Using Machine Learning to Generate Storm-Scale Probabilistic Guidance of Severe Weather Hazards in the Warn-on-Forecast System

MONTGOMERY L. FLORA,[a,b,c] COREY K. POTVIN,[c,a] PATRICK S. SKINNER,[b,c,a] SHAWN HANDLER,[b] AND AMY MCGOVERN[a,d]

[a] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[b] *Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma*
[c] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
[d] *School of Computer Science, University of Oklahoma, Norman, Oklahoma*

(Manuscript received 15 June 2020, in final form 17 February 2021)

ABSTRACT: A primary goal of the National Oceanic and Atmospheric Administration Warn-on-Forecast (WoF) project is to provide rapidly updating probabilistic guidance to human forecasters for short-term (e.g., 0–3 h) severe weather forecasts. Postprocessing is required to maximize the usefulness of probabilistic guidance from an ensemble of convection-allowing model forecasts. Machine learning (ML) models have become popular methods for postprocessing severe weather guidance since they can leverage numerous variables to discover useful patterns in complex datasets. In this study, we develop and evaluate a series of ML models to produce calibrated, probabilistic severe weather guidance from WoF System (WoFS) output. Our dataset includes WoFS ensemble forecasts available every 5 min out to 150 min of lead time from the 2017–19 NOAA Hazardous Weather Testbed Spring Forecasting Experiments (81 dates). Using a novel ensemble storm-track identification method, we extracted three sets of predictors from the WoFS forecasts: intrastorm state variables, near-storm environment variables, and morphological attributes of the ensemble storm tracks. We then trained random forests, gradient-boosted trees, and logistic regression algorithms to predict which WoFS 30-min ensemble storm tracks will overlap a tornado, severe hail, and/or severe wind report. To provide rigorous baselines against which to evaluate the skill of the ML models, we extracted the ensemble probabilities of hazard-relevant WoFS variables exceeding tuned thresholds from each ensemble storm track. The three ML algorithms discriminated well for all three hazards and produced more reliable probabilities than the baseline predictions. Overall, the results suggest that ML-based postprocessing of dynamical ensemble output can improve short-term, storm-scale severe weather probabilistic guidance.

KEYWORDS: Severe storms; Ensembles; Short-range prediction; Classification; Machine learning

1. Introduction

The National Oceanic and Atmospheric Administration (NOAA) Warn-on-Forecast program [WoF; Stensrud et al. 2009, 2013] is tasked with providing forecasters with reliable, probabilistic severe weather hazard guidance at very short lead times[1] (e.g., 0–3 h). Though operational convection-allowing models (CAMs) cannot fully resolve convective processes (Bryan et al. 2003), CAMs with ≤3-km horizontal grid spacing can partially resolve important storm-scale features (Potvin and Flora 2015), distinguish between severe convective modes (e.g., supercell versus mesoscale convective systems; Done et al. 2004; Weisman et al. 2008), and provide severe weather surrogates such as updraft helicity (UH) or parameterized prediction of hail size (Adams-Selin and Ziegler 2016; Snook et al. 2012; Labriola et al. 2017, 2019), and low-level wind gusts (Jirak et al. 2014; Hepper et al. 2016). UH is a model surrogate for supercell thunderstorms, which are prolific producers of severe weather hazards (Duda and Gallus 2010; Smith et al. 2012). Severe weather forecast algorithms based on UH have shown skill at both next-day (e.g., Sobash et al. 2011, 2016) and O(1 h) lead times (Snook et al. 2012; Yussouf et al. 2013a,b; Wheatley et al. 2015; Yussouf et al. 2015; Jones et al. 2016; Skinner et al. 2016, 2018; Jones et al. 2019; Flora et al. 2019; Yussouf et al. 2020). Parameterized predictions of hail size have performed well in test bed experiments, producing skill comparable to UH-based algorithms for predicting hail reports (Adams-Selin et al. 2019). Although CAM severe weather surrogates have demonstrated success at predicting severe weather hazards, there are a number of limitations. For example, UH is a poor predictor of severe, nonrotating thunderstorms (which are significant producers of severe wind gusts; Smith et al. 2012, 2013); the current resolutions of operational CAMs prevent explicit prediction of surface-based severe wind gusts (Bryan et al. 2003); and current parameterized hail predictions rely on poorly understood microphysical processes.

A growing alternative to using CAM severe weather surrogates are machine learning (ML) models capable of producing calibrated guidance from many input predictors (e.g., Gagne et al. 2017; Lagerquist et al. 2017; McGovern et al. 2017;

[1] Although forecast lead time is defined by the American Meteorological Society Glossary as the length of time between the issuance of a forecast and the occurrence of the phenomena that were predicted (see https://glossary.ametsoc.org/wiki/Forecast_lead_time), in the numerical weather prediction community it commonly refers to the interval between the forecast initialization and valid times.

Corresponding author: Montgomery L. Flora, monte.flora@noaa.gov

DOI: 10.1175/MWR-D-20-0194.1

Cintineo et al. 2014, 2018; Burke et al. 2020; McGovern et al. 2019b; Hill et al. 2020; Lagerquist et al. 2020; Cintineo et al. 2020; Loken et al. 2020; Sobash et al. 2020; Steinkruger et al. 2020). Studies adopting ML-based approaches range from nowcasting lead times (e.g., ≤1 h; Lagerquist et al. 2017; Cintineo et al. 2014, 2018; Lagerquist et al. 2020; Cintineo et al. 2020; Steinkruger et al. 2020) that leverage available observational and numerical weather prediction (NWP) data to next-day forecasts (e.g., lead times of 24–36 h) that use state-of-the-art CAM ensemble forecasts (e.g., Gagne et al. 2017; Burke et al. 2020; Hill et al. 2020; Loken et al. 2020; Sobash et al. 2020). In Lagerquist et al. (2017), ML models produced skillful probabilistic severe wind predictions for radar-observed storms. The operational NOAA/Cooperative Institute for Meteorological Satellite Studies (CIMSS) ProbSevere model (Cintineo et al. 2014, 2018) is a naïve Bayesian classifier that reliably predicts severe weather likelihood up to a lead time of 90 min. In a newer version, ProbSevere v2.0, the system can now produce probabilistic guidance for separate severe weather hazards (Cintineo et al. 2020). Using a convolutional neural network (CNN; LeCun et al. 1990), a deep learning technique, Lagerquist et al. (2020) produced a next-hour tornado prediction system with skill comparable to the ProbSevere system. In an idealized framework, Steinkruger et al. (2020) explored using ML methods to produce automated tornado warning guidance and found promising results. Random forests (Breiman 2001) have produced competitive next-day hail predictions (Gagne et al. 2017; Burke et al. 2020), reliable next-day severe weather hazard guidance (Loken et al. 2020), and even outperformed the Storm Prediction Center (SPC) Day 2 and 3 outlooks (Hill et al. 2020). Neural networks have also exhibited success in predicting next-day severe weather and were shown to be more skillful than a UH baseline in Sobash et al. (2020). A key advantage of ML models is their ability to leverage multiple input predictors and learn complex relationships to produce skillful, calibrated probabilistic guidance. An additional advantage for real-time operational settings is that once an ML model has been trained, making predictions on new data is computationally quick (≪1 s per example). One drawback is that the ML model will require refitting when the CAM configuration changes.

The goal of this study is to evaluate the skill and reliability of ML-generated severe weather probabilistic guidance using WoF System (WoFS) ensemble forecasts as inputs. To accomplish this goal, we trained gradient-boosted classification trees (Friedman 2002; Chen and Guestrin 2016), random forests, and logistic regression models on WoFS forecasts from the 2017–19 Hazardous Weather Testbed Spring Forecasting Experiments (HWT-SFE; Gallo et al. 2017) to determine which storms predicted by the WoFS will produce a tornado, severe hail, and/or severe wind report. These three ML algorithms are fairly common and have recently shown success in a variety of meteorological applications (e.g., Mecikalski et al. 2015; Erickson et al. 2016; Gagne et al. 2017; Lagerquist et al. 2017; Herman and Schumacher 2018a,b; Burke et al. 2020; Loken et al. 2019; McGovern et al. 2019a,b; Hill et al. 2020; Jergensen et al. 2020; Steinkruger et al. 2020).

Recent ML studies using real CAM ensemble output for severe weather prediction have been restricted to the next-day (24–36 h) paradigm and producing grid-based guidance (e.g., Gagne et al. 2017; Burke et al. 2020; Loken et al. 2019; Hill et al. 2020; Sobash et al. 2020). Next-day forecasting methods, however, operate on a larger spatial scale because of the limited intrinsic predictability of storms at those lead times (Lorenz 1969). In an early version of this work, we found that using a strictly grid-based approach produced overly smooth guidance for WoF-style forecasts, which are intended to provide probabilistic guidance for individual thunderstorms (Stensrud et al. 2009, 2013). That finding motivated the creation of the event-based framework developed in Flora et al. (2019), which is further adapted for this study. In this framework, we can develop ML-calibrated probabilistic guidance for individual thunderstorms that produces "event probabilities" or the likelihood of a storm producing an event within a neighborhood determined by the ensemble forecast envelope (i.e., the set of all possible storm locations predicted by the ensemble) rather than "spatial probabilities" or the probability of an event occurring within a prescribed radius of each model grid point (see Flora et al. 2019 for more on the distinction between event and spatial probabilities). We are also using the event-based approach since forecasters that use WoFS output focus on coherent regions of interest rather than strictly analyzing forecasts on a point-by-point basis (Wilson et al. 2019). However, as noted in Flora et al. (2019), a grid-based approach is the preferred method for non-thunderstorm-specific guidance, which is also being pursued using ML and WoFS forecast data (Clark et al. 2020).

We train the ML models to generate probabilistic forecasts for each severe hazard—tornado, hail, and wind—for each storm predicted by the WoFS. In evaluating the ML models, we use hazard-specific baselines generated from the WoFS forecasts of 2–5 km (midlevel) above ground level (AGL) UH, HAILCAST-based maximum hail diameter (Adams-Selin and Ziegler 2016), and 80-m AGL wind speed. For each of the three baselines, we compute the probability of exceeding a threshold (tuned per severe weather hazard) and extract the maximum probability from each ensemble storm track similar to Flora et al. (2019). The extracted probabilities were then calibrated using isotonic regression (Niculescu-Mizil and Caruana 2005; see section 4a) to improve their reliability. We hypothesize that the ML-generated probabilistic guidance should outperform the baseline predictions since the ML models can leverage more information from the CAM ensemble forecast output and provide flow-dependent corrections as opposed to using a fixed, single-threshold method.

The structure of the paper is as follows. Sections 2 and 3 describe the WoFS forecast datasets and the data processing procedures, respectively. Section 4 describes the ML models and methods used in this study. We present the results in section 5 with conclusions and limitations of the study discussed in section 6.

2. Description of the forecast data

The WoFS is an experimental multiphysics ensemble capable of producing rapidly updating severe weather guidance by frequently assimilating ongoing convection. The WoFS

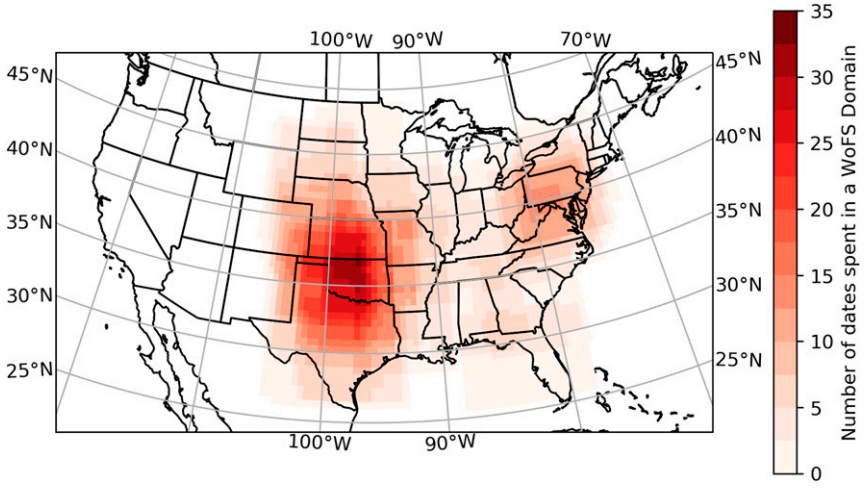

FIG. 1. Map of the number of times a 0.5° × 0.5° region was in a WoFS domain during the 2017–19 HWT-SFEs.

ensemble comprises 36 members at a 3-km horizontal grid spacing with the Advanced Research version of the Weather and Research Forecast Model (WRF-ARW; Skamarock et al. 2008) as the dynamic core. The physical parameterization configuration for the different ensemble members is provided in Skinner et al. (2018; their Table 1). The initial and lateral boundary conditions for the WoFS are provided by the experimental 3-km High-Resolution Rapid Refresh Ensemble (HRRRE; Dowell et al. 2016). The location of the WoFS domain changes daily and is centered over the region of the greatest severe weather potential. For the 2017 HWT-SFE the size of the domain was 750 km × 750 km, but for subsequent HWT-SFEs is 900 km × 900 km. Radial velocity, radar reflectivity, Geostationary Operational Environmental Satellite (*GOES-16*) cloud water path, and Oklahoma mesonet observations (when available) are assimilated every 15 min, with conventional observations assimilated hourly. During the 2017–18 HWT-SFEs, the ensemble adjustment Kalman filter included in the Data Assimilation Research Test bed (DART) software was used. During the 2019 HWT-SFE, data assimilation was performed using the Community Gridpoint Statistical Interpolation based ensemble Kalman square root filter (GSI-EnKF; Developmental Testbed Center (2017a,b). After five initial 15-min assimilation cycles, 18-member forecasts (a subset of the 36 analysis members) are issued every 30 min and provide forecast output every 5 min for up to 6 h of lead time. The reader can find additional details of the WoFS in Wheatley et al. (2015) and Jones et al. (2016, 2020).

This study uses 81 cases generated during the 2017–19 HWT-SFEs. During these experiments, WoFS domains were frequently centered over the Great Plains and mid-Atlantic with secondary focus on the Southeast and Midwest (Fig. 1). This is not surprising, because severe weather is most common over the Great Plains during the spring (severe weather has a less pronounced springtime maximum over the mid-Atlantic) and becomes more common elsewhere during the summer or cool season (Storm Prediction Center 2020). Overall, the dataset sufficiently samples environments relevant for springtime severe weather forecasting, but the trained ML algorithms may not be appropriate for year-round use.

To be consistent with recent WoFS verification studies (e.g., Skinner et al. 2018) and typical National Weather Service (NWS) warning lead times (Brooks and Correia 2018), the WoFS forecast data were aggregated into 30-min periods up to a lead time[2] of 150 min (e.g., 0–30, 5–35, . . . , 120–150 min). Given the rapid model error growth on spatiotemporal scales represented in WoFS forecasts, the whole dataset was split in two based on the forecast lead time, whereby 30-min forecast intervals beginning in the first hour (i.e., 0–30, 5–35, . . . , 60–90 min) are in one dataset (referred to as FIRST HOUR hereinafter) and forecast intervals beginning in the second hour are in a second dataset (i.e., 65–95, 70–100, . . . , 120–150 min; referred to as SECOND HOUR hereinafter). The different lead times within FIRST HOUR and SECOND HOUR are uniformly distributed (not shown). Splitting the dataset in this way allows the ML models to learn from the different forecast error characteristics in the two datasets (e.g., larger ensemble spread in SECOND HOUR than in FIRST HOUR), which should improve the models' skill. The predictability of individual storm-scale features greatly diminishes beyond 150-min lead times (Flora et al. 2018), and therefore forecasts at those lead times are not considered in this study.

3. Data preprocessing procedures

a. Ensemble storm-track identification and labeling

Object-based methods isolate important regions in a forecast space and are an effective method for reducing a large data

[2] It takes approximately 20–25 min to produce and disseminate the first two forecast hours of WoFS guidance to real-time users, so the effective lead time is shorter than the period since forecast initialization.

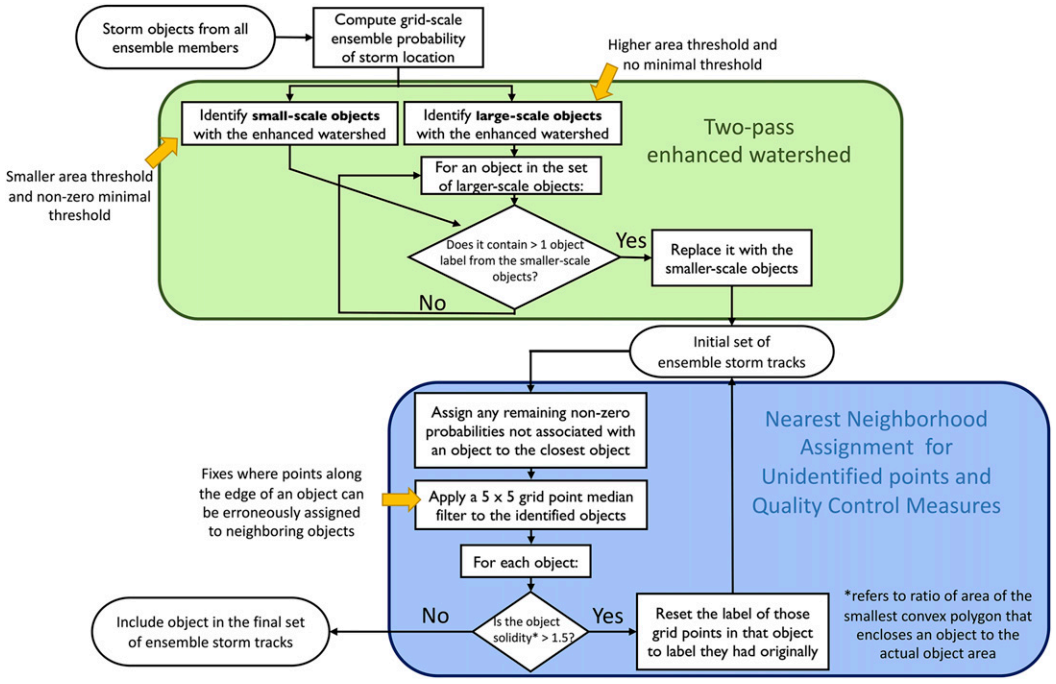

FIG. 2. Flowchart of the ensemble storm-track identification algorithm.

volume into manageable components. In past ML studies using CAM ensemble output, object-based methods have been used to extract data from individual ensemble members rather than from the ensemble as a whole (e.g., Gagne et al. 2017; Burke et al. 2020). However, there are limitations to extracting data from the individual ensemble members. First, applying an ML model to calibrate the individual member forecasts requires an additional procedure for combining the separate predictions into a single ensemble forecast (and potentially another round of calibration). Second, training ML models on the individual member forecasts neglects important ensemble attributes like the ensemble mean, which on average is a better prediction than any single deterministic forecast, and the ensemble spread (e.g., standard deviation), which can be a useful measure of forecast uncertainty. Past ML studies using CAM ensemble output have used ensemble statistics, but only in a grid-based framework (e.g., Loken et al. 2020). Therefore, we combined these past approaches by extracting ensemble information but within the event-based framework developed in Flora et al. (2019).

An ensemble storm track, conceptually, is a region bounded by ensemble forecast uncertainty in storm location. An ensemble storm track can be composed of a single ensemble member's storm track or some combination of up to all 18 ensemble members. Figures 2 and 3 show the ensemble storm-track identification algorithm and accompanying illustrations of the different steps of the procedure, respectively. First, per ensemble member, we identify storm tracks by taking peak column-maximum vertical velocity values composited over 30-min periods and thresholding them at 10 m s^{-1} (Fig. 3a). Storm tracks not meeting a 108-km^2 (12 grid cells) minimum area threshold are removed since such storms tend to be too small and/or short-lived to be likely to produce severe weather and were found to degrade the ensemble storm-track identification by producing too many objects. The ensemble probability of storm location (EP; Fig. 3b) at grid point i (based on N ensemble members) is calculated from the updraft tracks with the following equation:

$$\mathrm{EP}_i = \frac{1}{N}\sum_{j=1}^{N} \mathrm{BP}_{ij}, \tag{1}$$

where BP_{ij} (the binary probability at the ith grid point and jth ensemble member) is defined as

$$\mathrm{BP}_{ij} = \begin{cases} 1 & \text{if} \quad i \in S_j \\ 0 & \text{if} \quad i \notin S_j \end{cases}, \tag{2}$$

where S_j is the set of grid points within the updraft tracks for the jth ensemble member. The ensemble storm-track objects (Fig. 3c) are then identified from the EP field with the following procedure (Fig. 2):

1) Identify large-scale objects by applying the enhanced watershed algorithm (Lakshmanan et al. 2009; Gagne et al. 2016) with a large area threshold (3600 km^2 in this study) and no minimum threshold.
2) Identify smaller-scale objects by applying the enhanced watershed algorithm with a smaller area threshold (2700 km^2 in this study) and some minimum threshold. We choose a threshold of 5.5% (1 of 18 ensemble members) as setting the threshold higher than this causes excessive object break-up.
3) For each larger-scale object, if a larger-scale object contains multiple smaller-scale objects then replace it with the smaller-scale objects.

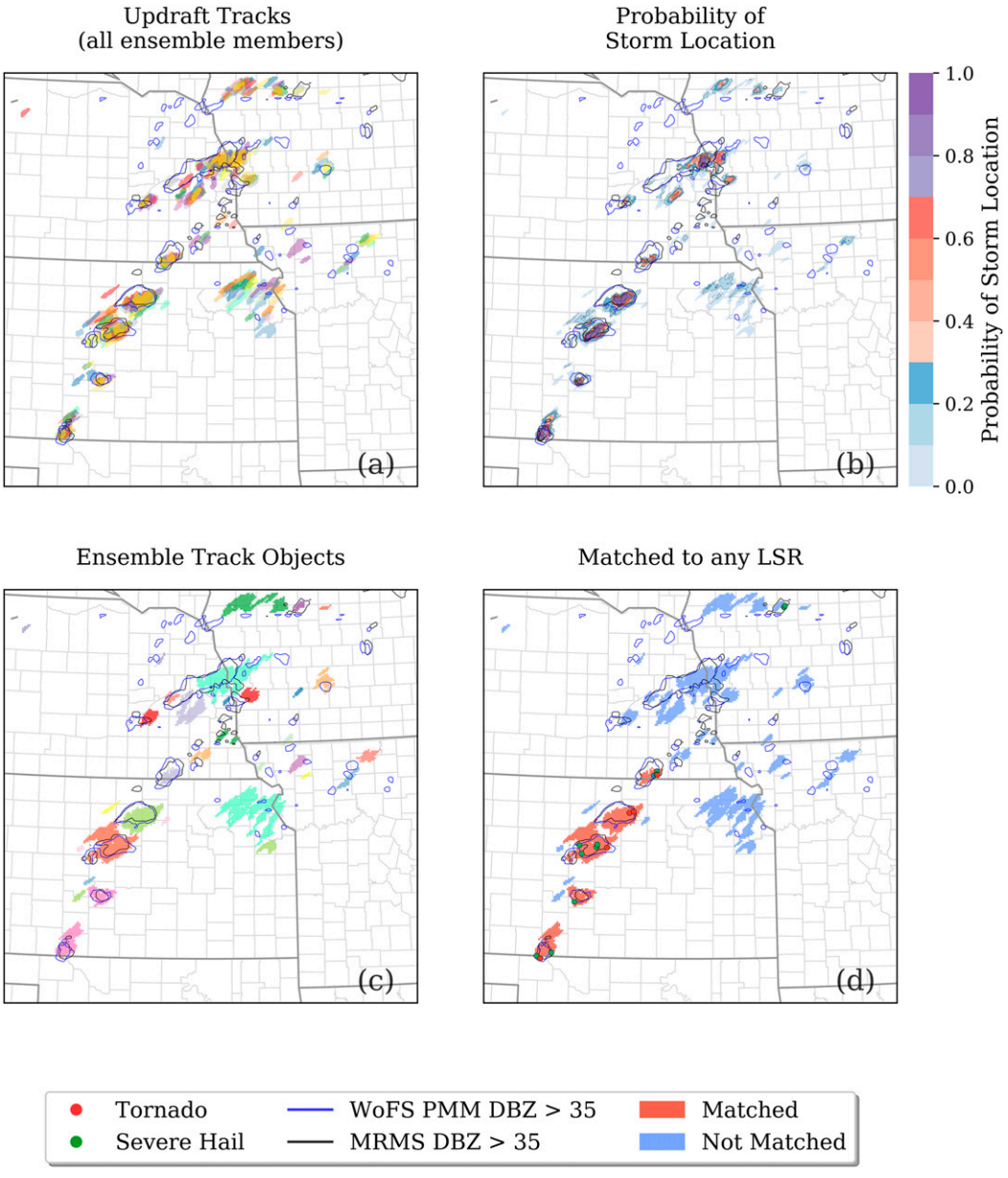

FIG. 3. Illustration of transforming individual ensemble member updraft tracks into ensemble storm tracks. (a) Paintball plot of updraft tracks identified from 30-min-maximum column-maximum vertical velocity and then quality controlled as described in section 3a. (b) Gridscale ensemble probability of storm location computed from the objects in (a). (c) Ensemble storm-track objects identified using the algorithm outlined in section 3a. (d) Ensemble storm-track objects containing a tornado (red dot) or severe hail (green dot) are shown in red (not matched are shown in blue). The technique is demonstrated using a 0–30-min forecast initialized at 2330 UTC 1 May 2018. For context, the 35-dBZ contour of the WoFS probability matched mean (blue) and Multi-Radar Multi-System (MRMS; black) composite reflectivity at forecast initialization time, respectively, are overlaid in each panel.

4) Assign any remaining nonzero probabilities not associated with an object to the closest object.
5) Apply a 5×5 gridpoint median filter to each grid point with nonzero probability (assigns it the object label that occurs most frequently within a two-gridpoint radius). This is necessary to quality control the previous step where points along the edge of an object can be erroneously assigned to neighboring objects.
6) For objects with a solidity [ratio of object area to convex area (area of the smallest convex polygon that encloses the region)] greater than a given threshold (1.5 in this study), reset the label of those grid points within that object to the

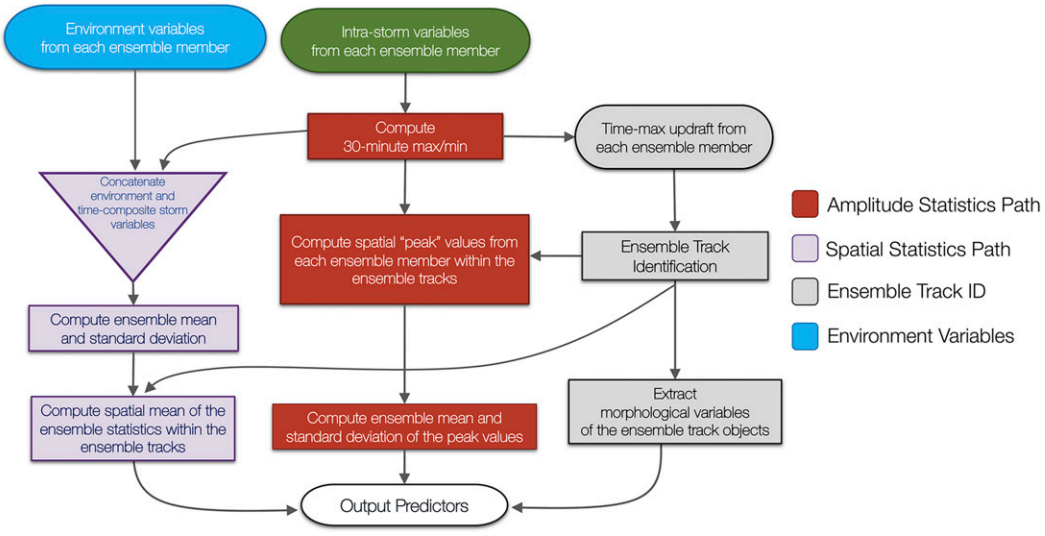

FIG. 4. Flowchart of the data preprocessing and predictor engineering used in this study. The three components are the ensemble storm-track object identification (shown in gray), the amplitude statistics (shown in red), and the spatial statistics [shown in purple (a combination of red and blue)]. Environmental variable input is shown in blue.

label they had originally (see Fig. 2). This quality control will "reset" an object if the previous steps produced an object with poor solidity.

7) Repeat steps 4–6 until no further changes occur.

The basis of the ensemble storm-track method is the enhanced watershed algorithm, which grows objects pixel-by-pixel from a set of local maxima until they reach a specified area or intensity criterion (Lakshmanan et al. 2009). Objects are restricted from growing into regions where intensity falls below the prescribed minimum threshold. Once an object is identified, it restricts additional objects from growing into the region surrounding pre-existing objects to maintain object separation (Lakshmanan et al. 2009). This two-pass procedure coupled with the nearest neighborhood assignment (step 4) addresses an issue raised in Flora et al. (2019): setting the enhanced watershed area threshold sufficiently low to prevent the merging of too many objects excessively reduced ensemble object size (see Fig. 3c in Flora et al. 2019). With this improved method, the enhanced watershed may grow objects to a greater size while maintaining object separation.

After we identify the ensemble storm tracks, we classify each according to whether it contains a tornado, severe hail, and/or severe windstorm report (Fig. 3d). To account for potential reporting time errors, we considered reports within ±15 min of either side of the 30-min forecast period (a 60-min window). Successfully predicting a severe weather hazard not only relies on correctly predicting the phenomena, but predicting the preceding convection itself. Sometimes, an observed storm may produce severe weather, but there is no corresponding forecast storm in the WoFS guidance. Since this issue is not controlled by the ML model, it does not undermine the goal of the ML prediction system, which is to predict which WoFS storms will become severe. However, the inability to account for missed storm reports where the WoFS does not predict the occurrence of a storm in a particular area highlights an important trade-off between the event-based prediction framework we use and the more traditional grid-based framework (which allows such misses to be included in the verification, but produces overly smooth forecasts). Moreover, restricting the ML model predictions to storms predicted by the WoFS mitigates the conflation of errors arising from the WoFS and from the ML models themselves, thereby facilitating verification and refinement of the ML models. Last, we recognize that local storm reports are error-prone (e.g., Brooks et al. 2003; Doswell et al. 2005; Trapp et al. 2006; Verbout et al. 2006; Cintineo et al. 2012; Potvin et al. 2019), but they are one of the best available verification databases for individual severe weather hazards, have been frequently used in past ML studies (e.g., Cintineo et al. 2014, 2018; Gagne et al. 2017; McGovern et al. 2017; Burke et al. 2020; Hill et al. 2020; Lagerquist et al. 2020; Sobash et al. 2020; Steinkruger et al. 2020), and are used in official evaluations of NWS warnings and SPC watches and outlooks.

b. *Predictor engineering*

Figure 4 depicts the data preprocessing and predictor engineering procedure. First, per ensemble member, the 30-min maximum (minimum) was calculated for the positively oriented (negatively oriented; denoted by asterisks in Table 1) intrastorm variables, and the environment variables were computed at the beginning of the valid forecast period to better sample the prestorm environment (see Table 1 for the input variables). Predictors subsequently generated from these fields are of two modes: spatial statistics (shown as the purple path in Fig. 4) or amplitude statistics (shown as the red path in Fig. 4). For the spatial statistics, we compute the ensemble mean and standard deviation at each grid point within the ensemble storm track, then spatially average them over the storm track.

TABLE 1. Input variables from the WoFS. The asterisk refers to negatively oriented variables. CAPE is convective available potential energy, CIN is convective inhibition, and LCL is the lifting condensation level. The midlevel lapse rate is computed over the 500–700-hPa layer, and the low-level lapse rate is computed over the 0–3-km layer. HAILCAST refers to maximum hail diameter from WRF-HAILCAST (Adams-Selin and Ziegler 2016; Adams-Selin et al. 2019). The buoyancy B is defined as $B = g(\theta'_{e,z=0}/\overline{\theta}_{e,z=0})$, where g is the acceleration due to gravity, $\overline{\theta}_{e,z=0}$ is the lowest-model-level average equivalent potential temperature, and $\theta'_{e,z=0}$ ($= \theta_{e,z=0} - \overline{\theta}_{e,z=0}$) is the perturbation equivalent potential temperature of the lowest model level. Values in parentheses indicate that those variables are extracted from different vertical levels or layers.

Intrastorm	Environment	Object properties
Updraft helicity (0–2 km; 2–5 km)	Storm-relative helicity (0–1 km,; 0–3 km)	Area
Cloud-top temperature*	75-hPa mixed-layer CAPE	Eccentricity
0–2-km avg vertical vorticity	75-hPa mixed-layer CIN	Orientation
Composite reflectivity	75-hPa mixed-layer LCL	Minor axis length
1–3-km max reflectivity	75-hPa mixed-layer equivalent potential temperature	Major axis length
3–5-km max reflectivity	U shear (0–6 km; 0–1 km)	Extent
80-m wind speed	V shear (0–6 km; 0–1 km)	Initialization time
10–500-m bulk wind shear	10-m U	
10-m divergence*	10-m V	
Column-max updraft	Midlevel lapse rate	
Column-min downdraft*	Low-level lapse rate	
Low-level updraft (1 km AGL)	Temperature (850, 700, and 500 hPa)	
HAILCAST max hail diameter	Dewpoint temperature (850, 700, and 500 hPa)	
Buoyancy*	Geopotential height (850, 700, and 500 hPa)	

We are only computing the spatial average (and not, e.g., the standard deviation within the storm track) to limit the number of predictors in favor of model interpretability over model complexity. We only compute amplitude statistics for the time-composite intrastorm variables. For the positively oriented (negatively oriented) intrastorm state variables, the spatial 90th (10th) percentile value (from grid points within an ensemble storm track) is computed from each ensemble member to produce an ensemble distribution of ''peak'' values. The 90th (10th) percentile is used as the ''peak value'' rather than maximum (minimum) since the maximum (minimum) value may be valid at only a single grid point, and therefore potentially unrepresentative. The ensemble mean and standard deviation are subsequently computed from each set of peak values to capture the expected amplitudes of storm features and the uncertainty therein. Reversing this procedure (i.e., computing the ensemble mean and standard deviation at each grid point and then finding the peak value) would have caused useful fine-scale details in the WoFS forecasts to be lost because of storm phase differences among ensemble members.

Last, we calculated a handful of properties describing the ensemble storm-track object morphology. These include area, eccentricity, major and minor axis length, and orientation. Altogether, there are 30 amplitude statistics, 76 spatial statistics, and 7 object properties for a total of 113 predictors.

4. Machine learning methods

a. *Machine learning models*

A linear regression model is a linear combination of learned weights β_i, predictors x_i, and a single bias term β_0:

$$z = \beta_0 + \sum_{i=1}^{N} \beta_i x_i, \tag{3}$$

where N is the number of predictors. For logistic regression, a logit transformation is applied to the output of the linear regression model:

$$p = \frac{1}{1 + \exp(-z)}, \tag{4}$$

where p is the model predictions [values between (0, 1)]. The weights are learned by minimizing the binary cross-entropy (also known as the log-loss) between the true binary labels y and model predictions with two additional terms for regularization (known together as the elastic net penalty):

$$C\sum_{k=0}^{K}[y_k \log_2(p_k) + (1 - y_k)\log_2(p_k)] + \frac{1-\alpha}{2}\sum_{k=0}^{K}\beta_k^2 + \alpha\sum_{k=0}^{K}|\beta_k|, \tag{5}$$

where K is the number of training examples, C {$=1/\lambda$, where $\lambda \in [0, \infty)$} is the inverse of the regularization parameter (adjusts the strength of the regularization terms relative to the log-loss), and $\alpha \in [0, 1]$ is a mixing parameter that adjusts the relative strength of the two regularization terms. The second term is known as the ''ridge'' penalty or L_2 error and it penalizes the model from heavily favoring predictors by encouraging the model to keep weights small. The last term is known as the ''lasso'' (least absolute shrinkage and selection operator) penalty or L_1 error and it allows weights to be zeroed out thereby removing predictors from the model. Since logistic regression explicitly combines predictors [see Eq. (3)] and the scale of the predictors can vary considerably, we normalize each training and testing set predictor by the training dataset mean and standard deviation. We did not normalize the predictors for the tree-based methods.

Tree-based methods are among the most common ML algorithms. A single classification tree recursively partitions a predictor space into a set of subregions using a series of

decision nodes where the splitting criterion favors increasing the "purity" (consisting of only one class) of these regions (Hastie et al. 2001). To prevent overfitting (restricting the subregions from becoming too narrowly defined) decision trees can be "pruned," for example, by imposing a maximum depth or removing final nodes (known as leaf nodes) below a minimum sample size. A classification random forest builds multiple, weakly correlated classification trees and merges their predictions to improve accuracy and stability over any individual decision tree (Breiman 2001). Random forests achieve the increased performance over a single decision tree by training each tree with a bootstrap resampling of the training examples and a small, random subset of predictors per split. The random forest prediction is the ensemble average of the event frequencies (from those examples in the leaf node) predicted by each individual classification tree (all trees are weighed equally). In contrast, an ensemble of decision trees can be combined using the statistical method known as gradient boosting where predictions are not made independently, but sequentially (Friedman 2002). The first tree is trained on the true targets, and then each additional tree is trained on the error residual of the previous tree. Conceptually, trees are added one at a time with each successive tree structure adjusted based on the results of the previous iteration. The final prediction of a gradient-boosted forest is the weighted sum of the predictions from the separate classification trees.

ML models may correctly rank predictions (i.e., predict the most probable class), yet produce uncalibrated probabilistic output. Isotonic regression is a nonparametric method for finding a nondecreasing (monotonic) approximation of a function and is commonly used for calibrating ML predictions (Niculescu-Mizil and Caruana 2005). Past studies in weather-based studies have found success using isotonic regression-based calibrations (Lagerquist et al. 2017; McGovern et al. 2019a; Burke et al. 2020). To compute calibrated probability estimates, isotonic regression seeks the best fit of the data that are consistent with the classifier's ranking. First, pairs of (p_i, y_i) are sorted based on p_i where p is the base classifier's uncalibrated predictions and y is the true binary labels. Starting with y_1, the algorithm moves to the right until it encounters a ranking violation ($y_i > y_{i+1}; 0 > 1$). Pairs (y_i, y_{i+1}) with ranking violations are replaced by their average and potentially averaged with previous points to maintain the monotonicity constraint. This process is repeated until all pairs are evaluated. The outcome is a model that relates a base classifier's prediction to a calibrated conditional event frequency (through the averaging of the rank violations).

In this study, we are using the random forest and logistic regression models available in the sci-kit learn package (Pedregosa et al. 2011). The gradient-boosted classification trees (XGBoost hereinafter) model comes from the open-source eXtreme Gradient Boosted (XGBoost) package (Chen and Guestrin 2016). The calibration model used is the isotonic regression model available in the sci-kit learn package (Pedregosa et al. 2011).

b. Developing baseline predictions from the WoFS

To provide baselines against which to test the ML model performance, we used WoFS forecasts of midlevel UH, HAILCAST-based maximum hail diameter, and 80-m AGL wind speed to predict which WoFS storms will produce a tornado, severe hail, and/or severe wind report, respectively. Midlevel UH has been frequently used as a baseline in other severe-weather-based ML studies (e.g., Gagne et al. 2017; Loken et al. 2020; Sobash et al. 2020) and has been used to predict tornadoes for WoFS-style forecasts (Wheatley et al. 2015; Jones et al. 2016; Yussouf et al. 2013b,a, 2016). The WRF-based HAILCAST has produced competitive next-hail day predictions (Adams-Selin et al. 2019) and 80-m AGL wind speed is a typical CAM product used by forecasters for severe wind prediction.

The baseline predictions are based on the ensemble probability of the hazard-specific variable exceeding a threshold where the ensemble probabilities are computed using Eq. (1), but the binary probability for the jth ensemble member at the ith grid point is defined as

$$\mathrm{BP}_{ij} = \begin{cases} 1 & \text{if} \quad \mathrm{f}_{ij} \geq q \\ 0 & \text{if} \quad \mathrm{f}_{ij} < q \end{cases}, \tag{6}$$

where q is the threshold and f_{ij} is the variable at the ith grid point for the jth ensemble member (Schwartz and Sobash 2017). We then set the event probability for a storm to the maximum ensemble probability within the ensemble storm track, similar to the method used in Flora et al. (2019). To tune the threshold for each severe weather hazard, we tested the baseline probabilities using fivefold cross validation on the training dataset (performance was evaluated on the five validation folds) and computed the cross-validation average performance for multiple metrics (Fig. 5). Changing the threshold for all three hazards reveals there is a trade-off between the ranking-based and calibration-based metrics (defined in section 5). Increasing the threshold improves reliability, but decreases the ability of the probabilities to discriminate between events and nonevents. For FIRST HOUR tornado prediction, we selected a threshold of midlevel UH $>180\,\mathrm{m^2\,s^{-2}}$ since a higher threshold degrades the ranking-based metrics, although reliability continues to improve (Fig. 5a). A similar argument can be made for the 1 in. (1 in. = 2.54 cm) and 40 kt ($1\,\mathrm{kt} \approx 0.51\,\mathrm{m\,s^{-1}}$) thresholds for severe hail and wind, respectively (Figs. 5c,e). A threshold near 1 in. is not unexpected for WRF-HAILCAST as it performed well against severe hail reports in past studies (Adams-Selin and Ziegler 2016; Adams-Selin et al. 2019). Given the inability to reliably produce near-surface wind speed $>$ 50 kt on a 3-km grid, it is also not surprising that the best threshold for severe wind is biased low. The results are similar in the SECOND HOUR dataset, and therefore we kept the optimal threshold the same for simplicity (Figs. 5b,d,f). Last, we found that the raw ensemble probabilities tended to be highly uncalibrated, producing substantial overforecasting biases for all three hazards (not shown). To calibrate the baseline probabilities, we trained an isotonic regression model per hazard on the probabilities produced from the training dataset.

c. Model tuning and evaluation

To assess expected model performance, both the FIRST HOUR and SECOND HOUR datasets were split into 64 dates for training and 17 dates for testing, respectively. Rather than

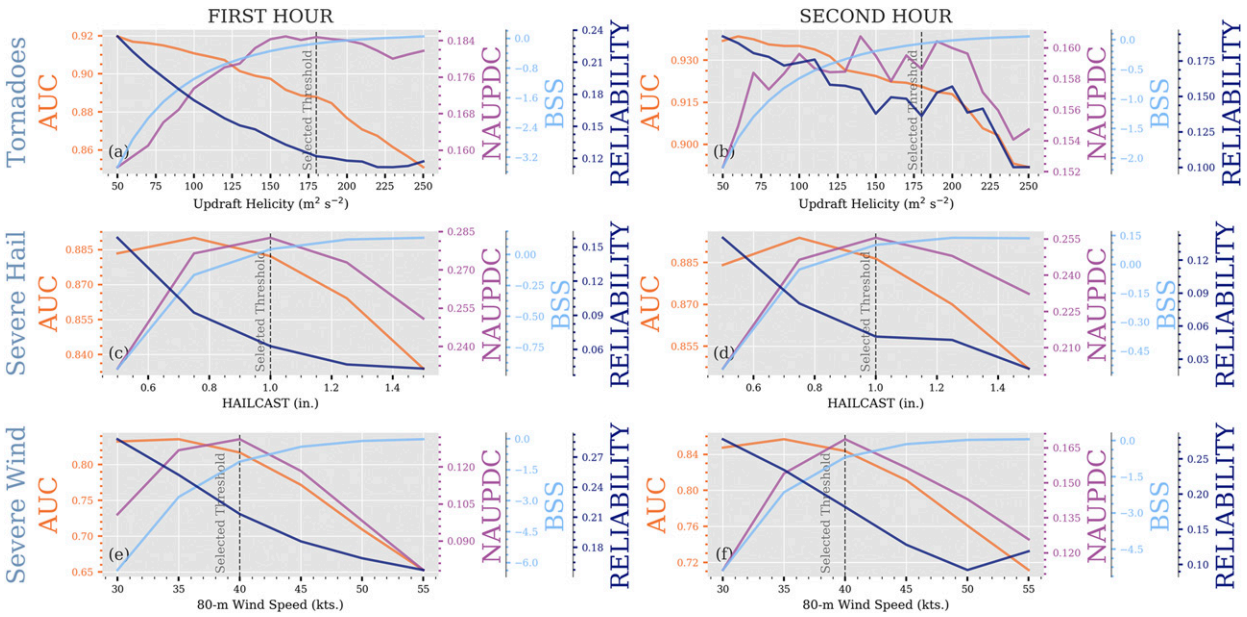

FIG. 5. Cross-validation average (within the training dataset) performance of the baseline probabilities as a function of a varying threshold for predicting (a),(b) tornadoes; (c),(d) severe hail; and (e),(f) severe wind, valid for (left) FIRST HOUR and (right) SECOND HOUR. Tornado, severe wind, and severe hail predictions are based on 2–5-km updraft helicity, 80-m wind speed, and WRF-HAILCAST maximum hail diameter from the WoFS forecast output, respectively. Metrics include AUC (orange), normalized AUPDC (NAUPDC; purple), Brier skill score (BSS; light blue), and the reliability component of the BSS (RELIABILITY; dark blue). The vertical dashed line labeled "selected threshold" indicates the threshold that optimizes certain metrics or limits trade-offs between the various metrics (see the text for details).

randomly separating the dates, we ensured that the ratio of dates with at least one event to the total number of dates was maintained for both the training and testing partitions. For example, if 40 of the 81 dates had a tornado (50%), then this ratio was approximately maintained in both the training and testing dataset. This simple approach helps ensure that the testing dataset is more representative of the training dataset, which limits bias in the assessment of model performance. We provide the number of examples in each training and testing dataset per hazard in Table 2.

Bayesian hyperparameter optimization (hyperopt; Bergstra et al. 2013) was used to identify the optimal hyperparameters for each model using fivefold cross validation over the training dataset. The hyperopt python package is based on a random search method but implements a Bayesian approach where performance on previous iterations helps determine the optimal hyperparameters. For this study, we are using the area under the performance diagram curve (defined in section 5c) as our optimization metric. The default stopping criterion in hyperopt is a user-set maximum number of evaluation rounds, so we implemented an early stopping criterion where a 1% improvement in performance must occur within a set number of rounds or else optimizing stops, which improves computational efficiency (we found that requiring said improvement at least every 10 rounds was sufficient). The hyperparameters and values used for each model are presented in Table 3. For those hyperparameters not listed, we used the default values in version 0.22 of the scikit-learn software (Pedregosa et al. 2011) and version 0.82 of the XGBoost software (Chen and Guestrin 2016). The optimal hyperparameter values for each model and severe weather hazard for the FIRST HOUR and SECOND HOUR dataset are provided in Tables 4 and 5, respectively.

Although the original model predictions were not significantly uncalibrated, we found that including isotonic regression improved the reliability of the ML probabilities (not shown). To prevent introducing bias, the isotonic regression is typically trained on the predictions and labels of the base model on a validation dataset. Rather than training on an independent validation dataset, we use the cross-validation approach from Platt (1999) where the base model is fit on each training fold and used to make predictions on the corresponding validation fold. The calibration model (e.g., isotonic regression) is then trained on the concatenation of the predictions from the different cross-validation folds. The base model can then be

TABLE 2. Numbers of examples in the training and testing datasets for the different severe weather hazards and lead time intervals.

	Training	Testing
	FIRST HOUR	
Tornado	346 341	82 750
Severe hail	349 508	79 583
Severe wind	330 840	98 251
	SECOND HOUR	
Tornado	262 878	82 483
Severe hail	258 270	87 091
Severe wind	258 991	86 370

TABLE 3. Hyperparameter values attempted for each model in the hyperparameter optimization.

Hyperparameter	Values
Random forest	
No. of trees	100, 250, 300, 500, 750, 1000, 1250, and 1500
Max depth	5, 10, 15, 20, 30, 40, and none
Min leaf node sample size	1, 5, 10, 15, 25, and 50
XGBoost	
No. of trees	100, 250, 300, 500, 750, 1000, 1250, and 1500
Min loss reduction γ	0, 0.001, 0.01, 0.3, 0.5, and 1
Max depth	2, 4,7, and 10
Learning rate η	10^{-1}, 10^{-2}, 10^{-3}, and 10^{-4}
Min child weight	1, 5, 10, 15, and 25
Ratio of predictors randomly selected per tree	0.7, 0.8, and 1.0
Subsample ratio of the examples	0.5, 0.6, 0.7, and 1.0
L_1 weight	0, 0.5, 1, 10, and 15
L_2 weight	0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1, and 1.0
Logistic regression	
C	0.0001, 0.001, 0.01, 0.1, 1.0
ρ (l1_ratio)	0.0001, 0.001, 0.01, 0.5, 1.0

TABLE 4. Optimal hyperparameter values for each model and severe weather hazard for the FIRST HOUR dataset.

Hyperparameter	Tornadoes	Severe hail	Severe Wind
Random forest			
No. of trees	100	1500	250
Max depth	40	40	20
Min leaf node sample size	10	1	1
XGBoost			
No. of trees	300	250	300
Min loss reduction (γ)	0.5	0	0
Max depth	10	10	7
Learning rate (η)	0.1	0.1	0.1
Min child weight	1	1	15
Ratio of predictors randomly selected per tree	0.7	0.8	0.8
Subsample ratio of the examples	1.0	0.6	1.0
L_1 weight	0.5	1	1
L_2 weight	0.001	0.0005	0.1
Logistic regression			
C	0.1	0.01	0.01
ρ (l1_ratio)	0.0001	0.01	0.001

refitted to the whole training dataset (with the optimal hyperparameters), while the calibration model is effectively fit on the whole training dataset without biasing the predictions.

For the final assessment, we evaluated the ML models and their respective baselines on the testing datasets. All metrics are bootstrapped ($N = 1000$) to produce confidence intervals for significance testing. For an unbiased measure of variance, the bootstrapping method requires independent samples, but our testing samples come from overlapping forecast ranges (0–30, 5–35, 10–40, etc.) and therefore are not independent from one another. We do not track the ensemble object in time, and therefore we cannot compute serial correlations on the full dataset. Based on a manual analysis of a small subset, however, we found that serial correlations for some predictors were not negligible (e.g., $r = 0.2$), but small enough that the confidence intervals should not markedly underestimate the true uncertainty of the various verification scores. The following verification results are aggregated over each dataset, FIRST HOUR and SECOND HOUR, respectively, but we found that performance for individual forecast lead times is fairly consistent (with some variance) within each dataset (not shown).

d. Sensitivity to class imbalance

The full dataset (combined FIRST HOUR and SECOND HOUR) used in this study is heavily imbalanced toward nonevents; 1.2%, 2.5%, and 4% of ensemble storm-track objects are matched to a tornado, severe hail, or severe wind report, respectively. ML algorithms often struggle to learn patterns and relationships from imbalanced datasets (Batista et al. 2004; Sun et al. 2009). A common method to reduce the effect of class imbalance is to randomly undersample the majority class (i.e., nonevents) to produce a balance of events and nonevents. However, for all three ML algorithms, we found that the model performance for each hazard was negligibly impacted by resampling the training dataset. Therefore, we did not resample to produce balanced classes prior to fitting the ML models in this study. The above result is not surprising if the class separation is sufficient to counteract the class imbalance. There are a significant number of ensemble storm tracks that are

TABLE 5. As in Table 4, but the SECOND HOUR dataset.

Hyperparameter	Tornadoes	Severe hail	Severe Wind
Random forest			
No. of trees	1250	1250	250
Max depth	20	20	40
Min leaf node sample size	50	5	5
XGBoost			
No. of trees	250	500	300
Min loss reduction (γ)	0	0	1.0
Max depth	10	10	10
Learning rate (η)	0.1	0.1	0.1
Min child weight	10	5	25
Ratio of predictors randomly selected per tree	0.7	1.0	0.8
Subsample ratio of the examples	0.7	1.0	0.7
L_1 weight (α)	1	0.5	10
L_2 weight (λ)	0.01	0.1	1.0
Logistic regression			
C	0.01	0.01	0.01
ρ (l1_ratio)	0.001	1.0	1.0

small (e.g., only composed of a single ensemble member's updraft track) and these are rarely matched to storm reports, making them easily distinguishable as nonevents. The ML algorithms are learning this distinction and are therefore better able to learn skillful relationships despite the training dataset having significant class imbalance (Flora 2020).

5. Results

The verification methods for this study include the receiver operating characteristic (ROC) curve (Metz 1978), performance diagram (Roebber 2009), and the attribute diagram (Hsu and Murphy 1986). The ROC curve and performance diagram are derived from converting forecast probabilities to a set of yes/no forecasts based on different probability thresholds and computing contingency table metrics. The four components of the contingency table are as follows:

1) "Hits": forecast "yes" for a given hazard and the ensemble storm track is matched to a corresponding LSR.
2) "Misses": forecast "no" for a given hazard, but the ensemble storm track is matched to a corresponding LSR.
3) "False alarms": forecast "yes" for a given hazard, but the ensemble storm track is not matched to a corresponding LSR.
4) "Correct negatives": forecast "no" for a given hazard and the ensemble storm track is not matched to a corresponding LSR.

The most common contingency metrics include probability of detection [POD; $a/(a + c)$], probability of false detection [POFD; $b/(b + d)$], success ratio [SR; $a/(a + b)$], false alarm ratio [FAR; $b/(a + b)$], critical success index [CSI; $a/(a + b + c)$], and frequency bias [$(a + b)/(a + c)$], where a, b, c, and d are the number of hits, false alarms, misses, and correct negatives, respectively.

a. Example forecasts

Figure 6 shows characteristic examples of good and poor forecasts from the random forest model; these represent the other models as well (not shown). These examples include high-confidence (probabilities closest to 1) forecasts matched and not matched to an event and low-confidence (probabilities closest to 0) forecasts matched to an event. The skill of the ML forecasts is largely driven by the ability of the WoFS to accurately analyze ongoing convection through data assimilation. The classification, as we will see, is sensitive to slight changes in object location/separation. There may be minimal subjective differences between a confident match and confident false alarm (high-confidence forecast not matched to the event), which is a limitation of the current method. For example, for high-confidence (higher probabilities) forecasts matched to an event, the convection is fairly organized, and the WoFS matches well with the observed reflectivity (Figs. 6a,d,g). Unfortunately, high-confidence forecasts not matched to an event can exhibit similar behavior (Figs. 6b,e,h). In Figs. 6a and 6b, storms in the Texas Panhandle have similar tornado probabilities despite only one of them producing tornado LSRs. It is possible that in this case the useful information for tornado forecasting in the WoFS was confined to larger spatial scales preventing discrimination of tornadic and nontornadic storms occurring in proximity to one another. Complicating the interpretation, some of these apparent forecast busts may in fact be associated with an unreported event. For example, Potvin et al. (2019) found that over 50% of tornadoes went unreported in the central United States from 1975 to 2016. For severe wind (Fig. 6h), the timing of the higher confidence forecast was premature as severe wind reports were eventually observed on the border of southern Ohio and northwest Kentucky (though the observed storms were outside the WoFS domain).

As we can see in Fig. 6, the ensemble storm tracks can be organized on a variety of spatial scales and properly identifying those features can be difficult for current object identification methods. One limitation of the current ensemble storm-track identification method is that in some cases it may not be able to isolate threats within convection organized on larger scales (see the linear convective modes in Figs. 6a,b,d,e). In future work, we will refine the ensemble track identification method to better identify storm tracks embedded within larger-scale storm tracks. This issue can also stem from the inability of the WoFS to reliably resolve isolated threats within larger-scale convection.

For low-confidence forecasts of severe hail and severe wind matched to an event, the convection is discrete and poorly organized (Fig. 6f) or disorganized and complex (Fig. 6i). For the first case, discrete, poorly organized convection suggests a weakly forced environment that has lower predictability and in which it is more difficult to produce an accurate ensemble analysis. For the second case the WoFS reflectivity generally agrees with the observed reflectivity, but the severe wind reports are associated with the weaker, isolated convection, which can have limited predictability as well (similar for tornadoes; Fig. 6c).

LSRs sometimes occur just outside of the boundaries of the ensemble storm tracks; see, for example, the severe hail report associated with the northernmost storm in Oklahoma in Fig. 6e. These missed reports may be unduly penalizing the ML model performance as they are likely associated with storm motion biases in the WoFS forecasts (Skinner et al. 2018; Flora et al. 2019), which is not controlled by the ML model. On the other hand, the ensemble storm-track areas are larger than a typical warning polygon and represent the WoFS's full range of storm location, and so our matching criterion is already relatively lenient. Given the impact of misses arising from small spatial errors in forecast storm tracks and spurious false alarms arising from missing reports, however, we argue that the following verification results likely underestimate the true skill of the ML models.

b. ROC diagrams

The ROC curve plots POD against POFD for a series of probability thresholds and, coupled with the area under the ROC curve (AUC), assesses the ability of the forecast system to discriminate between events and nonevents. An AUC = 0.5 indicates a no-skill prediction while a perfect discriminator will score an AUC = 1. All three ML models produced, on average, an AUC greater than 0.9 for all three severe weather hazards for both lead time sets (Fig. 7). While the ML model AUC

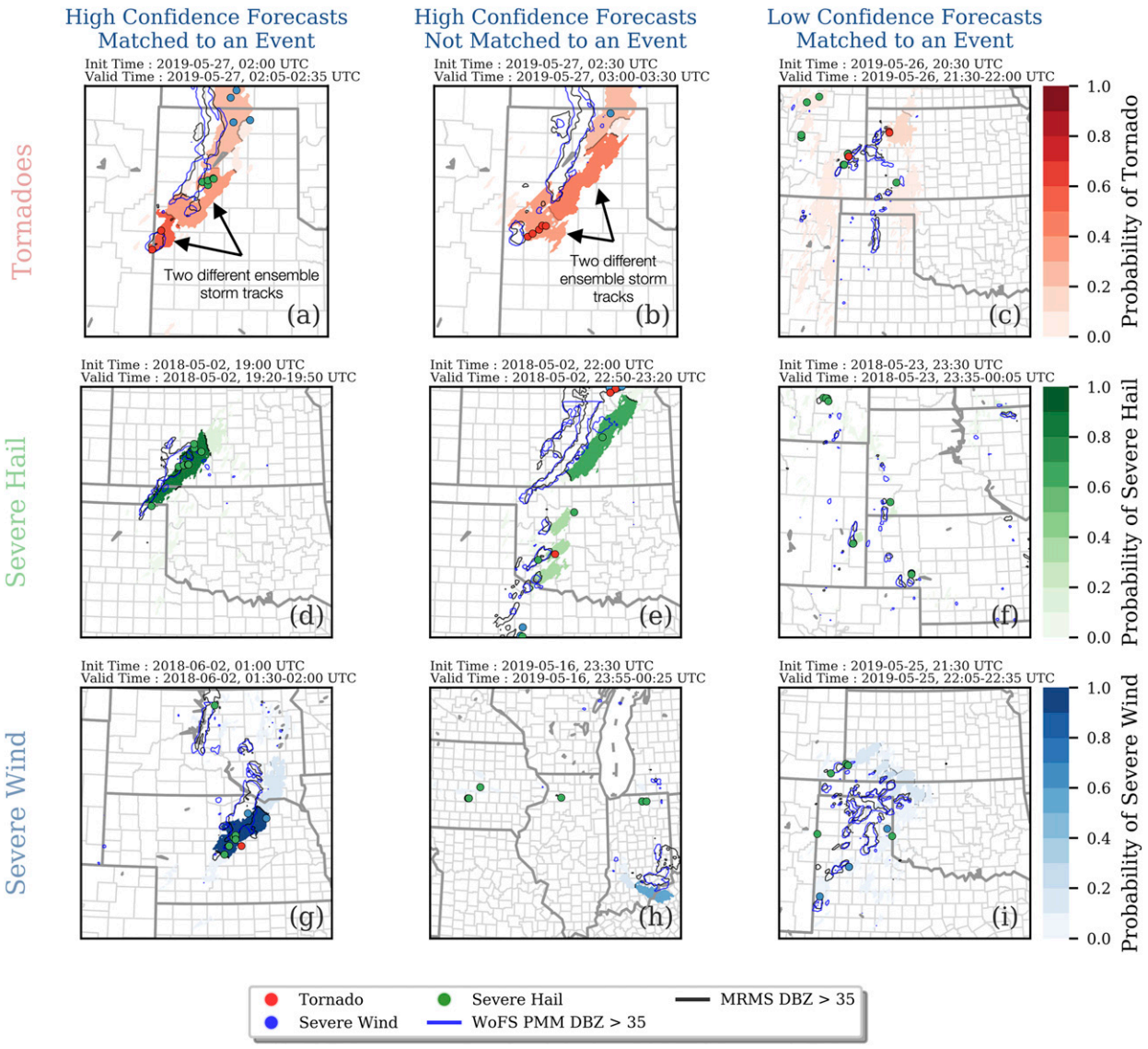

FIG. 6. Examples forecast from the random forest model predicting (a)–(c) tornadoes, (d)–(f) severe hail, and (g)–(i) severe wind. These forecasts are representative instances of (left) a high-confidence forecast matched to an event, (center) a high-confidence forecast not matched to an event, and (right) a low-confidence forecast matched to an event. For context, the 35-dBZ contour of the WoFS probability matched mean (blue) and MRMS (black) composite reflectivity at forecast initialization time, respectively, are overlaid in each panel. The forecast initialization and valid forecast period are provided in the top-left hand corner of each panel. Tornado, severe hail, and severe wind reports are shown as red, green, and blues circles, respectively. The tornado forecasts in (a) and (b) have been zoomed in to focus on the isolated supercell and the southern end of the MCS over the Texas Panhandle. The annotation highlights the two different ensemble storm tracks associated with two different observed storms.

scores were substantially better than those for the baseline predictions, the latter were near or above 0.9, suggesting that the WoFS guidance is already a fairly good discriminator for the three severe weather hazards. While the AUC is high, it is important to consider that this score is invariant to class imbalance and weighs event and nonevent examples equally. Thus, the AUC provides an overly optimistic assessment of discrimination in applications where less importance is placed on correctly predicting nonevents. For severe weather prediction, correct negatives are conditionally important in that it is only desirable to accurately predict nonevents in environments that favor severe weather (to reduce false alarms). However, a large number of ensemble storm tracks are easily distinguishable as nonevents (as mentioned in section 4d), which suggests that caution be exercised when interpreting the high AUC values in this study. This effect also explains why AUC increases for severe weather hazards with lower climatological event frequencies; for rarer events, the aforementioned ensemble storm tracks become even easier to identify as nonevents.

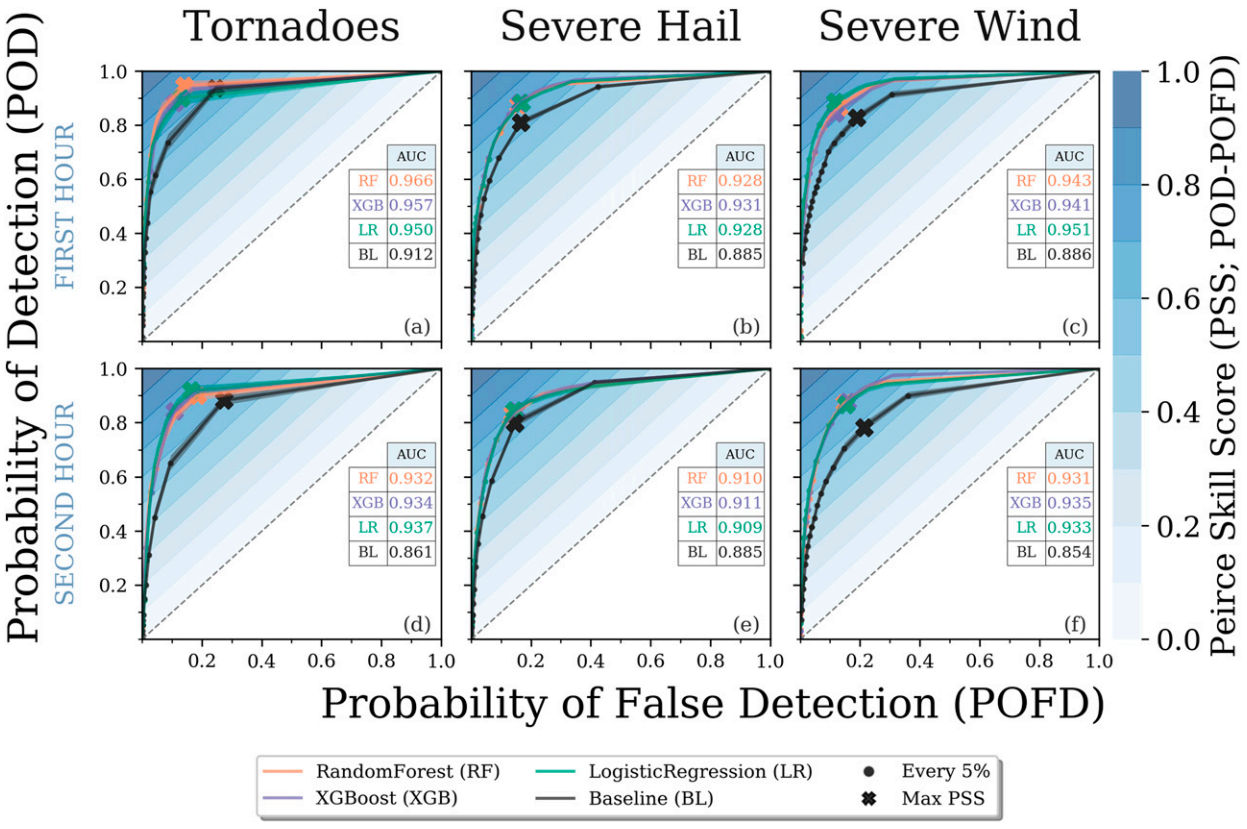

FIG. 7. ROC curves for the random forests (RF; light orange), gradient-boosted classifier trees [XGBoost (XGB); light purple], logistic regression (LR; green), and UH baseline (BL; black) predicting whether an ensemble storm track will contain a (left) tornado, (center) severe hail, or (right) severe wind report. Results are combined over 30-min predictions starting within the lead times (a)–(c) in the first hour (i.e., 0–30, 5–35, . . . , 60–90 min) and (d)–(f) in the second hour (i.e., 65–95, 70–100, . . . , 120–150 min). Each line (shaded area) is the mean (95% confidence interval), determined by bootstrapping the testing examples ($N = 1000$). Curves were calculated every 0.5%, with dots plotted every 5%. The diagonal dashed line indicates a random classifier (no skill). The mean AUC for each model is provided in the table on the right-hand side of each panel. The filled contours are the Pierce skill score (PSS; also known as the true skill score), which is defined as POD − POFD. The maximum PSS is indicated on each curve with an X.

c. *Performance diagrams*

The performance diagram[3] plots the SR against the POD for a series of probability thresholds and assesses the ability of the model to correctly predict an event while ignoring correct negatives (Roebber 2009). The performance diagram is complementary to the ROC curve, especially for imbalanced prediction problems (like severe weather forecasting) where it is more important to correctly predict events than nonevents (Davis and Goadrich 2006). CSI and frequency bias are functionally related to POD and SR and are also displayed on the performance diagram. A probabilistic forecast is considered to have perfect performance when the CSI and frequency bias are equal to 1 (corresponding to the upper right corner) for some probability threshold. However, for probabilistic forecasts of rare events, a maximum CSI of 1 is practically unachievable (Hitchens et al. 2013) and the maximum CSI tends to be associated with a frequency bias >1 (Baldwin and Kain 2006).

Similar to the ROC diagram, one can compute the area under the performance diagram curve (AUPDC[4]). Rather than computing the area through integration, which can be too optimistic, it is more robust to compute AUPDC from the weighted average of SR[5] (Boyd et al. 2013):

$$\text{AUPDC} = \sum_{k=1}^{K} (\text{POD}_k - \text{POD}_{k-1})\text{SR}_k, \tag{7}$$

where K is the number of probability thresholds used to calculate POD and SR. For this study, POD and SR were computed every 0.5% ($K = 200$). Unlike AUC, AUPDC is a function of class imbalance as changing the ratio of events to

[3] Commonly known as the precision-recall diagram (Manning and Schütze 1999) in the ML community where recall is POD and precision is SR.

[4] Also known as the area under the precision-recall curve, which is often acronymized as AUPRC or AUCPR.

[5] Known better by the term "average precision" where precision is synonymous with success ratio.

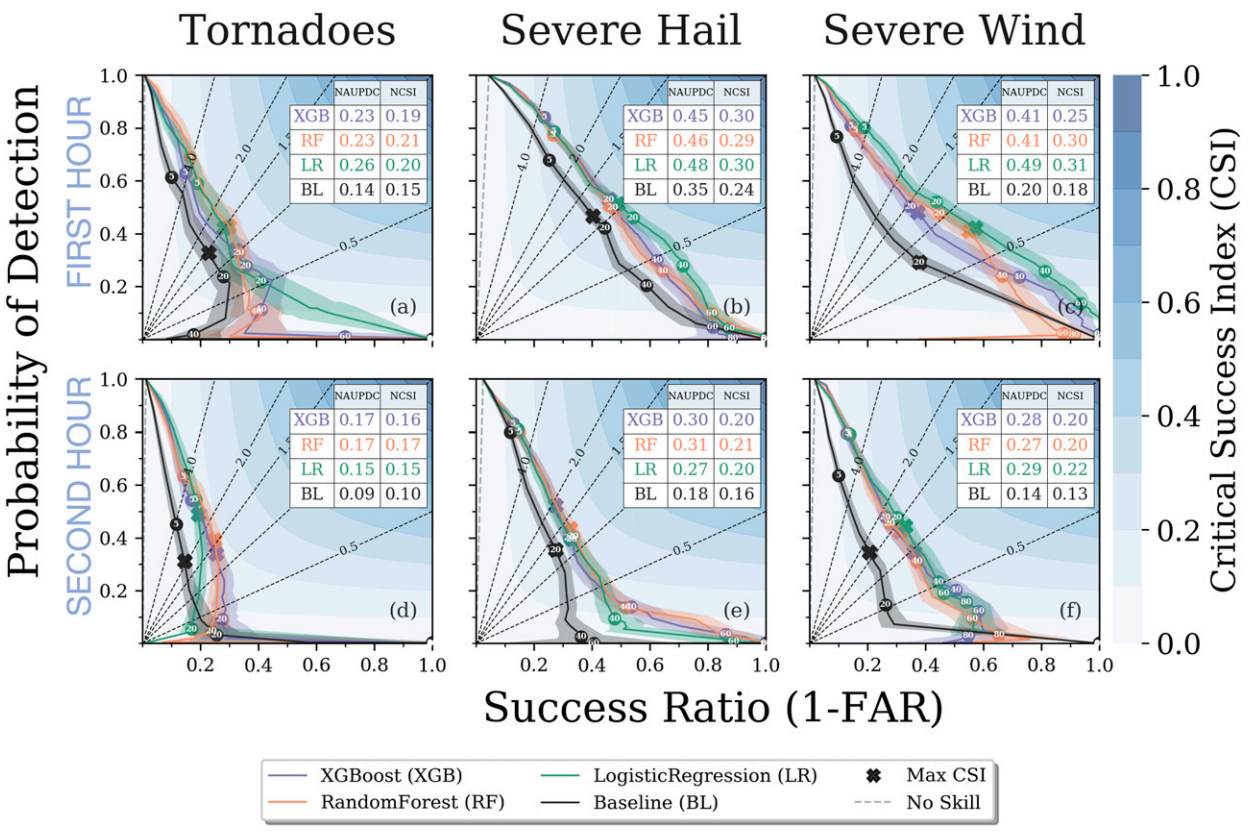

FIG. 8. As in Fig. 7, but for the performance diagram. The filled contours indicate the CSI, and the dashed diagonal lines are the frequency bias. The dashed gray line indicates a no-skill classifier defined by Eq. (8). The mean NAUPDC and NCSI for each model are provided in the table in the top-right corner of each panel. Points associated with the following probability thresholds are highlighted: 5%, 20%, 40%, 60%, and 80%. The maximum CSI is indicated on each curve with an X.

nonevents will alter the minimum possible SR, defined in Boyd et al. (2012) as

$$\mathrm{SR}_{\min} = \frac{c\mathrm{POD}}{1 - c + c\mathrm{POD}}, \tag{8}$$

where c is the climatological event frequency of the dataset (number of events divided by the total number of examples). If a curve lies along $\mathrm{SR}_{\min}$, the prediction system is considered to have no skill. Therefore, one can normalize AUDPC by the minimum possible AUPDC (Boyd et al. 2012), which facilitates comparing the model skill on datasets with different climatological event frequencies for a given hazard or comparing model performance for different hazards with different climatological event frequencies. The minimum AUPDC is

$$\mathrm{AUPDC}_{\min} = \frac{1}{\mathrm{pos}} \sum_{i=1}^{\mathrm{pos}} \frac{i}{i + \mathrm{neg}}, \tag{9}$$

where pos and neg are the number of event and nonevent examples in the testing dataset, respectively (Boyd et al. 2012). The normalized AUPDC (NAUPDC) is defined as

$$\mathrm{NAUPDC} = \frac{\mathrm{AUPDC} - \mathrm{AUPDC}_{\min}}{1 - \mathrm{AUPDC}_{\min}}. \tag{10}$$

Regardless of climatological event frequency, the best possible classifier will have an NAUPDC of 1 and the worst possible classifier will have an NAUPDC of 0. We can also normalize the maximum CSI by the maximum CSI of a no-skill system [equal to the climatological event frequency (c); derivation provided in the appendix] using a computation similar to Eq. (10) (hereinafter referred to as NCSI):

$$\mathrm{NCSI} = \frac{\mathrm{CSI}_{\max} - c}{1 - c}. \tag{11}$$

The performance diagrams are shown in Fig. 8. For the FIRST HOUR dataset (e.g., examples with a lead time of 0–30, 5–35, . . . , 60–90 min; Figs. 8a–c), the three ML models produced higher NAUPDC and maximum NCSI for severe hail and wind (Figs. 8b,c) than for tornadoes (Fig. 8a). Severe wind and hail events are more frequent than tornadoes, which gives the ML models more opportunities to learn from those examples. In addition, the processes governing hail growth and generation of strong near-surface winds are better resolved on a 3-km grid than the processes governing tornadogenesis, which is strongly influenced by small-scale processes in at least some cases (Coffer et al. 2017; Flournoy et al. 2020). For tornadoes and severe hail,

the NAUPDC and maximum NCSI of the three ML models were fairly indistinguishable from one another (Figs. 8a,b), but for severe wind (Fig. 8c), the random forest and logistic regression models produced substantially higher maximum NCSI than XGBoost. Other than for the severe wind random forest and logistic regression model, the frequency bias associated with maximum NCSI is greater than 1 (Figs. 8a,b), which matches expectations for rare events (Baldwin and Kain 2006).

All three ML models substantially outperformed their respective baselines, but the magnitude of improvement varied with severe weather hazard with the most substantial improvement for severe wind-based ML models. These results suggest that WoFS forecasts of 80-m AGL wind speed struggle to predict the strength of near-surface winds. WoFS has demonstrated success in capturing high wind events (e.g., 2020 Iowa Derecho), but it may be less successful in more marginal events where the predictability is lower. A definitive explanation for severe wind-based ML models having the largest improvement over the baseline prediction is beyond the scope of this paper, but warrants further exploration. Ultimately, these results highlight the ability of the ML models to leverage multiple predictors to produce the skillful guidance.

The performance curves were degraded for the SECOND HOUR dataset (e.g., examples with a lead time of 65–95, 70–100, . . . , 120–150 min; Figs. 8d–f). For probabilities ≤5%, the FAR remained relatively unchanged for tornadoes and the POD decreased, but for probabilities >5%, the FAR substantially increased, which decreased the NAUPDC and maximum NCSI. The increase in FAR also led to the maximum CSI occurring with an increased overforecasting frequency bias (especially for logistic regression). The predictability of storm-scale features relevant to tornado prediction (e.g., mid- and low-level mesocyclones) is greatly diminished at longer lead times (Flora et al. 2018) and therefore this degradation in skill is not surprising. For severe hail and wind (Figs. 8e,f), the changes in POD and FAR relative to FIRST HOUR compensated each other such that the maximum-CSI frequency bias increased to slightly above 1 (except the XGBoost model, which has a maximum-CSI frequency bias near 2.0). The major exception is the XGBoost severe wind model, which suffered from overforecasting bias in the FIRST HOUR dataset but in the SECOND HOUR dataset has a maximum-CSI frequency bias near 1 (1.08). The difference in performance between the baseline predictions and the three ML models is more pronounced in SECOND HOUR than FIRST HOUR suggesting that ML-based calibration of ensemble forecasts is more useful at longer lead times. This result suggests that the ML models are learning enough useful information from the ensemble statistics at these later lead times to partly compensate for the inevitable reduction in CAM forecast skill because of intrinsically limited storm-scale predictability.

For the FIRST HOUR dataset, the logistic regression models produced slightly higher mean NAUPDC values compared to the other ML models, which is associated with logistic regression producing higher SR (lower FAR) for all three severe hazards for probabilities ≥20%. To explain why the tree-based methods are producing more false alarms for higher confidence forecasts (i.e., >20%) than logistic regression, Fig. 9 illustrates how predictions from a random forest and logistic regression model compare for a simple noisy, imbalanced 2D dataset. A classic problem in ML is the trade-off between the bias and variance of a model. With a high-variance model, we risk overfitting to noisy or unrepresentative training data. In contrast, a high-bias model is typically simpler and tends to underfit the training data, failing to capture important regularities. By partitioning the predictor space into subregions, tree-based methods tend to produce highly complex decision surfaces (Fig. 9b). Tree-based methods derive their predictions from the local event frequencies in these subregions and if there is misclassification (e.g., ensemble storm tracks mislabeled as nonevents because of missing storm reports) or if the subregions have too few samples, then the local event frequencies can be biased. As a result, tree-based methods can struggle near decision boundaries or in poorly sampled regions of the predictor space. For example, near point $(X_1, X_2) = (-1, 1)$, the random forest probabilities do not reflect the uncertainty of the true labels and for points $X_2 > 2$, the predictions have high confidence, but instances of unrepresentative uncertainty [e.g., the probability of point $(X_1, X_2) = (2, 2.5)$ is 50%, but should be 100%]. It is well-documented that all three severe storms hazards suffer from significant reporting biases (Trapp et al. 2006; Allen and Tippett 2015; Potvin et al. 2019). The resulting misclassified storms coupled with poorly sampled phase spaces in our training dataset plausibly explain why the tree-based methods produce fewer higher-confidence forecasts than do the logistic regression models.

The logistic regression models, though, produced similar maximum NCSI as the other models and for the SECOND HOUR dataset, the overall difference in performance curves between the tree-based methods and logistic regression is fairly insignificant. Ultimately, logistic regression is a high bias model (which does not sufficiently generalize the data) and we suspect that with additional training data and an improved severe weather database that tree-based methods would outperform logistic regression.

d. Attribute diagrams

The attribute diagram plots forecast probabilities against their conditional event frequencies (Wilks 2011). Thus, the plot for a perfectly reliable forecast system will lie along the one-to-one line. Traditionally, the forecast probabilities are separated into equally spaced bins from which we compute the mean forecast probabilities and conditional event frequencies. The conditional event frequencies, however, can be sensitive to the bin interval, especially for smaller datasets. To address uncertainty in the conditional event frequencies, we computed the "consistency bars" from Bröcker and Smith (2007), which allows for an immediate interpretation of the confidence of the reliability of a prediction system. We can then assess reliability as the extent to which the conditional event frequencies fall within the consistency bars rather than strictly based on their distance from the diagonal. A common metric associated with

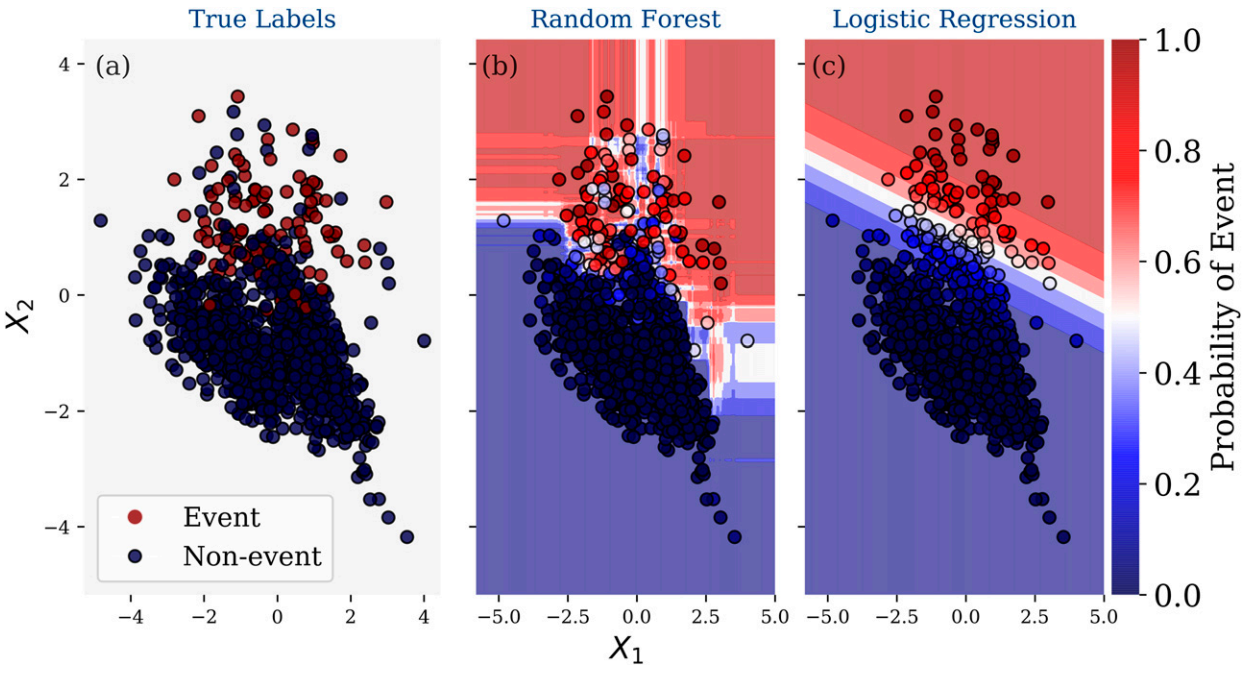

FIG. 9. Illustration of predictions for (a) a simple noisy imbalanced 2D dataset from (b) a random forest and (c) a logistic regression model. The filled contours show the predictions of the two models as decision surfaces.

the attribute diagram is the Brier skill score (BSS; Hsu and Murphy 1986) where regions of positive and negative BSS can be delimited on the attribute diagram based on the climatological event frequency. The Brier skill score is defined as

$$\mathrm{BSS} = \frac{\left[\frac{1}{K}\sum_{k=1}^{N} n_k(\overline{y_k} - \overline{y})^2\right] - \left[\frac{1}{N}\sum_{k=1}^{K} n_k(p_k - \overline{y_k})^2\right]}{\overline{y}(1-\overline{y})}, \qquad (12)$$

where p is the forecast probabilities, y is the binary target variable, K is the number of bins, N is the number of examples, n_k is the number of examples in the kth bin, $\overline{y_k}$ is the conditional event frequency in the kth bin, and $\overline{y}$ is the climatological event frequency. The two terms in the numerator (from left to right) are known as resolution and reliability, respectively, while the denominator is the uncertainty term. Reliability measures how well the forecast probabilities correspond with the conditional event frequencies while resolution measures how the conditional event frequencies differ from the climatological event frequency. The uncertainty term refers to uncertainty in the observations and is independent of forecast quality. A positive BSS (resolution > reliability) means that the model is better than the baseline prediction (climatological event frequency). BSS is sensitive to class imbalance, but the authors are unaware of any methods that attempt to normalize BSS by the climatological event frequency.

The attribute diagram results are shown in Fig. 10. For both lead time ranges, the severe hail- and severe wind-based ML models produced higher probabilities than the tornado-based models (cf Fig. 10b,c,e,f and Figs. 10a,d). The smaller forecast probabilities for tornadoes are not surprising for at least three reasons. First, as noted in the previous section, there are more severe hail and wind events than tornado events in the training dataset, which likely contributes to increased reliability by improving the local event frequencies for the tree-based methods and the coefficients of the linear model in logistic regression. Second, the processes governing tornadogenesis are not well represented on a 3-km grid and can include chaotic intrastorm processes such that weak tornadoes can form in environment otherwise characterized as nontornadic (Coffer et al. 2017, 2019; Flournoy et al. 2020), which lessens the signal-to-noise ratio and lowers ML model confidence. Third, storm-scale predictability limits (Flora et al. 2018) prevent greater confidence in tornado likelihood, especially at longer lead times.

For the FIRST HOUR dataset, all three ML models produced reliable severe wind probabilities up to 40%–50% with a modest underforecasting bias for higher probabilities (Fig. 10c). Severe hail probabilities for all three models were reliable up to 40% with a slight underforecasting bias for probabilities greater than 60% with probabilities up to 90% being produced (Fig. 10b). For severe hail and wind (Figs. 10b,c), the underforecasting bias was highest for logistic regression, which corresponds to the lower FAR at higher probabilities noted in the previous section.

For all severe weather hazards, reliability and resolution were degraded for the SECOND HOUR dataset. The tree-based tornado probabilities are arguably reliable, but all three ML models have a maximum probability between 30% and 40%, though these are fairly confident forecasts of such a rare event. For severe hail, the forecast probabilities below 60% were relatively reliable (an under forecasting bias for

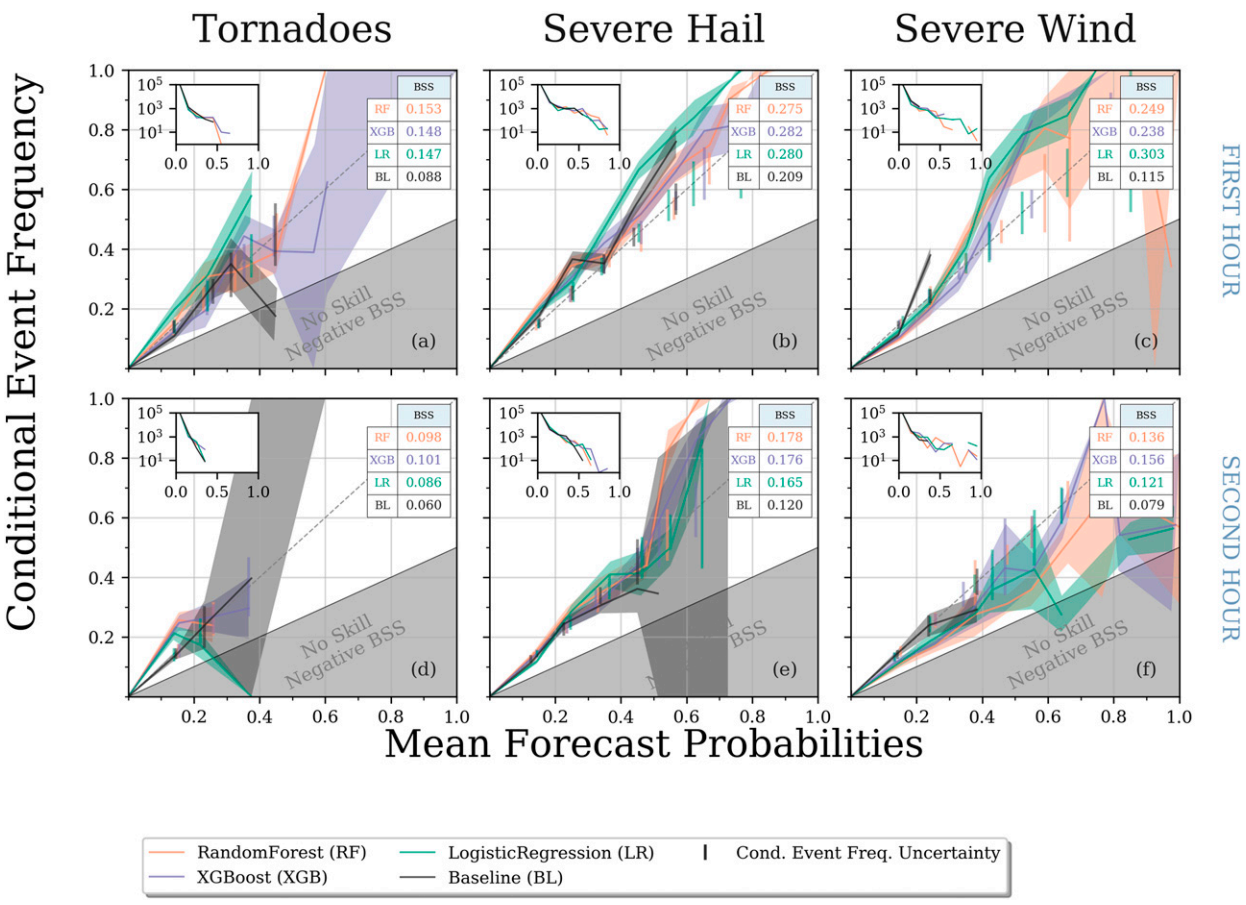

FIG. 10. As in Fig. 7, but for attribute diagrams. The bin increment of forecast probabilities is 10%. The inset figure is the forecast histogram for each model. The dashed line represents perfect reliability, and the gray region separates positive and negative Brier skill score (positive Brier skill score is above the gray area). The vertical lines along the diagonal are the error bars for the observed frequency for each model in each bin based on the method in Bröcker and Smith (2007). To limit figure crowding, error bars associated with an uncertainty of >50% for a given conditional observed frequency were omitted. The mean BSS for each model is provided in the table in the top-right corner of each panel.

higher probabilities) and the maximum forecast probability was modestly reduced compared to the FIRST HOUR dataset, which lowered the BSS (cf. Fig. 10b and Fig. 10e). The severe wind forecast probabilities for all three models became overconfident for probabilities ≥ 40% at longer lead times (cf. Fig. 10c and Fig. 10f).

For both lead time ranges and all three hazards, the baseline predictions were fairly reliable, but the ML models produced higher BSSs across the board. The severe hail- and severe wind-based ML models were capable of producing higher confidence forecasts than the baseline predictions, especially for severe wind-based models (Figs. 10c,f). The inability of the severe wind baseline to produce probabilities greater than 30%–40% highlights the ability of the ML models to incorporate complex forecast output to produce skillful severe wind forecasts even at higher probabilities. These results highlight that simple threshold methods are likely to overfit the training dataset and are suboptimal for capturing forecast uncertainty, which is consistent with the finding in Sobash et al. (2020). Surrogates methods also fail to leverage all available information from CAM ensemble forecast output, which will limit their potential accuracy.

6. Conclusions

The primary goal of Warn-on-Forecast is to provide human forecasters with short-term, storm-scale probabilistic severe weather guidance. Current CAM guidance can provide useful severe weather surrogates (e.g., updraft helicity), but it must be calibrated for individual severe weather hazards. An emerging approach to solving this problem are ML models, which can easily incorporate many predictors, are well suited for complex, noisy datasets, and have been shown to produce calibrated, skillful probabilistic guidance for a variety of meteorological phenomena.

In this study, gradient-boosted classification trees, random forests, and logistic regression models were trained on WoFS forecasts from the 2017–19 HWT-SFEs to predict which 30-min forecast storm tracks in the WoFS domain will produce a tornado, severe hail, and/or severe wind report up to lead times

of 150 min. A novel ensemble storm-track identification method inspired by Flora et al. (2019) was used to extract ensemble statistics of intrastorm and environmental parameters. We compared the ML predictions for tornadoes, severe hail, and severe wind against the probability of midlevel UH, WRF-HAILCAST maximum hail diameter, and 80-m AGL wind speed exceeding a threshold, respectively, with each threshold tuned to optimize performance. The primary conclusions are the following:

- The ML models produced substantially higher maximum Normalized Critical Success Indices [NCSIs; defined in Eq. (11)] and normalized area under the performance diagram than their respective baselines, especially at longer lead times. This latter result is especially encouraging since observation-based severe weather prediction methods rapidly degrade beyond nowcasting lead times.
- The ML models produced higher BSSs than their respective baselines. The most noticeable differences were for severe wind where the ML models produced BSSs nearly 2 times those of the predictions based on 80-m AGL wind speed.
- The ML models discriminated well (AUCs $>$ 0.9) for all three severe weather hazards up to a lead time of 150 min.
- For a given severe weather hazard, the contingency table metrics for the three ML algorithms were fairly similar. The severe hail predictions had the highest NCSI while tornado predictions had the lowest NCSI, especially at longer lead times.
- Depending on the hazard, the ML probabilities were fairly reliable up to 40%–60%. The severe wind and hail models produced higher probabilities than tornado-based models, but with an underconfidence bias. At longer lead times, severe hail forecast probabilities were reliable up to 50%–60% (depending on the model) while severe wind forecast probabilities became overconfident.

While these results are promising, there are some limitations to this study that should be considered. First, since we are operating in an event-based framework, we are not correcting for instances when the WoFS fails to accurately analyze ongoing convection or exhibits biases in storm location. In future studies, we plan to adopt a hybrid gridpoint-based/event-based framework that, in those circumstances, produces a complementary forecast that is largely based on WoFS environmental predictors. Second, the labeling of ensemble storm tracks was based on whether they contain a local storm report. We showed that because of small spatial errors in forecast storm tracks, reports may fall just outside the boundary of an ensemble storm track. Given these near misses, and the spurious false alarms arising from missing storm reports, the verification results likely underestimate the ML model skill. In an earlier version of this work, we attempted to use a buffer distance to account for storm motion biases in the WoFS forecasts, but this reduced the ML model skill. The skill reduction may have resulted from storm reports being matched to the wrong ensemble storm track in cases of multiple proximate storms. To properly label the forecast storms would require matching storm reports to observed storms and then matching those observed storms to the forecast storms. Such a sophisticated method was beyond the scope of this paper, but should be explored in future work. A third limitation of this study is that we did not evaluate the ML models for different geographic regions (e.g., Gagne et al. 2014; Herman and Schumacher 2018b; Sobash et al. 2020), diurnal times, or initialization times. The data in this study were largely sampled from the Great Plains (Fig. 1) so it will be important to assess the ML model performance in other regions. In future work, we plan to expand upon the verification of the ML predictions to highlight any potential failure modes.

There are additional potential extensions of this work. First, although the ML predictions outperformed competitive baselines, we did not compare with operational methods for predicting severe weather hazards (e.g., ProbSevere; Cintineo et al. 2014, 2018). To further assess the potential operational value of our prediction algorithms, and to increase forecaster trust in the algorithms, it will be necessary to evaluate the ML models against existing methods. Second, the labels used in this study are based on error-prone local storm reports. It will be crucial as a community to address these deficiencies in severe weather reporting. An alternative to storm reports would be to use radar-observed azimuthal shear (Smith and Elmore 2004; Miller et al. 2013; Smith et al. 2016; Mahalik et al. 2019) as a proxy for severe weather, but this approach has its own limitations. Third, the different ML algorithms were similarly skillful, but tended to over and underpredict in different situations. The best forecast may therefore be a weighted average of the different ML predictions, just as ensembles outperform deterministic forecasts in numerical weather prediction. Ensemble approaches can also provide estimates of forecast uncertainty, which can improve the trustworthiness of ML methods. Future work should therefore explore the use of ML model ensembles for severe weather prediction. Last, we only adopted a binary classification approach to predicting severe weather hazards (e.g., will a forecast storm produce a tornado?), but in future work, it is worth exploring multiclass approaches (e.g., will a forecast storm produce hail or a tornado or both?).

In addition to the more traditional ML algorithms used in this study, we also plan to apply CNNs (LeCun et al. 1990) to WoFS forecasts to predict severe weather. The primary advantage of CNNs is that they can learn from spatial data and do not require manual predictor engineering. CNNs have also showed success for a variety of meteorological applications (e.g., Gagne et al. 2019; Lagerquist et al. 2019; Wimmers et al. 2019; Lagerquist et al. 2020) and CNN interpretation techniques create metrics in the same space as the input spatial grids, making them easier to digest (McGovern et al. 2019b). Given that CNN can encode spatial information, CNN techniques may also prove useful in the aforementioned hybrid gridpoint-based/event-based framework, especially in the situations where the WoFS fails to predict an observed storm.

A thorough verification of a complex, end-to-end automated ML system is nearly impossible as one cannot possibly account for a complete list of failure modes (Doshi-Velez and Kim 2017). Therefore, automated guidance will require human forecaster input (known as the human in the loop paradigm). Recently, it has been shown that the combination of automated

guidance with human forecaster input has outperformed solely automated guidance for severe weather forecasting (Karstens et al. 2018). Thus, to build human forecasters' trust in ML predictions and maximize the use of automated guidance requires explaining the "why" of an ML model's prediction in understandable terms and creating real-time visualizations of these methods (Hoffman et al. 2017; Karstens et al. 2018). In ongoing research, we are using several ML interpretation methods to examine whether the algorithms are learning physical relationships and developing real-time visuals that explain ML model predictions using methods such as Shapley Additive Explanations (SHAP; Lundberg and Lee 2017).

Acknowledgments. Funding was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. We thank Vanna Chmielewski for informally reviewing an early version of the paper and three anonymous reviewers for their comments, which substantially improved the paper. Valuable local computing assistance was provided by Gerry Creager, Jesse Butler, Jeff Horn, Karen Cooper, and Carrie Langston.

Data availability statement. The experimental WoFS ensemble forecast data used in this study are not currently available in a publicly accessible repository. However, the data and code used to generate the results herein are available from the authors upon request.

APPENDIX

Derivation of Maximum Critical Success Index of a No-Skill System

From Roebber (2009), the critical success index can be defined as a function of success ratio s and probability of detection p:

$$\mathrm{CSI} = \frac{1}{s^{-1} + p^{-1} - 1}. \tag{A1}$$

Substituting the minimum success ratio for a no-skill system into Eq. (A1), we get

$$\mathrm{CSI} = 1 \Big/ \left(\frac{1 - c + cp}{cp} + \frac{1}{p} - 1 \right). \tag{A2}$$

We then multiply the numerator and denominator by cp (c is climatological event frequency):

$$\mathrm{CSI} = \frac{cp}{1 - c + cp + c - cp}, \tag{A3}$$

and then cancel the terms in the denominator to get the CSI of a no-skill system:

$$\mathrm{CSI} = cp. \tag{A4}$$

From Eq. (A4), the maximum CSI of a no-skill system occurs for $p = 1$ and is equal to climatological event frequency c.

REFERENCES

Adams-Selin, R. D., and C. L. Ziegler, 2016: Forecasting hail using a one-dimensional hail growth model within WRF. *Mon. Wea. Rev.*, **144**, 4919–4939, https://doi.org/10.1175/MWR-D-16-0027.1.

——, A. J. Clark, C. J. Melick, S. R. Dembek, I. L. Jirak, and C. L. Ziegler, 2019: Evolution of WRF-HAILCAST during the 2014–16 NOAA/Hazardous Weather Testbed Spring Forecasting Experiments. *Wea. Forecasting*, **34**, 61–79, https://doi.org/10.1175/WAF-D-18-0024.1.

Allen, J. T., and M. K. Tippett, 2015: The characteristics of United States hail reports: 1955–2014. *Electron. J. Severe Storms Meteor.*, **10** (3), https://ejssm.org/ojs/index.php/ejssm/article/viewArticle/149.

Baldwin, M. E., and J. S. Kain, 2006: Sensitivity of several performance measures to displacement error, bias, and event frequency. *Wea. Forecasting*, **21**, 636–648, https://doi.org/10.1175/WAF933.1.

Batista, G. E. A. P. A., R. C. Prati, and M. C. Monard, 2004: A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor.*, **6**, 20–29, https://doi.org/10.1145/1007730.1007735.

Bergstra, J., D. Yamins, and D. Cox, 2013: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *Proc. 30th Int. Conf. on Machine Learning*, Atlanta, GA, ICML, 115–123.

Boyd, K., V. S. Costa, J. Davis, and D. Page, 2012: Unachievable region in precision-recall space and its effect on empirical evaluation. https://arxiv.org/abs/1206.4667.

——, K. H. Eng, and C. D. Page, 2013: Area under the precision-recall curve: Point estimates and confidence intervals. *Machine Learning and Knowledge Discovery in Databases*, H. Blockeel et al., Eds., Springer, 451–466.

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

Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, https://doi.org/10.1175/WAF993.1.

Brooks, H. E., and J. Correia, 2018: Long-term performance metrics for National Weather Service tornado warnings. *Wea. Forecasting*, **33**, 1501–1511, https://doi.org/10.1175/WAF-D-18-0120.1.

——, C. A. Doswell, and M. P. Kay, 2003: Climatological estimates of local daily tornado probability for the United States. *Wea. Forecasting*, **18**, 626–640, https://doi.org/10.1175/1520-0434(2003)018<0626:CEOLDT>2.0.CO;2.

Bryan, G. H., J. C. Wyngaard, and J. M. Fritsch, 2003: Resolution requirements for the simulation of deep moist convection. *Mon. Wea. Rev.*, **131**, 2394–2416, https://doi.org/10.1175/1520-0493(2003)131<2394:RRFTSO>2.0.CO;2.

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.

Chen, T., and C. Guestrin, 2016: XGBoost: A scalable tree boosting system. *KDD'16: Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, Association for Computing Machinery, https://doi.org/10.1145/2939672.2939785.

Cintineo, J. L., T. M. Smith, V. Lakshmanan, H. E. Brooks, and K. L. Ortega, 2012: An objective high-resolution hail climatology of the contiguous United States. *Wea.*

Forecasting, **27**, 1235–1248, https://doi.org/10.1175/WAF-D-11-00151.1.

——, M. J. Pavolonis, J. M. Sieglaff, and D. T. Lindsey, 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.

——, and Coauthors, 2018: The NOAA/CIMSS ProbSevere Model' incorporation of total lightning and validation. *Wea. Forecasting*, **33**, 331–345, https://doi.org/10.1175/WAF-D-17-0099.1.

——, M. J. Pavolonis, J. M. Sieglaff, L. Cronce, and J. Brunner, 2020: NOAA ProbSevere v2.0—ProbHail, ProbWind, and ProbTor. *Wea. Forecasting*, **35**, 1523–1543, https://doi.org/10.1175/WAF-D-19-0242.1.

Clark, A. J., E. Loken, P. Skinner, and K. Knopfmeier, 2020: Machine-learning-derived severe weather probabilities from a warn-on-forecast system. *10th Conf. on Transition of Research to Operations*, Boston, MA, Amer. Meteor. Soc., 1483, https://ams.confex.com/ams/2020Annual/meetingapp.cgi/Paper/366971.

Coffer, B. E., M. D. Parker, J. M. L. Dahl, L. J. Wicker, and A. J. Clark, 2017: Volatility of tornadogenesis: An ensemble of simulated nontornadic and tornadic supercells in VORTEX2 environments. *Mon. Wea. Rev.*, **145**, 4605–4625, https://doi.org/10.1175/MWR-D-17-0152.1.

——, ——, 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.

Davis, J., and M. Goadrich, 2006: The relationship between precision-recall and ROC curves. *Proc. 23rd Int. Conf. on Machine Learning*, New York, NY, Association for Computing Machinery, 233–240, https://doi.org/10.1145/1143844.1143874.

Developmental Testbed Center, 2017a: Ensemble Kalman Filter (EnKF) user's guide for version 1.2. Developmental Testbed Center Doc., 86 pp., https://dtcenter.ucar.edu/EnKF/users/docs/enkf_users_guide/EnKF_UserGuide_v1.2.pdf.

——, 2017b: Gridpoint statistical interpolation user's guide version 3.6. Developmental Testbed Center Doc., 158 pp., https://dtcenter.org/com-GSI/users/docs/.

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.

Doshi-Velez, F., and B. Kim, 2017: Towards a rigorous science of interpretable machine learning. https://arxiv.org/abs/1702.08608.

Doswell, C. A., 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.

Dowell, D., and Coauthors, 2016: Development of a High-Resolution Rapid Refresh Ensemble (HRRRE) for severe weather forecasting. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 8B.2, https://ams.confex.com/ams/28SLS/webprogram/Paper301555.html.

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.

Erickson, M. J., J. J. Charney, and B. A. Colle, 2016: Development of a fire weather index using meteorological observations within the northeast United States. *J. Appl. Meteor. Climatol.*, **55**, 389–402, https://doi.org/10.1175/JAMC-D-15-0046.1.

Flora, M. L., 2020: Storm-scale ensemble-based severe weather guidance: Development of an object-based verification framework and applications of machine learning. Ph.D. thesis, University of Oklahoma, 193 pp.

——, C. K. Potvin, and L. J. Wicker, 2018: Practical predictability of supercells: Exploring ensemble forecast sensitivity to initial condition spread. *Mon. Wea. Rev.*, **146**, 2361–2379, https://doi.org/10.1175/MWR-D-17-0374.1.

——, P. S. Skinner, C. K. Potvin, A. E. Reinhart, T. A. Jones, N. Yussouf, and K. H. Knopfmeier, 2019: Object-based verification of short-term, storm-scale probabilistic mesocyclone guidance from an experimental warn-on-forecast system. *Wea. Forecasting*, **34**, 1721–1739, https://doi.org/10.1175/WAF-D-19-0094.1.

Flournoy, M. D., M. C. Coniglio, E. N. Rasmussen, J. C. Furtado, and B. E. Coffer, 2020: Modes of storm-scale variability and tornado potential in VORTEX2 near- and far-field tornadic environments. *Mon. Wea. Rev.*, **148**, 4185–4207, https://doi.org/10.1175/MWR-D-20-0147.1.

Friedman, J., 2002: Stochastic gradient boosting. *Comput. Stat. Data Anal.*, **38**, 367–378, https://doi.org/10.1016/S0167-9473(01)00065-2.

Gagne, D. J., 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.

——, ——, N. Snook, R. Sobash, J. Labriola, J. K. Williams, S. E. Haupt, and M. Xue, 2016: Hagelslag: Scalable object-based severe weather analysis and forecasting. *Sixth Symp. on Advances in Modeling and Analysis Using Python*, New Orleans, LA, Amer. Meteor. Soc., 447, https://ams.confex.com/ams/96Annual/webprogram/Paper280723.html.

——, ——, S. E. Haupt, R. A. Sobash, J. K. 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.

——, 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., 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.

Hastie, T., R. Tibshirani, and J. Friedman, 2001: *The Elements of Statistical Learning*. Springer, 533 pp.

Hepper, R. M., I. L. Jirak, and J. M. Milne, 2016: Assessing the skill of convection-allowing ensemble forecasts of severe MCS winds from the SSEO. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 16B.2, https://ams.confex.com/ams/28SLS/webprogram/Paper300134.html.

Herman, G. R., and R. S. Schumacher, 2018a: "Dendrology" in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Mon. Wea. Rev.*, **146**, 1785–1812, https://doi.org/10.1175/MWR-D-17-0307.1.

——, and ——, 2018b: 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.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Mon. Wea. Rev.*, **148**, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28**, 525–534, https://doi.org/10.1175/WAF-D-12-00113.1.

Hoffman, R. R., D. S. LaDue, H. M. Mogil, P. J. Roebber, and J. G. Trafton, 2017: *Minding the Weather: How Expert Forecasters Think*. The MIT Press, 488 pp.

Hsu, W., 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.

Jergensen, G. E., A. McGovern, R. Lagerquist, and T. Smith, 2020: Classifying convective storms using machine learning. *Wea. Forecasting*, **35**, 537–559, https://doi.org/10.1175/WAF-D-19-0170.1.

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.

Jones, T. A., K. Knopfmeier, D. Wheatley, G. Creager, P. Minnis, and R. Palikonda, 2016: Storm-scale data assimilation and ensemble forecasting with the NSSL experimental Warn-on-Forecast system Part II: Combined radar and satellite data experiments. *Wea. Forecasting*, **31**, 297–327, https://doi.org/10.1175/WAF-D-15-0107.1.

——, P. Skinner, N. Yussouf, K. Knopfmeier, A. Reinhart, and D. Dowell, 2019: Forecasting high-impact weather in landfalling tropical cyclones using a Warn-on-Forecast system. *Bull. Amer. Meteor. Soc.*, **100**, 1405–1417, https://doi.org/10.1175/BAMS-D-18-0203.1.

——, and Coauthors, 2020: Assimilation of *GOES-16* radiances and retrievals into the warn-on-forecast system. *Mon. Wea. Rev.*, **148**, 1829–1859, https://doi.org/10.1175/MWR-D-19-0379.1.

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.

Labriola, J., N. Snook, Y. Jung, B. Putnam, and M. Xue, 2017: Ensemble hail prediction for the storms of 10 May 2010 in south-central Oklahoma using single- and double-moment microphysical schemes. *Mon. Wea. Rev.*, **145**, 4911–4936, https://doi.org/10.1175/MWR-D-17-0039.1.

——, ——, ——, and M. Xue, 2019: Explicit ensemble prediction of hail in 19 May 2013 Oklahoma City thunderstorms and analysis of hail growth processes with several multimoment microphysics schemes. *Mon. Wea. Rev.*, **147**, 1193–1213, https://doi.org/10.1175/MWR-D-18-0266.1.

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.

——, ——, and D. J. Gagne II, 2019: Deep learning for spatially explicit prediction of synoptic-scale fronts. *Wea. Forecasting*, **34**, 1137–1160, https://doi.org/10.1175/WAF-D-18-0183.1.

——, ——, C. R. Homeyer, D. J. Gagne, and T. Smith, 2020: Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Mon. Wea. Rev.*, **148**, 2837–2861, https://doi.org/10.1175/MWR-D-19-0372.1.

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.

LeCun, Y., B. E. Boser, J. S. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, 1990: Handwritten digit recognition with a back-propagation network. *Advances in Neural Information Processing Systems*, D. Touetzky, Ed., MIT Press, 396–404.

Loken, E. D., A. J. Clark, A. McGovern, M. Flora, and K. Knopfmeier, 2019: 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.

——, ——, and C. D. Karstens, 2020: Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. *Wea. Forecasting*, **35**, 1605–1631, https://doi.org/10.1175/WAF-D-19-0258.1.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, https://doi.org/10.3402/tellusa.v21i3.10086.

Lundberg, S. M., and S.-I. Lee, 2017: A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30*, I. Guyon et al., Eds., Curran Associates, 4765–4774.

Mahalik, M. C., B. R. Smith, K. L. Elmore, D. M. Kingfield, K. L. Ortega, and T. M. Smith, 2019: Estimates of gradients in radar moments using a linear least squares derivative technique. *Wea. Forecasting*, **34**, 415–434, https://doi.org/10.1175/WAF-D-18-0095.1.

Manning, C., and H. Schütze, 1999: *Foundations of Statistical Natural Language Processing*. MIT Press, 680 pp.

McGovern, A., K. L. Elmore, J. I. David, S. 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.

——, C. D. Karstens, T. Smith, and R. Lagerquist, 2019a: Quasi-operational testing of real-time storm-longevity prediction via machine learning. *Wea. Forecasting*, **34**, 1437–1451, https://doi.org/10.1175/WAF-D-18-0141.1.

——, R. Lagerquist, D. J. Gagne II, E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019b: 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.

Mecikalski, J. R., J. K. Williams, C. P. Jewett, D. Ahijevych, A. LeRoy, and J. R. Walker, 2015: Probabilistic 0–1-h convective initiation nowcasts that combine geostationary satellite observations and numerical weather prediction model data. *J. Appl. Meteor. Climatol.*, **54**, 1039–1059, https://doi.org/10.1175/JAMC-D-14-0129.1.

Metz, C. E., 1978: Basic principles of ROC analysis. *Semin. Nucl. Med.*, **8**, 283–298, https://doi.org/10.1016/S0001-2998(78)80014-2.

Miller, M. L., V. Lakshmanan, and T. M. Smith, 2013: An automated method for depicting mesocyclone paths and intensities. *Wea. Forecasting*, **28**, 570–585, https://doi.org/10.1175/WAF-D-12-00065.1.

Niculescu-Mizil, A., and R. Caruana, 2005: Predicting good probabilities with supervised learning. *Proc. 22nd Int. Conf. on Machine Learning*, New York, NY, ACM, 625–632, https://doi.org/10.1145/1102351.1102430.

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

Platt, J. C., 1999: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, A. J. Smola et al., Eds., MIT Press, 61–74.

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

——, C. Broyles, P. S. Skinner, H. E. Brooks, and E. Rasmussen, 2019: A Bayesian hierarchical modeling framework for correcting reporting bias in the U.S. tornado database. *Wea. Forecasting*, **34**, 15–30, https://doi.org/10.1175/WAF-D-18-0137.1.

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

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.

Skamarock, W., 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.

——, and Coauthors, 2018: Object-based verification of a prototype Warn-on-Forecast system. *Wea. Forecasting*, **33**, 1225–1250, https://doi.org/10.1175/WAF-D-18-0020.1.

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.

——, 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.

Smith, T. M., and K. L. Elmore, 2004: The use of radial velocity derivatives to diagnose rotation and divergence. *11th Conf. on Aviation, Range, and Aerospace*, Hyannis, MA, Amer. Meteor. Soc., P5.6, https://ams.confex.com/ams/pdfpapers/81827.pdf.

——, and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) severe weather and aviation products: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 1617–1630, https://doi.org/10.1175/BAMS-D-14-00173.1.

Snook, N., M. Xue, and Y. Jung, 2012: Ensemble probabilistic forecasts of a tornadic mesoscale convective system from ensemble Kalman filter analyses using WSR-88D and CASA radar data. *Mon. Wea. Rev.*, **140**, 2126–2146, https://doi.org/10.1175/MWR-D-11-00117.1.

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

——, G. S. Romine, and C. S. Schwartz, 2020: A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived from a convection-allowing model. *Wea. Forecasting*, **35**, 1981–2000, https://doi.org/10.1175/WAF-D-20-0036.1.

Steinkruger, D., P. Markowski, and G. Young, 2020: An artificially intelligent system for the automated issuance of tornado warnings in simulated convective storms. *Wea. Forecasting*, **35**, 1939–1965, https://doi.org/10.1175/WAF-D-19-0249.1.

Stensrud, D. J., and Coauthors, 2009: Convective-scale Warn-on-Forecast system. *Bull. Amer. Meteor. Soc.*, **90**, 1487–1500, https://doi.org/10.1175/2009BAMS2795.1.

——, and Coauthors, 2013: Progress and challenges with warn-on-forecast. *Atmos. Res.*, **123**, 2–16, https://doi.org/10.1016/j.atmosres.2012.04.004.

Storm Prediction Center, 2020: NOAA. Accessed 16 April 2021, https://www.spc.noaa.gov/new/SVRclimo/climo.php?parm=anySvr.

Sun, Y., A. Wong, and M. S. Kamel, 2009: Classification of imbalanced data: A review. *Int. J. Pattern Recognit. Artif. Intell.*, **23**, 687–719, https://doi.org/10.1142/S0218001409007326.

Trapp, R. J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and R. Wolf, 2006: Buyer beware: Some words of caution on the use of severe wind reports in postevent assessment and research. *Wea. Forecasting*, **21**, 408–415, https://doi.org/10.1175/WAF925.1.

Verbout, S. M., H. E. Brooks, L. M. Leslie, and D. M. Schultz, 2006: Evolution of the U.S. tornado database: 1954–2003. *Wea. Forecasting*, **21**, 86–93, https://doi.org/10.1175/WAF910.1.

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW model. *Wea. Forecasting*, **23**, 407–437, https://doi.org/10.1175/2007WAF2007005.1.

Wheatley, D. M., K. H. Knopfmeier, T. A. Jones, and G. J. Creager, 2015: Storm-scale data assimilation and ensemble forecasting with the NSSL experimental Warn-on-Forecast system. Part I: Radar data experiments. *Wea. Forecasting*, **30**, 1795–1817, https://doi.org/10.1175/WAF-D-15-0043.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 288 pp.

Wilson, K. A., and Coauthors, 2019: Exploring applications of storm-scale probabilistic warn-on-forecast guidance in weather forecasting. *Virtual, Augmented and Mixed Reality, Applications and Case Studies*, J. Chen and G. Fragomeni, Eds., Springer, 557–572, https://doi.org/10.1007/978-3-030-21565-1_39.

Wimmers, A., C. Velden, and J. H. Cossuth, 2019: Using deep learning to estimate tropical cyclone intensity from satellite passive microwave imagery. *Mon. Wea. Rev.*, **147**, 2261–2282, https://doi.org/10.1175/MWR-D-18-0391.1.

Yussouf, N., J. Gao, D. J. Stensrud, and G. Ge, 2013a: The impact of mesoscale environmental uncertainty on the prediction of a tornadic supercell storm using ensemble data assimilation approach. *Adv. Meteor.*, **2013**, 1–15, https://doi.org/10.1155/2013/731647.

——, E. R. Mansell, L. J. Wicker, D. M. Wheatley, and D. J. Stensrud, 2013b: The ensemble Kalman filter analyses and forecasts of the 8 May 2003 Oklahoma City tornadic supercell storm using single- and double-moment microphysics schemes. *Mon. Wea. Rev.*, **141**, 3388–3412, https://doi.org/10.1175/MWR-D-12-00237.1.

——, D. C. Dowell, L. J. Wicker, K. H. Knopfmeier, and D. M. Wheatley, 2015: Storm-scale data assimilation and ensemble forecasts for the 27 April 2011 severe weather outbreak in Alabama. *Mon. Wea. Rev.*, **143**, 3044–3066, https://doi.org/10.1175/MWR-D-14-00268.1.

——, J. S. Kain, and A. J. Clark, 2016: Short-term probabilistic forecasts of the 31 May 2013 Oklahoma tornado and flash flood event using a continuous-update-cycle storm-scale ensemble system. *Wea. Forecasting*, **31**, 957–983, https://doi.org/10.1175/WAF-D-15-0160.1.

——, K. A. Wilson, S. M. Martinaitis, H. Vergara, P. L. Heinselman, and J. J. Gourley, 2020: The coupling of NSSL warn-on-forecast and FLASH systems for probabilistic flash flood prediction. *J. Hydrometeor.*, **21**, 123–141, https://doi.org/10.1175/JHM-D-19-0131.1.