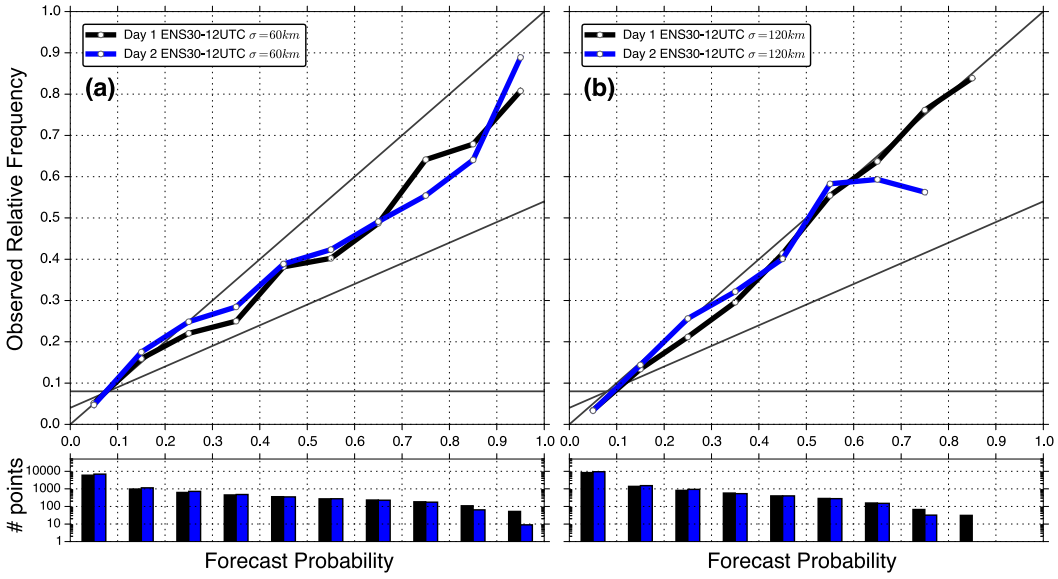

FIG. 8. As in Fig. 4, but for day 1 (black) and day 2 (blue) ENS30-12UTC SSPFs using σ = (a) 60 and (b) 120 km. The UH thresholds used for each forecast are indicated in Table 2.

the model spinup issues and poor resolution of mesoscale structures in the GFS initial conditions.

The lowest BS_{rel} and BSS for the day 2 E-SSPFs occurred at a slightly larger value of σ (~140–160 km) than the day 1 E-SSPFs (~120–140 km). There is also evidence of this in the reliability diagrams, with similarly overconfident predictions for small σ (Fig. 8a), yet for larger σ the day 2 E-SSPFs overpredict the highest probabilities (>60%). Thus, it may be necessary to increase the smoothing length scale to produce equivalent reliability for the day 2 E-SSPFs, at the expense of reducing the number of large forecast probabilities.

f. Effect of ensemble size on SSPF skill

Several studies have examined the effects of increasing ensemble size on precipitation forecast skill (Clark et al. 2011; Schwartz et al. 2014). For an ensemble configured similarly to that used in this work, Schwartz et al. (2014) found diminishing returns as the ensemble size approached 50 members, with little additional skill gained beyond 20–30 members. Here, we compare the results from the day 1 ENS30-00UTC forecasts to day 1 forecasts produced with only 10 ensemble members (ENS10-00UTC). To compare these two sets of forecasts, the FSS and attributes statistics for ENS10-00UTC were generated with 100 random 10-member ensembles, as in Schwartz et al. (2014). The distribution of FSS values from these one hundred 10-member ensembles is compared to the ENS30-00UTC results.

The ENS10-00UTC forecasts produce very similar FSSs when compared with the ENS30-00UTC forecasts, with only a small difference between the mean ENS10-00UTC FSS and the ENS30-00UTC FSS for all length scales (Fig. 9). Small differences were also noted in the reliability diagrams (not shown). The ENS10-00UTC FSS distribution is narrow, with the ENS30-00UTC FSS falling just above the range of ENS10-00UTC FSSs for small σ, but capturing the ENS30-00UTC FSS as σ increases. It appears that a substantial portion of the difference in FSS between the individual deterministic SSPFs from ENS30-00UTC and E-SSPFs produced by averaging the individual SSPFs (e.g., Fig. 5) can be achieved with an ensemble size of only 10 members, at least for the period and SSPFs examined here. This finding is promising, given that ensemble sizes > 10 members over CAM domains covering much of the CONUS require considerable computational resources, and current ensemble CAMs are using on the order of 10 members to produce ensemble predictions (e.g., Schwartz et al. 2015a). Using smaller verification time periods to construct E-SSPFs may result in more pronounced differences in 10- and 30-member ensembles, as the forecasts are penalized as a result of timing errors, unlike the present 24-h forecasts.

4. Comparison of E-SSPFs to operational severe weather forecasts

In addition to verification using OSRs, the E-SSPFs were compared to daily operational severe weather forecasts produced by the SPC. Here, we use only E-SSPFs from ENS30-00UTC with σ = 120 km and UH ≥ 75 $m^2 s^{-2}$ to compare to the SPC forecasts. These E-SSPFs

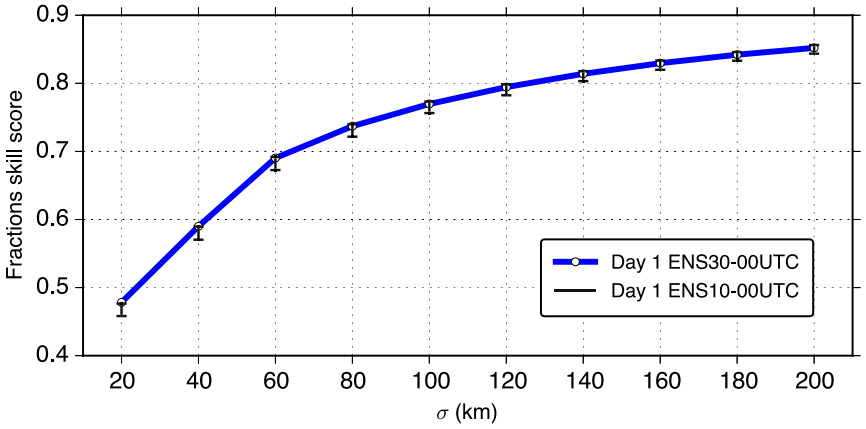

FIG. 9. As in Fig. 5 but for day 1 ENS30–00UTC FSS (blue) and the range (black interval) of FSS values from 100 realizations of ENS10-00UTC ensembles. Both forecasts use a UH threshold of 75 $m^2 s^{-2}$.

produced reliable probabilities with large FSSs and, overall, can be considered to be the best-performing SSPF. We compare the skill of these E-SSPFs to the skill of the SPC's "slight risk" areas contained in the day 1 convective weather outlooks, issued daily at 0600 UTC, and valid for the 1200–1200 UTC day 1 period, the same valid period for the E-SSPFs. The SPC slight risk area highlights the regions where SPC forecasters anticipate the occurrence of any form of severe weather on a particular day. Of the 32 0600 UTC SPC convective outlooks issued during the 32-day forecast period examined here, 31 contained a slight risk region within the verification domain, providing a large sample of SPC forecasts. As an additional point of reference, E-SSPF performance was also compared to the slight risk areas in the 32 1300 UTC SPC convective outlooks.

To produce a binary forecast of severe weather analogous to SPC slight risk areas from the continuous probability fields present in the E-SSPFs, a probability threshold was applied to each E-SSPF in 0.01 increments from 0 to 1, producing 100 severe weather forecast areas. Each of these areas are verified against the binary field of OSRs by computing a contingency table for each probability threshold, as in the ROC AUC calculation, then summing the contingency tables over the 32-day period (i.e., resulting in 100 contingency tables). The verification region was identical to that used in the earlier analyses (i.e., Fig. 1). The critical success index (CSI; also referred to as the threat score) was then computed for each table as follows:

$$\text{CSI} = \text{hits}/(\text{hits} + \text{misses} + \text{false alarms}).$$

The CSI has been consistently used to verify SPC categorical forecasts in previous work (e.g., Hitchens and Brooks 2014) and is a measure of the fraction of forecasts of severe weather that were correctly predicted (i.e., the score does not account for correct negatives). The two binary SPC slight risk areas (from the 0600 and 1300 UTC outlooks) were placed on the 80-km grid by flagging all grid points within the slight risk area. Using these grids, and comparing again to OSRs, the CSI was computed for the two SPC forecasts, using a contingency table summed over the 32 daily SPC outlooks.

As the probability threshold increases from 0, the CSI also increases, reaching a maximum of ~0.31 at an E-SSPF probability threshold of 29% (E-SSPF-29%; Fig. 10). This increase occurs as the number of false alarms decreases while the number of hits stays approximately the same. Beyond 29%, the CSI drops as the number of hits begins to decrease and overtakes additional gains by decreasing the number of false alarms. The E-SSPF-29% CSI is larger than both the CSI of the 0600 and 1300 UTC SPC forecasts (~0.27 and 0.29, respectively). One of the reasons for this is that the total number of grid points covered by E-SSPF-29% during the forecast period (1458) is smaller than the number of grid points forecast by the SPC (2597). In other words, the total number of grid boxes covered by E-SSPF-29% over the forecast period is 56% smaller than that of the SPC slight risk grid points. For individual events, the number of grid points using E-SSPF-29% is smaller than the corresponding SPC slight risk area on 28 of the 32 days (Fig. 11a).

The smaller number of grid points within the E-SSPF-29% area tends to increase the CSI of the E-SSPF-29% compared to the SPC forecasts. The SPC outlook areas tend to be drawn to prioritize hits over false alarms (Hitchens and Brooks 2014), resulting in areas that are larger than they would be if both were prioritized equally. To account for this effect when comparing SPC

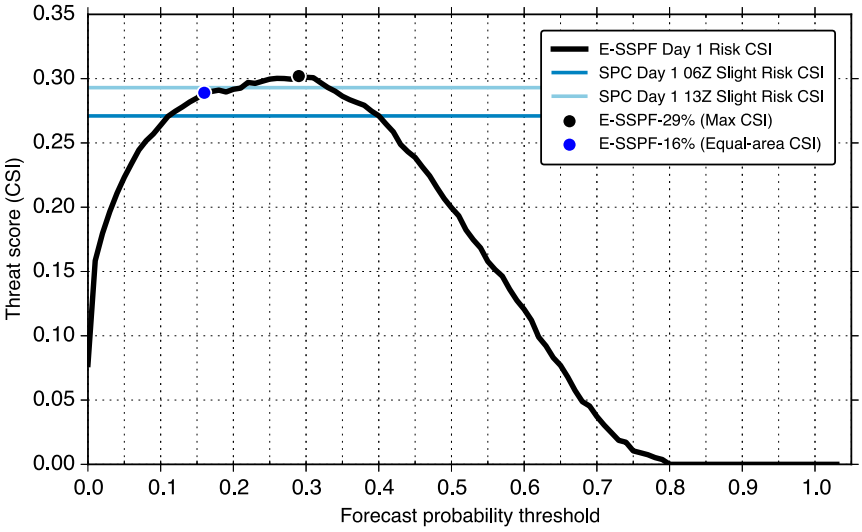

FIG. 10. CSI for ENS30-00UTC E-SSPFs using probability thresholds in 1% increments (black line). Shown for reference are the CSIs of the 0600 UTC (dark blue line) and 1300 UTC (light blue line) SPC slight risk areas, as well as the CSIs for the E-SSPF-16% (blue dot) and E-SSPF-29% areas (black dot).

outlooks to the E-SSPFs, the E-SSPF probability threshold that produces approximately the same number of forecast points over the 32-day period as the SPC outlooks was used as a second benchmark. This occurs at a probability threshold of 16% (Fig. 10). At this threshold, the individual E-SSPF areas and SPC outlook sizes become more uniform, with an approximately equal number of SPC outlooks that are larger than the E-SSPF areas and vice versa (Fig. 11b).

The CSI using E-SSPF-16% is 0.289 (Fig. 10), smaller than the CSI of E-SSPF-29% (0.31), but larger than the CSI of the 0600 UTC SPC outlooks, while falling slightly below the CSI of the 1300 UTC outlooks. Thus, when constrained by the typical size of SPC forecasts, the E-SSPF remains capable of providing guidance that has a similar level of skill to SPC outlooks. To examine the daily variations of skill, the CSI using E-SSPF-16% and the SPC outlook is computed for each day (Fig. 12).

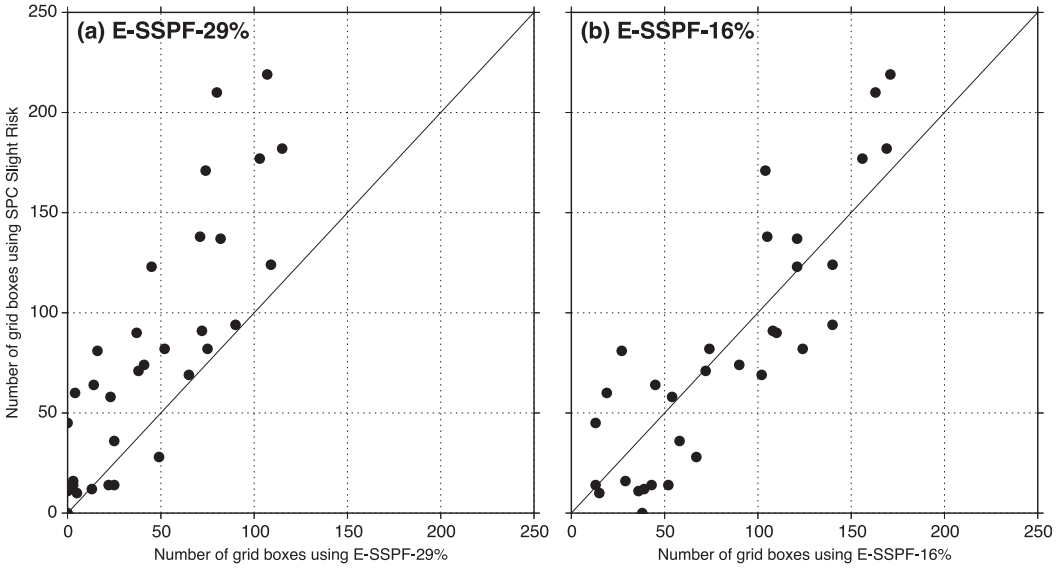

FIG. 11. Scatterplot of the number of grid boxes within the 32 SPC and the (a) E-SSPF-29% and (b) E-SSPF-16% forecast areas.

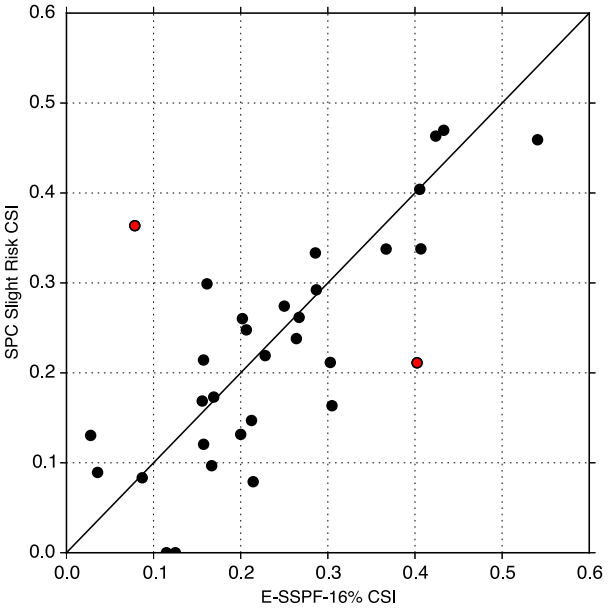

FIG. 12. Scatterplot of daily CSI scores for the 32 SPC and E-SSPF-16% forecasts. The two red dots indicate the days with the largest positive and negative differences in SPC CSI minus E-SSPF-16% CSI that are described in the text and shown in Fig. 13.

Of the 32 forecasts, E-SSPF-16% produced a larger CSI than the corresponding 0600 UTC SPC outlook on 18 days (56% of the days; Fig. 12). Days with larger E-SSPF-16% CSI tend to occur on days with large SPC CSI, indicating that the E-SSPF-16% skill is positively correlated with SPC forecast skill.

Several outliers exist where the difference in CSI was large between the SPC and E-SSPF-16% areas. On 30 May 2013, the E-SSPF-16% CSI was ~0.41 while the SPC forecast CSI was ~0.22 (Figs. 13a,b). This was a large-scale severe weather event, with reports occurring across much of the central United States. The E-SSPF-16% area extends farther to the east than the SPC forecast, correctly anticipating the occurrence of severe convection across Illinois and eastern Missouri, capturing reports in this area that were missed in the SPC forecast. On the other hand, the SPC forecast CSI was ~0.37 on 10 June 2013, while the E-SSPF-16% CSI was ~0.08 (Figs. 13c,d). On this day, severe thunderstorms formed in the afternoon across a confined region in central Tennessee, as a result of instability associated with an upper trough. While the SPC forecast correctly predicted these events across central Tennessee, the convection in the model was displaced to the north into central Kentucky and did not possess large UH values, so no SSRs and no E-SSPF-16% area were forecast. The model also produced many false alarms associated with a large E-SSPF-16% area in eastern Wyoming and western Nebraska where only a few OSRs occurred.

5. Summary and discussion

Probabilistic forecasts of severe convection for the day 1 and day 2 convective periods were created and verified using convection-allowing deterministic and ensemble forecast output from a 32-day period during the spring of 2013. The severe weather forecasts were produced by thresholding the model UH field to produce surrogate severe weather reports (SSRs) and smoothing the SSRs to produce surrogate severe probability forecasts (SSPFs). Ensemble SSPFs were created with forecasts initialized from 30 members of a mesoscale EnKF analysis system and initialized at both 0000 and 1200 UTC, while deterministic SSPFs were created with forecasts initialized by 1) GFS analyses at 0000 and 1200 UTC and 2) ensemble mean EnKF analyses at 0000 UTC.

Appropriate UH thresholds were determined primarily by the SSR bias compared to OSRs, which was near 1 for UH thresholds between 70 and 80 $m^2 s^{-2}$. Using a 75 $m^2 s^{-2}$ threshold, the ENS30-00UTC E-SSPFs possessed large FSSs (>0.5) at all length scales, although intermediate smoothing length scales (e.g., σ = 120 km) produced the most reliable E-SSPFs. The E-SSPFs outperformed the deterministic SSPFs in several ways. FSS values for ENS30-00UTC SSPFs were larger than both GFS-00UTC and EMEAN-00UTC SSPFs at small scales, while at larger scales, the E-SSPFs produced similar FSS scores as GFS-initialized SSPFs. At small scales, the E-SSPFs were more reliable than the deterministic SSPFs, especially for forecast probabilities $\geq$ 40%. As the scales increased, the differences in reliability decreased as the SSPFs became more similar, although differences remained at large probability values (>70%) where the E-SSPFs tended to produce more reliable forecasts than the deterministic SSPFs (although at the expense of forecast sharpness). Given the larger FSS values for a given length scale, the E-SSPFs do not have to be smoothed as much as the deterministic SSPFs to achieve a given FSS (e.g., FSS = 0.7).

SSPFs produced from forecasts initialized at 1200 UTC were slightly more skillful than their 0000 UTC counterparts, especially for the deterministic SSPFs. The E-SSPFs from ENS30-00UTC possessed larger FSSs at small scales than the later-initialized GFS-12UTC SSPFs, demonstrating that ensembles initialized at 0000 UTC can outperform 1200 UTC deterministic forecasts of severe weather during the day 1 period. For the day 2 period, the ENS30-12UTC E-SSPFs were more skillful than GFS-12UTC SSPFs at small scales, with FSS values converging at large scales, similar to the behavior of the day 1 forecasts. The day 2 ENS30-12UTC SSPFs were competitive

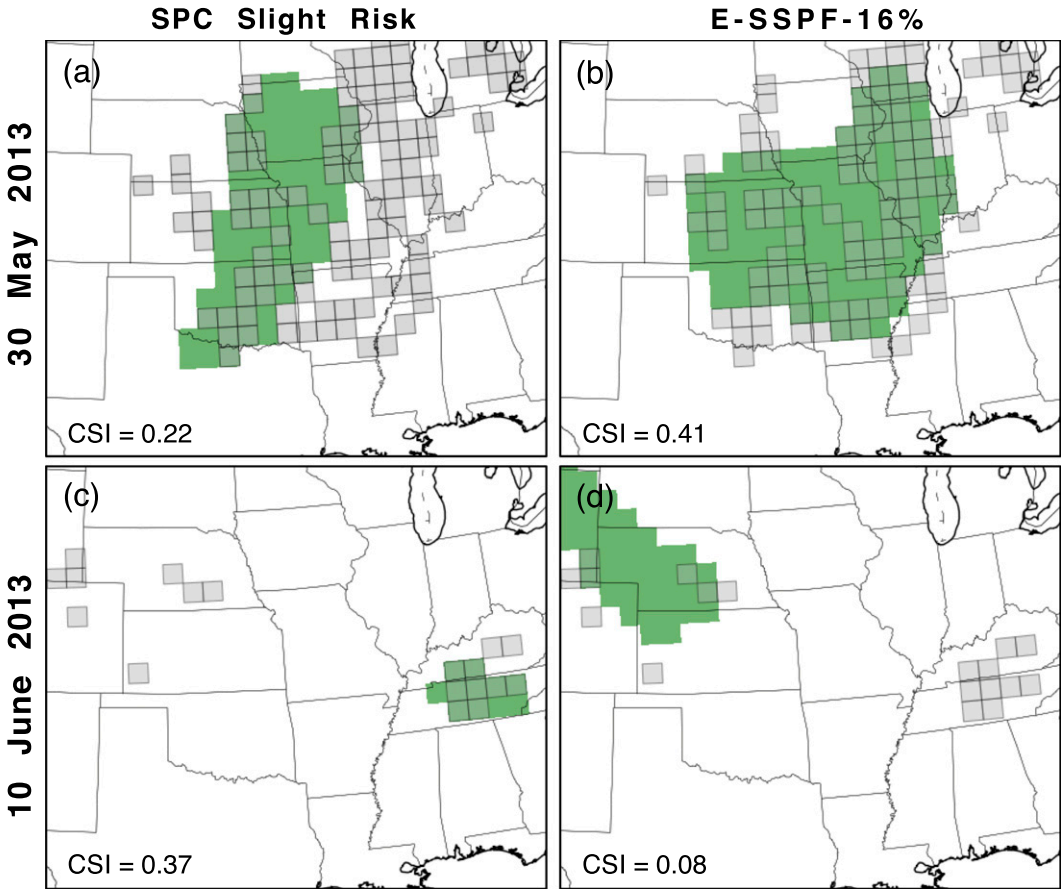

FIG. 13. (a),(c) SPC slight risk forecasts and (b),(d) E-SSPF-16% forecasts for (top) 30 May and (bottom) 10 Jun 2013 (green). The two cases had the largest difference between the SPC and E-SSPF-16% CSI scores. Shaded gray boxes indicate OSRs. Domain is restricted to the approximate verification region.

with the day 1 ENS30-12UTC SSPFs, although the day 2 SSPFs required slightly more smoothing to produce an optimal level of reliability, at the expense of forecast sharpness. Finally, 0000 UTC day 1 E-SSPFs using only 10 members performed similarly to the 30-member E-SSPFs, with slight degradation of skill at small scales.

The model climatology for surrogate fields such as UH will vary depending on the model resolution and initial conditions, as demonstrated here and in other work (e.g., Adlerman and Droegemeier 2002; Kain et al. 2008; S15). S11 determined that a UH threshold of $34\,m^2\,s^{-2}$ was optimal for their deterministic SSPFs, using a model with 4-km horizontal grid spacing. This UH threshold produced an SSR bias closest to 1 when placing the SSRs on the same grid. Thus, comparing SSPFs using different UH thresholds that produce similar SSR biases appears to be a suitable method for evaluating severe weather forecast skill from different convection-allowing models. Using fixed percentiles may also be appropriate when the forecast domain sizes are similar. As for σ, SSPFs with σ between 160 and 220 km produced reliable probabilities for the sets of deterministic forecasts examined here, as well as in S11 (cf. S11's Figs. 8b and 8c). The E-SSPFs required less smoothing to achieve the same level of reliability as the deterministic forecasts, suggesting that the ensemble accounted for part of the spatial uncertainty. While these values could serve as starting points for other deterministic or ensemble forecasts, σ ultimately depends on the scale dependence of the error in a particular modeling system and will require additional calibration to produce reliable guidance on smaller time scales (e.g., <24 h).

The SSPFs were also compared to SPC "slight risk" regions. To produce a binary forecast from the SSPF to compare with the slight risk area, two probability thresholds were chosen, one that produced the largest CSI and the other that produced the same number of forecast points as the SPC slight risk areas over the 32-day forecast period. In both instances, the area derived from the SSPF produced similar CSI scores to

those produced by the SPC forecast, suggesting that during this evaluation period and over the verification region considered, the E-SSPFs were comparable to the SPC forecasts at delineating the regions where severe weather occurred. Variability existed in the day-to-day performance of the SSPFs compared to the SPC forecasts; SSPFs for events with large-scale regions of severe weather appeared to perform similarly or outperform the SPC guidance, compared to events occurring with weaker forcing over smaller areas (e.g., Fig. 13). This suggests that the improvement that forecasters make over automated guidance, such as the SSPF, may be regime dependent. While not shown, the SSPF guidance outperformed the SPC guidance by a larger margin during the day 2 period, and future work is planned to understand differences in the day 2 period.

In the future, SSPFs could be used as a reference forecast for the skill of the SPC forecasts. To date, such a baseline does not exist for features directly resolved by numerical models for forecasts of severe weather (e.g., as is commonly applied to forecasts of tropical cyclone track and intensity). With knowledge of the performance characteristics of the model compared to human forecasts in anticipating severe weather events, forecasters can integrate model forecasts more effectively into their workflows. Creating SSPFs for individual severe weather hazards is the ultimate goal, with work ongoing to produce and calibrate this type of guidance using UH as well as other model surrogates.

Finally, the results here examined SSPF skill over a period dominated by strongly forced convective environments that were conducive to the development of traditional spring severe weather outbreaks. Because of this, the dataset is composed of a small subset of severe weather environments observed across the CONUS. As such, additional investigation of a larger collection of cases is needed to 1) quantify model skill at predicting severe weather hazards across a broader range of regions and seasons, such as nontraditional severe weather environments (e.g., low shear, high CAPE); 2) test additional surrogate fields that may be better suited to identifying severe convection in these situations; and 3) more fully examine and quantify the relationship between model performance and human forecast skill across a variety of predictability regimes.

Acknowledgments. This work was partially supported by National Oceanic and Atmospheric Administration Grant NA15OAR4590191. We would also like to acknowledge high-performance computing support from Yellowstone (ark:/85065/d7wd3xhc) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation.

REFERENCES

Adlerman, E. J., and K. K. Droegemeier, 2002: The sensitivity of numerically simulated cyclic mesocyclogenesis to variations in model physical and computational parameters. *Mon. Wea. Rev.*, **130**, 2671–2691, doi:10.1175/1520-0493(2002)130<2671:TSONSC>2.0.CO;2.

Anderson, J. L., T. Hoar, K. Raeder, H. Liu, N. Collins, R. Torn, and A. Arellano, 2009: The Data Assimilation Research Testbed: A community facility. *Bull. Amer. Meteor. Soc.*, **90**, 1283–1296, doi:10.1175/2009BAMS2618.1.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

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 M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, doi:10.1175/2010WAF2222404.1.

——, 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.

——, J. S. Kain, P. T. Marsh, J. Correia Jr., M. Xue, and F. Kong, 2012: 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.

Coniglio, M. C., K. L. Elmore, J. S. Kain, S. J. Weiss, M. Xue, and M. L. Weisman, 2010: Evaluation of WRF model output for severe weather forecasting from the 2008 NOAA Hazardous Weather Testbed Spring Experiment. *Wea. Forecasting*, **25**, 408–427, doi:10.1175/2009WAF2222258.1.

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

Duc, L., K. Saito, and H. Seko, 2013: Spatial–temporal fractions verification for high-resolution ensemble forecasts. *Tellus*, **65A**, 18 171, doi:10.3402/tellusa.v65i0.

Hitchens, N. M., and H. E. Brooks, 2014: Evaluation of the Storm Prediction Center's convective outlooks from day 3 through day 1. *Wea. Forecasting*, **29**, 1134–1142, doi:10.1175/WAF-D-13-00132.1.

Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181, doi:10.1175/WAF906.1.

——, 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.

——, and Coauthors, 2013: A feasibility study for probabilistic convection initiation forecasts based on explicit numerical guidance. *Bull. Amer. Meteor. Soc.*, **94**, 1213–1225, doi:10.1175/BAMS-D-11-00264.1.

Lean, H. W., P. A. Clark, M. Dixon, N. M. Roberts, A. Fitch, R. Forbes, and C. Halliwell, 2008: Characteristics of high-resolution versions of the Met Office Unified Model for forecasting convection over the United Kingdom. *Mon. Wea. Rev.*, **136**, 3408–3424, doi:10.1175/2008MWR2332.1.

Mittermaier, M., and N. Roberts, 2010: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**, 343–354, doi:10.1175/2009WAF2222260.1.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

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.

Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev.*, **142**, 4519–4541, doi:10.1175/MWR-D-14-00100.1.

Schumacher, R. S., and A. J. Clark, 2014: Evaluation of ensemble configurations for the analysis and prediction of heavy-rain-producing mesoscale convective systems. *Mon. Wea. Rev.*, **142**, 4108–4138, doi:10.1175/MWR-D-13-00357.1.

Schwartz, C. S., and Z. Liu, 2014: Convection-permitting forecasts initialized with continuously cycling limited-area 3DVAR, ensemble Kalman filter, and ''hybrid'' variational–ensemble data assimilation systems. *Mon. Wea. Rev.*, **142**, 716–738, doi:10.1175/MWR-D-13-00100.1.

——, 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.

——, and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, doi:10.1175/2009WAF2222267.1.

——, G. S. Romine, K. R. Fossell, 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.

——, ——, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015a: NCAR's experimental real-time convection-allowing ensemble prediction system. *Wea. Forecasting*, **30**, 1645–1654, doi:10.1175/WAF-D-15-0103.1.

——, ——, M. L. Weisman, R. A. Sobash, K. R. Fossell, K. W. Manning, and S. B. Trier, 2015b: A real-time convection-allowing ensemble prediction system initialized by mesoscale ensemble Kalman filter analyses. *Wea. Forecasting*, **30**, 1158–1181, doi:10.1175/WAF-D-15-0013.1.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 175 pp.

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.

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.

Torn, R. D., G. J. Hakim, and C. Snyder, 2006: Boundary conditions for limited-area ensemble Kalman filters. *Mon. Wea. Rev.*, **134**, 2490–2502, doi:10.1175/MWR3187.1.

Weisman, M. L., C. Evans, and L. Bosart, 2013: The 8 May 2009 superderecho: Analysis of a real-time explicit convective forecast. *Wea. Forecasting*, **28**, 863–892, doi:10.1175/WAF-D-12-00023.1.

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

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences: An Introduction*. 2nd ed. Academic Press, 467 pp.