Generating Probabilistic Next-Day Severe Weather Forecasts from Convection-Allowing Ensembles Using Random Forests

ERIC D. LOKEN

*Cooperative Institute for Mesoscale Meteorological Studies, and School of Meteorology, University of Oklahoma,
Norman, Oklahoma*

ADAM J. CLARK

School of Meteorology, University of Oklahoma, and NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma

CHRISTOPHER D. KARSTENS

NOAA/OAR/National Severe Storms Laboratory, and NOAA/NWS/Storm Prediction Center, Norman, Oklahoma

(Manuscript received 30 December 2019, in final form 18 May 2020)

ABSTRACT

Extracting explicit severe weather forecast guidance from convection-allowing ensembles (CAEs) is challenging since CAEs cannot directly simulate individual severe weather hazards. Currently, CAE-based severe weather probabilities must be inferred from one or more storm-related variables, which may require extensive calibration and/or contain limited information. Machine learning (ML) offers a way to obtain severe weather forecast probabilities from CAEs by relating CAE forecast variables to observed severe weather reports. This paper develops and verifies a random forest (RF)-based ML method for creating day 1 (1200–1200 UTC) severe weather hazard probabilities and categorical outlooks based on 0000 UTC Storm-Scale Ensemble of Opportunity (SSEO) forecast data and observed Storm Prediction Center (SPC) storm reports. RF forecast probabilities are compared against severe weather forecasts from calibrated SSEO 2–5-km updraft helicity (UH) forecasts and SPC convective outlooks issued at 0600 UTC. Continuous RF probabilities routinely have the highest Brier skill scores (BSSs), regardless of whether the forecasts are evaluated over the full domain or regional/seasonal subsets. Even when RF probabilities are truncated at the probability levels issued by the SPC, the RF forecasts often have BSSs better than or comparable to corresponding UH and SPC forecasts. Relative to the UH and SPC forecasts, the RF approach performs best for severe wind and hail prediction during the spring and summer (i.e., March–August). Overall, it is concluded that the RF method presented here provides skillful, reliable CAE-derived severe weather probabilities that may be useful to severe weather forecasters and decision-makers.

1. Introduction

With horizontal grid spacing less than approximately 4 km, convection-allowing models (CAMs) are important tools for severe weather forecasters, since they adequately resolve the dominant circulations of individual convective storms without convective parameterization (e.g., Weisman et al. 1997; Done et al. 2004). As a result, CAMs more accurately predict storm initiation, evolution, intensity, and mode compared to convection-parameterizing models (e.g., Kain et al. 2006, 2008). Depiction of storm mode is especially useful to severe

weather forecasters (e.g., Kain et al. 2006; Clark et al. 2012a) since a storm's morphology is related to its attendant hazards (e.g., Gallus et al. 2008; Duda and Gallus 2010; Schoen and Ashley 2011; Smith et al. 2012). However, CAMs currently lack horizontal grid spacing fine enough to explicitly simulate individual tornadoes, hailstones, or microscale severe wind events. Therefore, forecasters using CAM guidance must infer simulated severe weather occurrence from modeled storm attributes that are correlated with observed severe weather (e.g., Sobash et al. 2011).

An example of a commonly used simulated severe storm “surrogate” (Sobash et al. 2011, 2016, 2019), or proxy, is hourly maximum 2–5 km above ground level

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

DOI: 10.1175/WAF-D-19-0258.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](https://www.ametsoc.org/PUBSReuseLicenses) (www.ametsoc.org/PUBSReuseLicenses).

updraft helicity (hereafter, UH; e.g., [Kain et al. 2008, 2010](#); [Guyer and Jirak 2014](#); [Loken et al. 2017](#); [Sobash et al. 2011, 2016, 2019](#)). Large values of UH identify not only rotating updrafts associated with supercells, but also the sheared updrafts associated with severe mesoscale convective systems (MCSs; [Sobash et al. 2011](#)). As a result, UH has been found to be a skillful predictor of all-hazards severe weather ([Kain et al. 2008](#); [Sobash et al. 2011, 2016, 2019](#)). UH has also been used—generally in conjunction with simulated environmental variables—to forecast tornadoes ([Clark et al. 2013](#); [Guyer and Jirak 2014](#); [Gallo et al. 2016](#); [Sobash et al. 2019](#)) and severe wind and hail ([Jirak et al. 2014](#)). Other common simulated severe weather proxies include large values of hourly maximum upward vertical velocity (e.g., [Roberts et al. 2019](#)), low-level vertical vorticity (e.g., [Skinner et al. 2016](#); [Sobash et al. 2019](#)) and UH integrated from 0 to 1 km above the surface ([Sobash et al. 2019](#)).

One major drawback of these proxies is that they require extensive calibration to perform optimally. For example, [Sobash and Kain \(2017\)](#) demonstrated that the best UH threshold to use for all-hazards severe weather prediction varies by location and time of year. Moreover, if binary proxies are smoothed spatially to obtain probabilistic forecasts (e.g., [Sobash et al. 2011, 2016, 2019](#); [Loken et al. 2017](#)), the degree of spatial smoothing must be properly calibrated as well. Too little smoothing results in overforecasting bias, while too much can yield underforecasting and degrade sharpness and resolution (e.g., [Sobash et al. 2011, 2016](#); [Loken et al. 2017, 2019a,b](#)). Additionally, these calibrations are CAM and hazard dependent. For example, [Clark et al. \(2012b, 2013\)](#) used a larger UH threshold and smaller degree of spatial smoothing to forecast tornado pathlengths compared to that used by [Sobash et al. \(2011\)](#) to forecast all-hazards severe weather, while [Gagne et al. \(2017\)](#) used different UH thresholds to predict 25- and 50-mm diameter hail.

Another important drawback of simulated severe weather proxies is that they use limited information to determine the severe weather threat. For example, [Clark et al. \(2012b\)](#) and [Gallo et al. \(2016\)](#) noted that large values of UH may exist in environments that are not conducive to severe weather. However, even when proxies are filtered based on the simulated environment (e.g., [Clark et al. 2012b](#); [Jirak et al. 2014](#); [Gallo et al. 2016](#)), the resulting predictions may still be suboptimal since severe weather can still occur in locations with unfavorable simulated environments if the CAM has biases or is not representing the observed environment well. Moreover, the use of environment-based filtering does not mean the resulting prediction has considered all relevant forecast variables.

Another way to extract explicit severe weather guidance from CAMs is to statistically relate multivariate CAM output with the observed occurrence of severe weather. Indeed, this is the general approach of Model Output Statistics (MOS; [Glahn and Lowry 1972](#); [Klein and Glahn 1974](#)), which has shown promise for a variety of forecast fields, including: probability of precipitation, maximum and minimum temperatures, cloud coverage, near-surface wind, conditional probability of precipitation, and thunderstorms (e.g., [Glahn and Lowry 1972](#); [Klein and Glahn 1974](#); [Carter 1975](#); [Bermowitz 1975](#); [Schmeits et al. 2005](#); [Kang et al. 2011](#)). However, MOS relationships tend to be linear and based on regression while relationships between CAM forecast variables and observed severe weather are likely to be flow-dependent and nonlinear (e.g., [Legg and Mylne 2004](#); [Melhauser and Zhang 2012](#); [Torn and Romine 2015](#); [Trier et al. 2015](#)). Thus, machine learning (ML) techniques, which can model nonlinear relationships, may be more appropriate for diagnosing the severe weather threat conveyed by CAM or convection-allowing ensemble (CAE) guidance.

Indeed, recent studies have successfully used ML techniques to create probabilistic precipitation (e.g., [Gagne et al. 2014](#); [Herman and Schumacher 2018](#); [Loken et al. 2019a](#)) and severe weather (e.g., [Gagne et al. 2017](#); [Lagerquist et al. 2017](#); [Burke et al. 2020](#)) forecasts based partly or entirely on numerical weather prediction (NWP) predictors. For severe weather prediction, a common approach has been to use predictors associated with storm “objects,” which are identified by thresholding a certain simulated storm attribute (e.g., maximum hourly column total graupel mass in [Gagne et al. 2017](#); maximum hourly upward vertical velocity in [Burke et al. 2020](#)). Thus, the object identification process “filters out” areas of weaker or nonexistent simulated storms. Such an approach is efficient for ML training since it eliminates the need to consider predictors from all grid points but can underperform if there is poor correspondence between simulated and observed storms ([Gagne et al. 2017](#)). Conversely, when gridpoint-based predictors are used, training takes longer, but higher performance may be achieved when the CAE is imperfect, since the gridpoint predictors offer the ML algorithm more (and more relevant) information. Moreover, when gridpoint-based predictors and predictands are used, output probabilities are directly given in two-dimensional (rather than object) space, facilitating user interpretation of ML output.

While gridpoint-based methods have been used to obtain skillful probabilistic precipitation forecasts ([Herman and Schumacher 2018](#); [Loken et al. 2019a](#)), they are untested for severe weather prediction. Therefore, this study seeks to develop and evaluate an RF-based method

TABLE 1. SSEO initialization dates.

Month	2015	2016	2017	Total
January	—	1–9, 11–13, 15–31	3–27, 29–31	57
February	—	1–29	2–21, 23, 25–26, 28	53
March	—	1–7, 9–11, 13–19, 21–22, 25–31	1–7, 9–11, 13–20, 28–31	48
April	21–26, 28–30	1–10, 12–30	2–4, 6–27, 29–30	65
May	1–10, 12–19, 23–31	1, 3–10, 13–31	1–9, 12, 14–31	83
June	1–9, 11, 13–19, 21–30	1–8, 10–20, 22–23, 25–30	1–6, 8–11, 13–14, 16, 18–22, 24	73
July	1–6, 8–10, 12–16, 20–29	2, 8–14, 16–20, 23–26, 28, 30–31	1–4	48
August	1–3, 5–9, 11–14, 18–31	1–4, 6–10, 12–31	—	55
September	1–9, 11, 14, 19–21, 23–30	1, 3, 5–6, 9–13, 17–22, 24–26, 28	—	41
October	1–6, 9–31	1–2, 4–12, 14–18, 21–24, 27–31	—	54
November	1–3, 5–14, 16–18, 20–23, 25, 27–30	—	—	25
December	1, 4–8, 10, 12–31	—	—	27
Total	216	261	152	629

for creating individual-hazard day 1 (i.e., 1200–1200 UTC) severe weather probabilities from gridpoint-based CAE forecast output. Due to its skill (Jirak et al. 2016, 2018) and long data archive, the SPC's 7-member Storm-Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012, 2016, 2018) is used as the underlying dynamical forecast system. For evaluation against operationally relevant baselines, the RF-based severe weather forecasts are compared to SSEO UH-based probabilistic forecasts and SPC day 1 convective outlooks (COs) issued at 0600 UTC.¹ While multiple previous studies have applied ML to severe weather prediction, the RF method described herein is unique in that it uses gridpoint-based CAE forecast fields as predictors, produces probabilistic forecasts for multiple severe weather hazards over the full contiguous United States (CONUS), and is directly evaluated against top-performing human and NWP baselines.

The remainder of the paper is organized as follows: section 2 describes the methods, section 3 presents the results, section 4 analyzes two representative case studies, section 5 discusses key aspects of the results, and section 6 summarizes and concludes the paper.

2. Methods

a. Datasets

The forecast and observational datasets used herein span 629 days from late April 2015 to early July 2017 (Table 1). RF- and UH-based severe weather forecasts are derived from the SSEO (Jirak et al. 2012, 2016), a 7-member CAE with members that use different initial

and lateral boundary conditions, initialization times, and microphysics and turbulence parameterizations. Since SPC forecasters began using the SSEO in 2011 (Jirak et al. 2016), its convection-related forecasts have compared favorably with those from other experimental CAEs (Jirak et al. 2016). As a result, the SSEO was ultimately formalized as the High-Resolution Ensemble Forecast System Version 2 (HREFv2), which became the first operational CAE run by the National Oceanic and Atmospheric Administration's (NOAA's) Environmental Modeling Center in November 2017 (Jirak et al. 2018; Roberts et al. 2019; Loken et al. 2019b). All SSEO member forecasts are provided on a 4-km contiguous United States (CONUS) domain with 1199×799 points. Full SSEO specifications are summarized in Table 2.

SSEO forecasts are compared against SPC day 1 COs, which are issued daily by 0600 UTC and are valid from 1200 to 1200 UTC the following day. These COs include probabilistic forecasts of tornadoes, severe wind [i.e., wind speeds of at least 50 kt (58 mph)], and severe hail (i.e., a maximum hailstone diameter of 1 in. or greater), with probabilities valid for within 25 mi of a point (about a 40-km radius). The COs also denote locations with a 10% or greater probability of observing significant tornadoes [i.e., those with an enhanced Fujita (EF) rating of 2 or higher], significant severe wind [i.e., wind speeds at least 65 kt (75 mph)], and significant severe hail (i.e., a maximum hailstone diameter of 2 in. or greater) within 25 mi. Individual hazard probabilities are then used to determine a categorical outlook forecast based on the criteria in Table 3.

One limitation of the SPC day 1 COs is that only certain probability levels (i.e., 2%, 5%, 10%, 15%, 30%, 45%, and 60% for tornadoes; 5%, 15%, 30%, 45%, and 60% for severe wind and hail; and 10% for significant severe weather) are contoured. As a result, it is difficult

¹ 0600 UTC SPC COs are used because SPC forecasters, like the RFs, have access to 0000 UTC SSEO guidance during that forecast period.

TABLE 2. SSEO member specifications. Dynamic cores include those from the Advanced Research version of the Weather Research and Forecasting Model (WRF-ARW; Skamarock et al. 2008), the Weather Research and Forecasting Nonhydrostatic Mesoscale Model (WRF-NMM; Janjić et al. 2001; Janjić 2003), and the Nonhydrostatic Multiscale Model on the B grid (NMMB; Janjić and Gall 2012). Initial and lateral boundary conditions (ICs/LBCs) are taken from the North American Mesoscale Model (NAM; Janjić 2003), operational Rapid Refresh (RAP; Benjamin et al. 2016), and the National Centers for Environmental Prediction's Global Forecast System (GFS; Environmental Modeling Center 2003) as indicated. Microphysics parameterizations include the WRF single-moment 6-class (WSM6; Hong and Lim 2006), Ferrier et al. (2002), and Ferrier–Aligo (Aligo et al. 2018) schemes. Planetary boundary layer (PBL) parameterizations include the Mellor–Yamada–Janjić (MYJ; Janjić 2002) and Yonsei University (YSU; Hong et al. 2006) schemes. HRW refers to the High-Resolution Window model run.

Member	Dynamic core	ICs/LBCs	Microphysics	PBL	Initialization time
NSSL-WRF	WRF-ARW	NAM/NAM	WSM6	MYJ	0000 UTC
HRW ARW	WRF-ARW	RAP/GFS	WSM6	YSU	0000 UTC
HRW ARW (time-lagged)	WRF-ARW	RAP/GFS	WSM6	YSU	1200 UTC (12-h time lag)
HRW NMMB	NMMB	RAP/GFS	Ferrier	MYJ	0000 UTC
HRW NMMB (time-lagged)	NMMB	RAP/GFS	Ferrier	MYJ	1200 UTC (12-h time lag)
WRF-NMM	WRF-NMM	NAM/NAM	Ferrier	MYJ	0000 UTC
NAM NEST	NMMB	NAM/NAM	Ferrier–Aligo	MYJ	0000 UTC

to equitably compare SPC forecasts with the continuous RF- and UH-based forecasts from the SSEO. There are two potential remedies to this problem. The first is to truncate the SSEO-derived forecasts at the same probability levels as used by the SPC. The second is to spatially interpolate the SPC probabilities between contour levels (e.g., Herman et al. 2018). Both methods are used herein. However, in this study, continuous SPC probabilities are created using a method developed at the SPC (Karstens et al. 2019). Herein, raw SPC contours are filled/gridded using a top-hat distribution, such that all grid points enclosed by a contour are assigned that contour value. The gridding procedure is done using the General Meteorological Package (GEMPAK; desJardins et al. 1991) within a 1° expanded CONUS domain to negate chronic dampening of probabilities near the edges of the forecast domain. Next, unique probability areas are identified using watershed segmentation (e.g., Lakshmanan et al. 2009), and adjacent probability areas are bilinearly interpolated using a Euclidean distance transformation. Finally, the maximum probability level is assumed to be 25% greater than the maximum nonzero contoured probability level present in the forecast. Continuous SPC probabilities created using this method are henceforth referred to as “full” SPC probabilities, while the raw, discrete SPC probabilities are referred to as “original” SPC probabilities. Importantly, full SPC probabilities do not exist for significant severe weather forecasts, since the SPC only issues a 10% or greater probability contour for significant severe events. Additionally, the SPC does not issue day 1 outlook probabilities for all-hazards severe or significant severe weather.

Severe weather observations used for verification and RF training are taken from the SPC website (SPC 2019b) for wind and hail and the SPC Storm Events

Database (SSED; SPC 2019a) for tornadoes. The SSED was required for tornadoes since it displayed information about each tornado's EF rating, necessary for the prediction/verification of significant tornadoes. Unfiltered reports are used to account for all reported instances of severe weather.

b. UH-based forecasts

UH-based probability forecasts for each severe weather hazard are derived from the SSEO. These forecasts are created in the same manner described by Loken et al. (2017). Namely, the fraction of ensemble members exceeding a given UH threshold is noted at each grid point, and that fraction is smoothed using a two-dimensional isotropic Gaussian kernel density function. Therefore, the UH-based probability p at a given grid point can be expressed as

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

TABLE 3. SPC conversion table relating individual hazard probabilities to categorical day 1 COs (sig. = significant). (Adapted from http://www.spc.noaa.gov/misc/SPC_probotlk_info.html.)

Individual hazard probability	Tornado	Wind	Hail
≥2%	Marginal	—	—
≥5%	Slight	Marginal	Marginal
≥10%	Enhanced	—	—
≥10% and ≥10% sig.	Enhanced	—	—
≥15%	Enhanced	Slight	Slight
≥15% and ≥10% sig.	Moderate	Slight	Slight
≥30%	Moderate	Enhanced	Enhanced
≥30% and ≥10% sig.	High	Enhanced	Enhanced
≥45%	High	Enhanced	Enhanced
≥45% and ≥10% sig.	High	Moderate	Moderate
≥60%	High	Moderate	Moderate
≥60% and ≥10% sig.	High	High	Moderate

where f is the fraction of ensemble members exceeding some UH threshold, N is the number of points with at least one member exceeding the threshold, d_n is the distance between the current grid point and the n th point, and σ is the standard deviation of the Gaussian kernel. To determine the combination of UH threshold and σ to use for each hazard, the UH threshold is varied from 10 to 200 m² s⁻² in increments of 10 m² s⁻² while σ is varied from 30 to 210 km in increments of 30 km. The combination that optimizes the Brier skill score (BSS; e.g., Wilks 2011) for a given hazard over the entire dataset is used (right column of Fig. 1), with BSS measured relative to a constant forecast of observed hazard climatology during the 629-day dataset. The calibration is done on the 80-km verification grid (see below) rather than the native 4-km grid.

c. Random forest forecasts

1) RF METHOD OVERVIEW

A RF is an ensemble of decision trees (Breiman 2001). Individual decision trees (Breiman 1984) work by recursively splitting a dataset until a stopping criterion is reached (e.g., the tree reaches a specified maximum number of levels, the number of samples at a node falls below a specified threshold, etc.). Splitting criteria are determined by the algorithm during training. Specifically, at each node, the algorithm chooses the threshold and predictor variable that splits the data in a way that maximizes a dissimilarity metric (e.g., information gain, Gini impurity). Class predictions can then be made on unseen data by running a testing example through the tree and analyzing the training samples in the appropriate leaf node (i.e., terminal node). For example, class probabilities are expressed as the fraction of training examples associated with the given class in the leaf node containing the testing example.

Although individual decision trees are human-readable and relatively easy to interpret, they are prone to overfitting, such that small changes to a testing example's predictor variables can produce very different class predictions (e.g., Gagne et al. 2014). The RF algorithm helps remedy this overfitting tendency by growing multiple trees, which are made unique by 1) growing each tree based on a random subset of training examples, and 2) determining the best split at each node by considering a random subset of predictor variables (Breiman 2001). During testing, RF class probabilities are simply the mean probability from each tree in the RF. In this study, RFs and corresponding RF probabilities are created using random forest classifiers from the Python module Scikit-Learn (Pedregosa et al. 2011).

More information on RFs can be found in Loken et al. (2019a) and works cited therein.

2) PREDICTOR VARIABLES

The first step of creating RF-based probabilities is to determine which predictors (or input variables) the RF will consider. Here, predictor variables are based on SSEO forecast fields. However, only a small number of variables relevant to severe weather forecasting are originally stored within the SSEO data archive (i.e., the variables without asterisks in Table 4). To enhance RF skill, several predictor variables (i.e., those with asterisks in Table 4) are added to these original variables. For example, the product of most unstable convective available potential energy (MUCAPE) and 0–6-km wind shear is computed at each native 4-km grid point and stored as a predictor variable. Latitude, longitude, and smoothed UH probabilities are also added as predictors during preprocessing.

3) DATA PREPROCESSING

While the SSEO contains a limited number of archived forecast fields, there is originally an overwhelming amount of data potentially available to the RF, since each SSEO member forecasts each variable at 3-km grid spacing over the CONUS every hour. To make training the RF computationally feasible, the dimensionality of the SSEO dataset must be reduced through several steps of data preprocessing.

The first preprocessing step is to reduce the temporal dimension of the dataset. This is accomplished by taking a 24-h (1200–1200 UTC) temporal maximum (for the storm attribute variables; Table 4) or mean (for the environment-related variables) at each 4-km grid point. Next, these temporally aggregated forecast variables—as well as the observed storm reports—are remapped to an approximately 80-km grid (i.e., NCEP grid 211) to further reduce dataset dimensionality and to match the verification scales used by the SPC. For the storm attribute fields, remapping is done by selecting the maximum forecast value on the 4-km grid within each 80-km grid box. For the environment-related fields, remapping to the 80-km grid is done using a neighbor budget method (Accadia et al. 2003), which approximately conserves the remapped quantity. After remapping, the ensemble mean, maximum, minimum, and standard deviation values are computed for each forecast variable at every 80-km grid point. Additionally, smoothed UH probabilities (to be used as predictors) are derived based on the method in section 2b. However, the UH threshold and standard deviation of the Gaussian kernel combination used is that which maximizes area under the relative operating

FIG. 1. Heat maps showing how (left) area under the relative operating characteristics curve (AUC) and (right) Brier skill score (BSS) vary with the standard deviation of the Gaussian kernel and UH threshold for UH-based forecasts. Heat maps are for (row 1) any severe weather hazard, (row 2) any significant severe weather hazard, (row 3) any tornado, (row 4) significant tornadoes, (row 5) any

TABLE 4. Predictor variables. Asterisks denote variables that were added during preprocessing.

Storm attribute fields	Environment-related fields	Other
Max hourly simulated reflectivity	2-m temperature	Latitude*
Accumulated 1-h precipitation	2-m dewpoint temperature	Longitude*
Max hourly updraft speed	2-m relative humidity	Smoothed UH probabilities* (created by maximizing AUC)
Max hourly UH	MUCAPE	
	CIN	
	0–6-km shear	
	CAPE $\times$ shear*	

characteristic curve (AUC; e.g., Wilks 2011; left column of Fig. 1) rather than BSS, since AUC is a measure of potential skill after bias calibration (Wilks 2011) and RF outputs typically have low bias (e.g., Breiman 2001).

After preprocessing, a final set of predictors is obtained for input into the RF. Here, these predictors include the ensemble mean, maximum, minimum, and standard deviation of SSEO forecast fields as well as latitude, longitude, and UH-based probabilities (Table 4). For a given grid point prediction, the RF considers these quantities at the 25 closest 80-km grid points.

4) RF PREDICTIONS

The RF gives probabilistic predictions of whether a given 80-km grid box will experience the occurrence of at least one observed severe weather report (all hazards or individual hazard) over the 24-h day 1 CO period (i.e., 1200–1200 UTC). Separate RFs are used to predict the occurrence of: all-hazards severe weather, all-hazards significant severe weather, any tornadoes, significant tornadoes, any severe wind, significant severe wind, any severe hail, and significant severe hail. Finally, the predictions from these separate RFs are used to construct an RF-based day 1 categorical outlook using the same guidelines employed by the SPC (i.e., those in Table 3).

5) DISCRETE/TRUNCATED RF PROBABILITIES

To facilitate a fair comparison with the SPC day 1 outlooks, discrete RF probabilities are created for individual-hazards severe and significant severe weather forecasts using the same probability levels as the SPC (Table 3). Discrete RF probabilities (henceforth

referred to as truncated RF forecasts) are created by simply converting all continuous RF probabilities between discrete SPC probability levels to the lower probability. For example, continuous severe hail probabilities between 5% (inclusive) and 15% (exclusive) are converted to 5% probabilities, since they would all be contained within a 5% SPC contour. Similarly, for individual-hazard significant severe forecasts, truncated RF probabilities are 10% if the continuous RF probabilities meet or exceed 10% and 0% otherwise.

d. Verification

Probabilistic severe weather forecasts are evaluated over the entire CONUS (Fig. 2a) as well over the West, Midwest, and East (Fig. 2b), which are defined based on temperature and precipitation climatology and represent an aggregation of regions described in Bukovsky (2011). Forecasts are also analyzed seasonally, with winter, spring, summer, and fall defined as December–February, March–May, June–August, and September–November, respectively.

Forecasts are verified on the ~80-km NCEP grid 211 to approximately match the verification definitions used by the SPC, which evaluates the occurrence of severe weather within 40 km of a point, and to save computational expense during verification. Continuous RF, truncated RF, original SPC, full/continuous SPC, and (continuous) UH-based probabilities are evaluated and compared against each other whenever possible. Unfortunately, due to the limitations of the SPC forecasts, full SPC probabilities are not created for significant severe weather forecasts, and neither original nor full SPC probabilities exist for all-hazard severe or significant severe forecasts. Additionally, no quantitative

←

severe wind, (row 6) significant severe wind, (row 7) any severe hail, and (row 8) significant severe hail. In each case, the combination with the highest AUC or BSS is indicated by a white circle and noted below the plot. AUC is used for calibrating smoothed UH RF inputs, while BSS is used for calibrating the smoothed UH forecasts themselves.

FIG. 2. (a) Overall analysis domain (gray shading). (b) West (yellow), Midwest (blue), and East (purple) region analysis domains.

verification is performed on the RF- and SPC-based categorical outlooks, since these are not true probabilistic forecasts, but rather summary products that merge probability and intensity information. Forecast evaluation is done using 17-fold cross validation with 37 days per fold. A total of 17 folds are used here to balance the trade-off between computational expense and training set size and to provide an equal number of days (37) in each fold. As in [Loken et al. \(2019a\)](#), verification statistics are computed over the full set of 629 forecasts derived from each fold's testing set.

Metrics used for verification include: BSS, BS components (e.g., [Wilks 1995](#)), attributes diagrams (e.g., [Hsu and Murphy 1986](#)), and performance diagrams ([Roebber 2009](#)). While AUC is used to set the UH threshold and Gaussian kernel standard deviation for smoothed UH-based predictors, it is not used for forecast evaluation since it is not sensitive to bias and it tends to increase nonlinearly with increasing forecast skill such that two well-performing but differently skilled forecast systems may have similar AUC values near 1 ([Marzban 2004](#)).

The BS (e.g., [Wilks 1995](#)), which measures the magnitude of forecast probability errors, can be decomposed into reliability, resolution, and uncertainty components ([Murphy 1973](#); [Wilks 1995](#)), and is defined as

$$\begin{aligned} \text{BS} &= \frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2 \\ &= \frac{1}{N} \sum_{k=1}^K [n_k(p_k - \bar{o}_k)]^2 - \frac{1}{N} \sum_{k=1}^K [n_k(\bar{o}_k - \bar{o})]^2 + \bar{o}(1 - \bar{o}), \end{aligned} \quad (2)$$

where N is the total number of forecast/observation pairs, K is the number of forecast probability bins, p_i is the forecast probability at point i , o_i is the binary observation (i.e., 0 or 1) at point i , n_k is the number of forecasts in bin k , $\bar{o}_k$ is the mean observed relative frequency in bin k , and $\bar{o}$ is the overall sample climatological frequency. The three terms on the right of Eq. (2) represent the reliability, resolution, and uncertainty components of the BS, respectively. Meanwhile, the BSS compares the BS to that of a reference forecast, thus enabling a fair comparison for events with different climatological relative frequencies ([Wilks 1995](#)). Specifically, the BSS is defined as

$$\text{BSS} = \frac{\text{BS} - \text{BS}_{\text{ref}}}{0 - \text{BS}_{\text{ref}}} = 1 - \frac{\text{BS}}{\text{BS}_{\text{ref}}}, \quad (3)$$

where, herein, BS_{ref} is the BS resulting from always forecasting the observed climatology of the relevant dataset. A BSS of 1 (0) indicates perfect (no) skill relative to the reference forecast. In total, 95% confidence intervals (95CIs) for each forecast's BSS values are determined using resampling with replacement (i.e., bootstrapping; e.g., [Wilks 2011](#)). Specifically, 629 random samples (with replacement) are drawn from a given forecast's 629 individual-day BS values. The aggregate BS and BSS over the random sample are then computed and stored. After 10 000 iterations of this process, the 95% BSS confidence interval is noted by observing the 2.5- and 97.5-percentile values of the stored BSS distribution.

While the reliability component of the BS provides a single-number summary of how well forecast probabilities correspond with observations, attributes diagrams allow users to assess reliability separately for each of k probability bins. Herein, bins are defined by the following probability level ranges: [0%–1%), [1%–2%), [2%–5%), [5%–15%), [15%–25%), ..., [85%–95%), and [95%–100%]. Perfectly reliable forecasts fall along a line of slope 1 passing through the origin; over (under) forecasts fall below (above) this line. Attributes diagrams also contain horizontal and vertical lines plotted at the sample climatological relative frequency as well as a no-skill line located halfway between the horizontal climatology line and the perfect reliability line. Points above (below) the no-skill line contribute positively

FIG. 3. (a) CONUS-wide BSS for the full/continuous RF-based probabilities (dark red), truncated RF-based probabilities (yellow), original SPC probabilities (light blue), full/continuous SPC probabilities (dark blue), and UH-based probabilities determined using the optimal standard deviation and UH threshold combination for each hazard (gray). (b),(c) As in (a), but for the resolution and reliability components of the BS, respectively. Black bars denote 95% confidence intervals in (a). In (b) and (c), axes are scaled differently on either side of the breaks, allowing for easier interpretation of all data. Note that the SPC does not issue forecasts for all-hazards severe or significant severe weather and that full/continuous SPC probabilities are not available for the individual significant severe hazards.

(negatively) to the BSS when a reference forecast of climatology is used (Wilks 1995).

Performance diagrams (Roebber 2009) binarize probabilistic forecasts at specific probability levels (herein, 0%, 1%, 2%, 5%–95% in increments of 10%, and 100%) and display probability of detection (POD), success ratio (SR), bias, and critical success

index (CSI) on a single plot [e.g., see Eqs. (1)–(4) in Roebber (2009)]. Points falling closer to the top-right-hand corner of the diagram exhibit greater skill, since POD, SR, CSI, and bias are all optimized at a value of 1. Moreover, POD, SR, CSI, and bias are all independent of the number of correct negatives, making the performance diagram a good tool

for evaluating forecasts with many trivial correct negatives.

3. Results

a. Full-domain, full-period results

The continuous RF forecasts have the greatest overall BSS values for each of the hazards examined (Fig. 3a). Compared to the UH-based forecasts, the continuous RF forecasts give substantially better predictions for all hazards except significant tornadoes (Fig. 3a). This is an important result given that UH is a skillful predictor of severe weather (e.g., Kain et al. 2008; Sobash et al. 2011, 2016, 2019) and is widely used in test bed settings (e.g., Kain et al. 2008; Clark et al. 2012a; Guyer and Jirak 2014; Gallo et al. 2017; Roberts et al. 2019). The continuous RF forecasts always have better resolution (Fig. 3b) and frequently—though not always—have better reliability (Fig. 3c) than the UH forecasts. Of course, it is likely that the UH-based forecasts would obtain a higher BSS if a time- and space-varying UH threshold were used instead of a constant one (Sobash and Kain 2017). However, calibrating the UH threshold in space and time requires substantially more computational resources compared to a constant threshold calibration. While training a RF is also computationally intensive, the RF considers multiple variables, and its multivariate “calibration” occurs implicitly as the algorithm is run.

The continuous RF forecasts also perform substantially better than the full SPC forecasts for hail and wind but not tornado prediction (Fig. 3a), an unsurprising result given this study’s lack of tornado-specific predictors [e.g., significant tornado parameter (STP; Thompson et al. 2003), low-level storm relative helicity (e.g., Coffey et al. 2019), etc.]. Thus, it is possible that adding predictors with a stronger correlation to observed tornado and/or low-level mesocyclone occurrence could improve the RF tornado and significant tornado forecasts. However, even without tornado-specific predictors, the continuous RF forecasts have better resolution (Fig. 3b) and better (i.e., smaller) reliability values (Fig. 3c) than the continuous SPC forecasts for all hazards.

When the continuous RF forecasts are truncated at the probabilities used by the SPC, BSS values are, unsurprisingly, reduced (Fig. 3a). Much of this reduction comes from degraded reliability (Fig. 3c) rather than decreased resolution (Fig. 3b). However, the truncated RF probabilities still have substantially greater BSSs than the original SPC probabilities for severe wind (Fig. 3a). Truncated RFs also have higher BSSs relative to the original SPC forecasts for severe hail, with the

95CIs of the two forecasts just barely overlapping. For the significant severe hazards, the truncated RFs do not substantially outperform the original SPC forecasts. However, the continuous RF forecasts do have notably greater BSSs than the original SPC forecasts for significant severe wind and significant severe hail. This outperformance is due to the improved resolution (Fig. 3b) and reliability (Fig. 3c) that is possible with access to continuous rather than binary (i.e., $\geq 10\%$) forecast probabilities.

While the RF-based forecasts have the best resolution for all hazards (Fig. 3b), they do not necessarily have the best reliability (Fig. 3c); however, reliability among all forecasts for all hazards is generally very good (Figs. 4 and 5). Large deviations from perfect reliability are typically associated with small sample size in the relevant forecast probability bin(s) [e.g., the UH significant severe weather forecasts (Figs. 5a,c,e,g) at higher forecast probabilities and the RF and UH tornado probabilities greater than 30% (Fig. 4c)]. Interestingly, both the original and SPC probabilities underforecast tornadoes (Fig. 4c) and severe wind (Fig. 4e). For the original SPC forecasts, this underforecasting is at least partially due to their use of discrete probabilities (i.e., probabilities between two discrete levels are mapped to the lower level). However, the underforecasting may also reflect a general philosophy of the SPC to emphasize higher-end tornado and wind events, given that SPC categorical outlooks are directly dependent on forecast hazard probability (Table 3). For example, it is possible that forecasters may wish to convey a message other than “moderate” or “high risk” to emergency managers or other users when they anticipate higher probabilities of marginally severe wind [e.g., ~ 50 kt (58 mph)] or low-end (e.g., EF0) tornado reports. Similarly, the SPC may wish to have high POD—even at the expense of false alarm—for significant tornadoes and significant severe wind events, which could explain the SPC overforecasting for these hazards (Figs. 5c,e). The SPC does not have the same overforecasting bias for severe (Fig. 4g) and significant severe (Fig. 5g) hail, perhaps since these events have less potential for truly devastating impacts. Importantly, the UH and RF forecasts give equal weight to all observed storm reports and do not consider the potential societal impacts of observed severe weather.

As statistical methods, the UH and RF forecasts tend to struggle most for the rarest events, which have the least amount of data. For example, the UH forecasts have good reliability for most hazards but some overforecasting at higher probabilities for tornadoes (Fig. 4c) and significant severe weather hazards (Figs. 5a,c,e,g). Meanwhile, the continuous

FIG. 4. (a) Attributes diagrams for the full RF (dark red) and calibrated UH (gray) any severe weather forecasts. The black long-dashed line indicates perfect reliability, the solid black line indicates the “no skill” line, and the black short-dashed lines represent climatological relative frequency. (b) Number of forecasts in each probability bin for the forecasts in (a). (c),(d) As in (a) and (b), but for any tornado forecasts. Truncated RF (yellow), original SPC (light blue), and full SPC (dark blue) forecasts are shown in addition to the continuous RF (dark red) and calibrated UH (gray) forecasts. (e),(f) As in (c) and (d), but for any severe wind forecasts. (g),(h) As in (c) and (d), but for any severe hail forecasts.

RF forecasts tend to have either near-perfect reliability (e.g., Figs. 4g and 5a,e,g) or slight underforecasting (e.g., Figs. 4a,e) at most probability levels for most hazards. Unsurprisingly, the truncated RF forecasts tend to

underforecast relative to the continuous RF forecasts, since—like the original SPC forecasts—all continuous forecast probabilities less than a given discrete level are assigned to the next lowest level. Nevertheless, both the

FIG. 5. As in Fig. 4, but for (a),(b) any significant severe; (c),(d) significant tornado; (e),(f) significant severe wind; and (g),(h) significant severe hail forecasts. Note that, unlike in Fig. 4, the x and y axes stop at 0.5 and full SPC forecasts are not plotted.

continuous and truncated RF forecasts have excellent reliability for the prediction of all hazards at probabilities with a sufficiently large sample size.

Performance diagrams (Figs. 6a–h) generally corroborate the BSS-based results (Figs. 3a–c), showing a clear outperformance of the RF-based method for most hazards. For example, the continuous RF forecasts substantially outperform the UH forecasts for all-hazard severe (Fig. 6a) and significant severe (Fig. 6b) weather at all probability levels. The continuous and truncated

RF forecasts also clearly outperform both the SPC and UH forecasts for severe wind (Fig. 6e), significant severe wind (Fig. 6f), severe hail (Fig. 6g), and significant severe hail (Fig. 6h). Interestingly, for tornadoes, the RF-based forecasts perform as well as (for the lower forecast probabilities) or slightly worse than (for the higher forecast probabilities) those from the SPC, with the UH-based forecasts noticeably inferior (Fig. 6c). Again, the RF-based forecasts' worse performance for tornado prediction potentially reflects

FIG. 6. Performance diagrams for (a) any severe weather, (b) any significant severe weather, (c) any tornado, (d) significant tornado, (e) any severe wind, (f) significant severe wind, (g) any severe hail, and (h) significant severe hail forecasts. Note that only the full RF (dark red) and calibrated UH (gray) forecasts are shown in (a) and (b). All other panels additionally show original SPC (light blue) and truncated RF (yellow) forecasts. Full SPC (dark blue) forecasts are only shown in (c), (e), and (g).

FIG. 7. (a) BSS of full RF (dark red) and UH-based forecasts for any severe weather. UH forecasts use a Gaussian kernel standard deviation of 120 km and a UH threshold of 5 (dark purple), 10 (light purple), 20 (light blue), 30 (royal blue), 40 (dark blue), 50 (dark green), 60 (yellow), and $70 \text{ m}^2 \text{ s}^{-2}$ (orange), respectively. BSSs are computed over the winter (DJF), spring (MAM), summer (JJA), and fall (SON) seasons as well as over the entire year (All). (b) As in (a), but BSSs are computed over the West (W), Midwest (MW), and East (E) regions as well as over the full CONUS (All). (c) As in (a), but forecasts are for any significant severe weather and the UH-based forecasts use thresholds of 80 (light red), 90 (brown), 100 (yellow), 110 (tan), 120 (dark blue), 130 (blue), 140 (purple), and $150 \text{ m}^2 \text{ s}^{-2}$ (light purple). (d) As in (c), but BSSs are computed over the regions in (b).

the lack of tornado-specific predictors in this study. For significant severe hazards (Figs. 6d,f,h), skill is relatively low for all forecasts, but the RF forecasts have CSI values at least as high as those from SPC and UH forecasts.

In general, the continuous and truncated RF forecasts have similar CSI scores. There is one interesting exception, however: for the significant tornado forecasts, the truncated RF method is associated with a noticeably higher CSI (Fig. 6d). The likely cause is the poor reliability of the continuous RF forecasts at greater than 10% probabilities due to small sample

size (Figs. 5c,d). Because the continuous RF probabilities dramatically overforecast significant tornadoes above 10% probability, the truncation procedure dramatically improves reliability (Fig. 5c) and CSI (Fig. 6d) at the 10% level.

b. Seasonal and regional results

Consistent with Sobash and Kain (2017), it is found herein that the “best” UH threshold to use (i.e., the one that maximizes BSS) for all-hazards severe (Figs. 7a,b) and significant severe (Figs. 7c,d) weather prediction depends on season and region. The best-performing UH

threshold is particularly sensitive to region: values of 60, 40, and $30 \text{ m}^2 \text{ s}^{-2}$ (140, 110, and $130 \text{ m}^2 \text{ s}^{-2}$) are best for the West, Midwest, and East regions, respectively, for all-hazards severe (significant severe) weather (Figs. 7b,d). While the best UH threshold does not change much seasonally for the all-hazard severe weather forecasts (Fig. 7a), seasonal variations are more apparent for all-hazard significant severe weather forecasts (Fig. 7c). Importantly, the continuous RF always outperforms the best UH forecast over a given region or season (Figs. 7a–d).

When all forecasts are verified seasonally, a similar pattern emerges: with just one exception (i.e., fall tornado forecasts; Fig. 8a), the continuous RF forecasts have the highest BSSs for all hazards during all seasons (Figs. 8a–f). Both the continuous and truncated RF forecasts have substantially greater BSSs for summer severe wind prediction compared to either the UH or SPC forecasts (Fig. 8c). The continuous RF forecasts also dramatically outperform the best-performing SPC forecast for the prediction of spring and summer severe hail (Fig. 8e) and spring significant severe hail (Fig. 8f). Additionally, the continuous RF forecasts substantially outperform the UH-based forecasts—but not the continuous SPC forecasts—for spring severe wind (Fig. 8c) and winter tornadoes (Fig. 8a). However, it should be noted that using a spatiotemporally varying UH threshold would likely improve the BSSs of the UH forecasts (e.g., Sobash and Kain 2017), especially for the winter tornado forecasts. While the continuous RF forecasts generally exhibit noticeably larger BSSs than the other forecasts for significant severe hazards (e.g., Figs. 8b,d,f), the seasonal 95CIs are typically quite large for these hazards. Truncated RF forecast BSSs are generally higher—but not dramatically higher—than those from the original SPC forecasts for subsignificant severe weather prediction (i.e., Figs. 8a,c,e), although the truncated RF forecasts do have substantially better summer severe wind forecasts. For the significant severe hazards, the truncated RF probabilities have BSSs similar to the original SPC probabilities during each season.

When BSS is tabulated regionally, it is apparent that the RF method struggles at predicting tornadoes (Fig. 9a), significant tornadoes (Fig. 9b), and significant severe wind (Fig. 8d) in the West region. However, for all other hazards and regions (Figs. 9a–f), the continuous RF forecasts have the greatest BSSs. Regionally, the RF approach gives the greatest relative benefit for East severe wind prediction (Fig. 9c); both the continuous and truncated RF forecasts have substantially greater BSSs than the UH- or SPC-based forecasts. The continuous RF also noticeably outperforms the

UH and SPC forecasts for the prediction of West and East severe wind (Fig. 9c) and Midwest severe hail (Fig. 9e) and significant severe hail (Fig. 9f). As with the seasonal verification results, truncated RF significant severe probabilities (Figs. 9b,d,f) tend to have similar BSSs to original SPC probabilities for each region.

4. Case studies

a. 26–27 May 2015

At 1200 UTC 26 May, a midlevel trough and associated mesoscale convective complex (MCC) was centered over central Iowa. A line of thunderstorms extended along a surface front from the MCC southeast into eastern Mississippi. As the period progressed, the midlevel trough moved northeastward over the Great Lakes region and helped deepen an associated surface cyclone, ultimately leading to several tornado and severe wind reports in Illinois and Wisconsin before 1900 UTC. The cyclone's cold front also advanced eastward and helped force the development of severe-wind-producing thunderstorms over eastern Alabama, western Georgia, and the Ohio Valley. Farther west, storms initiated along a dryline extending from west-central Oklahoma southward into central Texas by 2300 UTC. These storms produced numerous reports of severe wind and hail, with multiple significant severe hail reports and one significant severe wind report.

The RF and SPC outlooks (Figs. 10a,b) had some notable differences on this day, including the RF outlook's use of the enhanced risk over two locations as well as the RF outlook's greater areal coverage of slight risk areas. In the Upper Midwest, the RF shifted the 2% and greater tornado probabilities westward compared to the SPC (Figs. 11a,b). As a result, the RF had better POD for tornadoes in eastern Iowa, southern Wisconsin, and northern Illinois. Along the Oklahoma–Texas border, the RF issued 10% tornado probabilities with 6% significant tornado probabilities. Ultimately, no significant tornadoes were observed in this region, although multiple tornado reports occurred near the RF's 10% tornado area. Unlike the SPC forecast, the RF forecast issued 30% severe wind probabilities in a region extending from the Ohio Valley to the western Florida Panhandle (Figs. 11c,d). Numerous severe wind reports were observed near these locations, giving the RF a better POD. The RF also moved the 15% severe wind area slightly southeastward into southern Oklahoma and northern Texas, which better captured some severe wind reports—including a significant severe wind report—in that region. Notably, the one

FIG. 8. (a) BSS for any tornado probabilistic forecasts from the full RF (dark red), truncated RF (yellow), original SPC (light blue), full SPC (dark blue), and spatially smoothed UH (gray). BSSs are computed over the winter (DJF), spring (MAM), summer (JJA), and fall (SON) seasons as well as over the entire year (All). Note that the UH-based

significant severe wind report fell near the RF's 2% contour for significant severe wind. Indeed, one advantage of the RF forecast is its ability to communicate nonzero (but still less than 10%) probabilities for significant severe weather. For severe hail, the RF forecasts gave a much larger 5% area than the SPC (Figs. 11e,f) but focused on a similar area for its 15% probabilities. However, unlike the SPC, the RF forecasts produced a large area of 30% severe hail probabilities and indicated a greater than 10% chance of significant severe hail in western Oklahoma and northern Texas. Ultimately, numerous severe and significant severe hail reports occurred in this region. Two significant severe hail reports in central Texas also fell outside of the RF's 10% "hatched area" for significant severe hail but within the RF's 2% significant severe hail contour. However, the RF forecast did have greater false alarm than the SPC in eastern Louisiana (where the RF issued 15% probabilities) and over a large area extending from central Wisconsin to the Gulf Coast (where the RF generally issued 5% probabilities). Nevertheless, the RF generally made improvements over the SPC forecast. A human forecaster with access to the RF probabilities on this day might have had more confidence in a Texas–Oklahoma significant severe hail event and a widespread severe wind event in the Ohio Valley and Southeast.

UH-based probabilities might have only communicated part of this story. For example, compared to RF all-hazard probabilities (Fig. 12a), UH-based probabilities (Fig. 12b) were much lower over the Ohio Valley and Southeast United States. However, UH all-hazards severe and significant severe probabilities (Figs. 11b,d) were generally similar to those from the RF (Figs. 12a,c) over the southern plains.

b. 18–19 May 2017

The SPC identified 18 May 2017 as a high-risk day in the southern plains (e.g., Fig. 13b), with their 0600 UTC outlook highlighting the potential for widespread long-track tornadoes in parts of Oklahoma and Kansas. At the surface, a cyclone was developing in the western Oklahoma Panhandle by 1200 UTC. Strong southerly winds throughout central Texas and Oklahoma advected rich low-level moisture into the southern plains, where strong deep-layer vertical wind shear

was in place. Storms began forming in the warm sector along the dryline in western Oklahoma and northern Texas by 1830 UTC and quickly became severe. Severe storms also formed along the warm front in central Kansas by 2130 UTC. Meanwhile, in the Northeast, severe hail- and wind-producing storms initiated ahead of a cold front in an unstable, sheared environment.

While the RF and SPC forecasts identified similar threat areas in their outlooks (Figs. 13a,b), they issued different maximum outlook categories, with the RF (SPC) issuing a moderate (high) risk in the southern plains and an enhanced (slight) risk in the Northeast. Interestingly, although the RF produced smaller tornado probability magnitudes in the southern plains (Fig. 14a), it gave larger areas of higher-end (i.e., >10%) tornado probabilities there. Indeed, most of the observed tornadoes occurred within these areas of higher-end RF probabilities. The RF tornado forecast also expanded its 2% tornado probabilities farther east compared to the SPC, enabling it to better capture the QLCS tornado reports in Missouri (Figs. 14a,b). While the RF and SPC agreed on the area with the largest significant tornado probability (i.e., southern Kansas and northern Oklahoma; Figs. 14a,b), most of the observed significant tornadoes fell outside of this region but within/near the RF's 2% significant tornado probability contour. The RF and SPC forecasts had very similar tornado forecasts in the Northeast, with the RF forecasts having slightly less false alarm area.

RF and SPC severe wind probability magnitudes were quite different on this day, with the RF having higher probabilities in both the eastern United States and the southern plains (Figs. 14c,d). These higher probabilities led to better POD for the RF in New York, northern Pennsylvania, and southern Oklahoma but greater false alarm in most of West Virginia and northern Texas. The RF also expanded the 15% probability area farther eastward compared to the SPC, giving it greater POD in Arkansas and Missouri.

RF and SPC hail forecasts were similar, although the RF extended the 30% probability area and 10% significant severe hail area farther south into central Texas, where severe and significant severe hail occurred (Figs. 14e,f). Additionally, the RF indicated

←

forecasts use the combination of standard deviation and UH threshold that produces the best BSS over the CONUS over the entire year. Black bars indicate 95% confidence intervals. (b) As in (a), but for significant tornadoes. (c) As in (a), but for any severe wind. (d) As in (a), but for significant severe wind. (e) As in (a), but for any severe hail. (f) As in (a), but for significant severe hail. Note that full SPC probabilities are not shown in the significant severe panels [i.e., (b), (d), and (f)].

FIG. 9. As in Fig. 8, but BSSs are computed over the West (W), Midwest (MW), and East (E) regions, as well as over the full CONUS (All).

FIG. 10. Day 1 categorical convective outlook from the (a) RF approach and (b) SPC 0600 UTC forecast, valid for the 24-h period ending at 1200 UTC 27 May 2015. Small red, blue, and green circles outlined in black represent observed tornado, severe wind, and severe hail reports, respectively. Observed significant tornado, significant severe wind, and significant severe hail reports are represented by white-outlined large red circles, black squares, and black triangles, respectively.

2% significant severe hail probabilities in New York and Kansas where significant severe hail was observed but fell outside of the RF or SPC 10% significant severe hail probabilities. Finally, the RF demonstrated better severe hail POD in Maryland (Figs. 14e,f). Overall, the RF-based outlook (Fig. 13a) and individual-hazard probabilities (Figs. 14a,c,e) compared favorably against the corresponding SPC forecasts on this day.

RF all-hazards severe and significant severe weather probabilities (Figs. 15a,c) also compared favorably with UH-based probabilities (Figs. 15b,d), especially in the Northeast, where the RF had better POD for severe and significant severe weather. In the southern plains, RF and UH forecasts were generally similar. However, it is noteworthy that the RF significant severe forecasts (Fig. 15c) shift the maximum probabilities southwest into western Oklahoma, close to a cluster of significant severe reports, while the UH probability maximum is in central Kansas, away from any such cluster (Fig. 15d).

5. Discussion

Compared to the SPC forecasts, the RF probabilities frequently highlighted similar areas for severe weather but gave different probability magnitudes. However, the RF forecasts herein occasionally assigned higher probabilities (e.g., greater than 15% or even 30%)

to areas outside of the SPC's marginal risk. When this happened, many times the areas with the higher RF probabilities did experience observed severe weather. This occurred most often for severe wind events in the East region. In these instances, it is possible that the differences between the SPC and RF forecasts could be partially explained by biases and nonmeteorological artifacts in the severe wind report observations, given the high ratio of estimated to measured severe wind reports in the eastern and southeastern United States (Edwards et al. 2018). While the RF algorithm views all observed storm reports equally (i.e., as unbiased, perfect observations) and does not consider storm coverage, density, intensity, or potential societal impacts when constructing its probabilities, SPC forecasters may be mindful of how their forecast probabilities equate to outlook categories (Table 3) and may emphasize higher-impact events that pose a greater threat to life and property.

The biggest advantage of the RF method described herein is its ability to create skillful CAE-derived severe weather guidance products analogous to those issued by the SPC. However, it must be emphasized that the goal in creating these RF-based products is not to replace human forecasters but to augment them. Indeed, this augmentation could potentially take a variety of (nonmutually exclusive) forms. First, RF-based forecasts could provide a skillful,

FIG. 11. (a) RF-based tornado probabilities (shaded) and significant tornado probabilities (contoured every 2% with $\geq 10\%$ probabilities hatched), valid for the 24-h period ending at 1200 UTC 27 May 2015. (b) As in (a), but for

FIG. 12. (a) RF- and (b) UH-based probabilities of all-hazards severe weather, valid for the 24-h period ending at 1200 UTC 27 May 2015. Observed severe weather reports are plotted as described in Fig. 10. (c),(d) As in (a) and (b), but all-hazards significant severe weather probabilities are plotted, and only significant severe observed reports are overlaid. Individual-day AUC and BS values are reported in the lower-right corner of each plot.

reliable first guess (e.g., Karstens et al. 2018) product, which forecasters could modify based on other data sources (e.g., satellite and radar trends, surface analyses, etc.) and their expertise. Such a product could increase forecaster efficiency and facilitate proper forecast calibration (Karstens et al. 2018). Used as a

first guess or “last check” product, the RF guidance may also identify a threat area that a forecaster might have overlooked for a given hazard (e.g., significant severe hail in the southern plains; Figs. 11e,f). The RF forecasts may also help simply by providing useful uncertainty information in challenging forecasting

←

SPC forecasts issued at 0600 UTC. (c),(d) As in (a) and (b), but for severe wind forecasts. (e),(f) As in (a) and (b), but for severe hail forecasts. For each hazard, corresponding observed severe weather reports are plotted as described in Fig. 10. Note that SPC forecasts do not have significant severe contours less than 10%. Individual-day AUC and BS values are given for each forecast, with overall hazard (significant hazard) metrics given in the bottom-right corner (at the bottom) of each panel.

FIG. 13. As in Fig. 10, but valid for the 24-h period ending at 1200 UTC 19 May 2017.

situations. Such uncertainty information may be especially valuable for more precisely quantifying the threat of significant severe weather, which is rare but extremely threatening to life and property. Finally, it is conceivable that the RF-based forecasts—when properly interrogated using ML interpretability metrics (e.g., [McGovern et al. 2019](#))—may give forecasters and researchers insight into ensemble biases or complex relationships between CAE forecast output and observed severe weather. Human forecasters learning from ML would not be unprecedented, as artificial intelligence techniques have recently provided new knowledge to human experts in other complex domains, such as the game of Go ([Silver et al. 2016, 2017](#)) and multiplayer no-limit Texas Hold'em poker ([Brown and Sandholm 2019](#)). A future study is planned to determine how and why RF-based severe weather probabilities differ from human and UH-based forecasts.

6. Summary and conclusions

This paper used a random forest (RF) to produce CONUS-wide 1200–1200 UTC day 1 convective outlooks (COs) and individual-hazard severe weather probabilities from Storm-Scale Ensemble of Opportunity (SSEO) forecast output. Temporally aggregated gridpoint-based forecast variables were used as predictors. The gridpoint-based approach is advantageous because it allows users to interpret RF output directly in two-dimensional space and does not require the assumption of perfect correspondence between simulated and observed storms.

Continuous and discrete (i.e., truncated) RF forecasts created herein were compared against calibrated, spatially smoothed 2–5-km updraft helicity (UH) forecasts as well as original and continuous (i.e., full) SPC day 1 COs issued at 0600 UTC. The continuous RF forecasts almost always produced the highest BSSs, both when the forecasts were verified over the entire dataset and when verification was performed regionally or seasonally. The truncated RF forecasts frequently had the second-highest BSSs and were often better—but never substantially worse—than the corresponding original SPC forecasts. In general, the RF method performed best relative to the SPC and UH forecasts for severe wind and hail prediction in the Midwest and East regions during the spring and summer. All forecasts—including the RF-based ones—generally had very good reliability, while the RF forecasts tended to have the best resolution.

Given the promising results of the RF technique described herein, it is important to evaluate its skill and value to forecasters in an operational environment. To this end, efforts are under way to apply the technique described herein to the operational HREFv2 with the goal of evaluating real-time RF forecasts in future Hazardous Weather Testbed Spring Forecasting Experiments (e.g., [Clark et al. 2012a](#); [Gallo et al. 2017](#)). While such formal evaluation is necessary to draw more robust conclusions, it is speculated that real-time RF-based guidance will aid human day 1 severe weather forecasts by providing forecasters with calibrated CAE-based severe hazard probabilities and outlooks.

FIG. 14. As in Fig. 11, but valid for the 24-h period ending at 1200 UTC 19 May 2017.

FIG. 15. As in Fig. 12, but valid for the 24-h period ending at 1200 UTC 19 May 2017.

Acknowledgments. Support for this work was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. Additional support was provided by the Developmental Testbed Center (DTC). The DTC Visitor Program is funded by the National Oceanic and Atmospheric Administration, the National Center for Atmospheric Research, and the National Science Foundation. We would also like to acknowledge high-performance computing support from Cheyenne ([doi:10.5065/D6RX99HX](https://doi.org/10.5065/D6RX99HX)) provided by NCAR’s Computational and Information Systems Laboratory, sponsored by the National Science Foundation. Additionally, we extend our thanks to two anonymous reviewers, whose feedback improved the quality of the manuscript. AJC and CDK contributed to this work as part

of regular duties at the federally funded NOAA/National Severe Storms Laboratory. The statements, findings, conclusions, and recommendations presented herein are those of the authors and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

REFERENCES

- Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932, [https://doi.org/10.1175/1520-0434\(2003\)018<0918:SOPFSS>2.0.CO;2](https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2).
- Aligo, E. A., B. Ferrier, and J. R. Carley, 2018: Modified NAM microphysics for forecasts of deep convective storms. *Mon. Wea. Rev.*, **146**, 4115–4153, <https://doi.org/10.1175/MWR-D-17-0277.1>.
- Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh.

- Mon. Wea. Rev.*, **144**, 1669–1694, <https://doi.org/10.1175/MWR-D-15-0242.1>.
- Bermowitz, R. J., 1975: An application of model output statistics to forecasting quantitative precipitation. *Mon. Wea. Rev.*, **103**, 149–153, [https://doi.org/10.1175/1520-0493\(1975\)103<0149:AAOMOS>2.0.CO;2](https://doi.org/10.1175/1520-0493(1975)103<0149:AAOMOS>2.0.CO;2).
- Breiman, L., 1984: *Classification and Regression Trees*. Wadsworth International Group, 358 pp.
- , 2001: Random forests. *Mach. Learn.*, **45**, 5–32, <https://doi.org/10.1023/A:1010933404324>.
- Brown, N., and T. Sandholm, 2019: Superhuman AI for multi-player poker. *Science*, **365**, 885–890, <https://doi.org/10.1126/science.aay2400>.
- Bukovsky, M. S., 2011: Masks for the Bukovsky regionalization of North America, Regional Integrated Sciences Collective. Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, accessed 1 August 2019, <http://www.narccap.ucar.edu/contrib/bukovsky/>.
- Burke, A., N. Snook, D. J. Gagne, 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>.
- Carter, G. M., 1975: Automated prediction of surface wind from numerical model output. *Mon. Wea. Rev.*, **103**, 866–873, [https://doi.org/10.1175/1520-0493\(1975\)103<0866:APOSWF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1975)103<0866:APOSWF>2.0.CO;2).
- Clark, A. J., and Coauthors, 2012a: An overview of the 2010 Hazardous Weather Testbed experimental forecast program spring experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, <https://doi.org/10.1175/BAMS-D-11-00040.1>.
- , J. S. Kain, P. T. Marsh, J. Correia Jr., M. Xue, and F. Kong, 2012b: Forecasting tornado pathlengths using a three-dimensional object identification algorithm applied to convection-allowing forecasts. *Wea. Forecasting*, **27**, 1090–1113, <https://doi.org/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, <https://doi.org/10.1175/WAF-D-12-00038.1>.
- Coffer, B. E., M. D. Parker, R. L. Thompson, B. T. Smith, and R. E. Jewell, 2019: Using near-ground storm relative helicity in supercell tornado forecasting. *Wea. Forecasting*, **34**, 1417–1435, <https://doi.org/10.1175/WAF-D-19-0115.1>.
- desJardins, M. L., K. F. Brill, and S. S. Schotz, 1991: Use of GEMPAK on UNIX workstations. *Proc. Seventh Int. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology*, New Orleans, LA, Amer. Meteor. Soc., 449–453.
- Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) Model. *Atmos. Sci. Lett.*, **5**, 110–117, <https://doi.org/10.1002/asl.72>.
- Duda, J. D., and W. A. Gallus, 2010: Spring and summer Midwestern severe weather reports in supercells compared to other morphologies. *Wea. Forecasting*, **25**, 190–206, <https://doi.org/10.1175/2009WAF2222338.1>.
- Edwards, R., J. T. Allen, and G. W. Carbin, 2018: Reliability and climatological impacts of convective wind estimations. *J. Appl. Meteor. Climatol.*, **57**, 1825–1845, <https://doi.org/10.1175/JAMC-D-17-0306.1>.
- Environmental Modeling Center, 2003: The GFS atmospheric model. NCEP Office Note 442, 14 pp., <http://www.lib.ncep.noaa.gov/ncepoftenotes/files/on442.pdf>.
- Ferrier, B. S., Y. Jin, Y. Lin, T. Black, E. Rogers, and G. DiMego, 2002: Implementation of a new grid-scale cloud and rainfall scheme in the NCEP Eta model. Preprints, *19th Conf. on Weather Analysis and Forecasting/15th Conf. on Numerical Weather Prediction*, San Antonio, TX, Amer. Meteor. Soc., 10.1, https://ams.confex.com/ams/SLS_WAF_NWP/techprogram/paper_47241.htm.
- Gagne, D., A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043, <https://doi.org/10.1175/WAF-D-13-00108.1>.
- , —, S. Haupt, R. Sobash, J. Williams, and M. Xue, 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>.
- 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>.
- , and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed spring forecasting experiment. *Wea. Forecasting*, **32**, 1541–1568, <https://doi.org/10.1175/WAF-D-16-0178.1>.
- Gallus, W. A., N. A. Snook, and E. V. Johnson, 2008: Spring and summer severe weather reports over the Midwest as a function of convective mode: A preliminary study. *Wea. Forecasting*, **23**, 101–113, <https://doi.org/10.1175/2007WAF2006120.1>.
- Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, [https://doi.org/10.1175/1520-0450\(1972\)011<1203:TUOMOS>2.0.CO;2](https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2).
- 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 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>.
- , E. R. Nielsen, and R. S. Schumacher, 2018: Probabilistic verification of Storm Prediction Center convective outlooks. *Wea. Forecasting*, **33**, 161–184, <https://doi.org/10.1175/WAF-D-17-0104.1>.
- Hong, S.-Y., and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc.*, **42**, 129–151.
- , Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, <https://doi.org/10.1175/MWR3199.1>.
- Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecasting*, **2**, 285–293, [https://doi.org/10.1016/0169-2070\(86\)90048-8](https://doi.org/10.1016/0169-2070(86)90048-8).
- Janjić, Z. I., 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, 61 pp., <http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf>.
- , 2003: A nonhydrostatic model based on a new approach. *Meteor. Atmos. Phys.*, **82**, 271–285, <https://doi.org/10.1007/s00703-001-0587-6>.
- , and R. Gall, 2012: Scientific documentation of the NCEP nonhydrostatic multiscale model on the B grid (NMMB).

- Part 1 Dynamics. NCAR Tech. Note NCAR/TN-489+STR, 75 pp., <https://doi.org/10.5065/D6WH2MZX>.
- , J. P. Gerrity Jr., and S. Nickovic, 2001: An alternative approach to modeling. *Mon. Wea. Rev.*, **129**, 1164–1178, [https://doi.org/10.1175/1520-0493\(2001\)129<1164:AAATNM>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<1164:AAATNM>2.0.CO;2).
- Jirak, I. L., S. J. Weiss, and C. J. Melick, 2012: The SPC storm-scale ensemble of opportunity: Overview and results from the 2012 Hazardous Weather Testbed spring forecasting experiment. *26th Conf. on Severe Local Storms*, Nashville, TN, Amer. Meteor. Soc., 137, <https://ams.confex.com/ams/26SLS/webprogram/Paper211729.html>.
- , 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>.
- , —, and —, 2016: Comparison of the SPC storm-scale ensemble of opportunity to other convection-allowing ensembles for severe weather forecasting. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 102, <https://ams.confex.com/ams/28SLS/webprogram/Paper300910.html>.
- , A. J. Clark, B. Roberts, B. T. Gallo, and S. J. Weiss, 2018: Exploring the optimal configuration of the High Resolution Ensemble Forecast System. *25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 14B.6, <https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345640.html>.
- 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, <https://doi.org/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, <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>.
- Kang, J.-H., M.-S. Suh, K.-O. Hong, and C. Kim, 2011: Development of updateable model output statistics (UMOS) system for air temperature over South Korea. *Asia-Pac. J. Atmos. Sci.*, **47**, 199–211, <https://doi.org/10.1007/s13143-011-0009-8>.
- Karstens, C. D., and Coauthors, 2018: Development of a human-machine mix for forecasting severe convective events. *Wea. Forecasting*, **33**, 715–737, <https://doi.org/10.1175/WAF-D-17-0188.1>.
- , R. Clark III, I. L. Jirak, P. T. Marsh, R. Schneider, and S. J. Weiss, 2019: Enhancements to Storm Prediction Center convective outlooks. *Ninth Conf. on Transition of Research to Operations*, Phoenix, AZ, Amer. Meteor. Soc., J7.3, <https://ams.confex.com/ams/2019Annual/webprogram/Paper355037.html>.
- Klein, W. H., and H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. *Bull. Amer. Meteor. Soc.*, **55**, 1217–1227, [https://doi.org/10.1175/1520-0477\(1974\)055<1217:FLWBMO>2.0.CO;2](https://doi.org/10.1175/1520-0477(1974)055<1217:FLWBMO>2.0.CO;2).
- 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>.
- Lakshmanan, V., K. Hondl, and R. Rabin, 2009: An efficient, general-purpose technique for identifying storm cells in geospatial images. *J. Atmos. Oceanic Technol.*, **26**, 523–537, <https://doi.org/10.1175/2008JTECHA1153.1>.
- Legg, T., and K. Mylne, 2004: Early warnings of severe weather from ensemble forecast information. *Wea. Forecasting*, **19**, 891–906, [https://doi.org/10.1175/1520-0434\(2004\)019<0891:EWOSWF>2.0.CO;2](https://doi.org/10.1175/1520-0434(2004)019<0891:EWOSWF>2.0.CO;2).
- 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>.
- , —, A. McGovern, M. Flora, and K. Knopfmeier, 2019a: Postprocessing next-day ensemble probabilistic precipitation forecasts using random forests. *Wea. Forecasting*, **34**, 2017–2044, <https://doi.org/10.1175/WAF-D-19-0109.1>.
- , —, M. Xue, and F. Kong, 2019b: Spread and skill in mixed- and single-physics convection-allowing ensembles. *Wea. Forecasting*, **34**, 305–330, <https://doi.org/10.1175/WAF-D-18-0078.1>.
- Marzban, C., 2004: The ROC curve and the area under it as performance measures. *Wea. Forecasting*, **19**, 1106–1114, <https://doi.org/10.1175/825.1>.
- McGovern, A., R. Lagerquist, D. John Gagne, 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>.
- Melhauser, C., and F. Zhang, 2012: Practical and intrinsic predictability of severe and convective weather at the mesoscales. *J. Atmos. Sci.*, **69**, 3350–3371, <https://doi.org/10.1175/JAS-D-11-0315.1>.
- Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, [https://doi.org/10.1175/1520-0450\(1973\)012<0595:ANVPOT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2).
- Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
- Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensembles. *Bull. Amer. Meteor. Soc.*, **100**, 1245–1258, <https://doi.org/10.1175/BAMS-D-18-0041.1>.
- Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, <https://doi.org/10.1175/2008WAF2222159.1>.
- Schmeits, M. J., K. J. Kok, and D. H. P. Vorgelezang, 2005: Probabilistic forecasting of (severe) thunderstorms in the Netherlands using model output statistics. *Wea. Forecasting*, **20**, 134–148, <https://doi.org/10.1175/WAF840.1>.
- Schoen, J., and W. S. Ashley, 2011: A climatology of fatal convective wind events by storm type. *Wea. Forecasting*, **26**, 109–121, <https://doi.org/10.1175/2010WAF2222428.1>.
- Silver, D., and Coauthors, 2016: Mastering the game of go with deep neural networks and tree search. *Nature*, **529**, 484–489, <https://doi.org/10.1038/nature16961>.
- , and Coauthors, 2017: Mastering the game of go without human knowledge. *Nature*, **550**, 354–359, <https://doi.org/10.1038/nature24270>.
- 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>.
- Skinner, P. S., L. J. Wicker, D. M. Wheatley, and K. H. Knopfmeier, 2016: Application of two spatial verification

- methods to ensemble forecasts of low-level rotation. *Wea. Forecasting*, **31**, 713–735, <https://doi.org/10.1175/WAF-D-15-0129.1>.
- Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part I: Storm classification and climatology. *Wea. Forecasting*, **27**, 1114–1135, <https://doi.org/10.1175/WAF-D-11-00115.1>.
- 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>.
- , C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, <https://doi.org/10.1175/WAF-D-15-0138.1>.
- , —, —, and M. L. Weisman, 2019: Next-day prediction of tornadoes using convection-allowing models with 1-km horizontal grid spacing. *Wea. Forecasting*, **34**, 1117–1135, <https://doi.org/10.1175/WAF-D-19-0044.1>.
- SPC, 2019a: Storm Prediction Center WCM page: Severe weather database files (1950–2017). Accessed 16 December 2019, <https://www.spc.noaa.gov/wcm/>.
- , 2019b: Severe weather event summaries: NWS local storm reports. Accessed 16 December 2019, <https://www.spc.noaa.gov/climo/online/>.
- Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243–1261, [https://doi.org/10.1175/1520-0434\(2003\)018<1243:CPSWSE>2.0.CO;2](https://doi.org/10.1175/1520-0434(2003)018<1243:CPSWSE>2.0.CO;2).
- Torn, R. D., and G. S. Romine, 2015: Sensitivity of central Oklahoma convection forecasts to upstream potential vorticity anomalies during two strongly forced cases during MPEX. *Mon. Wea. Rev.*, **143**, 4064–4087, <https://doi.org/10.1175/MWR-D-15-0085.1>.
- Trier, S. B., G. S. Romine, D. A. Ahijevych, R. J. Trapp, R. S. Schumacher, M. C. Coniglio, and D. J. Stensrud, 2015: Mesoscale thermodynamic influences on convection initiation near a surface dryline in a convection-permitting ensemble. *Mon. Wea. Rev.*, **143**, 3726–3753, <https://doi.org/10.1175/MWR-D-15-0133.1>.
- Weisman, M. L., and W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.*, **125**, 527–548, [https://doi.org/10.1175/1520-0493\(1997\)125<0527:TRDOEM>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<0527:TRDOEM>2.0.CO;2).
- Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. International Geophysics Series, Vol. 59, Elsevier, 467 pp.
- , 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.