A Comparison of Neural-Network and Surrogate-Severe Probabilistic Convective Hazard Guidance Derived from a Convection-Allowing Model

RYAN A. SOBASH, GLEN S. ROMINE, AND CRAIG S. SCHWARTZ

National Center for Atmospheric Research, Boulder, Colorado

(Manuscript received 27 February 2020, in final form 20 July 2020)

ABSTRACT

A feed-forward neural network (NN) was trained to produce gridded probabilistic convective hazard predictions over the contiguous United States. Input fields to the NN included 174 predictors, derived from 38 variables output by 497 convection-allowing model forecasts, with observed severe storm reports used for training and verification. These NN probability forecasts (NNPFs) were compared to surrogate-severe probability forecasts (SSPFs), generated by smoothing a field of surrogate reports derived with updraft helicity (UH). NNPFs and SSPFs were produced each forecast hour on an 80-km grid, with forecasts valid for the occurrence of any severe weather report within 40 or 120 km, and 2 h, of each 80-km grid box. NNPFs were superior to SSPFs, producing statistically significant improvements in forecast reliability and resolution. Additionally, NNPFs retained more large magnitude probabilities (>50%) compared to SSPFs since NNPFs did not use spatial smoothing, improving forecast sharpness. NNPFs were most skillful relative to SSPFs when predicting hazards on larger scales (e.g., 120 vs 40 km) and in situations where using UH was detrimental to forecast skill. These included model spinup, nocturnal periods, and regions and environments where supercells were less common, such as the western and eastern United States and high-shear, low-CAPE regimes. NNPFs trained with fewer predictors were more skillful than SSPFs, but not as skillful as the full-predictor NNPFs, with predictor importance being a function of forecast lead time. Placing NNPF skill in the context of existing baselines is a first step toward integrating machine learning–based forecasts into the operational forecasting process.

1. Introduction

Convection-allowing models (CAMs), i.e., numerical weather prediction (NWP) models configured with horizontal grid spacings ≤ 4 km, are routinely used to provide forecast guidance for convective storms. While partially resolving convective systems, CAMs do not resolve most hazards (i.e., tornadoes or hail ≥ 2.54 cm in diameter). Thus, an extensive body of research has been devoted to developing methods to extract hazard information from CAMs to improve hazard forecasting, primarily using diagnostics as surrogates, or proxies, for the occurrence of a convective hazard (e.g., Kain et al. 2008; Sobash et al. 2011, 2016, 2019; Clark et al. 2013; Loken et al. 2017; Gallo et al. 2016, 2018, 2019b).

The majority of studies using CAM diagnostics as surrogates determine the locations of severe convective hazards by thresholding the diagnostic field and smoothing the resulting binary forecast to account for spatial errors in hazard location (henceforth referred to as the "surrogate-severe" framework). When using deterministic CAM output, a quasi-probabilistic forecast of severe weather is produced; for an ensemble, the individual smoothed deterministic hazard forecasts can be averaged to produce a surrogate-severe ensemble-based hazard forecast (e.g., Sobash et al. 2016). Studies have used surrogate-severe forecasts to produce hazard guidance for CAMs, including individual hazards such as large hail (Gagne et al. 2017) and tornadoes (Clark et al. 2013; Gallo et al. 2016, 2018, 2019b; Sobash et al. 2019), as well as evaluate differences in the skill of severe hazard predictions among different CAM configurations (e.g., Gallo et al. 2019a).

The 2–5 km above ground level updraft helicity (UH) diagnostic (Kain et al. 2008) has been the most utilized diagnostic, as it can indicate the presence of supercells and intense squall lines in CAM output, which regularly produce convective hazards. Sobash and Kain (2017) demonstrated that surrogate-severe forecasts are most skillful when the UH threshold varies as a function of season and region, in part due to the diversity of convective environments and modes that generate severe

Corresponding author: Dr. Ryan A. Sobash, sobash@ucar.edu

DOI: 10.1175/WAF-D-20-0036.1

weather hazards. Variations in appropriate UH thresholds have also been noted in case studies of cool-season convection (e.g., Guyer and Jirak 2014). However, varying the UH threshold is difficult to implement without a large set of forecasts that can be used for calibration.

Other challenges with using the surrogate-severe framework include the need for multiple diagnostics to capture different processes responsible for different severe weather hazards (e.g., UH for mesocyclones versus column maximum graupel for hail) and the rigidity of a diagnostic threshold value (i.e., a storm with intensity slightly below the threshold is viewed as nonsevere). Further, the smoothed probabilities derived from a deterministic forecast are based solely on the spatial density of points exceeding the diagnostic threshold, often with fixed smoothing parameters. Constructing an optimal set of diagnostics, thresholds, and smoothing parameters quickly becomes onerous within the surrogate-severe framework as one attempts to anticipate the variety of convective modes, processes, and diagnostic magnitudes that are related to severe weather hazards. While model configuration also plays a role in how these parameters are chosen (e.g., Potvin et al. 2019), the sensitivity to convective environment and mode are most relevant, since model configuration can presumably be accounted for a priori, while the convective environment and scenario can vary substantially from day to day.

One potential path forward to improve severe weather hazard guidance with CAMs is to apply machine-learning (ML) algorithms to learn the relationships between CAM diagnostics, environmental properties, and diagnostic magnitudes associated with the potential for severe weather hazards. While ML techniques have been applied to a variety of high-impact weather prediction problems (Marzban and Stumpf 1998; Cintineo et al. 2014; Lagerquist et al. 2017; McGovern et al. 2017; Herman and Schumacher 2018), few have used CAMs as input into ML models to produce convective hazard guidance. For example, Gagne et al. (2017) and Burke et al. (2020) both trained ML models by combining remotely sensed hail size observations and forecast CAM hail size diagnostics, and then used the trained models to predict and calibrate hail size using real-time CAMs as input. While their ML hail forecasts exhibited skill, some verification metrics indicated that the surrogate-severe hail forecasts generated with UH[1] were competitive with the ML-based guidance, although the ML forecasts were less biased.

Other studies have used ML techniques to classify simulated storms but did not produce hazard guidance. For example, Robinson et al. (2013) used a neural network to classify simulated CAM storms as severe or nonsevere in CAM simulations downscaled from regional reanalyses, while, Gagne et al. (2019) used CAMs together with a convolutional neural network to identify different storm modes. A common limitation of many of these studies is the reliance on a storm-based framework, which constrained ML predictions to locations where storms were present in CAM output, neglecting storm initiation and placement biases in CAMs.

To move beyond the surrogate-severe framework of generating convective hazard guidance, this study evaluates the skill of output from a feed-forward neural network (NN) that was designed to provide grid-based probabilistic predictions of convective hazards using CAM diagnostics as input fields. A feed-forward NN was chosen due to its simplicity, and makes training with a large database of forecasts more tractable. Compared to storm-based ML guidance (e.g., Burke et al. 2020), the grid-based NN predictions are not constrained to produce predictions where simulated storms occur, allowing the NNs to potentially learn biases associated with CAM storm initiation or placement. The NN probabilistic forecasts were compared to quasi-probabilistic forecasts generated with UH using the surrogate-severe framework to understand the locations, times, and environments in which ML-based forecasts provided added benefit over the UH-based surrogate-severe forecasts for convective hazard prediction. Finally, the sensitivity of the NN predictions to different subsets of predictors is documented to better understand what variables contribute most to NN forecast skill.

2. Methodology

a. 3-km WRF forecast dataset

The forecast dataset consisted of 497, 36-h, forecasts produced with version 3.6.1 of the WRF Model (Skamarock et al. 2008) and used NOAA 0.5° Global Forecast System (GFS) initial and boundary conditions. The forecasts had 3-km horizontal grid spacing and spanned the entire contiguous United States (CONUS; Fig. 1). Physics schemes are listed in Table 1. Events were selected from the NOAA Storm Prediction Center (SPC) severe weather event archive. Many criteria determined which events are included in the SPC archive, including the number of observed storm reports on a given day and the maximum categorical threat level (e.g., moderate risk). The forecast dataset consisted of

[1] Both Wendt et al. (2016) and Adams-Selin et al. (2019) noted that UH is one of the most skillful CAM diagnostics for anticipating severe hail occurrence, especially for hail ≥ 2 in. in diameter.

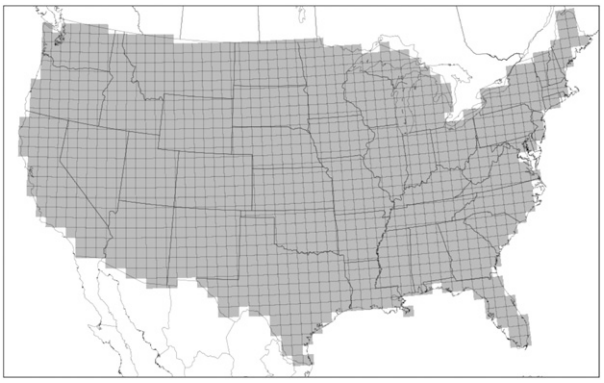

FIG. 1. WRF computational domain. The 80-km grid boxes used for verification are shaded.

events between 15 October 2010 and 15 July 2017, excluding 15 July–15 October each year. The selection strategy was designed to identify high-impact warm- and cool-season severe weather events east of the Rockies, while neglecting events in the western United States and in association with landfalling tropical cyclones. This reforecast dataset was used in previous work to study next-day precipitation (Schwartz and Sobash 2019) and tornado (Sobash et al. 2019) forecast skill. The full list of events, additional details about the selection criteria, and configuration choices for WRF are available in Sobash et al. (2019).

b. *Preprocessing CAM diagnostic fields*

Each of the 497 3-km WRF forecasts produced a set of 38 diagnostics, which were preprocessed to produce input into the NN. The diagnostics included environmental, upper-air, and explicit surrogate fields (Table 2). Some diagnostics were derived from combinations of other diagnostics [e.g., the significant tornado parameter (STP) or the 700–500-hPa lapse rate; Table 2]. To reduce the dimensionality of the raw model output, the diagnostics were upscaled onto an 80-km grid. For the upper-air and environmental fields, each 80-km grid point was assigned the mean value of the 3-km gridpoint values within each 80-km grid box at each forecast hour. For the surrogate fields, each 80-km grid point was assigned the maximum value over all 3-km grid points within each 80-km grid box. Since the surrogate fields were computed as hourly maximum values (Kain et al. 2010), using the gridbox maximum ensures that extremes in the output were retained.

In addition to the 38 diagnostics computed for each forecast grid point, larger spatial and temporal averages and maxima were computed to account for spatial and temporal errors in storm placement. Specifically, predictors were constructed by averaging the environmental fields and computing the maximum of the explicit fields within one or two grid boxes (in each direction) and temporal 0-, 1-, or 2-h windows of each 80-km grid box, producing six additional predictors for each environmental and explicit diagnostic. Forecast hour, gridpoint latitude and longitude, and day of year were also used as predictors. Combined, 174 predictors were used as input for each 80-km grid box at each forecast hour (Table 2; Fig. 2). Since observed severe weather reports were only available within the CONUS, 80-km grid points outside the CONUS were not used for training. The end result of the preprocessing procedure was a collection of 1298 80-km grid boxes, for each forecast hour and forecast, producing a total of ~21 million grid boxes, each with 174 predictors.

c. *Training and verification*

To reduce the computational burden of training the NN with ~21 million grid boxes, the preprocessed WRF dataset was split in half using a random selection of grid boxes, leaving ~11 million grid points for training. Using this thinned dataset, six NNs were trained for each year between 2011 and 2016, with the grid boxes for that year removed. For example, forecasts for 2012 were produced using a NN trained with forecast grid boxes occurring in 2010–11 and 2013–17. NNs were not trained for the 2010 and 2017 forecasts to ensure that each year's NN had a similar size of training data (2010 and 2017 only had 9 and 19 WRF forecasts, respectively). The final training dataset for each year included ~9 million grid boxes (~400 forecasts) and ~2 million

TABLE 1. Physical parameterization schemes used for WRF Model forecasts.

Parameterization type	Scheme	Reference
Microphysics	Thompson	Thompson et al. (2008)
Longwave and shortwave radiation	Rapid Radiative Transfer Model for Global Climate Models (RRTMG) with ozone and aerosol climatologies	Mlawer et al. (1997) Iacono et al. (2008) Tegen et al. (1997)
Planetary boundary layer	Mellor–Yamada–Janjić (MYJ)	Mellor and Yamada (1982); Janjić (1994, 2001)
Land surface model	Noah	Chen and Dudhia (2001)
Cumulus parameterization	None	None

TABLE 2. List of 42 base predictors used to train the NNs. The mean of the environmental and upper-air fields, and the maximum of the explicit fields, within each 80-km grid box, was used as input into the NNs. In addition to the 42 base predictors, 132 neighborhood predictors were constructed by taking larger spatial and temporal means and maximums of the 15 environmental and 7 explicit fields, as described in the text, resulting in a final set of 174 predictors used as input into the NNs.

Base predictor	Type
Forecast hour	Static
Day of year	Static
Latitude	Static
Longitude	Static
Surface-based convective available potential energy	Environment
Most-unstable convective available potential energy	Environment
Surface-based convective inhibition	Environment
Mixed-layer convective inhibition	Environment
0–6-km bulk wind difference	Environment
Mixed-layer lifted condensation level	Environment
0–1-km bulk wind difference	Environment
0–1-km storm-relative helicity	Environment
0–3-km storm-relative helicity	Environment
2-m temperature	Environment
2-m dewpoint temperature	Environment
Surface pressure	Environment
Product of most-unstable convective available potential energy and 0–6-km bulk wind difference	Environment
Significant tornado parameter	Environment
700–500-hPa lapse rate	Environment
Hourly max 2–5-km updraft helicity	Explicit
Hourly max 0–3-km updraft helicity	Explicit
Hourly max 0–1-km updraft helicity	Explicit
Hourly max updraft speed below 400 hPa	Explicit
Hourly max downdraft speed below 400 hPa	Explicit
Hourly max 10-m wind speed	Explicit
Hourly precipitation accumulation	Explicit
925-, 850-, 700-, and 500-hPa zonal wind speed	Upper air
925-, 850-, 700-, and 500-hPa meridional wind speed	Upper air
925-, 850-, 700-, and 500-hPa temperature	Upper air
925-, 850-, 700-, and 500-hPa dewpoint temperature	Upper air

grid boxes evaluation (~100 forecasts). Prior to training, the preprocessed data were normalized based on the full dataset distribution mean and standard deviation for each predictor.

The NN configuration used for training was chosen based on previous work that applied NNs to meteorological data (e.g., Gagne et al. 2019). In general, a NN is configured with multiple layers of neurons, including an input, output, and one or more hidden layers. Each neuron takes a linear weighted combination of inputs from the previous layer and computes an output using an activation function that is fed into the subsequent layer (Fig. 2). Here, the input layer consisted of 174 neurons, one for each predictor, while the output layer consists of 1 neuron, providing a probability of any severe weather hazard associated with the combined set of input predictors. The hidden layer included 1024 neurons, other sizes of the hidden layer produced NNs with similar performance characteristics. Other hyperparameter choices are provided in Table 3. Finally, the weights within the NN were initialized randomly, meaning each retraining of the NN produced slightly different results, but these variations in skill were small and did not impact the conclusions. Since the results were largely insensitive to several of the hyperparameters, we feel that the NN configuration here is robust and not designed specifically to optimize predictions from this set of forecasts. Generalization of the NNs beyond the CAM output used here is discussed in section 8.

SPC storm reports[2] (Schaefer and Edwards 1999) were used for training two NNs and for verifying the predictions. The two NNs were identical except for the spatial radius (40 and 120 km) used to label grid boxes where any severe report occurred; the temporal tolerance of 2 h was identical for both NNs. The choice of a spatial radius of 40 km was made to match the SPC probabilistic forecast definition and to be consistent with the 80-km grid size. The 120-km length scale was used to examine forecast predictability on larger spatial scales. The 2-h temporal tolerance was based on the desire of SPC to produce 4-h probabilistic severe weather guidance in the future (Krocak and Brooks 2020). Based on these choices, the NNs output probabilities of any severe weather report occurring within 40 or 120 km in space and 2 h in time for each grid box and forecast hour. The output from the NNs will be denoted as neural network probability forecasts (NNPFs).

NNPFs were verified with the binary storm report fields for each of the 469 events between 2011 and 2016 using the Brier skill score (BSS; Wilks 2006) and the relative operating characteristic area under the curve (ROCA; Mason 1982; Marzban 2004). Additionally, reliability diagrams (Wilks 2006) were computed to assess forecast reliability. As in Sobash and Kain (2017; SK17), instead of computing the BSS with the full sample climatology as a reference forecast, a spatially and temporally varying 30-yr severe weather climatology was computed using all severe reports occurring between 1986 and 2015. This climatology was computed for each 80-km grid box, day of the year, and hour of the

[2] Including all reports of tornadoes, hail ≥ 1 in. in diameter, and measured or estimated wind gusts ≥ 50 kt (including reports with an unknown wind gust magnitude). Reports retrieved from SPC storm report archive available at https://www.spc.noaa.gov/wcm/. Section 8 discusses some of the issues associated with using storm reports for NN training and verification.

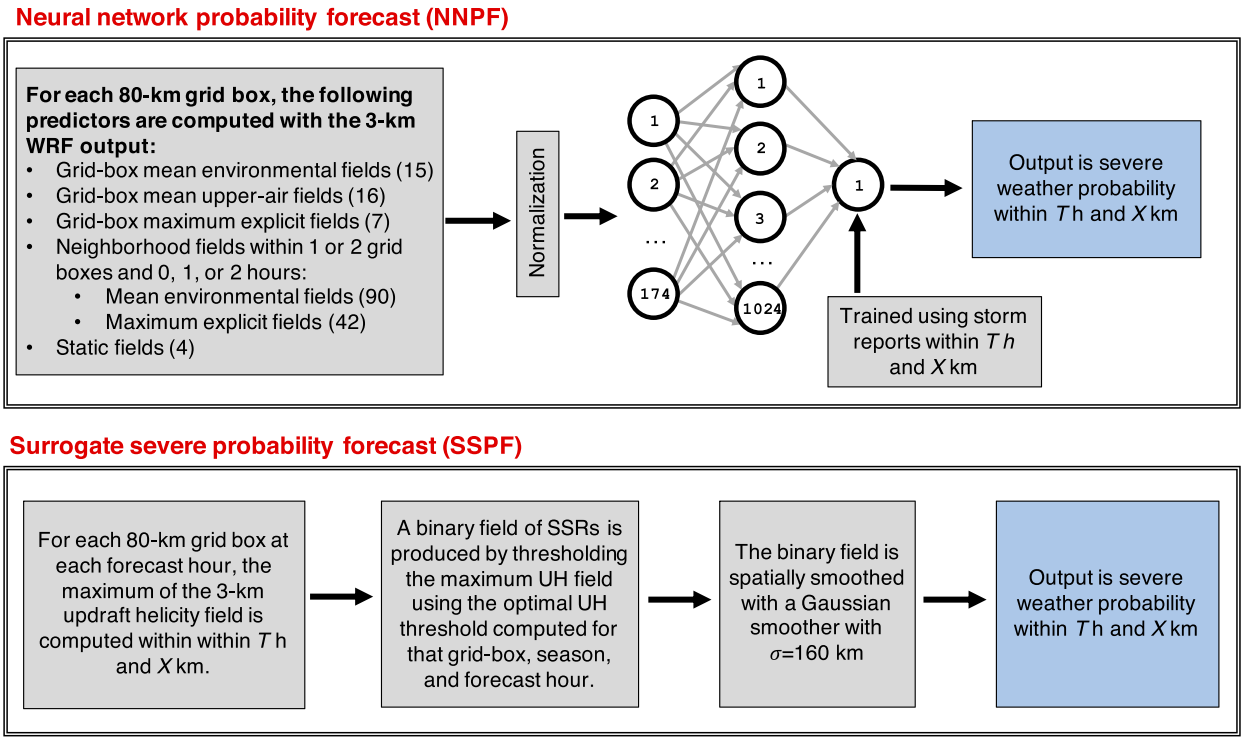

FIG. 2. Summary of WRF preprocessing and procedure to generate SSPFs and NNPFs, including the neural network configuration; T and X represent the time (2 h) and space (40 or 120 km) windows used in the paper to train the NNs.

day, by aggregating reports within space and time windows consistent with how the probabilities were defined when training the NNs. For example, the baseline climatological forecast when computing the BSS for the 120-km, 2-h probabilities was a climatology of severe weather occurring within 120 km and 2 h of an 80-km grid box. To produce a smoothly varying climatology, a Gaussian smoother was applied with a standard deviation of 15 days, 1.5 grid boxes, and 1.5 h. The ROCA was computed with the scikit-learn (Pedregosa et al. 2011) function ''roc_auc_score,'' which computes the ROCA using a trapezoidal method with all unique probability values as decision points. To reduce the potential impact of small probabilities when computing the BSS and ROCA, probabilities $< 0.1\%$ were set to 0.0 in the NNPFs.

d. Creation of surrogate-severe probability forecasts

Surrogate-severe probability forecasts (SSPFs) were used as a baseline to assess the added value of the NNPFs. The procedure to produce the SSPFs was similar to that outlined in Sobash et al. (2011) and SK17 who both used the UH diagnostic to identify locations where surrogate-severe reports (SSRs) occurred in the model and produced quasi-probabilistic forecasts by smoothing the binary SSR fields. While Sobash et al. (2011) used a fixed UH threshold to determine SSR locations, SK17 used a varying UH threshold based on latitude, longitude, and day of the year, demonstrating that doing so led to improvements in forecast skill. To produce the most skillful UH guidance possible, we apply the methods of SK17 here, using ''optimal'' UH thresholds that lead to SSR biases near one when compared to a field of observed storm reports (OSRs) computed analogously to the SSRs

TABLE 3. Settings used to construct and train the neural networks. The neural networks were trained using the keras python package and employed graphics processing units (GPUs) to accelerate the training process.

Hyperparameter	Value
No. of hidden layers	1
No. of neurons in hidden layer	1024
Dropout rate	0.1
Learning rate	0.001
No. of training epochs	10
Hidden layer activation function	Rectified linear unit
Output layer activation function	Sigmoid
Optimizer	Stochastic gradient descent
Loss function	Binary cross-entropy
Batch size	1024
Regularization	L2
Batch normalization	On

(i.e., a binary 80-km grid indicating locations where at least one report occurred).

Several modifications to the SK17 approach were required. First, the discontinuous nature of the current WRF forecasts prevented optimal thresholds from being computed for each day of the year (i.e., weeks often exist in the dataset where no WRF forecasts were produced). Instead, two sets of optimal thresholds were computed, one using all warm-season (March–July) and another using all cool-season (October–February) forecasts. Second, optimal UH thresholds were computed for each forecast hour, rather than daily in SK17, since SSPFs and NNPFs were produced for each 4-h period centered on each forecast hour, rather than the 24-h SSPFs in SK17. Third, a 2-h temporal neighborhood was used to aggregate OSRs (e.g., for the 1000 UTC climatology, all OSRs occurring within 0800–1200 UTC were considered when computing the bias). Finally, if fewer than 25 OSRs occurred in a given 80-km grid box, after spatial aggregation, then the optimal UH threshold for that grid box was set to that using all SSRs and OSRs within the domain. Other aspects of selecting the UH thresholds were identical to SK17, including using all SSRs and OSRs within a two gridbox spatial neighborhood to compute the biases.

The result of the UH calibration process were fields of optimal UH thresholds, computed for each 80-km grid box, that varied by season, latitude, longitude, and forecast hour (Fig. 3). All 497 forecasts were used to compute the optimal UH thresholds, which gives an advantage to the SSPFs over the NNPFs, since the optimal UH thresholds were determined based partly on the verifying OSRs. Due to this, the skill of the UH guidance presented here is likely an upper bound, reflecting the highest possible baseline for the NNPFs to exceed. The variability of the optimal UH thresholds was similar to that noted in SK17, including smaller optimal UH thresholds during the cool season than in the warm season and smaller optimal UH thresholds in the eastern United States compared to the central Plains during the warm season (Fig. 3).

The optimal UH thresholds were used to produce 40- and 120-km, 4-h SSPFs with output from each WRF forecast. The 40-km (120-km), 4-h SSPFs were based on SSRs computed where the optimal UH threshold was exceeded within 40 km (or 120 km) and 2 h at each 80-km grid box. Thus, the probabilistic event definitions for SSPFs and NNPFs were equivalent (i.e., the probability of an event within 40 or 120 km and 2 h of an 80-km grid box). The SSPFs were produced by smoothing the binary SSRs using a Gaussian smoother at each hour with a Gaussian standard deviation σ of 160 km. This choice of σ was informed by previous work (e.g., Sobash et al. 2011) and produced SSPFs that maximized the BSS for most locations and times. Since the SSPFs were based on the maximum UH occurring within specified space and time windows, they can be considered smoothed neighborhood-maximum ensemble probabilities (NMEPs), as defined in Schwartz and Sobash (2017). While SSPFs were produced using spatial smoothing of the SSRs, the NNPFs were not spatially smoothed. In fact, spatial smoothing decreased NNPF skill (further discussion of smoothing and its impact on probability magnitudes is provided in the next section). As with the NNPFs, probabilities $< 0.1\%$ were set to 0.0 for the SSPFs and verified using the metrics described in section 2c.

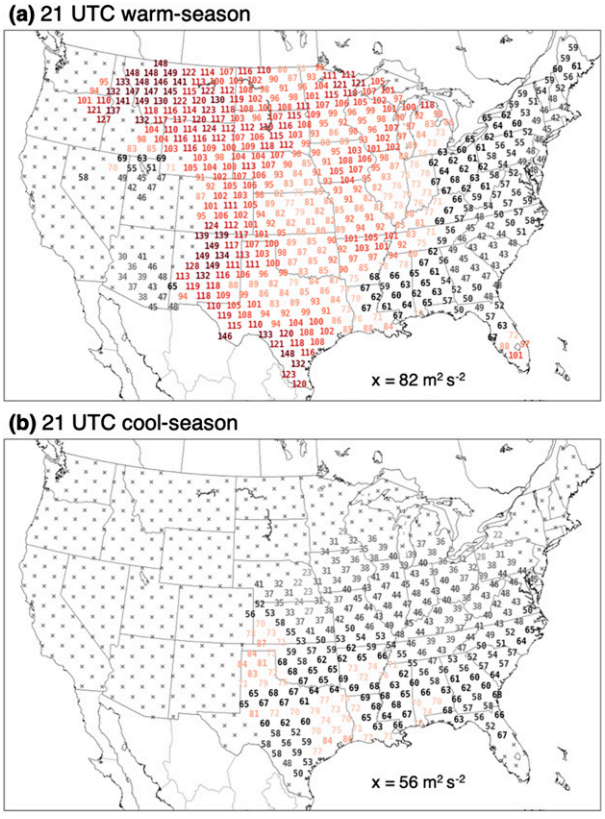

FIG. 3. Optimal UH thresholds for 2100 UTC for the (a) warm season and (b) cool season. Optimal UH thresholds computed as described in the methodology. Locations where fewer than 25 observed severe weather reports occurred (denoted with x) were assigned an optimal UH threshold using a bias of 1 computed with all grid boxes.

3. Comparison of NNPF and SSPF probability distributions

To assess the difference in the probabilities produced by the NNPFs and the SSPFs, the frequency of 40- and

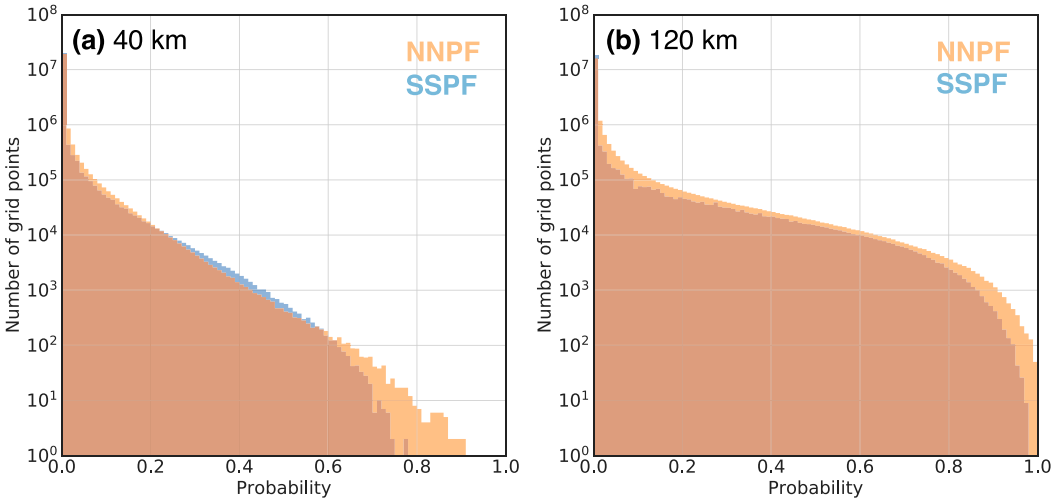

FIG. 4. Histogram of (a) 40- and (b) 120-km NNPF and SSPF probabilities for all grid points, forecast hours, and events.

120-km 2-h forecast probability values over all forecast hours and grid points was examined. While both the 120-km SSPFs and NNPFs covered nearly all probabilities, the 40-km forecasts rarely produced probability values > 75% (Fig. 4). This behavior was expected, since the probabilities in the 40-km forecasts are defined for an event occurring over a smaller spatial scale than in the 120-km forecasts. The most notable differences in the SSPF and NNPF distributions occurred at small and large probability values. The NNPFs produced fewer probability values of zero, and a larger number of grid points with probability values > 70% (Fig. 4). The largest relative differences between the forecasts tended to occur at probabilities between 1% and 20%, as well as probabilities > 70%, where the NNPFs produced as many at 600% more grid points. For example, probability magnitudes >~95% were produced about 10 times as often in the 120-km NNPFs compared to the 120-km SSPFs (Fig. 4b); this was similar in the 40-km forecasts, although at smaller probability values (Fig. 4a). For intermediate probability values between 30% and 60%, the number of grid points were more similar, with the 40-km (120-km) SSPFs producing a slightly larger (smaller) number of grid points than the NNPFs.

The larger number of NNPF probability values > 60%–70% is partly a result of the spatial smoothing used to create the SSPFs, which reduced the magnitude of the SSPF probabilities. While reducing the smoothing length scale would produce SSPFs with a similar number of grid points with forecast probabilities > 60% compared to NNPFs, this generally decreased SSPF skill and produced larger differences within smaller probability ranges (not shown). For the SSPFs, the spatial smoothing procedure is the mechanism that produces the probabilistic uncertainty estimates, based solely on the spatial distribution of UH points exceeding the given threshold. On the other hand, the NNs have learned this uncertainty from previous observations of severe weather events, providing a better representation of the underlying uncertainty within the NNPFs. Thus, it appears that the NNPFs possess probability distributions that have a similar character to the SSPFs, but retain the highest probability magnitudes due to the lack of spatial smoothing. Whether these differences occur concomitant with improved forecast skill will be examined in the next section.

4. Daily verification of NNPFs and SSPFs

Daily BSSs were computed for each event by aggregating the skill of all 4-h forecasts across all forecast hours and grid boxes to isolate events when large skill differences occurred. While the daily skill between the SSPFs and NNPFs was strongly correlated, likely due to both being based on the same underlying WRF forecast, daily NNPF BSSs were consistently larger than the corresponding daily SSPF BSSs for both the 40- and 120-km forecasts, but especially the 120-km forecasts (Fig. 5). In fact, the SSPFs outperformed the NNPFs for only 20 of the events at 120 km (Fig. 5b). The average BSS difference was ~0.04 for the 40-km forecasts, and ~0.11 for the 120-km forecasts. The larger difference in skill at the 120-km length scale may be due to enhanced predictability on large scales, with the 120-km forecasts better able to make use of the large-scale and environmental information compared to the SSPFs (at 40 km, both forecasts may be equally impacted by reduced predictability).

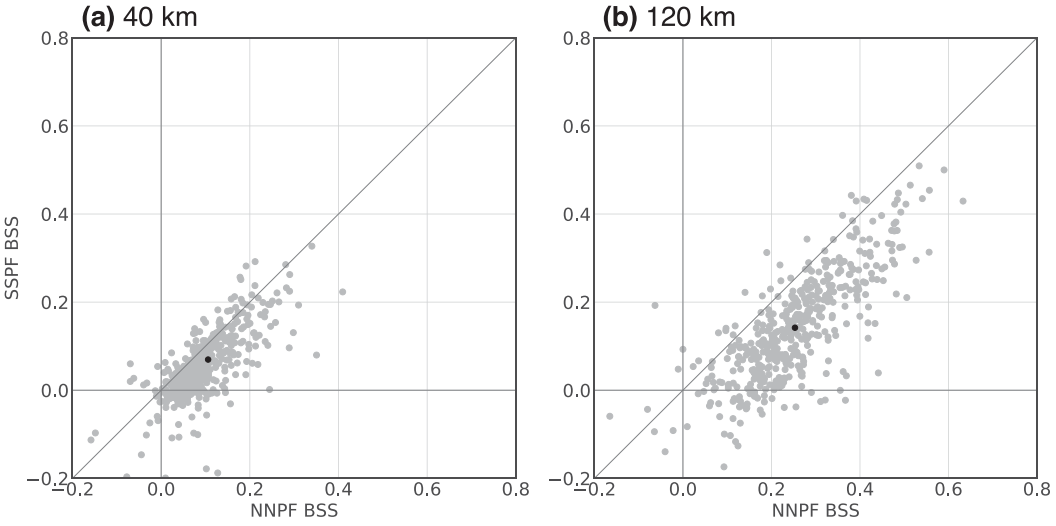

FIG. 5. Scatterplot of 469 daily BSSs for the (a) 40- and (b) 120-km SSPFs and NNPFs shown as filled gray circles. Aggregate BSS for all forecasts shown as filled black circles. The daily BSS was computed with all individual 4-h forecasts for all forecast hours and grid boxes over the CONUS. The BSS was computed using a temporally and spatially varying 30-yr severe weather climatology as the reference forecast, as described in the text.

SSPFs and NNPFs notably differed during an overnight convective event between 0000 and 1200 UTC 20 December 2012. During this event, a convective line formed in eastern Oklahoma and Texas, moving eastward into the southeastern United States, producing 163 wind, 3 tornado, and 6 hail reports (Fig. 6a). The 0000 UTC 19 December 2012 WRF forecast correctly predicted the location and timing of this line ~33 h in advance, although the character of the convective cores was more cellular in the model than in observations (Fig. 6b). Given the limited UH magnitudes (e.g., $<20\,m^2\,s^{-2}$) associated with the convective line, SSPFs were low, with the maximum SSPF magnitudes displaced to the north of where reports were observed, leading to BSSs near zero at both spatial scales (Figs. 7a,b). On the other hand, the NNPF magnitudes were larger with a maximum shifted to the south, in much better agreement with reports, leading to positive BSSs (Figs. 7c,d).

SSPFs were produced for fixed UH thresholds of 10, 20, and $30\,m^2\,s^{-2}$ to test if a threshold determined a posteriori would lead to better guidance rather than relying on the optimal calibrated UH threshold (Fig. 8). For this event, SSPFs using fixed UH thresholds of 10 and $20\,m^2\,s^{-2}$ were more skillful than those using the optimal calibrated UH threshold, but even so, the BSS for these fixed-UH SSPFs remained less than the

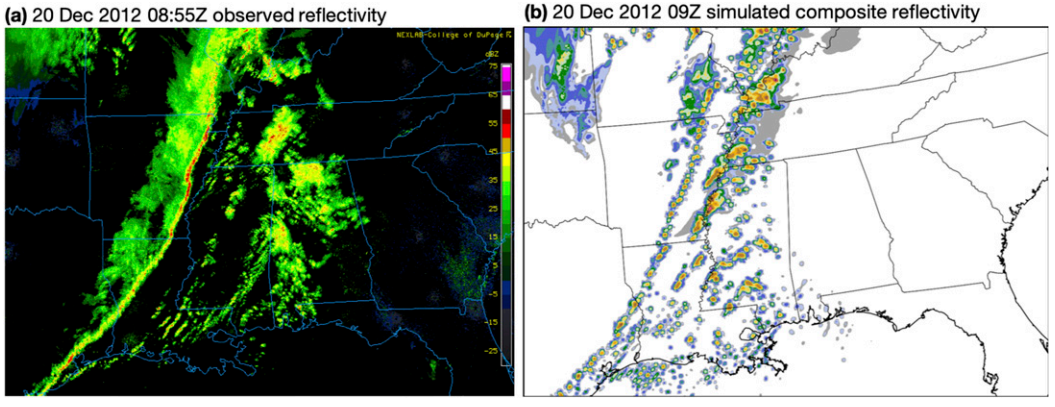

FIG. 6. (a) 0855 UTC observed and (b) 0900 UTC simulated composite reflectivity for the 20 Dec 2012 convective weather event. Simulated reflectivity generated from deterministic WRF 33-h forecast initialized at 0000 UTC 19 Dec 2012.

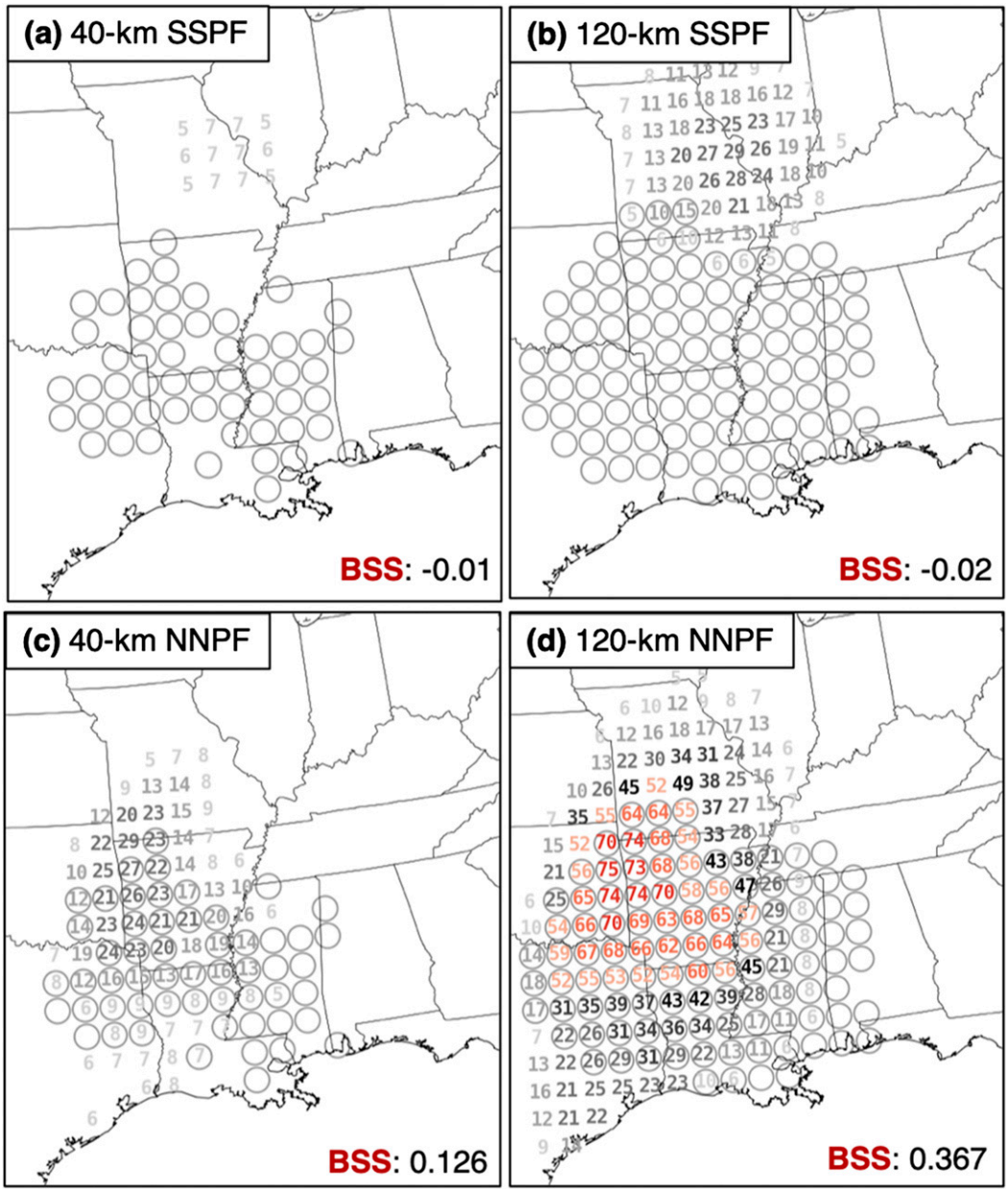

FIG. 7. Maximum 4-h severe hazard (tornado, hail, or wind) probability produced over all forecast hours between 0000 UTC 19 Dec 2012 and 1200 UTC 20 Dec 2012 for (a) 40-km SSPFs, (b) 120-km SSPFs, (c) 40-km NNPFs, and (d) 120-km NNPFs. SSPFs and NNPFs were derived from the deterministic WRF forecast initialized on 0000 UTC 19 Dec 2012. Circles represent grid boxes where at least one severe weather report was received within (left) 40 or (right) 120 km during the 36-h forecast.

NNPFs. That is, even a UH threshold that was chosen after the event, to maximize SSPF skill, was not as skillful as the NNPFs. Additionally, the fixed-UH SSPF probabilities, as well as the optimal UH SSPFs, were shifted spatially with respect to the reports, with a probability maximum near St. Louis, Missouri, while the NNPFs produced a maximum in western Arkansas where many wind reports were observed. For this event, the NNPFs provided added value beyond what was possible by carefully calibrating the UH magnitude, either through the usage of climatological optimal threshold or an a posteriori fixed UH threshold that was selected to maximize forecast skill. Given the challenge of anticipating cool-season severe weather, especially those occurring in environments of limited instability where wind damage is the predominant threat, it is

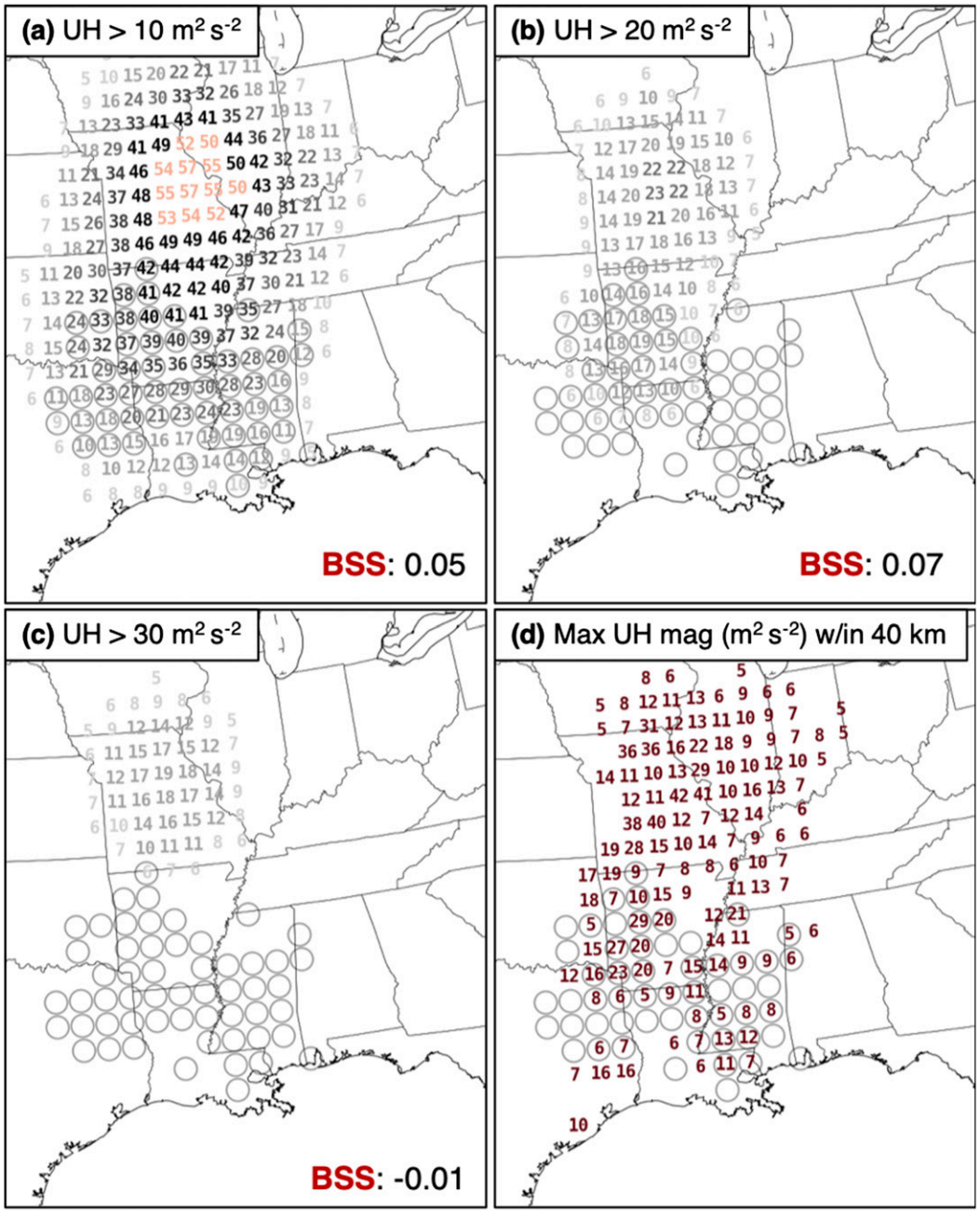

FIG. 8. (a)–(c) As in Fig. 7, but for 40-km SSPFs produced using a fixed UH threshold, rather than the optimal UH threshold, of 10, 20, and 30 $m^2 s^{-2}$ in (a)–(c), respectively. (d) Maximum UH magnitude within 40 km of each 80-km grid box.

promising that the NNPFs were able to provide superior severe weather guidance compared to the SSPFs.

5. Aggregate verification of NNPF and SSPFs

In aggregate, NNPFs produced larger BSS, ROCA, and better reliability than SSPFs at both length scales (Fig. 9). Similar to the daily BSS results, the BSS differences were larger at 120 km than 40 km. At 40-km, NNPFs were more reliable at probabilities > 50%, but still suffered from overforecasting. At 120 km, overforecasting was reduced, and NNPFs produced almost perfect reliability for all probabilities, while SSPFs slightly overforecasted at probabilities between 20% and 50%. To provide additional detail on when and where differences in skill occurred between the NNPFs

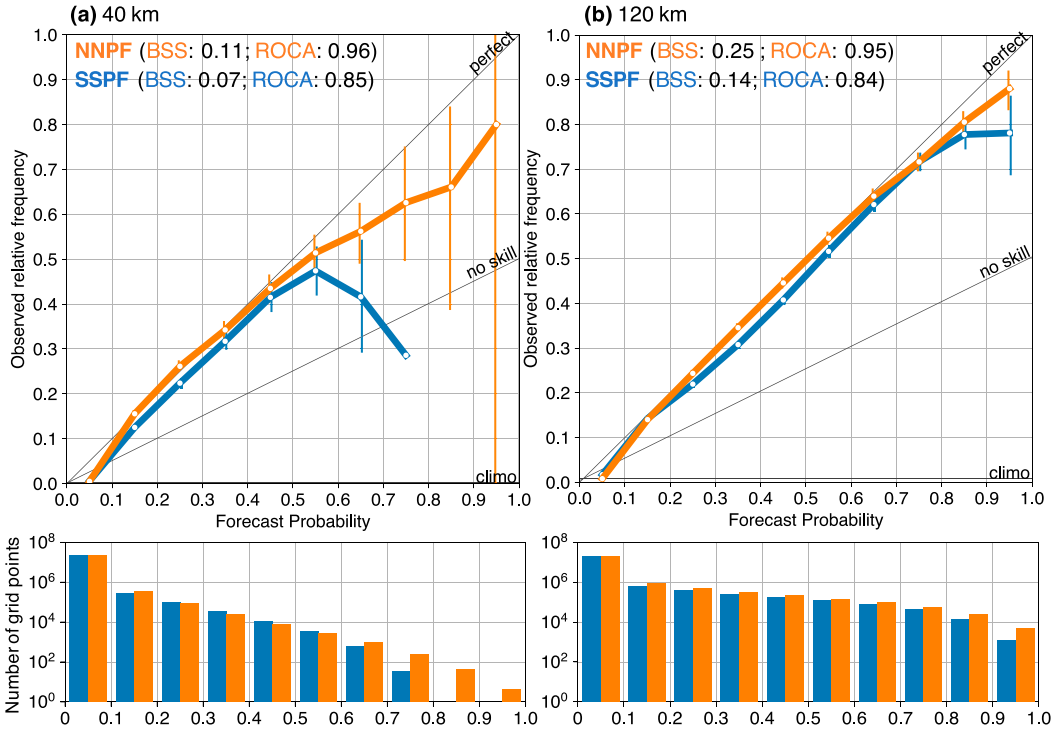

FIG. 9. (top) Reliability diagram and (bottom) frequency histogram for (a) 40- and (b) 120-km NNPFs and SSPFs, aggregated over all forecasts and forecast hours. Vertical lines in the reliability diagram indicate bootstrapped 90% confidence intervals. Bins are 0%–<10%, 10%–<20%, etc.

and SSPFs, the verification results were further aggregated by forecast hour, grid box, and environment.

a. Verification by forecast hour

While both SSPFs and NNPFs possessed skill relative to climatology at all forecast hours, NNPFs statistically significantly outperformed SSPFs for both neighborhood sizes (Fig. 10). Additionally, BSS differences were accompanied by improvements in ROCA, with ROCA as much as 0.2 larger for the NNPFs (Fig. 11). The magnitude of BSS differences between the NNPFs and SSPFs was partly a function of neighborhood size, with less advantage of the NNPFs at 40 km than 120 km (Fig. 10). The scale dependence of the BSS differences could be due to small-scale uncertainty impacting the skill of both the SSPFs and NNPFs at 40 km. Conversely, ROCA magnitudes for both NNPFs and SSPFs were largely insensitive to neighborhood size (Fig. 11).

Magnitudes of the BSS and ROCA differences between NNPFs and SSPFs exhibited some diurnal variability. While the SSPFs and NNPFs had similar diurnal cycle of skill, with BSS maximized during the peak of the diurnal cycle and minimized during the overnight and early morning, the biggest differences in skill occurred early in the forecast, during model spinup (Fig. 10). During the first few hours of integration, no spinup was observed for the NNPFs, with 1–6-h NNPFs exhibiting fairly constant BSS and ROCA, while the BSS and ROCA of the SSPFs increase during the first 6 h as convection and the associated UH field spins up (Figs. 10 and 11). The ability of the ML forecasts to account for spinup during the first few hours of the forecast is useful, likely relying on larger-scale fields and weighting UH and other high-resolution fields less. This hypothesis will be examined in section 6.

Differences in BSS also maximized between forecast hours 18–20, during the peak period of convection initiation (Fig. 10). Here, the NN may have learned biases related to delayed forecast initiation relative to observations. (e.g., Kain et al. 2013). This is also partly reflected in the earlier timing of the peak in skill of the ML forecasts (~2200 UTC), compared to the SSPFs (~0000 UTC). The ROCA differences decreased slightly during this period (Fig. 11), indicating that improvement in skill was mainly related to improved forecast reliability. Future work should investigate the ability of the NNs to adjust for biases in forecast initiation. Finally, ROCA differences were maximized overnight when the ROCA dropped for the SSPFs (Fig. 11), which may be related to issues with UH being a poor predictor of severe weather associated with elevated nocturnal convective systems.

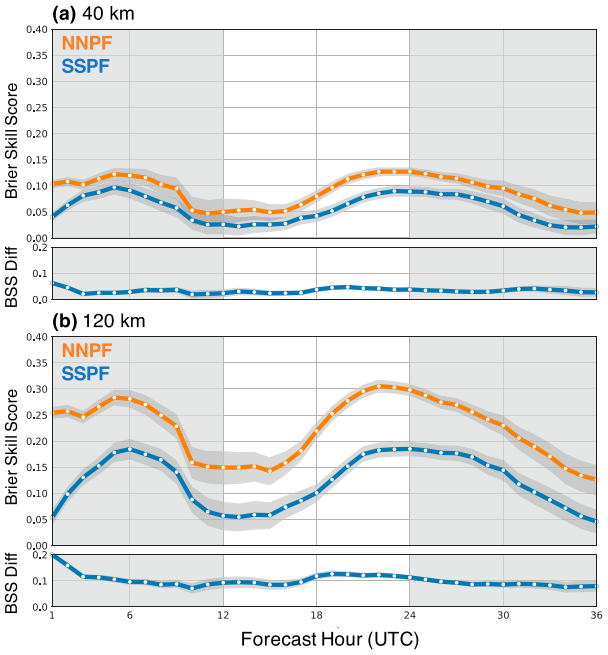

FIG. 10. Brier skill score and BSS difference aggregated by forecast hour for (a) 40- and (b) 120-km SSPFs and NNPFs. Shading along curves represents bootstrapped 90% confidence intervals.

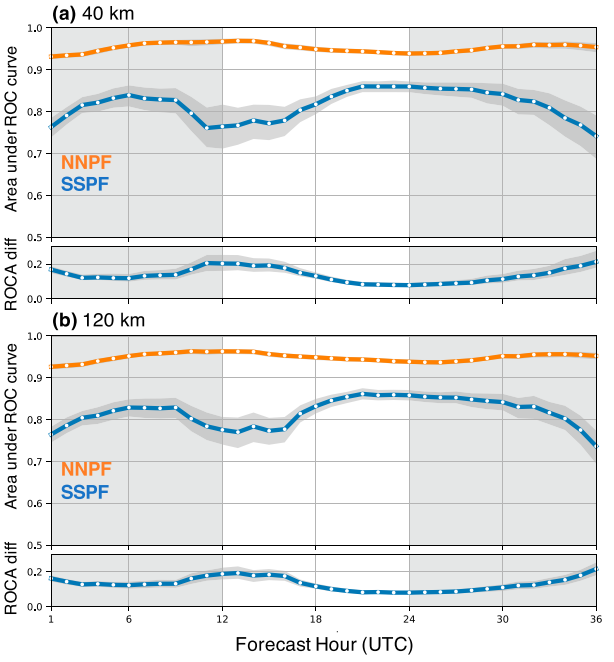

FIG. 11. As in Fig. 10, but for ROCA.

b. *Verification by grid box*

BSSs were computed for each 80-km grid box to reveal spatial variations in forecast skill. For each grid box, forecasts for all forecast hours and the surrounding eight grid boxes were included to increase the sample size. SSPF skill was maximized across the central Plains, with decreased BSS toward the southern and southeastern United States (Figs. 12a,b). On the other hand, the NNPF BSS maxima occurred in the northeastern United States, central Ohio River Valley, and central Plains (Figs. 12c,d). These spatial patterns held for both neighborhood sizes.

Differences in NNPF and SSPF skill were maximized across the eastern United States for both neighborhood sizes (Figs. 12e,f). BSS increases of >0.05 (for the 40-km forecasts; Fig. 12e) and >0.15 (for the 120-km forecasts; Fig. 12f) occurred across much of the southeastern United States and in several areas of the western United States. The smallest BSS differences occurred across the central United States, with several grid boxes in western Nebraska having differences near zero meaning UH alone provided enough information to produce skillful forecasts of severe convective hazards.

Spatial patterns of skill suggest NNPFs can substantially improve upon SSPFs for severe weather prediction across the eastern and western United States, with smaller improvements over the central United States. Across the east, increased NNPF skill is likely due to the decreasing utility of UH as a surrogate for the most common severe weather hazards in that region. Severe reports are often obtained from nonsupercellular convective modes in the eastern CONUS (Ashley et al. 2019). Combined with the abundance of severe wind reports that may not exceed strict severe criteria [i.e., wind gust ≥ 50 kt ($1\text{ kt} \approx 0.51\text{ m s}^{-1}$)], UH alone, even if carefully calibrated, is a poor predictor of whether or not a thunderstorm will produce severe hazards. Across the western CONUS, the relative rarity of severe weather events presents a challenge for NNPFs, given the lack of many training examples over this region. Even so, NNPFs substantially outperform SSPFs in areas where severe weather reports occur with some regularity. One of these areas is in southern Arizona, where severe weather often occurs in association with monsoon thunderstorms during the summer. Many of these events are driven by intense downburst winds in environments with high LCLs, moderate CAPE, and weak deep-layer shear (Carlaw et al. 2017). As in the eastern United States, Arizona events usually do not consist of supercells, thus UH is an insufficient severe weather surrogate. Other diagnostic fields that are incorporated into the NN (e.g., 10-m wind speed, LCL height, etc.) may be providing more useful information in these regions where convective wind reports are common.

Southern Florida and southwest Texas are two regions where NNPFs do not appreciably outperform SSPFs, and in the latter actually underperform SSPFs. In both of these regions, storm reports are rarely

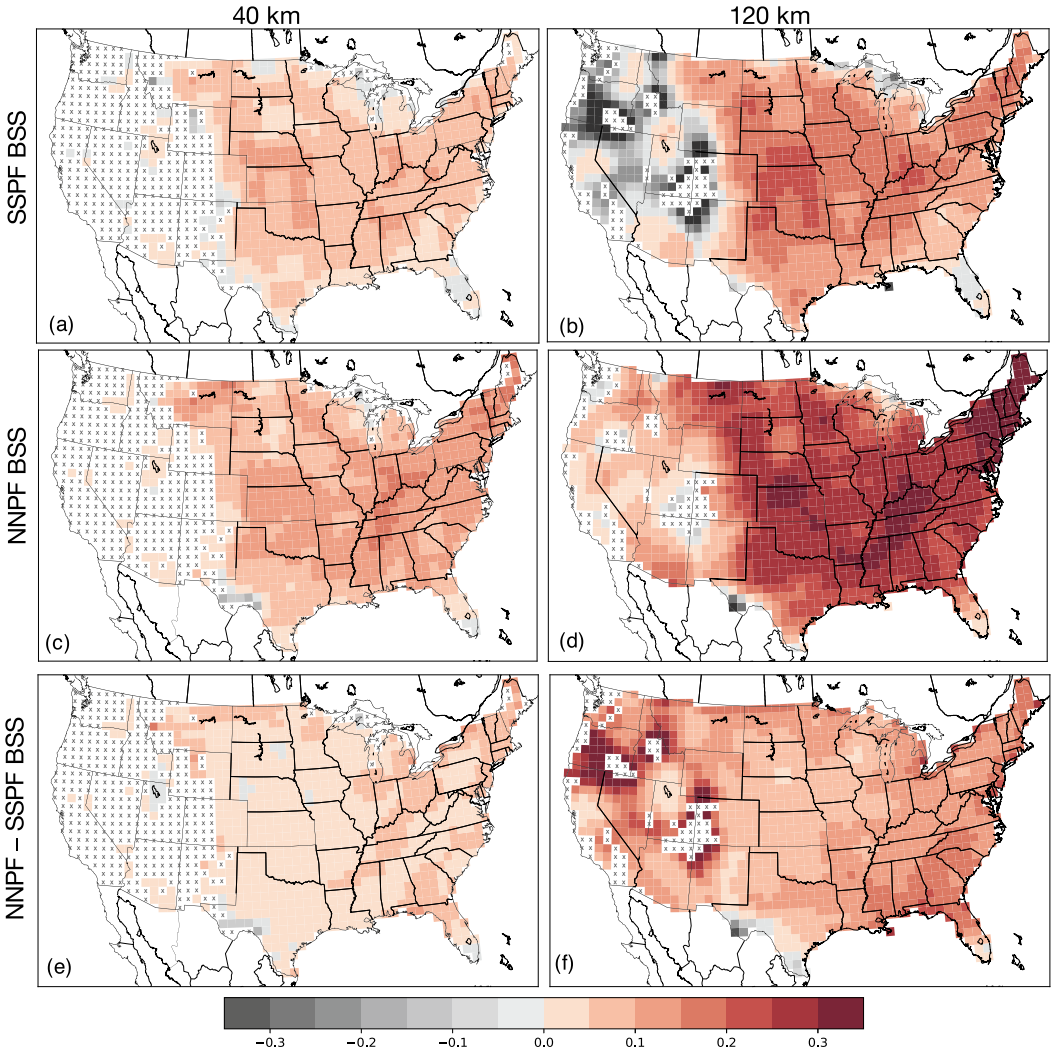

FIG. 12. Brier skill score computed for each 80-km grid box for all (a),(b) SSPFs and (c),(d) NNPFs at the (left) 40- and (right) 120-km length scales. The BSS was computed using all forecasts within one grid box to increase sample size and reduce small-scale spatial variations in the BSS. (e),(f) BSS differences between the NNPF and SSPFs. The BSS is not shown at locations where <25 total observed storm reports occurred within the one gridbox neighborhood, denoted by an "x."

received, although convection is common. As a result of the difficulty of getting storm reports, the optimal UH thresholds are large in both areas during the warm season ($>100\,m^2\,s^{-2}$; Fig. 3a). While the NN should be able to learn that reports do not often occur in these areas (through the latitude and longitude fields) and adjust probabilities accordingly, the relatively small areas where these reporting biases exist may make it difficult for the NN to sufficiently modify forecasts, leading to overpredictions in the NNPFs.

c. *Verification by environment*

To isolate the convective regimes where NNPFs were able to outperform SSPFs, BSS and ROCA values were computed for forecast grid points within specific most-unstable convective available potential energy (MUCAPE) and 0–6-km deep-layer shear (SHR06) bins. The MUCAPE and SHR06 magnitudes used for aggregation were the spatial averages within each 80-km grid box. Verification results using the 120-km neighborhood are provided here; regimes with positive skill were similar when using the 40-km neighborhood, although magnitudes of BSS were reduced.

For both the SSPFs and NNPFs, the maximum BSS occurred in regimes where either moderate amounts of MUCAPE and SHR06, or both, were present (Figs. 13a,c). In regimes with weak SHR06 and low MUCAPE, skill was reduced. In fact, the SSPFs in the

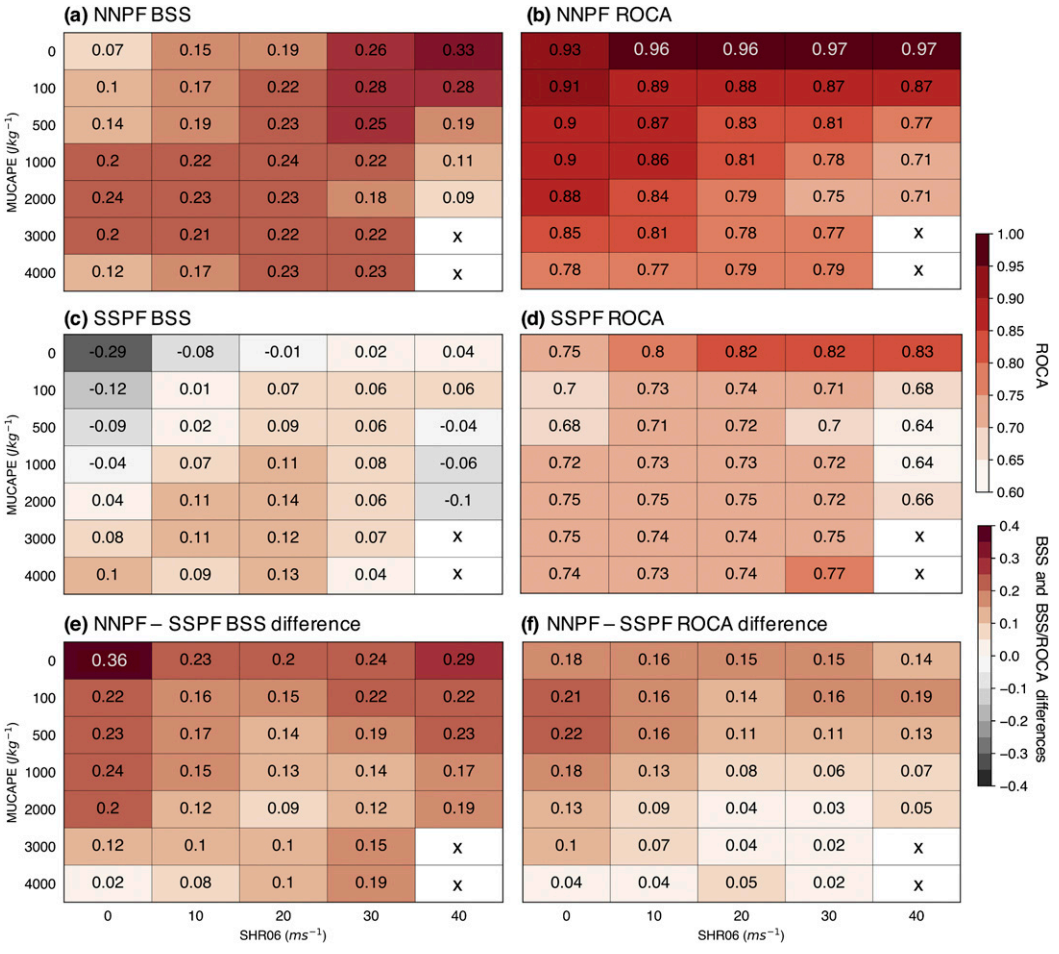

FIG. 13. (left) BSS and (right) ROCA aggregated by MUCAPE and SHR06 magnitudes for the 120-km (a),(b) NNPFs and (c),(d) SSPFs. (e) BSS and (f) ROCA differences between the 120-km NNPFs and 120-km SSPFs. Bin edges for MUCAPE and SHR06 are shown in each panel. Shading indicates larger magnitudes or differences. Scores not shown in bins comprising <100 forecast grid boxes (indicated by "x").

low MUCAPE/low SHR06 regimes performed worse than climatology. The reliance on UH to produce SSPFs is clearly detrimental in these regimes, since UH is typically produced in environments with at least modest MUCAPE and SHR06 (hence, the maxima observed in these regimes). Poor forecast skill also occurred in regimes with very large SHR06 ($>40\,m\,s^{-1}$) and moderate MUCAPE, although the sample size of forecast points in these regimes was small.

ROCA tended to behave opposite to BSS, with ROCA maximized in low MUCAPE/high SHR06 regimes (Figs. 13b,d). This behavior could be due to the underlying predictability of events, since severe weather events occurring with weak MUCAPE and moderate to high SHR06 [i.e., high-shear, low-CAPE (HSLC), regimes; Sherburn and Parker 2014; Sherburn et al. 2016] often occur with robust amounts of large-scale forcing, leading to enhanced predictability. Yet, the slightly reduced BSS magnitudes suggest that the reliability of the SSPFs in the HSLC regime was worse than SSPFs in the moderate-to-high CAPE regimes. It may be that given the enhanced predictability in HSLC regimes, that the smoothing length scale of 160 km for the SSPFs was too broad, leading to reduced reliability.

While positive benefits were noted across the entire MUCAPE/SHR06 phase space, benefits of NNPFs compared to SSPFs were maximized in regimes where UH was a poor predictor, namely in HSLC environments (Figs. 13e,f). For these HSLC grid boxes, improvements in BSS of >0.2 were common, with corresponding large improvements in ROCA (>0.15–0.20). This combination of BSS and ROCA indicates that NNPFs improved the underlying ability to discriminate between severe and nonsevere events in HSLC regimes, which often consist of events that are challenging to anticipate, especially solely with UH. In the "supercell" regime, ROCA differences

were much smaller (e.g., 0.02–0.04 for MUCAPE > 2000 J kg^{-1} and SHR06 > ~20 m s^{-1}), indicating that the NNPFs were not able to improve forecast discrimination as effectively as in the HSLC regime, but since BSS differences were positive, forecast reliability was improved in the NNPFs.

6. Sensitivity of NNPF skill to predictor choices

Given the potential correlation between the 174 predictors, a more limited set of predictors could possibly produce equally skillful forecasts. Reducing the number of predictors is desirable to both minimize the computational burden of training the NNs and to improve interpretation of the trained NNs. Here, four NNs were trained with subsets of the full 174 predictors to determine the role of categories of predictors in producing skillful NNPFs, specifically training only with the midlevel UH predictors (11 predictors; UHonly-NNPF), training without the spatial mean and maximum neighborhood fields (using only the 42 base predictors; NoNeighbor-NNPF), without the explicit convection-related fields (113 predictors; NoExplicit-NNPF), and without the upper-air fields (158 predictors; NoUpperAir-NNPF). All four NNs used the four static predictors (Table 2) and were trained only for the 120-km spatial neighborhood.

The removal of subsets of predictors when training the NNs resulted in NNPFs that were less skillful than the original NNPFs (Fig. 14). The largest reduction in skill occurred for the UHonly-NNPFs, yet the BSS (Fig. 14a) and ROCA (Fig. 14b) were superior to the SSPFs, even though both relied solely on the same midlevel UH diagnostic as input. The skill difference between the UHonly-NNPFs and SSPFs may be due to the improved estimate of uncertainty that is learned by the NNs. Additionally, the NNs do not have a rigid optimal UH threshold and can learn more complex nonlinear relationships between the UH magnitude, static predictors such as latitude and longitude, and the likelihood of severe weather. The UHonly-NNPF for the 20 December 2012 convective event appears much more similar to the SSPFs than the NNPFs (cf. Figs. 15a,b). Thus, given only UH information, the NNs were unable to correctly shift the forecast probabilities southward closer to the observations.

Eliminating the upper-air fields had a negligible impact on the BSS and ROCA, with larger decreases in BSS and ROCA when the neighborhood and explicit diagnostics were removed (Fig. 14). During many forecast hours, especially overnight, the NoNeighbor-NNPFs had a higher BSS than the NoExplicit-NNPFs, even though the NoNeighbor-NNPFs used only 42 predictors compared to 113 for the NoExplicit-NNPFs.

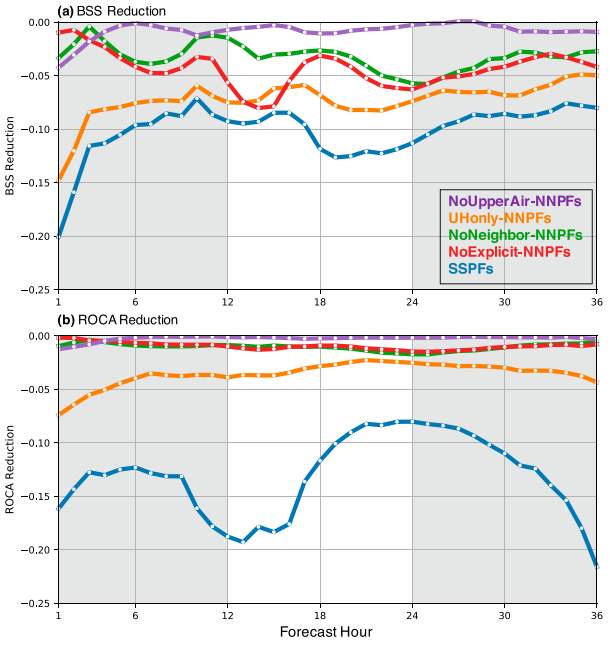

FIG. 14. As in Fig. 10b, but for (a) BSS and (b) ROCA reduction for 120-km NNPFs trained with subsets of the input predictors relative to NNPFs trained with all predictors.

More predictors did not necessarily result in better forecasts, suggesting that overfitting may be an issue for some combinations of predictors. For the 20 December 2012 event, the NoUpperAir-NNPF and the NoExplicit-NNPF were the most skillful of the four reduced-predictor forecasts and was comparable to the NNPF using the full set of predictors, correctly shifting the probabilities southward compared to the UHonly-NNPF and the SSPF. Interestingly, the explicit diagnostics such as UH, updraft speed, etc., were not important for this event, since their inclusion only marginally improved the BSS (Figs. 15a,f). The lack of sensitivity to the explicit predictors for this case may reflect deficiencies in the representation of finescale convective lines during cool-season severe weather events, since most of the explicit diagnostics did not produce robust signatures (not shown), limiting their utility.

Finally, large differences in BSS and ROCA occurred during model spinup (i.e., forecast hours 0–6; Fig. 14). While the NoNeighbor-NNPFs and NoExplicit-NNPFs, produced BSS and ROCA values only slightly smaller than the original NNPFs during these forecast hours, the UHonly-NNPFs and SSPFs both had larger reductions in BSS and ROCA. Additionally, the largest reductions in skill for the NoUpperAir-NNPFs, occurred during model spinup. The environmental information within the NNPFs may be more valuable than the explicit or neighborhood predictors during the first 6 forecast

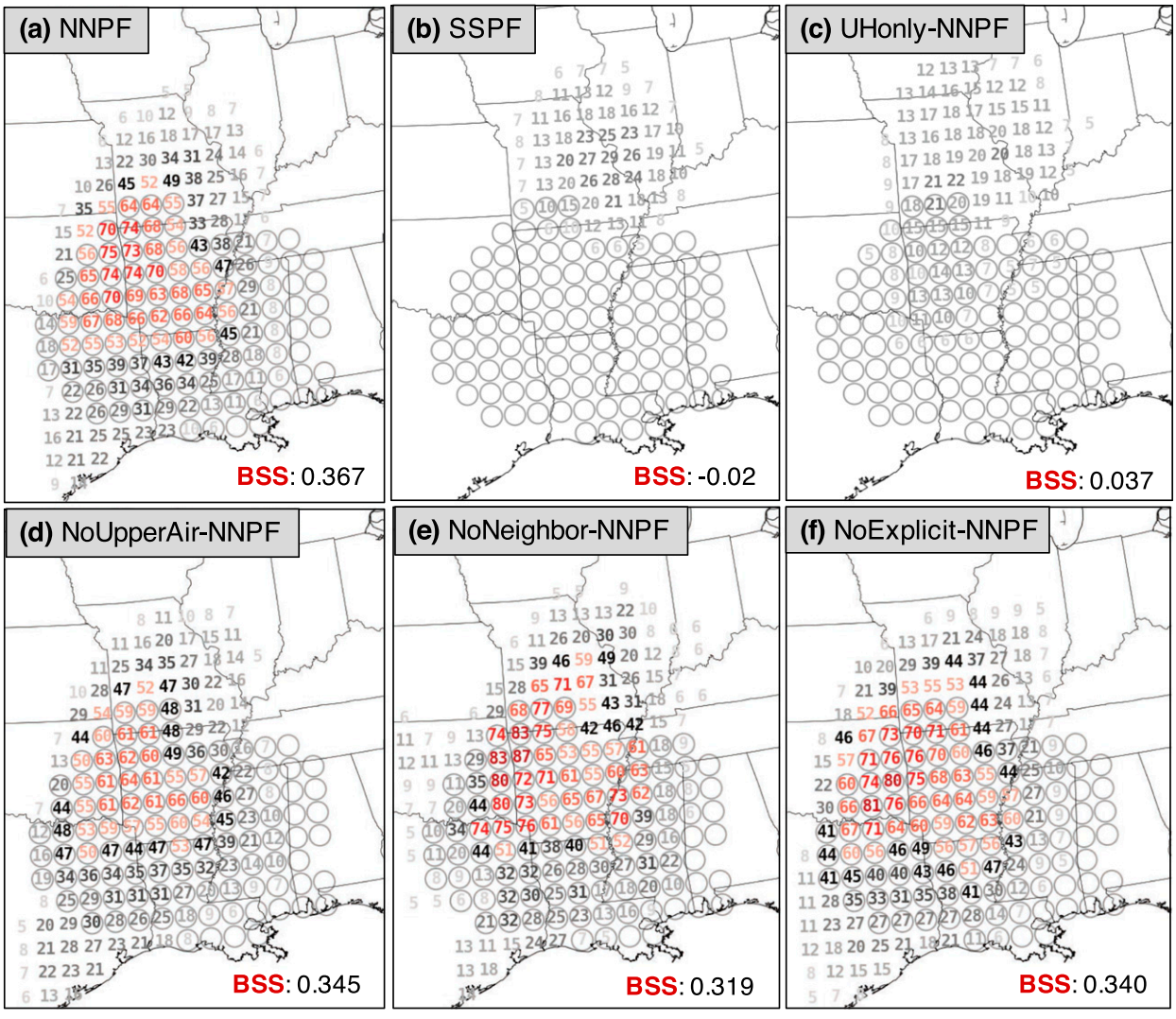

FIG. 15. As in Fig. 8, but for the 120-km (a) NNPF, (b) SSPF, (c) UHonly-NNPF, (d) NoUpperAir-NNPF, (d) NoNeighbor-NNPF, and (e) NoExplicit-NNPF derived from the 0000 UTC 19 Dec 2012 forecast.

hours, since removal of either of these two predictor sets did not result in appreciable reductions in skill. In other words, NNs trained with environmental information alone were able to make skillful predictions during the first few hours, reducing the impact of model spinup that plagued both the SSPFs and the UHonly-NNPFs.

7. Summary

To determine if ML algorithms can improve upon the skill of surrogate-severe guidance based on UH, two sets of probabilistic forecasts of severe weather hazards were generated for 462 severe weather events between 2010 and 2017 over the CONUS using output from deterministic WRF-based CAM forecasts. These included a surrogate-severe UH-based probability forecast (SSPF) and a neural network–based probability forecast (NNPF). The SSPFs were constructed by applying a UH threshold to the hourly maximum midlevel UH field and smoothing the resulting binary output. The SSPF UH threshold was calibrated based on time of day, season, and spatial location to produce the most skillful guidance. The NNPFs were trained with environmental and surrogate diagnostics and designed to predict the probability of any severe weather report occurring within specified time and space windows. Both the SSPFs and NNPFs were generated for 4-h windows, centered on each forecast hour within the 36-h CAM forecast, at two spatial scales (40 and 120 km), and were verified with SPC storm reports using the BSS and ROCA.

In aggregate, NNPFs more frequently produced larger probability values compared to SSPFs, primarily as a

result of the smoothing required to produce SSPFs. These differences in probability distributions were associated with differences in BSS and ROCA, as NNPFs possessed larger BSSs than SSPFs at both spatial scales, for most individual events, and were statistically significantly more skillful at both spatial scales at all forecast hours. ROCA differences were largest overnight, suggesting improved ability to discriminate between nocturnal events in the NNPFs. While BSS differences between the NNPFs and SSPFs were larger for the 120-km forecasts, ROCA differences were not a function of spatial scale, suggesting that the added benefit of NNPFs over SSPFs at 120-km was a function of better calibration and not an inherent difference in the ability to discriminate between events and nonevents. BSS differences were largest during the first hours of the forecast, when model spinup hampered the utility of the UH output, as well as during the beginning of the first diurnal cycle (i.e., 1600–2000 UTC), when NNPFs potentially accounted for convection initiation biases inherent within the SSPFs.

The largest NNPF–SSPF BSS differences occurred in the western and eastern United States, where SSPF skill was reduced. In these regions, severe weather environments are often not supportive of supercells and may involve other modes such as quasi-linear convective systems. The difference in skill as a function of environment was supported when verifying based on convective regime with MUCAPE and SHR06. In environments supportive of supercells, SSPFs produced skillful guidance, and the skill gap between NNPFs and SSPFs was reduced compared to environments not supportive of supercells (i.e., those with small MUCAPE and/or small SHR06 magnitudes). In the nonsupercellular regimes, SSPF skill was poor, and often worse than climatology, while NNPFs were substantially superior to SSPFs. Large improvements in BSS and ROCA also occurred in HSLC environments, where marginal instability and less robust CAM surrogate diagnostic signatures often lead to poor operational forecasts (e.g., Guyer and Jirak 2014).

Finally, sensitivity tests were undertaken to determine the impact of removing various sets of predictors from the NN training. The environmental predictors were more valuable than the explicit predictors during the first few hours of the forecast, when the latter were spinning up, while the explicit predictors were more valuable overnight, potentially providing useful guidance on the longevity of overnight mesoscale convective systems. Finally, NNs trained with only midlevel UH information (and static fields) outperformed SSPFs, suggesting that even without additional diagnostics, NNs can learn useful relationships about forecast uncertainty and the behavior of the UH diagnostic in different seasons and regions better than accounting for these variations by computing optimal UH thresholds.

8. Discussion

Some questions regarding the optimal configuration choices for a ML-based algorithm went unaddressed in this work, including the sensitivity of forecast skill to many of the NN hyperparameters and choice of ML algorithm (e.g., using a random forest [RF; Breiman 2001] instead of a NN). Regarding the optimal NN configuration, the NN model trained with data from 2010 to 2015 was applied to produce real-time forecasts during the spring of 2020. Preliminary results show that this NN configuration remains capable of producing similarly skillful forecasts to those documented here, providing confidence in the ability of the NN configuration to generalize beyond the present training dataset. Regarding the choice of algorithm, we initially trained a RF to produce the forecast probabilities using the same preprocessed input as the NNs; these forecasts were slightly less skillful than the NNPFs, but were still more skillful than the SSPFs (i.e., the RF forecasts were more similar to the NNPFs than the SSPFs). Optimizing the RF hyperparameters (e.g., number of trees) may have produced forecasts with similar skill to the NNPFs, but the NNs were faster to train given the availability of graphics processing units (GPUs) and produced smaller output files, since only the NN weights and biases need to be stored, rather than each decision tree within a RF. Our subjective impression is that the gains in skill achieved by using NNPFs rather than SSPFs appear to be insensitive to the choice of using a RF or NN.

Since the NNPFs are generated from NNs that use observed storm reports, the NNPFs inherit several of the biases that exist within the storm report database. These biases include the presence of an abundance of wind damage reports that are not associated with severe wind gusts, especially in areas of the eastern CONUS (Weiss and Vescio 1998; Weiss et al. 2002; Doswell et al. 2005; Smith et al. 2013; Edwards et al. 2018; Bunkers et al. 2020). For example, the 20 December 2012 event presented in section 4 consisted of mostly estimated wind gusts based on wind damage reports, with only a few measured gusts, although the guidance was accurate in depicting the likelihood of severe weather reports. Such guidance may still be useful for forecasters in anticipating impacts even though wind speeds may have not reached severe criteria. While the usage of an 80-km grid likely reduces some of these overreporting biases, it cannot account for underreporting biases in areas with low population density, such as large swaths of the western CONUS (Weiss et al. 2002). Additionally, our training

dataset does not include many events occurring in the western CONUS, or marginally severe events in other regions. In these scenarios, NNPFs may be prone to overprediction. In the future, we hope to include additional events and verification datasets, such as radar-estimated hail sizes, in the training procedure, and to produce distinct probabilities for measured and estimated wind gusts.

Other simple postprocessing baselines should be considered for comparison to ML-based guidance to justify the added complexity and computational costs of ML algorithms. One example is probabilistic forecasts derived using historical frequencies of reports given explicit and environmental parameters, such as calibrated probabilistic guidance generated by the SPC using output from the NOAA High-Resolution Ensemble Forecast (HREF) and Short-Range Ensemble Forecast (SREF) systems (Jirak et al. 2014). These calibrated forecasts may perform more skillfully than SSPFs since the historical information informs the probability magnitudes, while SSPF magnitudes are solely a function of the spatial density of points where the UH threshold is exceeded within chosen space and time windows.

Given the robust improvements in skill of NNPFs across a broad range of environments, forecast hours, and regions, especially in environments where SSPF skill was poor, this work supports the inclusion of ML-based severe weather guidance in the forecasting process to assist in the identification of severe weather hazards. That said, our forecasts were not designed to mimic current operational forecasting guidance, such as SPC Convective Outlooks, as in prior work that produced and verified SSPFs (e.g., Sobash et al. 2011, 2016; Loken et al. 2017). As implemented here, SSPFs and NNPFs are more aligned with efforts to produce rapidly updating finescale probabilistic hazard guidance, such as that envisioned within the NOAA Forecasting a Continuum of Environmental Threats (FACETs) paradigm (Rothfusz et al. 2018), where the spatial and temporal scales of the guidance may vary based on forecast lead time and the underlying predictability of each hazard. Additionally, future SPC guidance products will likely provide more temporal specificity, and efforts to produce subdaily probabilistic forecasts are underway within NOAA (Krocak and Brooks 2020; I. Jirak 2020, personal communication). Using ML-based algorithms to produce first-guess or final-check guidance products could form the basis for these next-generation probabilistic convective weather postprocessing systems, yet better understanding the internals of the trained NNs will be necessary to elucidate the most important input fields and to reduce their complexity before they are used operationally (e.g., McGovern et al. 2019).

Acknowledgments. We thank David Gagne, Adam Clark, Eric Loken, and Dave Ahijevych for helping to refine early versions of this work. Additionally, we thank Victor Gensini and two anonymous reviewers, whose feedback improved the manuscript. This work was partially supported by NOAA OAR Grants NA17OAR4590114 and NA19OAR4590128, as well as the NCAR Short-term Explicit Prediction program. We would like to acknowledge high-performance computing support from Cheyenne (Computational and Information Systems Laboratory 2019) provided by NCAR's Computational and Information Systems Laboratory. The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Data availability statement. The data that support the findings of this work are available from the corresponding author upon request.

REFERENCES

Adams-Selin, R. D., A. J. Clark, C. J. Melick, S. R. Dembek, I. L. Jirak, and C. L. Ziegler, 2019: Evolution of WRF-HAILCAST during the 2014–2016 NOAA/Hazardous Weather Testbed spring forecasting experiments. *Wea. Forecasting*, **34**, 61–79, https://doi.org/10.1175/WAF-D-18-0024.1.

Ashley, W. S., A. M. Haberlie, and J. Strohm, 2019: A climatology of quasi-linear convective systems and their hazards in the United States. *Wea. Forecasting*, **34**, 1605–1631, https://doi.org/10.1175/WAF-D-19-0014.1.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, https://doi.org/10.1023/A:1010933404324.

Bunkers, M. J., S. R. Fleegel, T. Grafenauer, C. J. Schultz, and P. N. Schumacher, 2020: Observations of hail-wind ratios from convective storm reports across the continental United States. *Wea. Forecasting*, **35**, 635–656, https://doi.org/10.1175/WAF-D-19-0136.1.

Burke, A., N. Snook, D. J. Gagne II, S. McCorkle, and A. McGovern, 2020: Calibration of machine learning-based probabilistic hail predictions for operational forecasting. *Wea. Forecasting*, **35**, 149–168, https://doi.org/10.1175/WAF-D-19-0105.1.

Carlaw, L. B., A. E. Cohen, and J. W. Rogers, 2017: Synoptic and mesoscale environment of convection during the North American monsoon across central and southern Arizona. *Wea. Forecasting*, **32**, 361–375, https://doi.org/10.1175/WAF-D-15-0098.1.

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

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, and D. T. Lindsay, 2014: An empirical model for assessing the severe weather potential of developing convection. *Wea. Forecasting*, **29**, 639–653, https://doi.org/10.1175/WAF-D-13-00113.1.

Clark, A. J., 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, https://doi.org/10.1175/WAF-D-12-00038.1.

Computational and Information Systems Laboratory, 2019: Cheyenne: HPE/SGI ICE XA System (NCAR Community Computing).

National Center for Atmospheric Research, Boulder, CO, accessed 1 May 2020, doi:10.5065/D6RX99HX.

Doswell, C. A., III, H. E. Brooks, and M. P. Kay, 2005: Climatological estimates of daily local nontornadic severe thunderstorm probability for the United States. *Wea. Forecasting*, **20**, 577–595, https://doi.org/10.1175/WAF866.1.

Edwards, R., J. T. Allen, and G. W. Carbin, 2018: Reliability and climatological impacts of convective wind estimations. *J. Climate Appl. Meteor.*, **57**, 1825–1845, https://doi.org/10.1175/JAMC-D-17-0306.1.

Gagne, D. J., II, A. McGovern, J. B. Basara, and R. A. Brown, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

——, S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Mon. Wea. Rev.*, **147**, 2827–2845, https://doi.org/10.1175/MWR-D-18-0316.1.

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

——, ——, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2018: Blended probabilistic tornado forecasts: Combining climatological frequencies with NSSL-WRF ensemble forecasts. *Wea. Forecasting*, **33**, 443–460, https://doi.org/10.1175/WAF-D-17-0132.1.

——, and Coauthors, 2019a: Initial development and testing of a convection-allowing model scorecard. *Bull. Amer. Meteor. Soc.*, **100**, ES367–ES384, https://doi.org/10.1175/BAMS-D-18-0218.1.

——, A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2019b: Incorporating UH occurrence time to ensemble-derived tornado probabilities. *Wea. Forecasting*, **34**, 151–164, https://doi.org/10.1175/WAF-D-18-0108.1.

Guyer, J. L., and I. L. Jirak, 2014: The utility of convection-allowing ensemble forecasts of cool season severe weather events from the SPC perspective. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 37, https://ams.confex.com/ams/27SLS/webprogram/Paper254640.html.

Herman, G. R., and R. S. Schumacher, 2018: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

Iacono, M. J., J. S. Delamere, E. J. Mlawer, M. W. Shephard, S. A. Clough, and W. D. Collins, 2008: Radiative forcing by long-lived greenhouse gases: Calculations with the AER radiative transfer models. *J. Geophys. Res.*, **113**, D13103, https://doi.org/10.1029/2008JD009944.

Janjić, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.

——, 2001: Nonsingular implementation of the Mellor-Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, 61 pp., http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf.

Jirak, I. L., C. J. Melick, and S. J. Weiss, 2014: Combining probabilistic ensemble information from the environment with simulated storm attributes to generate calibrated probabilities of severe weather hazards. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 2.5, https://ams.confex.com/ams/27SLS/webprogram/Paper254649.html.

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, https://doi.org/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, https://doi.org/10.1175/2010WAF2222430.1.

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

Krocak, M. J., and H. E. Brooks, 2020: An analysis of subdaily severe thunderstorm probabilities for the United States. *Wea. Forecasting*, **35**, 107–112, https://doi.org/10.1175/WAF-D-19-0145.1.

Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine learning for real-time prediction of damaging straight-line convective wind. *Wea. Forecasting*, **32**, 2175–2193, https://doi.org/10.1175/WAF-D-17-0038.1.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2017: Comparison of next-day probabilistic severe weather forecasts from coarse- and fine-resolution CAMs and a convection-allowing ensemble. *Wea. Forecasting*, **32**, 1403–1421, https://doi.org/10.1175/WAF-D-16-0200.1.

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

——, and G. J. Stumpf, 1998: A neural network for damaging wind prediction. *Wea. Forecasting*, **13**, 151–163, https://doi.org/10.1175/1520-0434(1998)013<0151:ANNFDW>2.0.CO;2.

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

McGovern, A., K. L. Elmore, D. J. Gagne II, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith, and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision-making for high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

——, R. Lagerquist, D. J. Gagne II, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

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

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res.*, **102**, 16 663–16 682, https://doi.org/10.1029/97JD00237.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Potvin, C. K., and Coauthors, 2019: Systematic comparison of convection-allowing models during the 2017 NOAA HWT spring forecasting experiment. *Wea. Forecasting*, **34**, 1395–1416, https://doi.org/10.1175/WAF-D-19-0056.1.

Robinson, E. D., R. J. Trapp, and M. E. Baldwin, 2013: The geospatial and temporal distributions of severe thunderstorms from high-resolution dynamical downscaling. *J. Appl. Meteor. Climatol.*, **52**, 2147–2161, https://doi.org/10.1175/JAMC-D-12-0131.1.

Rothfusz, L. P., R. Schneider, D. Novak, K. Klockow-McClain, A. E. Gerard, C. Karstens, G. J. Stumpf, and T. M. Smith, 2018: FACETs: A proposed next-generation paradigm for

high-impact weather forecasting. *Bull. Amer. Meteor. Soc.*, **99**, 2025–2043, https://doi.org/10.1175/BAMS-D-16-0100.1.

Schaefer, J. T., and R. Edwards, 1999: The SPC tornado/severe thunderstorm database. Preprints, *11th Conf. on Applied Climatology*, Dallas, TX, Amer. Meteor. Soc., 215–220.

Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

——, and ——, 2019: Revisiting sensitivity to horizontal grid spacing in convection-allowing models over the central and eastern United States. *Mon. Wea. Rev.*, **147**, 4411–4435, https://doi.org/10.1175/MWR-D-19-0115.1.

Sherburn, K. D., and M. D. Parker, 2014: Climatology and ingredients of significant severe convection in high-shear, low-CAPE environments. *Wea. Forecasting*, **29**, 854–877, https://doi.org/10.1175/WAF-D-13-00041.1.

——, ——, J. R. King, and G. M. Lackmann, 2016: Composite environments of severe and nonsevere high-shear, low-CAPE convective events. *Wea. Forecasting*, **31**, 1899–1927, https://doi.org/10.1175/WAF-D-16-0086.1.

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

Smith, B. T., T. E. Castellanos, A. C. Winters, C. M. Mead, A. R. Dean, and R. L. Thompson, 2013: Measured severe convective wind climatology and associated convective modes of thunderstorms in the contiguous United States, 2003–09. *Wea. Forecasting*, **28**, 229–236, https://doi.org/10.1175/WAF-D-12-00096.1.

Sobash, R. A., and J. S. Kain, 2017: Seasonal variations in severe weather forecast skill in an experimental convection-allowing model. *Wea. Forecasting*, **32**, 1885–1902, https://doi.org/10.1175/WAF-D-17-0043.1.

——, ——, 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, https://doi.org/10.1175/WAF-D-10-05046.1.

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

——, C. S. Schwartz, M. L. Weisman, and G. S. Romine, 2019: Next-day prediction of tornadoes using convection-allowing models with 1-km grid spacing. *Wea. Forecasting*, **34**, 1117–1135, https://doi.org/10.1175/WAF-D-19-0044.1.

Tegen, I., P. Hollrig, M. Chin, I. Fung, D. Jacob, and J. Penner, 1997: Contribution of different aerosol species to the global aerosol extinction optical thickness: Estimates from model results. *J. Geophys. Res.*, **102**, 23 895–23 915, https://doi.org/10.1029/97JD01864.

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

Weiss, S. J., and M. D. Vescio, 1998: Severe local storm climatology 1955–1996: Analysis of reporting trends and implications for NWS operations. Preprints, *19th Conf. on Severe Local Storms*, Minneapolis, MN, Amer. Meteor. Soc., 536–539.

——, J. A. Hart, and P. R. Janish, 2002: An examination of severe thunderstorm wind report climatology: 1970–1999. Preprints, *21st Conf. on Severe Local Storms*, San Antonio, TX, Amer. Meteor. Soc., 446–449.

Wendt, N. A., I. L. Jirak, and C. J. Melick, 2016: Verification of severe weather proxies from the NSSL-WRF for hail forecasting. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 110, https://ams.confex.com/ams/28SLS/webprogram/Manuscript/Paper300913/nawendt_sls28_ext_abstract.pdf.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.