# The Prediction of Potential Tornado Damage Intensity Using Machine Learning

MICHAEL F. SESSA[a] AND ROBERT J. TRAPP[a]

[a] *Department of Atmospheric Sciences, University of Illinois Urbana–Champaign, Urbana, Illinois*

(Manuscript received 22 December 2023, in final form 23 March 2025, accepted 7 April 2025)

ABSTRACT: This study uses nine classification machine learning algorithms to examine their skill in making short-fused, storm-based predictions of significant or nonsignificant tornado damage intensity, conditioned upon tornadogenesis, using pretornadic mesocyclone characteristics and the near-storm environment. Radar predictors are from approximately 30 min before tornadogenesis, while environmental predictors are from the model-analysis hour nearest but before the time of tornadogenesis. The most-skilled classifiers are logistic regression, random forests, and gradient boosting as measured by each model's cross-validated accuracy (≈89%), precision (≈93%), and recall (≈73%) and other binary classification metrics. Learning curves indicate adequate training of models, and calibration curves reveal the reliability of predicted probabilities, with random forests being the most reliable. Also, permutation tests demonstrate the statistical significance of the cross-validated model accuracy. Out of the four radar predictors included in this study, radar-derived pretornadic mesocyclone width and differential velocity are the most important over convective mode and distance from the radar, followed by environmental vertical wind shear and composite parameters. Specifically, wider and stronger pretornadic mesocyclones in environments characterized by larger values of vertical wind shear and composite parameters increase the likelihood of significant tornadoes. The model results could build forecaster confidence in the anticipation of tornado damage intensity and aid forecasters in making informed impact-based warning tag decisions. This could better protect life and property by providing a summary of data relevant to potential tornado damage rating before tornado formation. Important future work includes the addition of other radar-based predictors and the development of a more diverse and realistic sample of tornadic events.

SIGNIFICANCE STATEMENT: Given that the majority of tornado damage and fatalities are due to strong-to-violent tornadoes (EF2+), the purpose of this study is to explore tornado damage intensity prediction for ongoing thunderstorms using binary classification machine learning applied to pretornadic radar data and the near-storm environment. Our results show a skilled prediction of potential tornado damage intensity, conditioned upon tornadogenesis, across multiple models demonstrated through the correct prediction of 73% of EF2+ tornadoes included in this study from the top-performing model. Applying this study in an operational setting may aid in better short-fused anticipation of significant tornado events.

KEYWORDS: Mesocyclones; Severe storms; Tornadoes; Storm environments; Classification

## 1. Introduction

Recently, there has been a rapidly increasing interest in artificial intelligence and machine learning (ML) in the atmospheric sciences. Several studies have shown how applying ML techniques, along with a physical understanding of the environment and the phenomena of concern, can substantially improve the predictions of many types of high-impact weather (e.g., McGovern et al. 2014, 2017; Karstens et al. 2018; Cintineo et al. 2020b). ML has provided several advantages, such as incorporating diverse data sources (e.g., satellite, radar, numerical weather prediction, in situ observations) and handling large datasets with many predictor variables in an efficient manner. Additional benefits include integrating physical understanding into models, removing subjective analysis, and gaining new knowledge by uncovering signals in complex datasets that would be difficult and time consuming to discover by manual analysis (e.g., Gagne et al. 2019; Schlef et al. 2019; Miller et al. 2020; McGuire and Moore 2022). The processing time of ML model predictions also tends to be faster than alternative methods.

Incorporating multiple data sources in an ML application can help address the predictive challenges of severe convective storms. Individually, near-storm environmental information, Doppler radar data, and satellite data have limitations in predicting severe convective storms and their hazards. For example, the WSR-88D network has limited coverage and ability to consistently resolve features of concern, and the spatial and temporal evolution of the volatile near-storm environment can be difficult to represent with available observations and numerical model output. However, the unification of their contributions creates a more useful tool for anticipating severe convective storm hazards. In particular, the benefit of combining these data sources in an ML environment to assess severe convective storm potential has been demonstrated (e.g., Cintineo et al. 2014; Mecikalski et al. 2015; Czernecki et al. 2019; Cintineo et al. 2020b; Leinonen et al. 2022).

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/AIES-D-23-0113.s1.

*Corresponding author*: Michael F. Sessa, msessa2@illinois.edu

DOI: 10.1175/AIES-D-23-0113.1 e230113

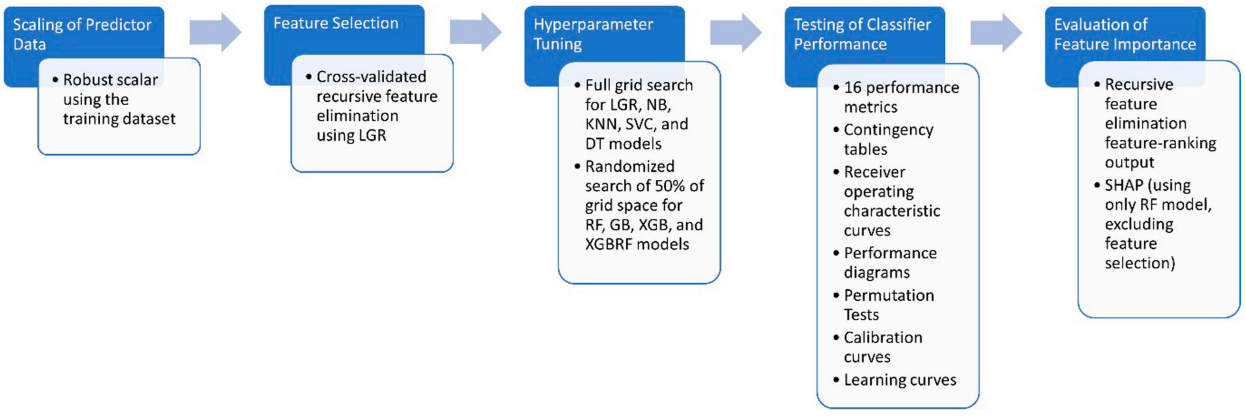


FIG. 1. Schematic illustrating the modeling procedure used.

Additionally, ML applications such as neural networks, support vector machines, decision trees, and logistic regression have been used to improve severe weather forecasting (e.g., Hill et al. 2020; Loken et al. 2020), which includes the prediction of tornadoes (e.g., Marzban and Stumpf 1996; Adrianto et al. 2009; Lagerquist et al. 2020; Sobash et al. 2019; Steinkruger et al. 2020), downdraft winds (e.g., Marzban and Stumpf 1998; Lagerquist et al. 2017), hail (e.g., Manzato 2013; Czernecki et al. 2019; Gagne et al. 2019; Burke et al. 2020), and lightning (Mostajabi et al. 2019; Zhou et al. 2020). ML has also been used to recognize and track thunderstorms and their features (e.g., Ashley et al. 2019; Cintineo et al. 2020a; Jergensen et al. 2019), forecast convective initiation (Mecikalski et al. 2015; Ahijevych et al. 2016), and identify environments supporting supercells (Shield and Houston 2022) using various data sources.

For tornadoes specifically, ML applications have mainly focused on the prediction of tornadogenesis (e.g., Gagne et al. 2012; Lagerquist et al. 2020; Steinkruger et al. 2020; Coffer et al. 2021). Other ML applications (Sandmæl et al. 2023) and statistical studies (e.g., Thompson et al. 2017; Smith et al. 2020a,b) have developed methods to detect or characterize ongoing tornadoes using characteristics of the tornadic circulation itself (e.g., rotational velocity, tornadic debris signature) as well as near-storm environmental information (e.g., composite parameters) and nonmeteorological factors such as population density.

Although strong-to-violent tornadoes cause most tornado-related fatalities and damages (Ashley 2007; Simmons and Sutter 2011; Anderson-Frey and Brooks 2019), few ML applications focus on the prediction of tornado damage intensity before a tornado forms. Using components of the significant tornado parameter (STP; Thompson et al. 2002) from Rapid Update Cycle analysis soundings, Togstad et al. (2011, hereafter $T_{2011}$) used a multivariate logistic regression equation to show its greater skill in discriminating between nontornadic and significantly tornadic [enhanced Fujita scale 2+ (EF2+); Wind Science and Engineering Center 2006] supercell environments over the use of composite parameters (e.g., STP). Using environmental parameters from Rapid Update Cycle analysis soundings, Nowotarski and Jensen (2013) demonstrated skill using a self-organizing map algorithm to classify soundings as nonsupercell, weak-tornadic supercell, or significant tornadic supercell, finding that kinematic variables (especially ground-relative winds) performed better than thermodynamic parameters. More recently, Gensini et al. (2021, hereafter $G_{2021}$) used model reanalysis-derived vertical profiles at the grid point nearest to the tornado report to understand environmental controls on tornado damage intensity.

The additional use of Doppler radar data of ongoing storms is lacking in these ML applications of tornado damage intensity prediction. Sessa and Trapp (2020, 2023, hereinafter ST20 and ST23, respectively), who showed robust linear correlations between pretornadic mesocyclone width and the EF rating of the associated tornado, demonstrate the potential advantages of such data. ST23 also discusses other statistical studies that use pretornadic radar data for the prediction of tornado damage intensity (e.g., Gibbs and Bowers 2019; French and Kingfield 2021). ST23 focused explicitly on pretornadic mesocyclone characteristics and further investigated Trapp et al. (2017)'s hypothesis, initially explored through idealized supercell simulations, that intense tornadoes should form more readily out of wide, rotating updrafts rather than narrow ones. The pretornadic focus eliminated the effects of the tornado on the mesocyclone characteristics and thus allowed for the exploration of any relationships between the pretornadic mesocyclone and tornado damage intensity. The 13-km Rapid Refresh (RAP) version 3 (Benjamin et al. 2016) analysis data provided the pretornadic environment of these cases to explore relationships with tornado damage intensity (see ST23). The strong relationships linking the pretornadic radar characteristics of mesocyclone width and intensity $\Delta V$ and the near-storm environment with tornado damage intensity, shown by ST20 and ST23, motivate the use of ML to explore these relationships further.

Since 2016, National Weather Service (NWS) local forecast offices have used impact-based tornado warnings (IBWs; Casteel 2016) to communicate the potential impacts to life and property through "considerable" or "catastrophic" damage threat tags for imminent or ongoing EF2+ damage or the confirmed presence of a violent tornado, respectively (Warning Decision Training Division 2021). The decision to use these

TABLE 1. List of environmental variables sampled or calculated (from ST23).

| Parameter | Units | Abbreviation | Method |
|---|---|---|---|
| 100-mb mixed-layer CAPE | J $kg^{-1}$ | MLCAPE | MetPy (May et al. 2022) |
| 100-mb mixed-layer CIN | J $kg^{-1}$ | MLCIN | MetPy |
| Surface-based CAPE | J $kg^{-1}$ | SBCAPE | MetPy |
| Surface-based CIN | J $kg^{-1}$ | SBCIN | MetPy |
| Most unstable CAPE | J $kg^{-1}$ | MUCAPE | MetPy |
| Most unstable CIN | J $kg^{-1}$ | MUCIN | MetPy |
| 0–3-km CAPE | J $kg^{-1}$ | CAPE03 | MetPy |
| Lifted index | ° | LI | MetPy |
| 0–8-km bulk shear | m $s^{-1}$ | S08 | MetPy |
| 0–6-km bulk shear | m $s^{-1}$ | S06 | MetPy |
| 0–3-km bulk shear | m $s^{-1}$ | S03 | MetPy |
| 0–1-km bulk shear | m $s^{-1}$ | S01 | MetPy |
| 0–500-m bulk shear | m $s^{-1}$ | S500 | MetPy |
| Effective bulk shear | m $s^{-1}$ | EBS | Manually calculated from Thompson et al. (2007, hereafter T07) |
| Bunkers right storm motion | m $s^{-1}$ | BR | MetPy |
| 0–6-km mean flow | m $s^{-1}$ | 06Mean | MetPy |
| 0–1-km storm-relative helicity | $m^2$ $s^{-2}$ | 01SRH | MetPy |
| 0–3-km storm-relative helicity | $m^2$ $s^{-2}$ | 03SRH | MetPy |
| 0–500-m storm-relative helicity | $m^2$ $s^{-2}$ | 0500SRH | RAP analysis variable |
| Effective storm-relative helicity | $m^2$ $s^{-2}$ | ESRH | T07 |
| Effective layer base height | m | ELB | T07 |
| Effective layer top height | m | ELT | T07 |
| Effective layer depth | m | ELD | T07 |
| 0–2-km storm-relative wind | m $s^{-1}$ | 02SRW | MetPy |
| 4–6-km storm-relative wind | m $s^{-1}$ | 46SRW | MetPy |
| 9–11-km storm-relative wind | m $s^{-1}$ | 911SRW | MetPy |
| 0–3-km lapse rate | °C $km^{-1}$ | 03LR | MetPy |
| 3–6-km lapse rate | °C $km^{-1}$ | 36LR | MetPy |
| Lifting condensation level height | m | LCLh | MetPy |
| Level of free convection height | m | LFCh | MetPy |
| LCL-to-LFC layer depth | m | LCL–LFC | LFCh–LCLh |
| 0–3-km relative humidity | % | 03RH | MetPy |
| 3–6-km relative humidity | % | 36RH | MetPy |
| LCL-to-LFC relative humidity | % | LCL–LFC-RH | Mean RH between LCL and LFC |
| 0–1-km energy helicity index | None | 01EHI | Manually calculated from Thompson et al. (2003, hereafter T03) |
| 0–3-km energy helicity index | None | 03EHI | T03 |
| Fixed-layer supercell composite parameter | None | SCPf | T03 |
| Effective layer supercell composite parameter | None | SCPe | MetPy |
| Fixed-layer significant tornado parameter | None | STPf | MetPy |
| 0–1-km tornadic energy helicity index | None | torEHI | Manually calculated (https://www.spc.noaa.gov/exper/mesoanalysis/help/help_tehi.html) |
| Tornadic tilting and stretching parameter | None | TTS | Manually calculated (Bothwell et al. 2002; https://www.spc.noaa.gov/exper/mesoanalysis/help/help_tts.html) |
| Critical angle | ° | CA | MetPy (Esterheld and Giuliano 2021) |

tags is determined once a tornado is ongoing with a combination of radar signatures, environmental parameters, and population density (e.g., Smith et al. 2020b) or storm-spotter reports to confirm the presence of a tornado. After further operational development, the information provided by the ML applications in this study could help an operational forecaster predict real-time potential tornado damage intensity during the pretornadic stages of ongoing thunderstorms, which would allow for the earlier issuance of impact-based warning tags about tornado damage intensity to better protect life and property.

The two main research questions addressed in this study are as follows: Can pretornadic mesocyclone characteristics (up to ~2 km), such as width and intensity, be used to predict real-time potential tornado damage intensity, conditional on tornadogenesis, using ML? How does the addition of the pretornadic near-storm environment change the predictive skill? This study shows how ML could be applied to help process large amounts of near-storm environmental and pretornadic radar data together to reveal relationships between these variables and tornado damage intensity.

TABLE 2. List of predictors from the pretornadic radar dataset used in the ML models (from ST23).

| Predictor | Description |
|---|---|
| Mean pretornadic mesocyclone width (km) | The mean mesocyclone width over the lowest three elevation angles and all volume scans analyzed during the pretornadic period |
| Peak pretornadic mesocyclone intensity $\Delta V$ (m s$^{-1}$) | The maximum $\Delta V$ over the lowest three elevation angles and all volume scans analyzed during the pretornadic period |
| Distance from radar (km) | The distance from the radar to the center of the radial velocity couplet using the 0.5° elevation angle scan closest in time to tornadogenesis |
| Convective mode | Either discrete supercell, QLCS, or multicell; converted to a binary numeric format using one-hot encoding ("1" for the present mode and "0" for the remaining modes) |

An outline of this study is as follows. Section 2 describes the methods used in the preprocessing of data and the tuning, training, and evaluation of several classification ML models. Section 3 presents the results of the models, which show a skilled prediction of potential tornado damage intensity. This section also discusses model interpretation and weaknesses and a detailed feature importance analysis. Section 4 reviews future operational implementation and a case study using the final versions of three different models, while highlighting caveats and additional work needed to develop an operationally useful product. Section 5 provides a summary and conclusions, while outlining important future work. Finally, the supplement includes a discussion of the adequacy of model training, the significance of the results, and the reliability of modeled probabilities.

## 2. Methods

### a. Data

For this study and its companion (ST23), the original dataset of 102 tornado cases from ST20 was expanded to 300 (see Fig. 1 from ST23) using data primarily from 2019 and 2020. The cases were selected using full tornado track statistics from the Storm Prediction Center (SPC) "One Tornado (ONETOR)" dataset, applying the basic case-selection methodology of ST20. The expanded tornadic dataset is composed of 78 EF0, 116 EF1, 59 EF2, 33 EF3, 10 EF4, and 4 EF5 tornadoes, of which 130 are associated with discrete supercells (DSCs; 69 EF0 − 1 and 61 EF2+), 124 with quasilinear convective systems (QLCSs; 92 EF0 − 1 and 32 EF2+), and 46 with multicells (MULs; 33 EF0 − 1 and 13 EF2+). Only four violent tornadoes came from the expanded analysis in ST23 with the remaining violent tornadoes included coming from the original ST20 dataset, which required a consideration of a large range of years (2011–19). The supplemental material confirms the adequacy of training using a dataset of this size. The need to manually analyze WSR-88D level-II data for each case limits the dataset size. The analysis of archived radar data[1] was performed using the Gibson Ridge radar software version 3[2] (GR2Analyst 3.0), which allowed for the determination of the width and intensity $\Delta V$ of pretornadic mesocyclones using Doppler velocity data. Pretornadic mesocyclone width is defined as the linear distance between the inbound and outbound velocity peaks in the vortex couplet, and the pretornadic mesocyclone intensity is defined as the differential velocity $\Delta V$ computed using the inbound and outbound velocity peaks. These were calculated over the lowest three elevation angles using all volume scans available up to 30 min before tornadogenesis (see ST20 and ST23 for details). This would include up to four full-volume scans plus additional scans of the lowest three elevation angles, depending on the volume coverage pattern. The presence of a mesocyclone required a peak $\Delta V \geq 10$ m s$^{-1}$ over a horizontal distance of $<7$ km, over the depth of the three lowest radar elevation angles. To summarize other restrictions from ST23, all tornadoes were within $\leq$100 km of a WSR-88D, each tornado had to be the first produced by its respective parent storm, and no tornado within 1 h and 80 km of another tornado was included in the dataset if its EF rating was more than one EF category lower than the peak EF rating within this time and space range.

The 13-km RAP analysis from the NOAA Thematic Real-Time Environmental Distributed Data Services (THREDDS; https://www.ncei.noaa.gov/thredds/catalog/model/rucrap.html) server provided the data for the exploration of the tornadic environments of each case. A 10 × 10-gridcell box (approximately 130 km × 130 km), roughly centered on the location of tornadogenesis, bounded the environmental sampling region. The last analysis time before tornadogenesis supplied the data. As described in ST23, a mask removed unrepresentative grid points with a 1 km AGL simulated reflectivity $>$ 20 dBZ, most unstable convective available potential energy (MUCAPE) $<$ 75 J kg$^{-1}$, or any parameter value equal to 0. Then, the spatial means of all selected environmental variables (Table 1) were calculated within this box. ST23 further describes the methodology employed to complete the environmental analysis, including the computation of the environmental variables.

The basis for the ML algorithms used in this study was the scikit-learn Python package (Pedregosa et al. 2011). Nine different classification ML algorithms were explored to perform binary predictions of potential tornado damage intensity, defined herein as the maximum EF rating of a tornado, using radar-derived pretornadic mesocyclone characteristics and the pretornadic near-storm environment as predictors. The binary classification was of nonsignificant (EF0–1) versus significant

[1] The Amazon Web Services radar archive is available at https://s3.amazonaws.com/noaa-nexrad-level2/index.html.

[2] These details of this software are available at https://www.grlevelx.com/gr2analyst_3/.

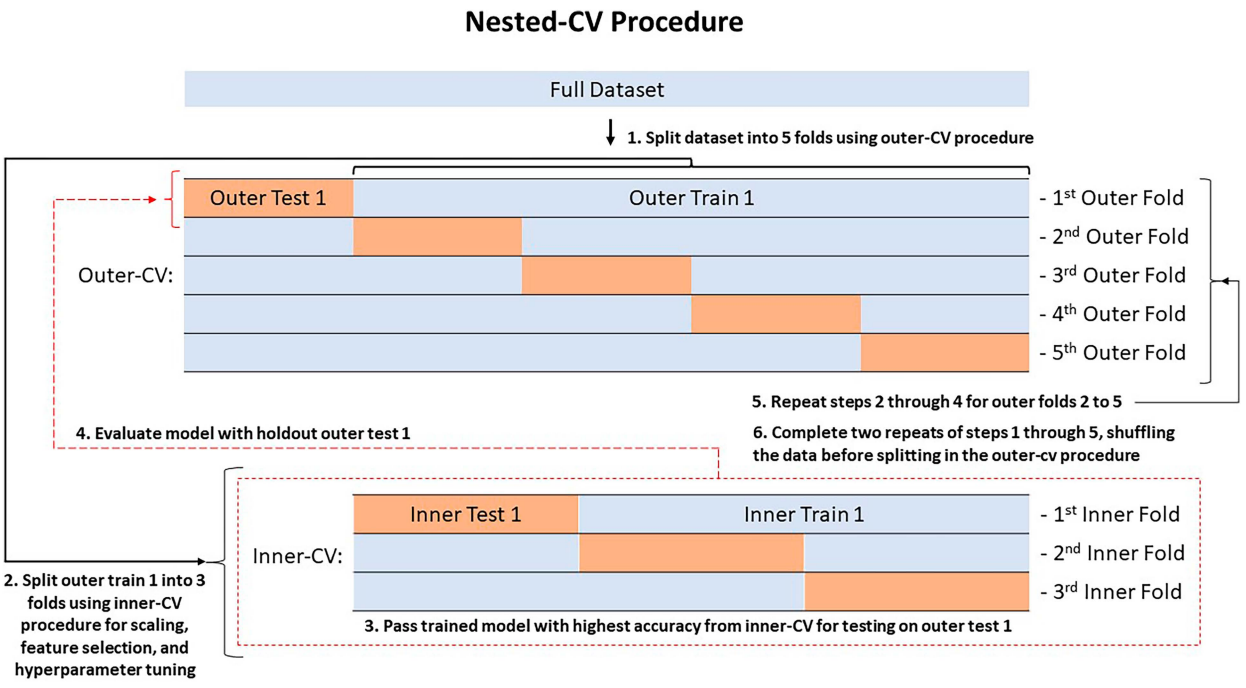


FIG. 2. Schematic illustrating the nested-CV procedure used, which places the inner-fold CV procedure used for feature selection, hyperparameter tuning, and training inside the outer-fold CV procedure for testing the model on unseen data.

(EF2+) tornado damage intensity, with the default probability threshold of 50% used to determine the predictions. A binary classification was chosen over individual EF ratings as a probabilistic prediction separating nonsignificant from significant tornado damage would be most useful to an operational forecaster and, ultimately, to the protection of people and property. Also, it alleviates some of the uncertainty in making predictions of individual damage-based ratings as the EF scale is based on damage indicators and produces a subjective estimate of tornadic wind speeds. The limitations of the EF scale and its dependence on damage indicators are well known (e.g., Doswell and Burgess 1988; Edwards et al. 2013). This study uses the EF scale to discriminate nonsignificant tornadoes (i.e., those with EF0–1 ratings) from significant tornadoes (i.e., those with EF2–5 ratings). The radar predictors used were the mean pretornadic mesocyclone width, convective mode (DSC, QLCS, or MUL), peak pretornadic mesocyclone $\Delta V$, and distance of the storm from the radar (Table 2). The environmental predictors consisted of spatial means of the environmental variables listed in Table 1. Refer to ST20 and ST23 for further exploration of the variables from the pretornadic radar and environmental datasets beyond what Tables 1 and 2 provide.

### b. Models

The nine classification models used were logistic regression (LGR; Kleinbaum et al. 2002), K-nearest neighbor (KNN; Gron 2017a), decision trees (DTs; Mitchell 1997a), naïve Bayes (NB; Mitchell 1997b), random forests (RFs; Breiman 2001), support vector machines (SVCs; Gron 2017b), gradient boosting (GB; Friedman 2001), extreme gradient boosting (XGB; Chen and Guestrin 2016), and extreme gradient-boosted RFs (XGBRFs; Chen and Guestrin 2016). In the presentation of the results of this study, the discussion focuses on the two top-performing all-predictor (LGR and GB), radar-only (RF and GB), and environment-only (RF and NB) models according to Cohen's kappa statistic (Cohen 1960), known within meteorology as the Heidke skill score (Doswell et al. 1990; Wilks 2019). Each classifier uses a pipeline or linear sequence (Pedregosa et al. 2011; Haddad et al. 2022) to complete the entire classification procedure, which included data scaling, feature selection, hyperparameter tuning, and model testing to ensure a consistent procedure, reproducibility of

TABLE 3. Hyperparameter-tuning grids used for each model of focus. Bold values indicate the final model selections that produced the highest accuracy.

| Model | Hyperparameter-tuning grid |
|---|---|
| LGR | Solvers: [newton-cg, lbfgs, **liblinear**]<br>*C* (inverse of regularization strength): [100, 10, 1, 0.1, **0.01**] |
| NB | No hyperparameter tuning completed |
| RF | N_estimators: [4, 8, 16, **32**, 64, 100, 200]<br>Max_depth: [1, 4, 7, **10**, 13]<br>Min_samples_split: [2, 3, **4**, 5, 6, 7, 8]<br>Min_samples_leaf: [**1**, 2, 3, 4, 5] |
| GB | N_estimators: [4, 8, 16, **32**, 64, 100]<br>Learning_rate: [0.5, 0.25, 0.1, **0.05**, 0.01]<br>Min_samples_split: [0.1, **0.3**, 0.5, 0.7, 0.9]<br>Min_samples_leaf: [**0.1**, 0.3, 0.5]<br>Max_depth: [1, **4**, 7, 10, 13] |

TABLE 4. Performance metric descriptions. The description provides a range of values for each predictor and a baseline value where appropriate.

| Metric | Description | Range and baseline |
|---|---|---|
| Accuracy | $\frac{TP + TN}{TP + FP + TN + FN}$<br>Percentage of correct predictions | Range: 0 (bad)–1.0 (good)<br>Baseline (only negative class predictions): 0.646 |
| Brier score | (Probability of positive class – observed label)$^2$<br>Measure of how far the predictions lie from the observed labels | Range: 0 (good)–1.0 (bad)<br>Baseline (based on the climatological frequency of significant tornadoes): 0.42 |
| Brier skill score (BSS) | $\frac{\text{Brier score} - \text{baseline Brier score}}{\text{Baseline Brier score}}$<br>Relative skill of the probability prediction over the climatological forecast | Range: $-\infty$ (bad)–1.0 (good)<br>Baseline (no improvement over climatology): 0 |
| Cohen's kappa or Heidke skill score (HSS) | $\frac{\text{Accuracy} - \text{chance accuracy}}{\text{1–chance accuracy}}$<br>or<br>$2 \times \frac{TP \times TN - FN \times FP}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)}$<br>Level of agreement between and within predictions from the original classifier and those from a random classifier; chance accuracy is the accuracy of the random classifier | Range: −1.0 (bad)–0 (no skill)–1.0 (good)<br>Baseline (no improvement over climatology): 0 |
| False positive rate (FPR) | $\frac{FP}{FP + TN}$<br>Ratio of incorrect positive predictions to all negative observations | Range: 0 (good)–1.0 (bad) |
| Log loss | For one observation: −[Observed label × log (probability of positive class) + (1 − observed label) × log (1 – probability of positive class)]<br>Difference between observations and predictions weighted by probabilities of the positive class | Range: 0 (good)–$\infty$ (bad)<br>Baseline (based on the climatological frequency of significant tornadoes): 0.63 |
| Matthews correlation coefficient (MCC) | $\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$<br>Correlation between the predicted class and the actual class, considering all four contingency-table categories | Range: −1.0 (bad)–0 (no skill)–1.0 (good)<br>Baseline (no improvement over climatology): 0 |
| Negative predictive value (NPV) | $\frac{TN}{TN + FN}$<br>Precision of negative predictions or % correct of negative predictions | Range: 0 (bad)–1.0 (good)<br>Baseline (only negative class predictions): 0.647 |
| Precision or success ratio | $\frac{TP}{TP + FP}$<br>Ratio of correct positive predictions to all positive predictions | Range: 0 (bad)–1.0 (good)<br>Baseline (only positive class predictions): 0.353 |
| Precision–recall area under the curve (PR AUC) | Area under the curve representing precision (success ratio) against recall (POD) for different probability thresholds | Range: 0 (bad)–1.0 (good)Baseline: 0 represents no skill |
| Recall or probability of detection (POD) | $\frac{TP}{TP + FN}$<br>Detection of positive observations | Range: 0 (bad)–1.0 (good) |
| Receiver operating characteristics area under the curve (ROC AUC) | Area under the curve representing the true-positive rate against the false-positive rate for different probability thresholds | Range: 0 (bad)–0.5 (bad)–1.0 (good)<br>Baseline: <0.5 represents no skill |

results, and the prevention of data leakage (when information from outside the training dataset is used to create the model). Figure 1 shows an outline of the modeling procedure.

### c. *Nested cross-validation procedure*

Importantly, data preparation steps use a cross-validation-separated training dataset, which is not included in the test data, to prevent any data leakage, overfitting of the model, and inflation of model results. To achieve this, the pipeline used nested cross validation (CV; e.g., Drosdowsky and Chambers 2001; Martens et al. 2018), which does not require an independent test set for model evaluation. Specifically, the nested procedure used $k$-fold CV (e.g., Loken et al. 2022; Shield and Houston 2022), which is particularly effective when using a smaller data sample, like in this study. The general CV procedure first shuffles and then splits the dataset into "$k$" groups. The inner-CV configuration in this study used stratified threefold CV, wherein each fold had the same

TABLE 5. Classification metrics for the nine all-predictor classifiers used averaged over each CV fold, with the standard deviation given in parentheses. The best score for each metric is in bold text, and the models are ordered from highest to lowest HSS.

| | GB | LGR | NB | XGBRF | XGB | RF | SVC | KNN | DT |
|---|---|---|---|---|---|---|---|---|---|
| Accuracy | **0.888 (0.026)** | 0.886 (0.024) | 0.884 (0.029) | 0.884 (0.03) | 0.883 (0.028) | 0.879 (0.026) | 0.878 (0.031) | 0.873 (0.029) | 0.839 (0.071) |
| Brier score | 0.103 (0.019) | 0.12 (0.029) | **0.1 (0.019)** | 0.212 (0.036) | 0.12 (0.039) | 0.12 (0.03) | 0.103 (0.016) | 0.107 (0.014) | 0.132 (0.051) |
| BSS | 0.754 (0.021) | 0.713 (0.028) | **0.76 (0.025)** | 0.493 (0.024) | 0.714 (0.019) | 0.714 (0.021) | 0.754 (0.022) | 0.743 (0.029) | 0.684 (0.021) |
| Cohen's kappa (or HSS) | **0.741 (0.064)** | 0.737 (0.057) | 0.733 (0.069) | 0.733 (0.073) | 0.729 (0.069) | 0.721 (0.064) | 0.718 (0.071) | 0.709 (0.069) | 0.621 (0.2) |
| FPR | **0.027 (0.022)** | 0.031 (0.028) | 0.029 (0.025) | 0.029 (0.023) | **0.027 (0.022)** | 0.036 (0.021) | 0.031 (0.041) | 0.045 (0.046) | 0.067 (0.07) |
| Log loss | 0.369 (0.072) | 0.401 (0.075) | **0.346 (0.067)** | 0.616 (0.075) | 0.401 (0.095) | 0.425 (0.153) | 0.361 (0.058) | 0.901 (0.465) | 1.498 (1.664) |
| MCC | **0.755 (0.059)** | 0.75 (0.056) | 0.747 (0.065) | 0.747 (0.069) | 0.746 (0.062) | 0.734 (0.058) | 0.735 (0.071) | 0.725 (0.064) | 0.634 (0.202) |
| NPV | **0.871 (0.032)** | 0.87 (0.029) | 0.868 (0.031) | 0.868 (0.034) | 0.866 (0.035) | 0.866 (0.034) | 0.861 (0.027) | 0.866 (0.036) | 0.845 (0.069) |
| Precision (or success ratio) | **0.939 (0.048)** | 0.933 (0.06) | 0.935 (0.055) | 0.934 (0.052) | 0.938 (0.048) | 0.92 (0.044) | 0.935 (0.08) | 0.911 (0.078) | 0.806 (0.243) |
| PR AUC | 0.844 (0.061) | **0.893 (0.036)** | 0.884 (0.036) | 0.877 (0.038) | 0.842 (0.061) | 0.861 (0.047) | 0.892 (0.029) | 0.859 (0.035) | 0.752 (0.136) |
| Recall (or POD) | **0.733 (0.077)** | **0.733 (0.067)** | 0.727 (0.074) | 0.727 (0.078) | 0.72 (0.085) | 0.724 (0.081) | 0.711 (0.063) | 0.723 (0.092) | 0.667 (0.21) |
| ROC AUC | 0.881 (0.039) | 0.895 (0.038) | 0.884 (0.041) | **0.899 (0.025)** | 0.877 (0.041) | 0.887 (0.035) | 0.896 (0.031) | 0.889 (0.03) | 0.832 (0.1) |

proportion of nonsignificant and significant tornadoes as the original dataset. The outer-CV configuration used a stratified fivefold CV, repeated three times, with the data shuffled before each repetition. The repetition with different samplings of the unseen test data provides more robust results than a single iteration, as the performance metrics from each repetition are averaged together (e.g., Krstajic et al. 2014). Figure 2 provides a schematic of the nested-CV procedure.

### *d. Scaling*

The first step in the modeling pipeline was scaling the predictor data within the inner-CV procedure to normalize their range (e.g., Flora et al. 2021; Chase et al. 2022). Scaling was necessary to reduce the nonphysical inflated influence from large-range predictors for algorithms such as KNN, SVC, and LGR and to increase model-training efficiency for all classifiers. Using the training dataset, a scalar robust to outliers removed the median and scaled the data according to the range between the first and third quartiles for each predictor. This allowed for outliers to remain in the dataset and prevented the need to remove cases from an already small sample.

### *e. Feature selection*

The second step in the modeling pipeline was feature selection within the inner-CV procedure. Performing feature selection before training and testing the classifiers helps reduce overfitting and improve interpretability by removing redundant data, improve model accuracy by removing misleading data, and shorten training time due to the presence of less data. This study uses recursive feature elimination (RFE; Guyon et al. 2002; $G_{2021}$) for feature selection. The procedure fits a chosen classifying core to the training data and ranks features based on the importance provided by the classifier. The RFE algorithm then removes the least important features and refits the model on the remaining features through successive iterations until the desired number of features remains. An automated, cross-validated RFE algorithm, which automatically determines the best combination of predictors, was used to achieve the best-performing (using accuracy) and final model. In this study, RFE used LGR as its core for each model, which was chosen based on performance (i.e., model performance varied slightly as the classifier within RFE was changed) and time efficiency.

Correlated predictors can bias the feature ranking and, therefore, the removal of features, so an exploration of the impact of correlated predictors was completed. As was determined using the variance inflation factor (VIF), several predictors in this study exhibit multicollinearity (VIF $\geq$ 10). Multicollinearity was particularly present within the environmental dataset, especially between predictors of the same category, such as within CAPE parameters and low-level shear parameters. Correlated predictors may become misleadingly insignificant or produce incorrect feature coefficients by providing similar information about the target variable. RFE, Shapley additive explanations (SHAP; Lundberg and Lee 2017; Lundberg et al. 2020), and other sensitivity tests addressed multicollinearity concerns. The RFE procedure specifically reduced

TABLE 6. Contingency table showing summed values across all CV folds for the (a) all-predictor LGR model and GB model, (b) the radar-only RF model and GB model, and (c) the environment-only RF model and NB model.

| (a) All predictor | | (b) Radar only | | (c) Environment only | |
|---|---|---|---|---|---|
| True negatives | False positives | True negatives | False positives | True negatives | False positives |
| LGR: 564<br>GB: 566 | LGR:18<br>GB: 16 | RF: 549<br>GB: 564 | RF: 33<br>GB: 18 | RF: 500<br>NB: 479 | RF: 82<br>NB: 103 |
| False negatives | True positives | False negatives | True positives | False negatives | True positives |
| LGR: 85<br>GB: 85 | LGR: 233<br>GB: 233 | RF: 75<br>GB: 86 | RF: 243<br>GB: 232 | RF: 181<br>NB: 163 | RF: 137<br>NB: 155 |

the issue of multicollinearity among the environmental variables. In a separate investigation, correlated variables were removed successively to evaluate further the potential effect of variables showing large multicollinearity (VIF $\geq$ 10) on model skill or feature importance rankings. This process revealed little change in feature importance rankings and model performance, except when higher importance variables were removed and were not included in final model iterations.

### *f. Hyperparameter tuning*

The third step in the modeling pipeline was hyperparameter tuning (e.g., Gagne et al. 2019; Chase et al. 2022) within the inner-CV procedure, which allows for the configuration of variables used to control model training. Grid-search CV and randomized-search CV (Pedregosa et al. 2011) determined the optimal set of hyperparameters for each model that resulted in the best performance using a scoring method of accuracy. The grid-search method evaluates every combination of hyperparameters from a provided grid of values, while a randomized search only evaluates a defined number of combinations. The grid search can be very time consuming, depending on the number of combinations evaluated. Therefore, a full grid search of hyperparameters was only completed for LGR, NB, KNN, SVC, and DT models. For RF, GB, XGB, and XGBRF models, a randomized search of 50% of the specified grid space was completed. Table 3 shows the grids of hyperparameters used for the models of focus, with the remaining provided in Table S1 in the online supplemental material. A one-dimensional search outward from a default starting value determined the range of values iterated over for each hyperparameter. This search found a reasonable upper and lower bound, while also considering computation time. Within the range of values used, model performance did not change drastically. Although still limited, tree-based models showed the greatest sensitivity to hyperparameters. Scikit-learn documentation provides detailed descriptions of all hyperparameters (Pedregosa et al. 2011).

### *g. Model evaluation and feature importance*

In the final step of the modeling pipeline, several performance metrics were used to evaluate model testing in the outer CV and were selected to consider multiple perspectives in determining model performance (Table 4). These include multiple probability thresholds, imbalanced data, and equal consideration of both classes to avoid biases and provide a complete picture of each classifier's performance. Metrics from each test fold were averaged to evaluate model performance on unseen data, and testing of models on multiple different groups of unseen data provides more robust results. Metrics include the false positive rate (FPR) and the negative predictive value (NPV), which focus on the negative class (<50% probability of EF2+ tornadoes), and metrics such as recall (i.e., POD) and precision–recall area under the curve (PR AUC; Saito and Rehmsmeier 2015), which focus on the positive class (>50% probability of EF2+ tornadoes). Also, metrics such as Cohen's kappa statistic, Matthews correlation coefficient (MCC; Matthews 1975; Chicco and Jurman 2020), and PR AUC are not influenced by the slightly imbalanced dataset used in this study. MCC and Cohen's kappa statistic also help reveal the skill of each classifier over a random classifier or one that makes predictions based upon class frequencies, consider all four values in the contingency table, and are not biased or weighted to the positive or negative class. Brier score, Brier skill score (BSS; Brier 1950), and log loss evaluate forecasted probabilities. In this study, equal attention should be given to evaluating the models' ability to successfully predict significant tornadoes and limit false alarms, as both are important for creating a useful and trustworthy tool for anticipating tornado damage intensity. The modeling pipeline used a random state of four to ensure the reproducibility of results.

Additionally, learning curves evaluated the training of the models, representativeness of the dataset, and over- or underfitting of the models. Due to the relevance of and desire for the direct use of the probabilities output by the models, calibration curves were used to explore the reliability of the modeled probabilities and ensure the modeled probabilities aligned with the actual likelihood of events (Hsu and Murphy 1986). Permutation tests were also completed for each model to help evaluate the statistical significance of the results (Ojala and Garriga 2010). The supplemental material provides an evaluation and discussion of model training, the statistical significance of results, and the reliability of model-output probabilities (Figs. S1–S9).

The feature ranking output from the RFE algorithm revealed the features most often ranked as important to the classification predictions across the CV folds. There can be biases in interpreting feature importance from RFE, as the rankings can be sensitive to the order in which variables are omitted. Thus, other model interpretation and feature importance methods were explored.

TABLE 7. Performance of the all-predictor LGR model (GB model) broken down by EF rating summed over all CV folds. Sample size by EF rating includes CV repetitions.

| EF rating | Sample size | FP source | FN source | Correct | % Correct |
|---|---|---|---|---|---|
| 0 | 234 | 3 (3) | — | 231 (231) | 99% (99%) |
| 1 | 348 | 15 (13) | — | 333 (335) | 96% (96%) |
| 2 | 177 | — | 82 (83) | 95 (94) | 54% (53%) |
| 3 | 99 | — | 3 (2) | 96 (97) | 97% (98%) |
| 4 | 30 | — | 0 (0) | 30 (30) | 100% (100%) |
| 5 | 12 | — | 0 (0) | 12 (12) | 100% (100%) |

SHAP allowed for further exploration of the predictions made by the classification models in an interpretable and reliable manner by demonstrating how each predictor changed the probability of nonsignificant versus significant tornadoes. This provided a means for understanding the influence of each predictor on the model classifications outside of potential biases from multicollinearity. For consistency and efficiency, the SHAP analysis only used the RF model, and the modeling pipeline did not include RFE to be able to explore the SHAP values of all predictors. SHAP analyses with other tree-based models produced similar results. SHAP values (Štrumbelj and Kononenko 2014; Shapley 2016) represent the contribution of each variable to the model output, which in this study used probability. One can deeply explore the classification of individual cases (local model explanation) and the effect of a predictor across the entire dataset (global model explanation), allowing more understanding of the model's decision-making process to increase confidence in the predictions' interpretability and trustworthiness.

## 3. Machine learning results and feature importance

This section focuses on the performance of the LGR and GB models using the radar- and environment-based predictors ("all predictor"). When averaged over all CV folds, the accuracy of these two models was 89%, with a standard deviation of ≈0.025 (Table 5). The low standard deviation in accuracy across the CV folds shows the consistency of the results across different divisions of the dataset (Table 5). For reference, the DT model had the lowest accuracy at 84%, with a standard deviation of 0.071 (Table 5). The use of standard deviation across CV folds is meant to provide a general idea of the range of model performance and does not allow for an assessment of the statistical significance of intermodel differences.

In this study, the baseline values of log loss and Brier score are approximately 0.63 and 0.42, respectively, which models would need to outperform to beat predictions based upon climatology. This is a result of using the climatological frequency of EF2+ tornadoes from the training dataset (35.3%) with the corresponding observed label of 1 in the log loss and Brier score equations in Table 4. For both models of focus, log loss (LGR: 0.401 and GB: 0.369), Brier score (LGR: 0.120 and GB: 0.103), and BSS (LGR: 0.713 and GB: 0.754) provide evidence for a skilled model. Although there is good performance in the correct classification of both potential nonsignificant

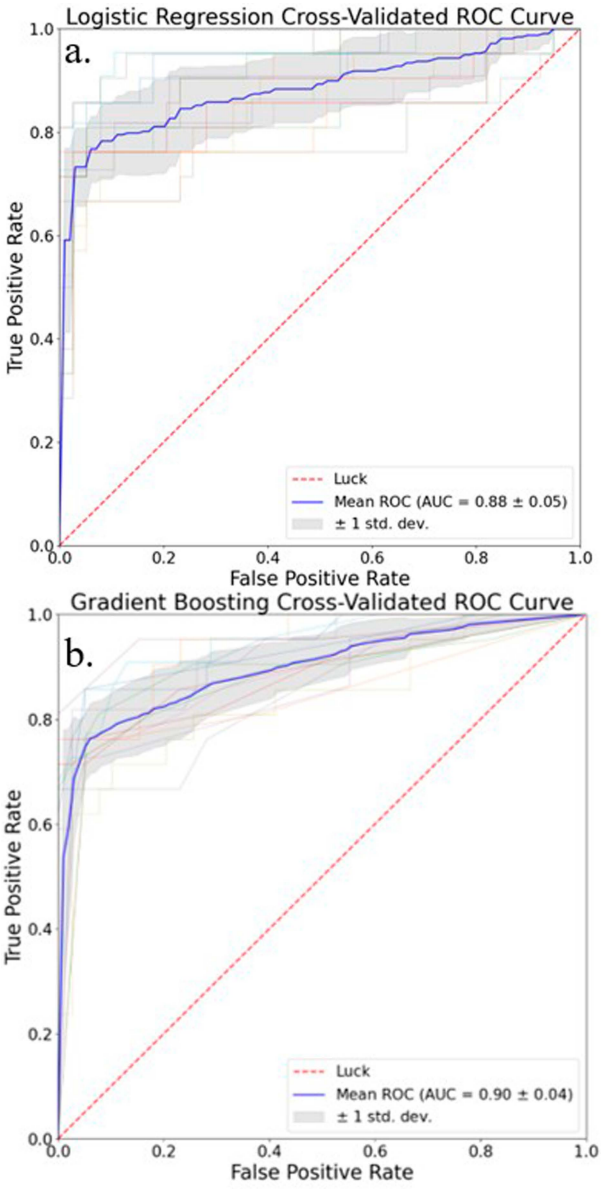


FIG. 3. ROC curves for each CV fold, as well as the mean ROC curve and AUC across all CV folds (solid blue) from the all-predictor (a) LGR model and (b) GB model. The red dashed line represents the no-skill line, and the gray, shaded region is plus-or-minus one standard deviation, showing the spread of the ROC curves across the 15 CV folds. A skilled model will have an ROC curve bowed to the top-left corner of the plot, maximizing true negatives (TNs) and true positives (TPs).

TABLE 8. Performance of the all-predictor LGR model (GB model) broken down by convective mode summed over all CV folds. Sample size by mode includes CV repetitions.

| Mode | Sample size | FP source | FN source | Correct | % Correct |
|---|---|---|---|---|---|
| QLCS | 372 | 3 (3) | 30 (30) | 339 (339) | 91% (91%) |
| DSC | 390 | 10 (9) | 35 (36) | 345 (345) | 88% (88%) |
| MUL | 138 | 5 (4) | 20 (19) | 113 (115) | 82% (83%) |

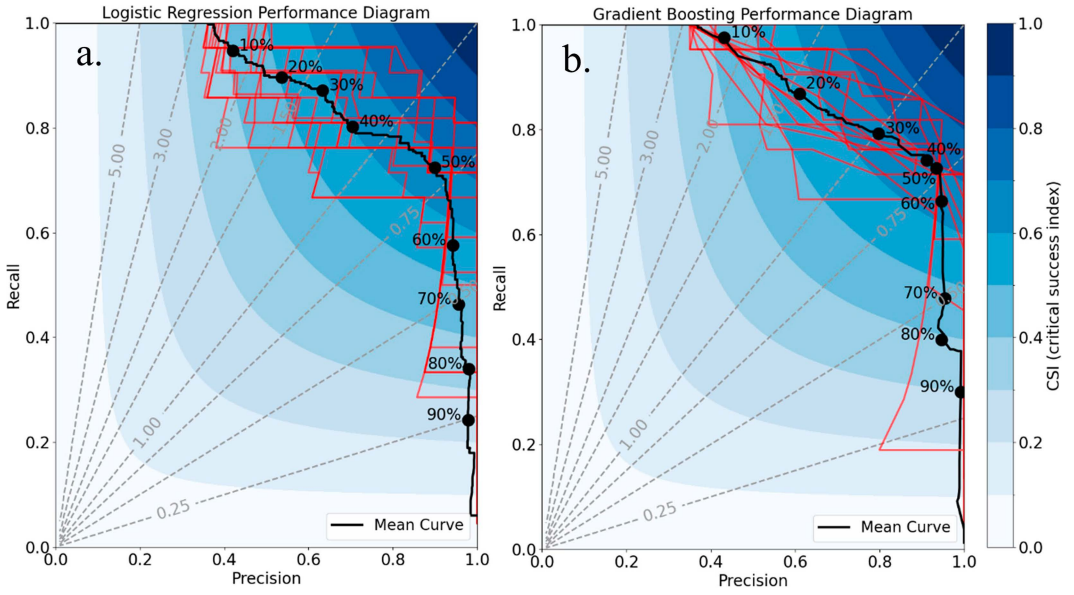


FIG. 4. Performance diagram curves for each CV fold, as well as the mean curve across all CV folds (solid black) for the all-predictor (a) LGR model and (b) GB model. The blue-shaded region represents CSI, and the gray dashed lines indicate frequency bias. The curve is plotted at 1% probability threshold intervals, with the probability threshold labeled every 10%.

and significant tornadoes, precision (i.e., success ratio) and recall (i.e., POD) reveal a tendency for more misclassifications of EF2+ tornadoes [false negatives (FNs)] than EF0 − 1 tornadoes [false positives (FPs)] resulting in a lower recall score (LGR: 0.733 and GB: 0.733) and a higher precision score (LGR: 0.93 and GB: 0.939). The better classification of potential nonsignificant tornadoes is also seen in the contingency tables for each model, from which many of the metrics used are determined (Table 6). Cohen's kappa statistic (LGR: 0.737 and GB: 0.741) and MCC (LGR: 0.75 and GB: 0.755) both provide evidence of skilled predictions for both classes.

When performing the same metric analysis for the models only using radar-based predictors ("radar only"), a skilled prediction of potential tornado damage intensity is still shown across the nine classifiers, although with slightly weaker performance metrics. The RF and GB models showed the best performance (e.g., mean accuracy of 86.0% and 85.4% for the RF and GB models, respectively; Table S2). Additionally, the pattern of more FNs than FPs remains (Table 6). For the models only using environment-based predictors ("environment only"), a less skilled classification of potential tornado damage intensity is reflected in the reduced performance metrics across the nine classifiers (e.g., mean accuracy of 70.8% and 70.4% for the top-performing models of RF and NB, respectively; Table S3). This shows that anticipating potential tornado damage intensity for ongoing thunderstorms using environmental predictors alone is difficult and stresses the importance of combining the radar characteristics with the environmental parameters. However, the environment-only models could provide helpful information if radar metrics from ongoing storms are unavailable.

Returning our focus to the all-predictor LGR and GB models, to better understand the contributors to misclassifications, the occurrence of FPs and FNs was sorted by EF rating (Table 7) and convective mode (Table 8). First, when viewing the breakdown of misclassifications by EF rating for both models of focus, most misclassifications are from FNs resulting from the incorrect classification of EF2 events as nonsignificant. Both models correctly classify just over 50% of EF2s as significant (Table 7). The other EF categories are correctly classified more than 95% of the time, with all EF4 and EF5 events correctly classified as significant. One must note the small sample of EF4+ tornadoes (14 cases) due to their rare occurrence and exercise caution when drawing conclusions and applying model results related to the EF4+ cases in this dataset. When viewing the three convective modes present in this dataset, there is not as distinct a pattern in misclassifications (Table 8). The accuracy for each mode is greater than 80%, with the lowest being 82% and 83% for MUL for the LGR and GB models, respectively. This is

TABLE 9. Average relative feature importance for the top 10 predictors, calculated using RFE, with the all-predictor LGR model.

| Position | Predictor | Average ranking |
|---|---|---|
| 1 | Pretornadic mesocyclone width | 1 |
| 2 | QLCS mode | 2.4 |
| 3 | 0–1-km bulk shear | 4.667 |
| 4 | 0–8-km bulk shear | 6.333 |
| 5 | Pretornadic mesocyclone intensity | 7.667 |
| 6 | 0–3-km relative humidity | 8.2 |
| 7 | LCL height | 8.267 |
| 8 | Effective bulk shear | 8.4 |
| 9 | Tornadic tilting and stretching | 9.133 |
| 10 | 0–500-m SRH | 10.4 |

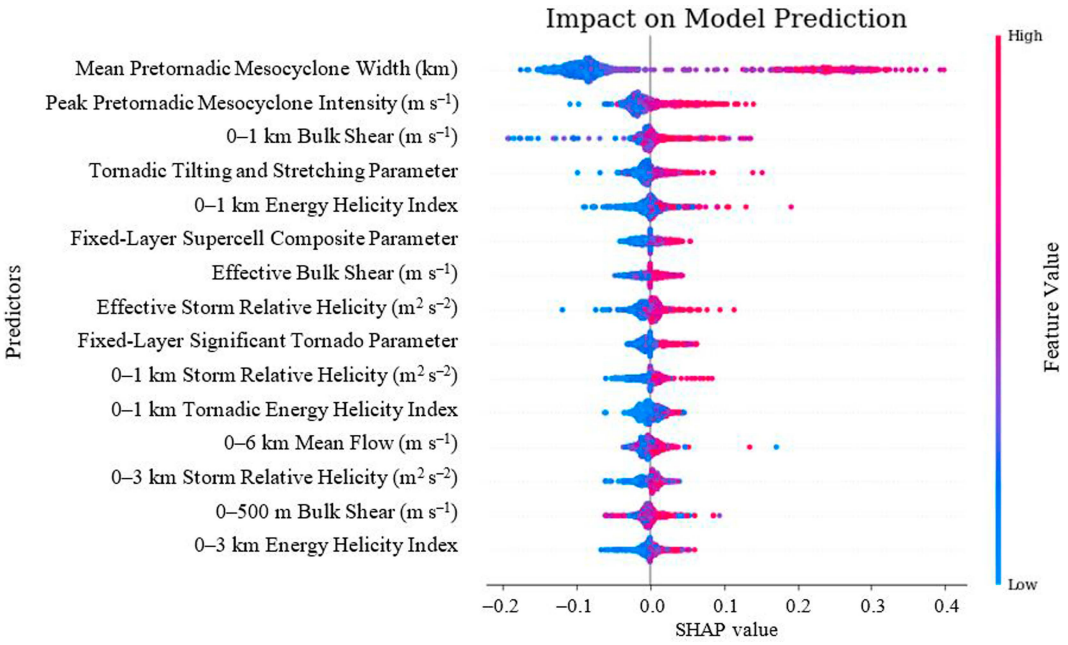


FIG. 5. SHAP values of each feature for every sample from the all-predictor RF model, sorted from most important to least important, based upon the mean absolute value of the SHAP values for each feature. Warmer colors indicate larger feature values, and colder colors represent smaller ones. Wider sections of the violin plots indicate a greater density of points.

undoubtedly influenced by the small number (46) of MUL cases, with 45 misclassified DSC cases and 33 misclassified QLCS cases for both models, the majority being FNs. The supplemental material provides the EF rating and convective mode breakdown for the radar- and environment-only models.

Receiver operating characteristic (ROC) (Wilks 2019) curves and performance diagrams (Roebber 2009) all provide another means of assessing model performance across different probability threshold values. When averaged across all CV folds, ROC AUC of 0.895 and 0.881 and PR AUC of 0.893 and 0.844 for the LGR and GB models, respectively, provide evidence of a skilled prediction of potential tornado damage intensity (Fig. 3 and Table 5). The performance diagrams further reveal a skilled model, with the mean curve demonstrating high critical success index (CSI) and low bias for several probability thresholds (Fig. 4). These plots could also be the basis for adjusting the probability thresholds to increase the detection of significant events at the cost of misclassifying more nonsignificant events. For radar-only models, the ROC curves and performance diagrams show similar performance to all-predictor models (Figs. S10 and S11), and for environment-only models, they show reduced performance (Figs. S12 and S13).

A vital part of model interpretation is understanding feature importance and how each feature influences the model classifications. To address the former in part and to better understand the RFE that determined what features were used to make predictions, the average relative feature importance from the all-predictor LGR RFE rankings is used to order the features from most important to least important. Within the RFE feature-importance ranking, features selected in a particular CV fold are ranked 1, with all remaining unselected features ranked based on their importance, which is determined from the predictor's coefficients from LGR. The ranking reveals the features most selected by the RFE algorithm for inclusion in the final model. This perspective of average feature importance across all CV folds reveals that the top 10 predictors includethe pretornadic mesocyclone width and intensity, the QLCS storm mode, and an array of environmental parameters, including low-level and deep-layer shear parameters, moisture parameters, and composite parameters (Table 9). The top 20 included other composite parameters, such as effective layer supercell composite parameter (SCPe) and 0–3-km energy helicity index (03EHI), and thermodynamic parameters, such as the LCL-to-LFC depth, 3–6-km lapse rate (36LR), and surface-based CAPE (SBCAPE). The RFE procedure revealed that the most-skilled predictions could be achieved with a relatively simple model using only about the top 10 predictors, which most often consisted of three radar and seven environmental predictors. The RFE feature-importance ranking for radar- and environment-only models showed consistent relationships, with the pretornadic mesocyclone width, pretornadic mesocyclone intensity, and QLCS storm mode being the most important for the radar-only models, and kinematic and composite parameters being the most important for the environment-only models (Table S6). For the radar-only models, RFE only kept the top four radar predictors on average. In comparison, the environment-only models, on average, required the top 15 predictors to achieve the most-skilled predictions.

To further understand and explore feature importance, SHAP was used to view how each feature influences the

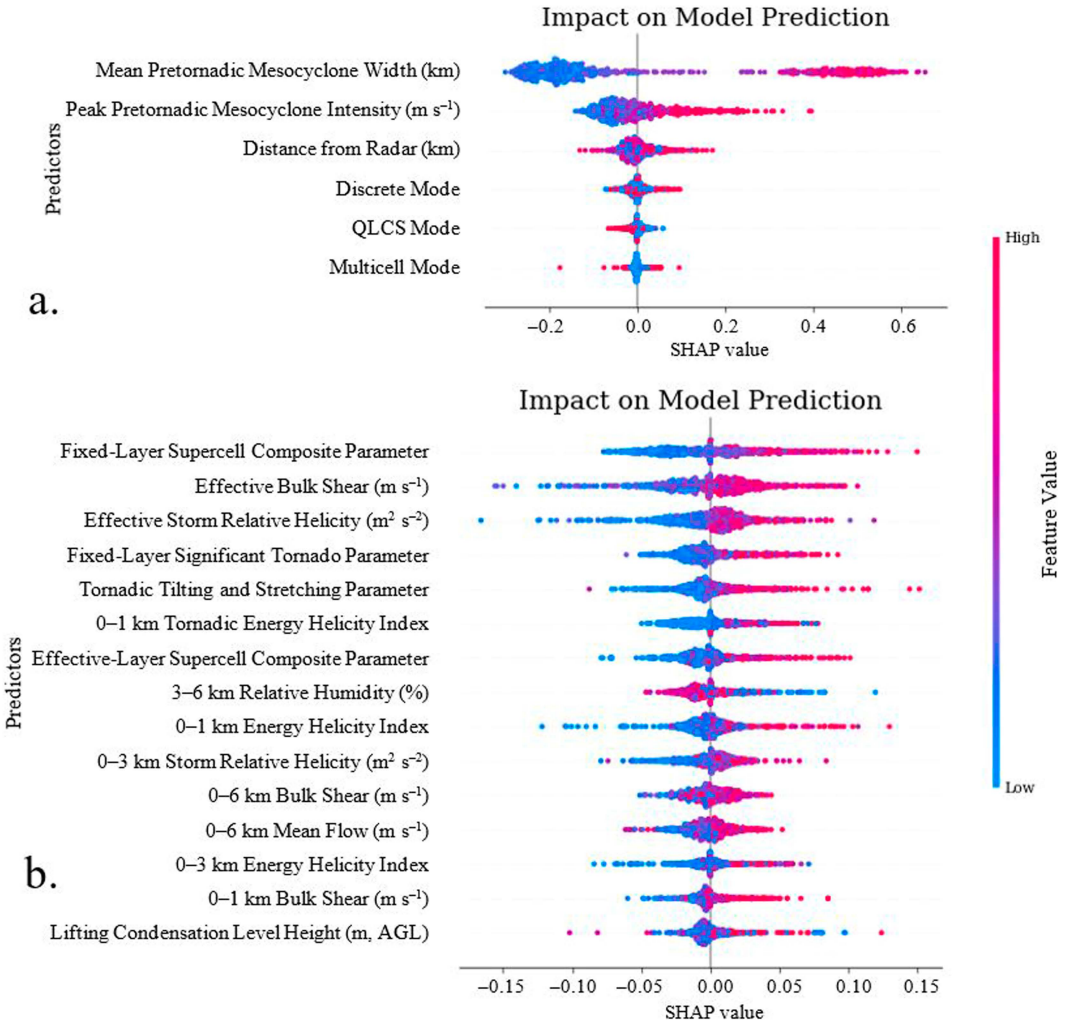


FIG. 6. As in Fig. 5, but when only (a) radar predictors and (b) environmental predictors are used.

model predictions for individual cases and the entire dataset. Given the known limitations of the RFE feature ranking described earlier, the SHAP analysis provides a more complete and reliable interpretation of feature importance. Here, SHAP values represent the contribution of each predictor toward pushing the probability of significant tornadoes away from the expected value, which is based on the climatology of the dataset used (35% chance of significant tornadoes). All SHAP analyses average the SHAP values across all CV folds. Figure 5 plots the SHAP values of the top 15 predictors for each case in the dataset for the all-predictor RF model, sorting the features based on the mean absolute value of the SHAP values. This figure shows the distribution of each predictor's impacts on the model output. The larger the magnitude of the SHAP value, the greater the influence of the predictor on the prediction of potential tornado damage intensity. The top two features are average pretornadic mesocyclone width and peak pretornadic mesocyclone intensity. However, the influence of mesocyclone intensity is comparable to that of the most influential environmental parameters. Wider and stronger pretornadic mesocyclones push the model predictions toward significant tornadoes. The remainder of the top 15 predictors are kinematic and composite environmental parameters, including low-level shear and storm-relative helicity (SRH), effective bulk shear (EBS), effective SRH (ESRH), tornadic tilting and stretching (TTS), 0–1-km EHI (01EHI), SCP, and STP, with larger values increasing the probability of significant tornadoes.

Figure 6a reveals consistent results for the radar-only model SHAP analysis. The pretornadic mesocyclone width is most important, followed by intensity, again showing that wide and strong pretornadic mesocyclones increase the probability of significant tornadoes. Each with a smaller impact, distance from the radar and the DSC, QLCS, and MUL storm modes round out the radar predictor ranking. Storms farther from the radar slightly increase the probability of a significant tornado. The increase could be due to the models learning the relationship between radar-resolution degradation and increasing beamwidth with increasing distance. This can make rotation signatures appear weaker and wider than they are, with the wide bias dominating this effect. A DSC storm mode

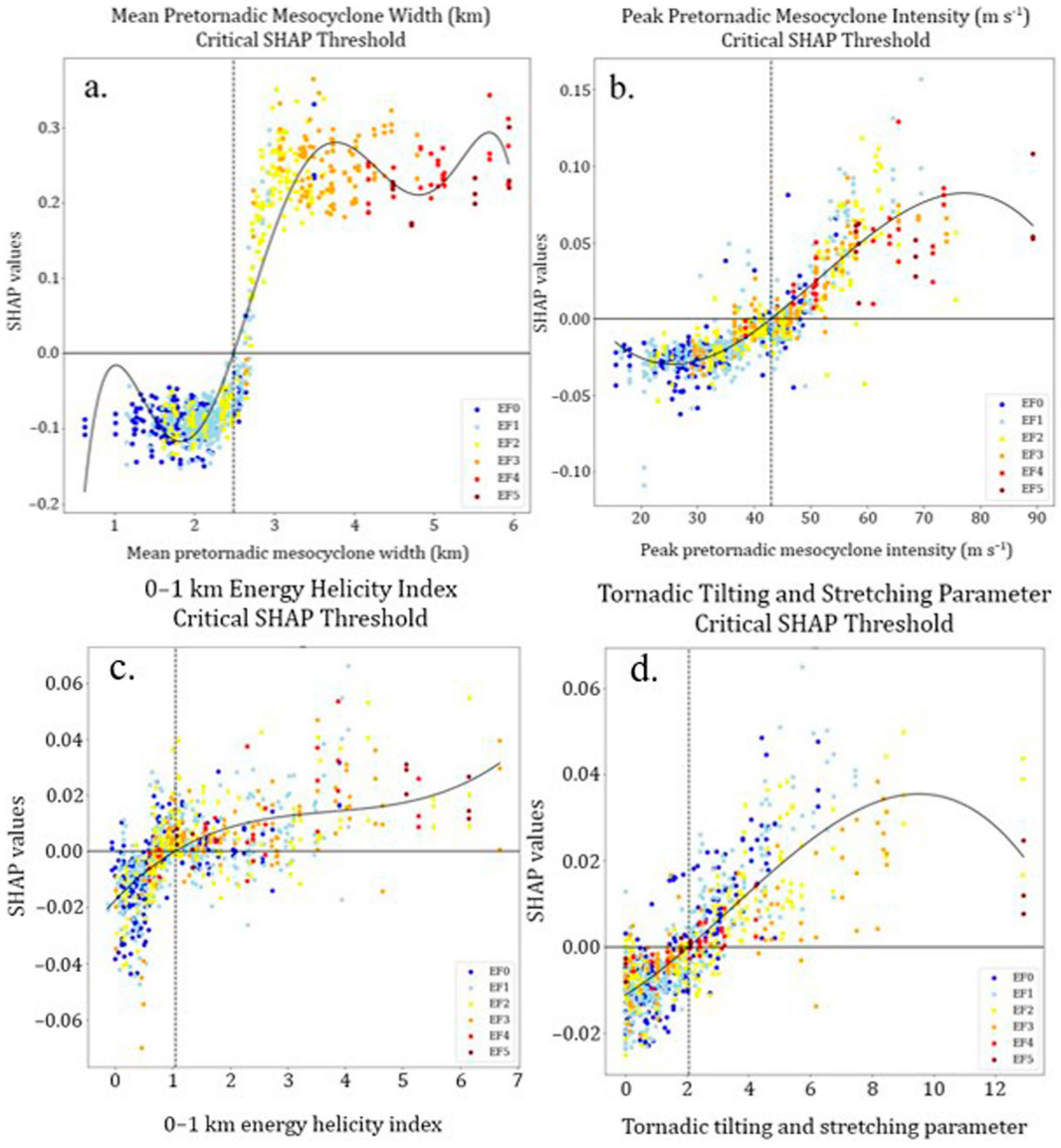


FIG. 7. SHAP values for (a) the mean pretornadic mesocyclone width (km), (b) the peak pretornadic mesocyclone intensity (m $s^{-1}$), (c) EHI01, and (d) TTS for all cases color coded by EF rating from the all-predictor RF model. The black curve is the SHAP = $f(x)$ polynomial. The vertical black dashed line shows the intersection between the SHAP = $f(x)$ polynomial and the 0 SHAP value.

slightly increases the probability of a significant tornado. A QLCS storm mode slightly decreases the probability of a significant tornado, and the MUL storm mode has no apparent impact on model output.

Figure 6b plots the SHAP values of the top 15 predictors from the environment-only models and reveals consistent results in the all-predictor models. Composite parameters, EBS, and ESRH are the most important predictors, and larger values increase the probability of significant tornadoes. 3–6-km relative humidity (36RH) is in the top 10, with drier midlevel air associated with an increased probability of significant tornadoes. Other important environmental relationships revealed in the top 15 predictors include stronger deep-layer shear and faster storm motion increasing the probability of significant tornadoes. SHAP analyses isolating certain groups of environmental parameters (e.g., thermodynamic, wind, and composite) were also completed to better understand the environmental relationships the models are using to predict potential tornado damage intensity, and this revealed a few interesting results (not shown). A main takeaway from this additional analysis is that using the subsets of environmental parameters slightly reduced model performance, showing the importance of using the most important predictors from each subset to achieve the best-performing model. Additionally, when composite parameters were removed, EBS and ESRH were the most important wind shear predictors, with 36RH

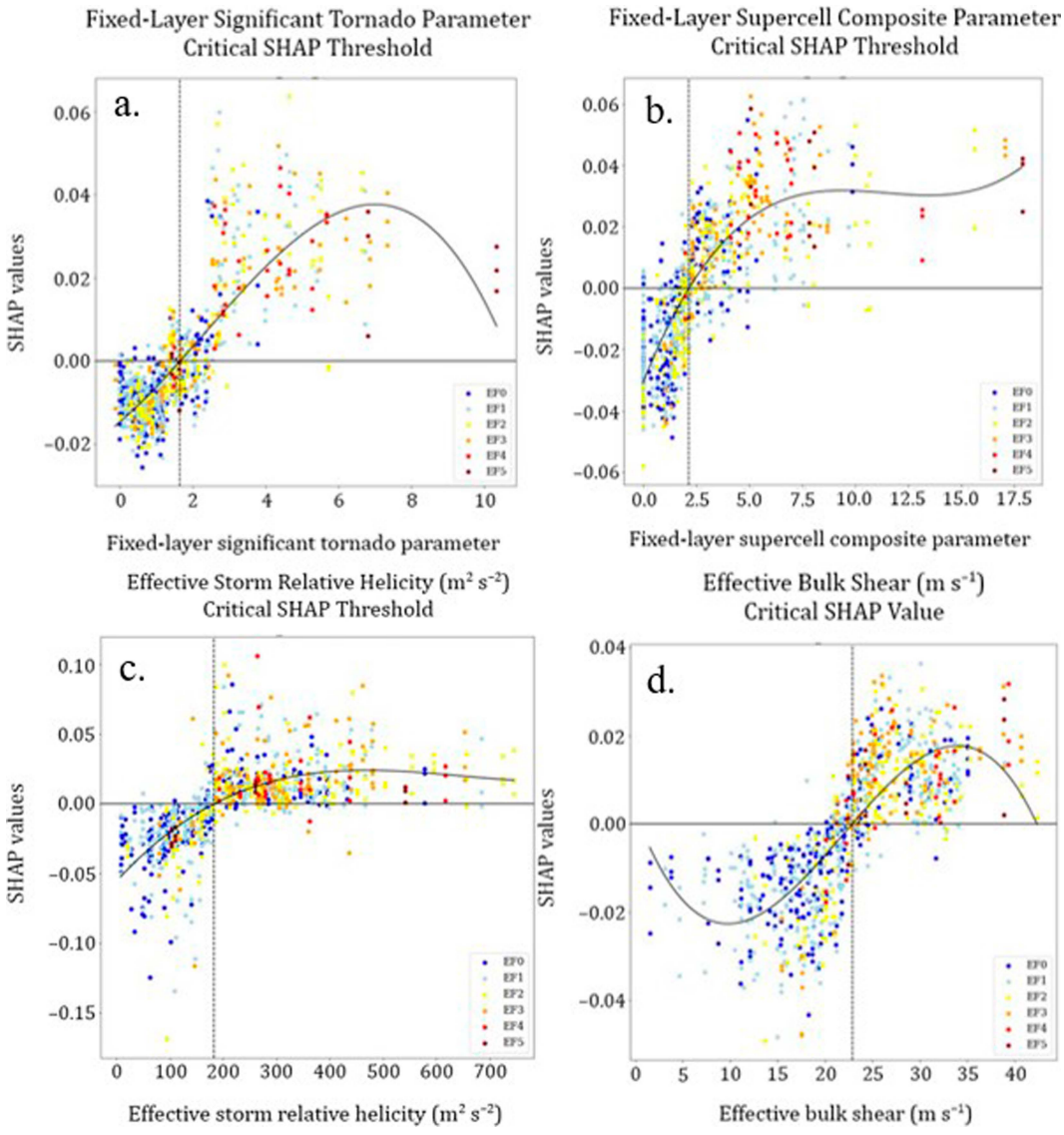


FIG. 8. As in Fig. 7, but for (a) STPf, (b) SCPf, (c) ESRH ($m^2 s^{-2}$), and (d) EBS ($m s^{-1}$).

and mixed-layer CAPE (MLCAPE) being the most important thermodynamic predictors, consistent with the full environmental model.

One can fit a polynomial to the distribution of SHAP values for a particular predictor and evaluate where the polynomial crosses the zero line. This point is where there is a directional change in the contribution of the predictor toward a classification of potential tornado damage intensity. Further, it reveals the importance and influence of each feature. The SHAP polynomial plots are shown for a variety of the top radar and environmental predictors when all predictors are used, with the points color coded by EF rating to show which predictors separate the cases the greatest across their SHAP values (note the differing *y* axes to better visualize the spread of the distributions and the shape of the polynomial; Figs. 7–9). For example, the transitional SHAP values for the pretornadic mesocyclone width and intensity are 2.5 km and 42 m $s^{-1}$, respectively (Fig. 7). When viewing some of the top environmental parameters, the transitional values for fixed-layer SCP (SCPf), fixed-layer STP (STPf), ESRH, EBS, and S01 are 2.3, 1.7, 180 $m^2 s^{-2}$, 23, and 16.5 m $s^{-1}$, respectively.

The structure and degree of the best-fit polynomial are dependent on the variable. For example, SRH parameters, SCPf, and the radar predictors show the greatest change in SHAP values near the transitional value. However, the influence on the probability of a significant tornado does not increase beyond that. Once the pretornadic mesocyclone width reaches about 3.75 km or ESRH reaches 350 $m^2 s^{-2}$, there is little change in each predictor's contribution toward the probability of significant tornadoes. Other predictors like S01, TTS, or STPf show a more linear increase in their SHAP values with increasing parameter values. When considering whether the transitional SHAP value depends on the convective mode, no substantial differences exist for any top

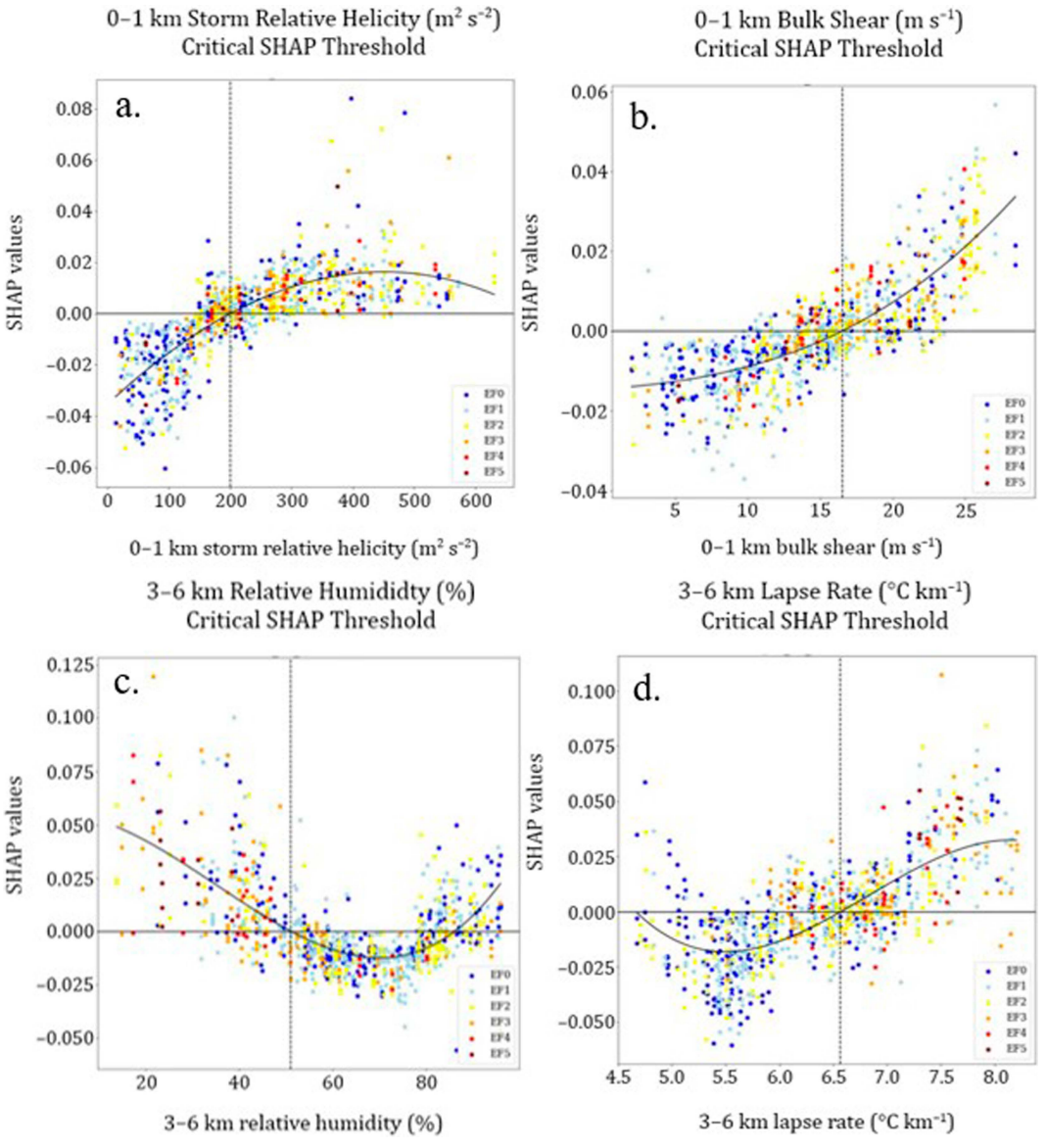


FIG. 9. As in Fig. 7, but for (a) SRH01 ($m^2 s^{-2}$), (b) S01 ($m s^{-1}$), (c) 36RH (%), and (d) 36LR (°C $km^{-1}$).

predictors (not shown). While nonsignificant and significant tornadoes occur on both sides of the transitional SHAP values, they still demonstrate useful transitional regions of predictors of tornado damage intensity.

To briefly compare the classification skill of the environment-only models to that of previous work (given that there are no previous studies that use radar data from ongoing thunderstorms to anticipate tornado damage intensity), the LGR models of $T_{2011}$ and $G_{2021}$ are considered. When viewing these comparisons, one must consider that these studies use different data sources for model training and testing. The performance of the environment-only LGR model in this study replicates the performance of these two studies when viewing the mean ROC AUC of 0.73 ($T_{2011}$ = 0.72 and $G_{2021}$ = 0.74). The $G_{2021}$ also used an RF model and achieved a higher mean ROC AUC of 0.79 (0.73 in this study). When considering predictor importance, $T_{2011}$ only explored the components of STP in their environmental models for supercell storms. In contrast, although without composite parameters, $G_{2021}$ explored a greater variety of environmental parameters comparable to this study for supercell and discrete nonsupercell storms. Overall, there is overlap in the most important predictors highlighted in each study, with kinematic parameters (mainly low-level shear) being the most important, followed secondarily by thermodynamic parameters.

## 4. Example application and suggested future development

The 300-case dataset was used to train the models for application to three tornadic events encompassing a range of

TABLE 10. The 14 cases selected from the 11 Dec 2021, 19 Jun 2023, and 8 Jun 2022 tornadic events for testing on final versions of models.

| Case number | Date | State | County | Mode | Genesis time | Genesis location (lat, lon) | EF rating |
|---|---|---|---|---|---|---|---|
| 1 | 11 Dec 2021 | Missouri | Webster | DSC | 0013 UTC | 37.345°, −92.75° | 1 |
| 2 | 11 Dec 2021 | Arkansas | Jonesboro | DSC | 0107 UTC | 35.79°, −90.55° | 4 |
| 3 | 11 Dec 2021 | Missouri | Saint Charles | QLCS | 0135 UTC | 38.6°, −90.91° | 3 |
| 4 | 11 Dec 2021 | Illinois | Cass | MUL | 0147 UTC | 39.9°, −90.24° | 2 |
| 5 | 11 Dec 2021 | Tennessee | Dyer | MUL | 0432 UTC | 36.12°, −89.26° | 3 |
| 6 | 11 Dec 2021 | Tennessee | Smith | QLCS | 1004 UTC | 36.24°, −85.95° | 0 |
| 7 | 11 Dec 2021 | Tennessee | Grundy | QLCS | 1339 UTC | 35.35°, −85.71° | 1 |
| 8 | 8 Jun 2022 | Nebraska | Nuckolls | QLCS | 0242 UTC | 40.15°, −97.89° | 0 |
| 9 | 8 Jun 2022 | Kansas | Johnson | QLCS | 0610 UTC | 38.96°, −94.79° | 1 |
| 10 | 8 Jun 2022 | Missouri | Jackson | QLCS | 0637 UTC | 39.12°, −94.31° | 2 |
| 11 | 19 Jun 2023 | Mississippi | Scott | DSC | 0104 UTC | 32.46°, −89.62° | 1 |
| 12 | 19 Jun 2023 | Mississippi | Rankin | DSC | 0205 UTC | 32.12°, −90.13 | 2 |
| 13 | 19 Jun 2023 | Mississippi | Rankin | DSC | 0226 UTC | 32.28°, −89.68° | 1 |
| 14 | 19 Jun 2023 | Mississippi | Jasper | DSC | 0430 UTC | 32.03°, −89.28° | 3 |

tornado damage intensities and durations from different storm modes: 11 December 2021 in the central and southeastern United States, 19 June 2023 across Mississippi, and 8 June 2022 across the central Great Plains. This application serves primarily as a validation case study of the ML methods explored in this study and demonstrates how operational meteorologists could use these ML models. In the future, a more representative validation dataset should be developed to fully understand how the techniques used within this study would generalize to the larger, more diverse, and more realistic sample of tornadic events experienced in an operational setting.

Recalling that, for this study, the radar-based predictors require manual analysis, a sample of seven tornadoes (one EF0, two EF1s, one EF2, two EF3s, and one EF4) from 11 December 2021, four tornadoes (two EF1s, one EF2, and one EF3) from 19 June 2023, and three tornadoes (one EF0, one EF1, and one EF2) from 8 June 2022 was selected for consideration (Table 10). The selection criteria outlined in ST20 and ST23 were used. As ST23 demonstrated the ability to apply the pretornadic relationships to not only the first tornado produced by a storm but also subsequent tornadoes, a subsequent EF4 tornado from 11 December was included, which was preceded by two very brief (duration $< 5$ min) nonsignificant tornadoes. Thus, the mesocyclone characteristics preceding the initially nonsignificant tornadoes were used to anticipate the subsequent EF4 tornado, which had the highest damage rating produced by that storm (see ST23 for discussion related to subsequent tornadoes). The feature importance analyses completed earlier in this study were the basis for the predictors used (Table 11). The LGR, RF, and GB models (the models that tended to have the best performance) were each run using all the top-selected predictors, the radar-only predictors, and all the top-selected environmental-only predictors. To use all data available, models were trained on the entire 300-case dataset, tuned using a grid search of hyperparameters, and evaluated on the 14 selected cases using the best model from the grid search.

Table 12 shows the details of the models, as well as the performance on the 14 cases. When considering overall performance, models perform well, with misclassifications coming from EF2–3 cases. In the radar-only models, each model misclassified the same EF2 and EF3 cases as nonsignificant (Table S7a). There is decreased performance when using only the top environmental predictors, with mainly nonsignificant cases being misclassified (Table S7b).

To better understand the classifications, Figs. 10 and 11 show the influence of each predictor on the probability of significanttornadoes for each 11 December 2021 case using SHAP values, separated by nonsignificant and significant cases. Figures 12 and 13 show the same for the environment-only models. This analysis uses the RF model utilizing the top chosen predictors. The plots show how each feature influences the predicted probability of significant tornadoes from the base climatological probability $E[f(x)]$ from the training dataset to a final probability $f(x)$. The influence of the predictors in this case study is consistent with the relationships shown earlier, with larger (smaller) values of pretornadic mesocyclone width and intensity and composite and kinematic environmental parameters increasing (decreasing) the probability of significant tornadoes. These cases provide examples of where the top five predictors contribute to an increased probability of increased tornado damage intensity (e.g., Fig. 11a)

TABLE 11. Predictors chosen for the case study.

| Predictors chosen |
|---|
| Pretornadic mesocyclone width |
| Pretornadic mesocyclone intensity |
| QLCS storm mode |
| DSC storm mode |
| MUL storm mode |
| Distance from radar |
| Fixed-layer supercell parameter |
| Fixed-layer significant tornado parameter |
| Tornadic tilting and stretching |
| Effective bulk shear |
| Effective storm-relative helicity |
| 0–1-km bulk shear |
| 3–6-km relative humidity |
| Mixed-layer CAPE |

TABLE 12. Training hyperparameters and accuracy when using the 300-case dataset and case study classification accuracy and misclassified cases when all top predictors are used.

| Model | Hyperparameters | Training accuracy | Accuracy | Misclassified cases |
|---|---|---|---|---|
| LGR | "*C*": 10<br>"penalty": "l2"<br>"solver": "newton" | 0.877 | 13/14 | Case 12 (EF2) |
| RF | "max_depth": 4<br>"min_samples_leaf": 1<br>"min_samples_split": 2<br>"n_estimators": 8 | 0.903 | 12/14 | Case 3 (EF3), case 10 (EF2) |
| GB | "learning_rate": 0.5<br>"max_depth": 1<br>"min_samples_leaf": 0.1<br>"min_samples_split": 0.1<br>"n_estimators": 4 | 0.890 | 12/14 | Case 3 (EF3), case 10 (EF2) |

and when the influence of the environmental parameters can cancel (e.g., Figs. 10c and 12b).

When more closely considering the incorrectly classified case #3 (EF3 case), the mesocyclone width was consistently around 2.5–2.6 km, and the mesocyclone intensity varied from 30 to 35 m $s^{-1}$ leading up to tornadogenesis. The narrower pretornadic mesocyclone width and weaker intensity had a negative influence (21% decrease in significant tornado probability), which was stronger than the positive influence from environmental predictors (TTS of 4.96, S01 of 22.1 m $s^{-1}$, and ESRH of 439.8 $m^2$ $s^{-2}$ resulting in 8% increase in significant tornado probability) leading to a 22% probability of significant tornado damage intensity (Fig. 11c). For the environment-only iteration, which produced a correct prediction, Fig. 13c more clearly reveals the positive influence of the environment, leading to a 60% chance of a significant tornado. More specifically, this is an example of how the transitional values identified by the SHAP analysis are not clear dividing lines, as the mean pretornadic mesocyclone width of 2.62 km was just above the threshold of 2.5 km and yet decreased the probability of significant tornadoes. This highlights a limiting factor when heavily weighted predictors are near their transitional thresholds. When viewing the correctly classified case 6 (EF0), the pretornadic mesocyclone width ranged from 1.8 to 2.0 km, which decreased the probability of significant tornado damage intensity (15% decrease in significant tornado probability). Outside of a 6% increase in significant tornado probability from S01 (22.5 m $s^{-1}$) and ESRH (227.8 $m^2$ $s^{-2}$), there is a negative influence from the environmental predictors, leading to the 15.5% probability of a significant tornado damage rating (Fig. 10b). Figure 12b reveals a cancellation of the influence of environmental predictors in the environment-only iteration, leading to a 35% chance of a significant tornado and a correct prediction.

These results and patterns are consistent across the radar-only (not shown) and environment-only predictions (Figs. 12 and 13). One notable result from the environment-only iteration is that the nonsignificant cases had predictor influence largely cancel out, or several predictors had values near transitional thresholds, resulting in smaller SHAP values and an essentially unchanged value from $E[f(X)]$. For the significant cases, all the top predictors contribute to an increased probability of significant tornadoes, leading to their correct classification.

The successful application of the ML models to unseen cases demonstrates the future potential utility of this ML application in an operational setting and encourages further validation from a more representative dataset, including outbreak and localized tornadic events across an array of storm environments. Additionally, implementing the ML models for tornado damage intensity prediction for ongoing thunderstorms would require extensive development and testing of automated predictor calculation. After the ingestion of predictors, the predictions from the models take seconds to compute. As described in ST23, environmental parameters would be the easiest to obtain in real time from observed soundings, operational model grids, or the NOAA SPC mesoanalysis (Bothwell et al. 2002). One would need to develop an automated method for extracting the spatial means of the environmental variables centered on tracked mesocyclones.Also, this method would need to quantify the radar-based predictors in real time using automated approaches, as the manual calculation of the radar variables would be overly demanding of forecasters during real-time warning operations. The original mesocyclone detection algorithm (MDA; Stumpf et al. 1998) computes mesocyclone diameter at each elevation angle (using the same width definition used in this study) and several intensity metrics, including differential velocity, which could be extracted in real time. The MDA also identifies and tracks individual mesocyclones in time and produces a nine-scan time series of all metrics. To allow for operational use, an automated method for extracting a moving time average of the width and intensity of identified mesocyclones would need to be developed to ingest real-time calculations into a model. Pretornadic mesocyclones are only sometimes resolvable with the current WSR-88D network (due to radar range, contaminated data, etc.), which is a limiting factor to the number of storms the model could be applied to.

While automated methods are needed for potential future operationalization, the use of data collected from automated methods could introduce noise to the predictors and potentially impact model performance. For example, when considering the environmental parameters, the use of a different model would introduce new model strengths and biases into the predictors. When considering the radar predictors, while the MDA definitions for mesocyclone width and intensity are the same as what is used in this study, the identification of

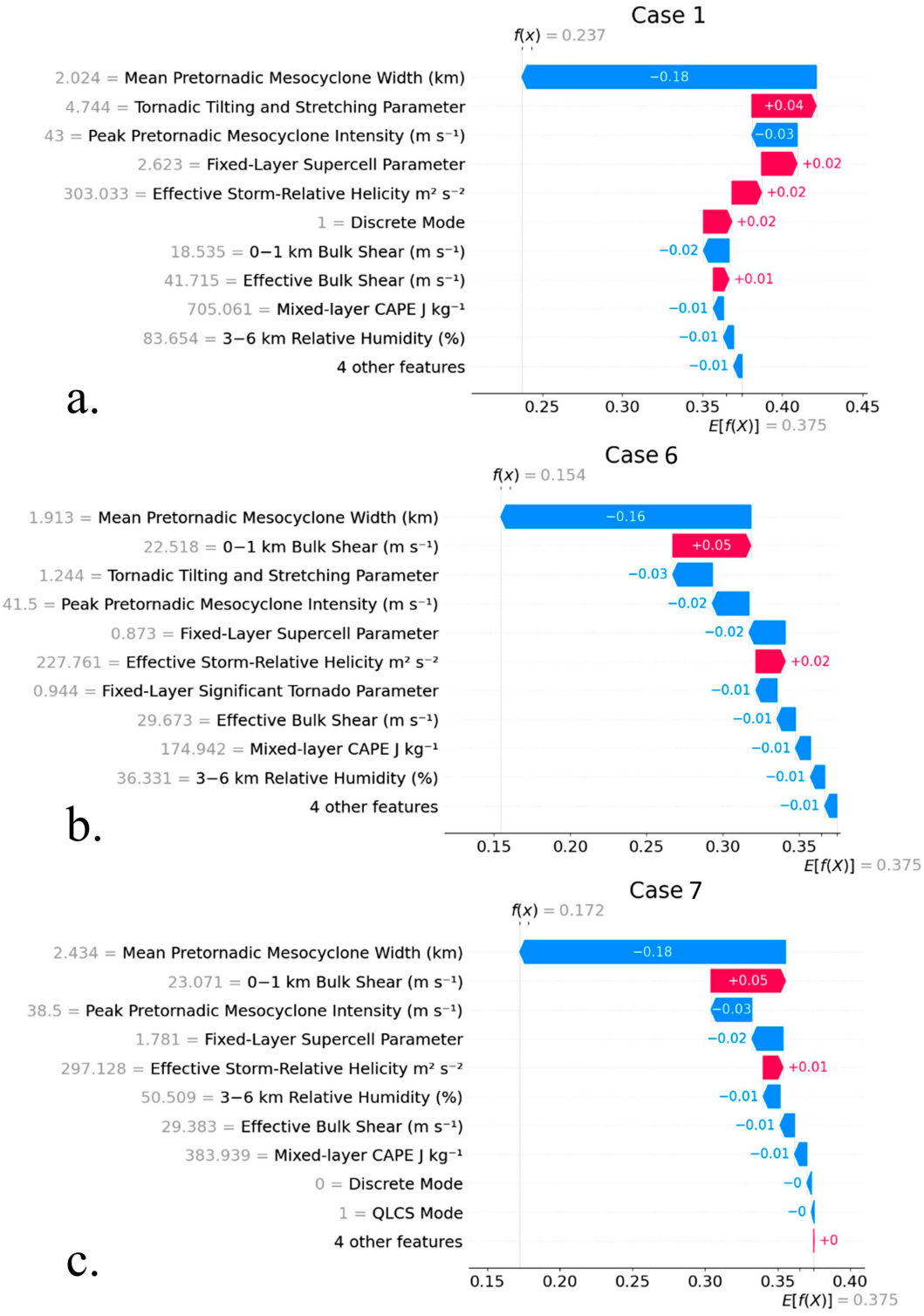


FIG. 10. SHAP waterfall plots of the nonsignificant cases (EF0 − 1) from the December case study highlighting the influence of the top 10 predictors on the probability of significant tornadoes when all selected predictors are used for (a) case 1, (b) case 6, and (c) case 7 (see Table 10). The values of each predictor are listed to the left of the predictor name. The $E[f(X)]$ is the base climatology of the probability of significant tornadoes, while $f(x)$ represents the predicted probability of significant tornadoes. Colored bars represent the impact of each predictor on the predicted probability of significant tornadoes, with features increasing the probability of significant tornadoes shown in red and features decreasing the probability shown in blue. Each predictor's SHAP value or individual influence on the final probability of significant tornadoes $f(x)$ is shown on or next to each bar. Note that the last bar is the four least impactful predictors combined.

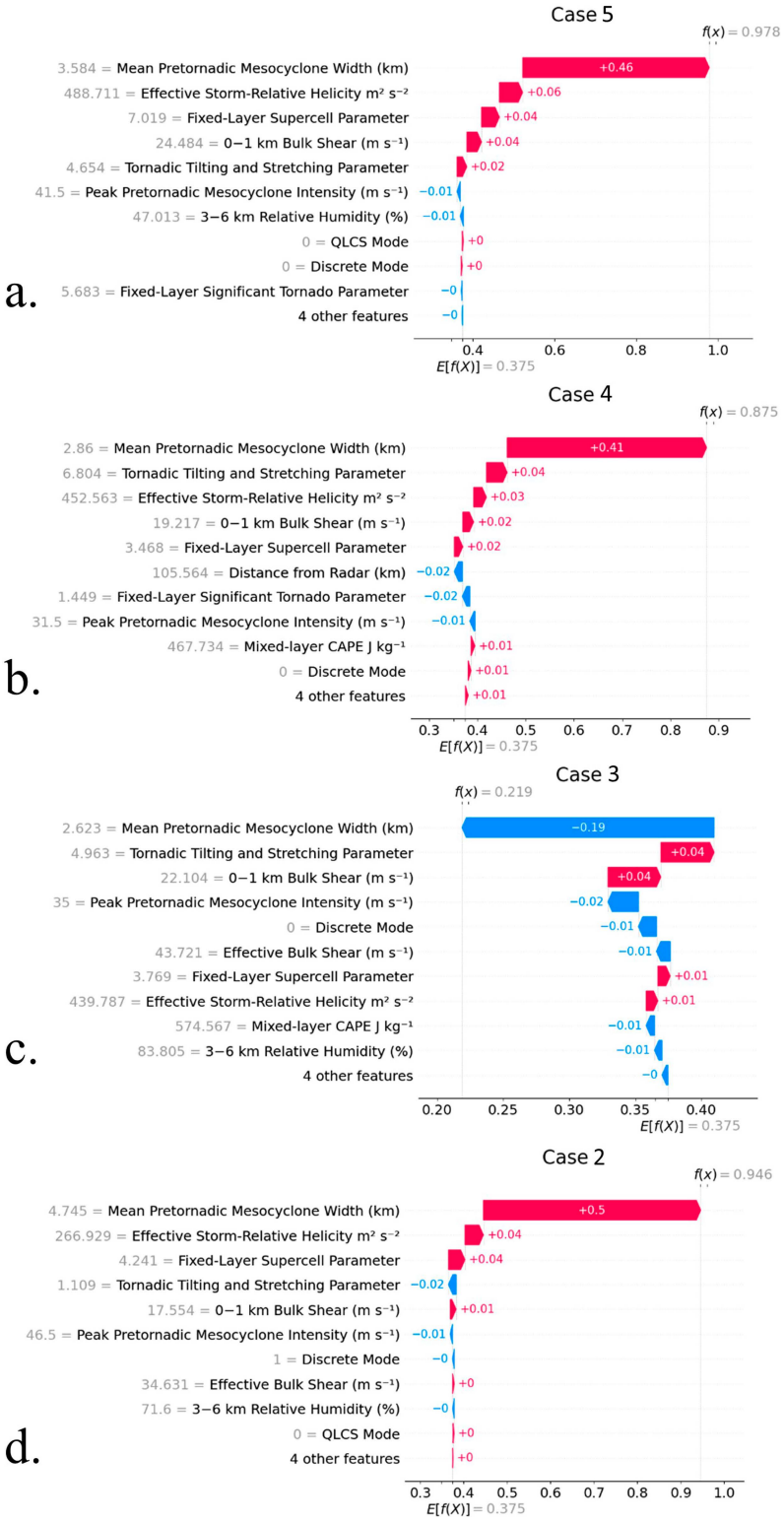


FIG. 11. As in Fig. 10, but for the significant (EF2–4) cases for (a) case 5, (b) case 4, (c) case 3, and (d) case 2.

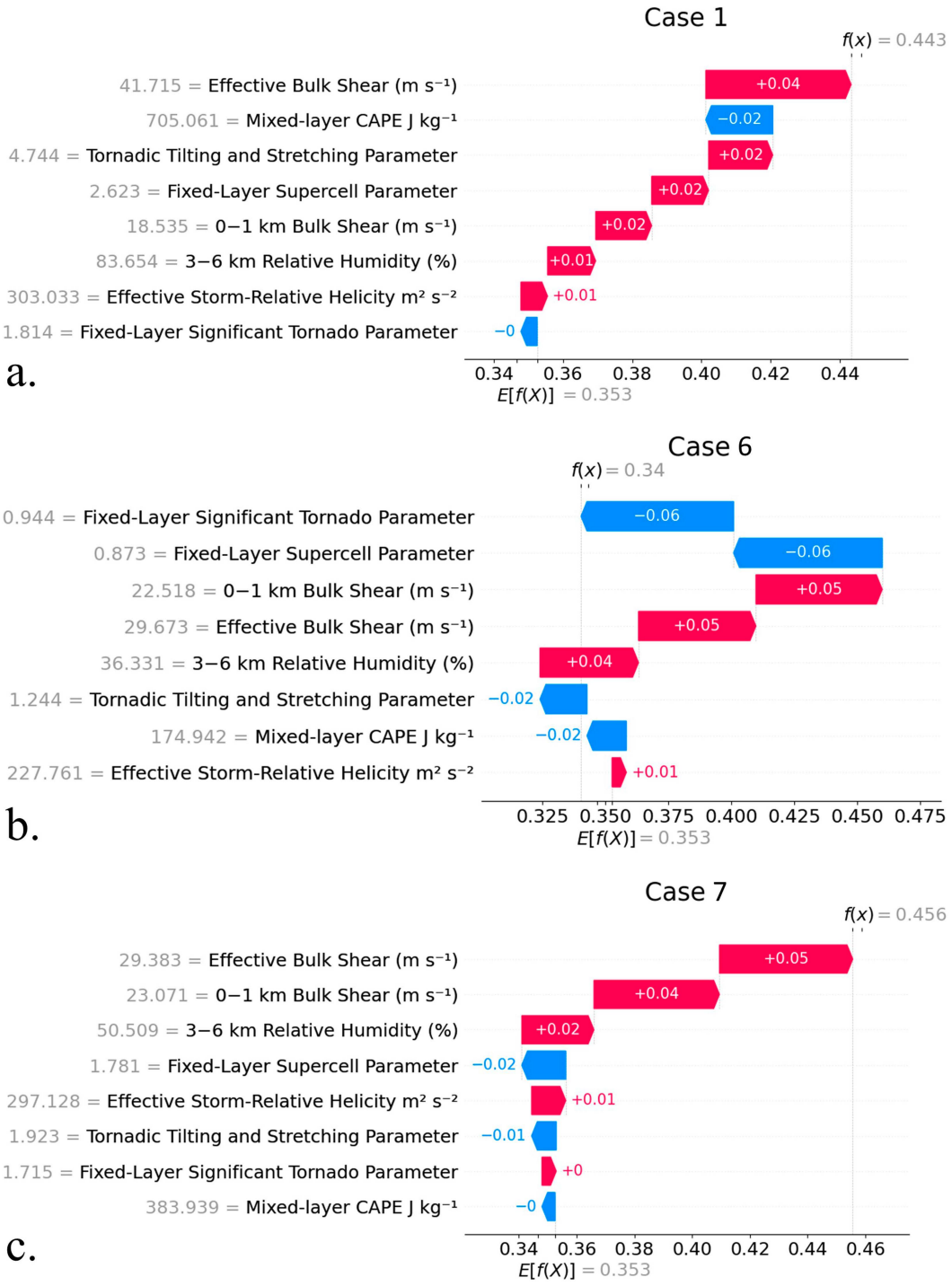


FIG. 12. As in Fig. 10, but for the environment-only classifiers.

rotation signatures and the peak inbound and outbound values within them from a manual analysis compared to an automated one could produce different results. This raises uncertainty around using mesocyclone predictors from the MDA, for example, with a model trained on data from manual analysis. An automated calculation of predictors could also impact feature importance and the transitional values of predictors identified from the SHAP analysis. For these reasons, it may be worthwhile in the future to develop a large dataset of the radar metrics used in this study using an automated method, such as the MDA. The dataset from automated calculations could then be used to test the models trained on manually calculated predictors or to fully

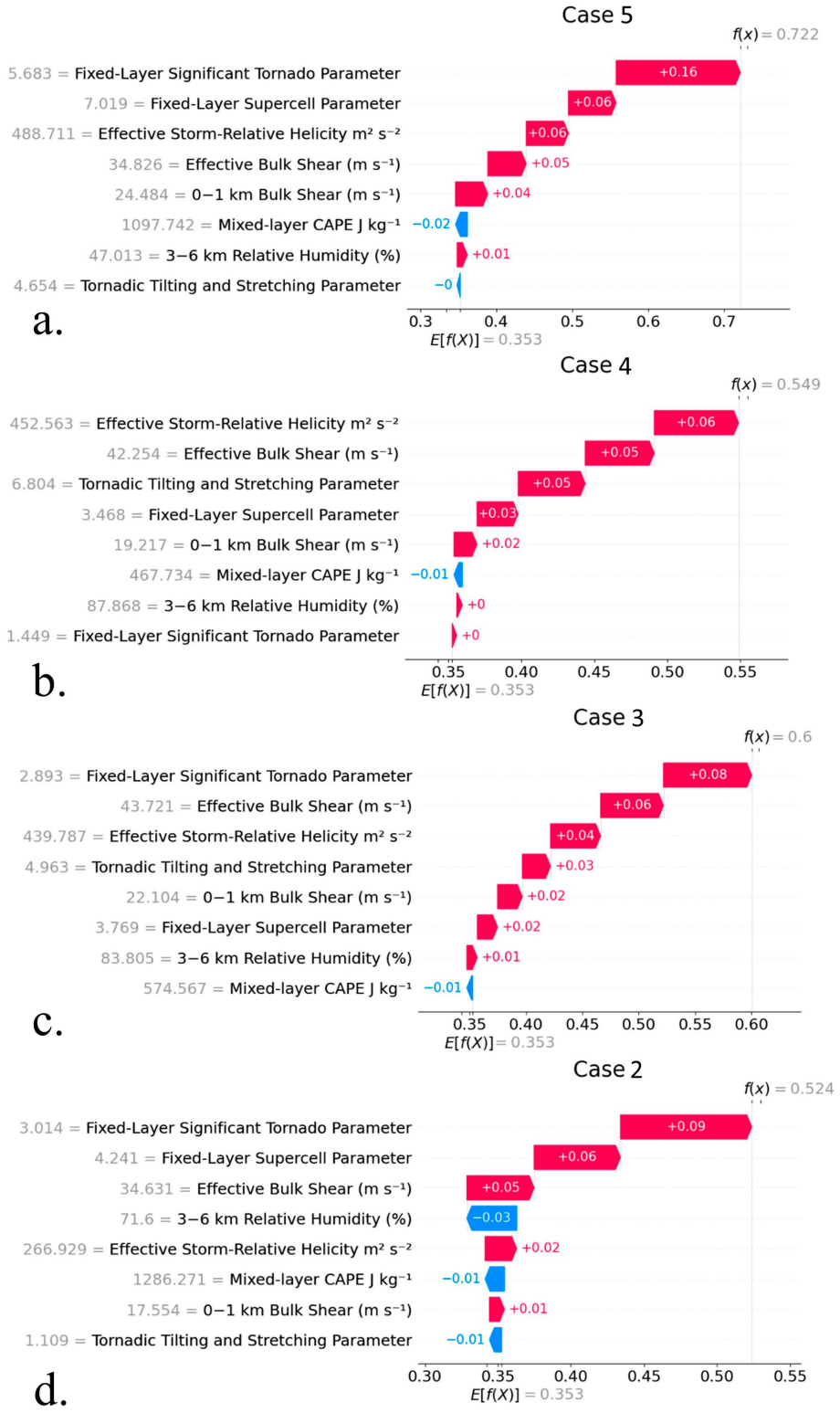


FIG. 13. As in Fig. 10. but the environment-only classifiers.

retrain and test models. This would allow for an understanding of the potential impact of automated predictor calculations on model performance.

As the ML applications applied in this study are intended to predict tornado damage intensity, conditional on tornadogenesis, the intensity prediction should be coupled to a tornadogenesis prediction. Recent attempts at utilizing ML for the prediction of tornadogenesis (e.g., Cintineo et al. 2020b; Lagerquist et al. 2020; Steinkruger et al. 2020) involve using radar-observed storm attributes, convection-allowing storm-surrogate diagnostics such as updraft helicity, and near-storm environmental parameters. Radar signatures such as the $Z_{DR}$ arc and column and $K_{DP}$ foot (Crowe et al. 2012; Homeyer et al. 2020; Van Den Broeke 2020) and mesocyclone characteristics such as width, depth, and intensity and their trends (Gibbs and Bowers 2019; Sandmæl et al. 2019) are examples of simple observational tools outside of ML applied to the anticipation of tornadogenesis. Importantly, an operational forecaster would likely not utilize a tornado damage intensity model like those outlined here if a particular environment is determined to not support tornadogenesis, via the analysis of observations and numerical-model grids, using previously established environment–tornadogenesis relationships (e.g., Thompson et al. 2007, 2012; Nowotarski and Jensen 2013; Sobash et al. 2016, 2019; Coniglio and Parker 2020).

It is essential to discuss the limitations of this study to contribute to the goal of responsible and reliable use of artificial intelligence in the environmental sciences (e.g., McGovern et al. 2022) and to highlight where to focus further development of ML applications for tornado damage intensity anticipation. One limitation is the time constraints of the pretornadic period, particularly for QLCS tornadoes but even for some discrete storms. For example, there may only be one pretornadic volume scan where mesocyclonic rotation is present to supply data to the ML model. Additionally, the restrictions imposed during case selection in ST23 also introduce bias and limit the current operational usefulness of the models. Currently, these models are trained on the first tornado produced by storms occurring within 100 km of radar. From 2011 to 2022, 35% of all tornadoes and 33% of EF2+ tornadoes occurred beyond 100 km of a radar site. Range corrections could be applied in the future to allow for more accurate radar characteristics beyond 100 km (e.g., Newman et al. 2013). While ST23 demonstrates how subsequent tornadoes could be included, they need to be incorporated in model training and testing to understand their impact on model performance. Also, the inclusion of only the strongest tornadoes in a 1 h and 80 km time and space range focuses the model predictions on the strongest tornado a particular storm and environment are capable of producing. It would be beneficial to remove this restriction in the future to include weaker tornadoes in significant tornado environments to explore the impact on model performance. Furthermore, there are only four radar-based predictors included. Additional limitations related to the presence and resolvability of the radar predictors used in this study, as well as the U.S. tornado report climatology and the EF scale, are discussed in ST23. One must keep these limitations in mind when interpreting, applying, and further developing this study.

## 5. Conclusions

This study examined the use of several classification ML algorithms in a binary classification framework to predict the potential for nonsignificant or significant tornadoes within an ongoing storm. Pretornadic radar characteristics and near-storm environmental parameters composed the predictors. The results demonstrated a skilled binary classification of potential tornado damage intensity, conditional on tornadogenesis, especially when models used both the radar and near-storm environmental predictors. Using the radar or environmental predictors alone still resulted in a skilled prediction, although the environment-only models show a noticeable drop in performance. LGR, RF, and GB were the most-skilled classifiers as measured by several cross-validated binary-classification metrics, ROC curves, and performance diagrams. Learning curves showed training adequacy, and calibration curves demonstrated that the probabilities output by the classifiers tended to be reliable, especially for the all-predictor and environment-only models. Permutation tests revealed the significance of the results and that the skill of the models comes from real dependency between the predictors and tornado damage intensity. RFE and SHAP analyses explored feature importance and the model's decision-making process. They revealed a more physical understanding of the model performance and relationships between the predictors and tornado damage intensity, including establishing transitional thresholds of predictors in the models. The pretornadic radar predictors of mesocyclone width and intensity were the most important, followed by vertical wind shear and composite environmental parameters. Additionally, the nonsignificant/significant "boundary" cases of EF1s and EF2s contributed the most to misclassifications, while EF0s and EF3–EF5s tended to be classified correctly. Finally, case studies showed the potential utility of the ML models in operational forecasting after further development, including the need for and benefit of automated calculations of predictors for use in real time.

Importantly, the results presented here show that there is no clear dividing line when determining potential tornado damage intensity. The transitional values of predictors identified from the SHAP analysis should not be used as hard cutoffs when determining the real-time potential for significant tornadoes as there are exceptions to these relationships, and the realization of the maximum damage intensity of a tornado is more complex than the specific values of individual predictors. However, these results show the benefit of utilizing ML to combine data sources to aid in anticipating tornado damage intensity.

Future work should involve expanding the training dataset and further developing a validation dataset to grow confidence in model performance and how well the model would generalize to a more extensive and diverse sample of tornadic events. This should include removing the more restrictive case-selection criteria (e.g., 100-km radar range, first tornado produced, and strongest tornadoes in certain time and space range). A larger dataset would also allow for the subset of cases that have similar damage indicators or well-constrained damage surveys to understand the potential impact of biases from the damage-based EF scale on model performance.

Additionally, other predictors should be explored, such as from satellite data (e.g., overshooting-top area; Marion et al. 2019), from other radar metrics (e.g., $Z_{DR}$ column area, mesocyclone depth, mesocyclone width, and intensity over more than the lowest three elevationangles; Gibbs and Bowers 2019; French and Kingfield 2021), or from other environmental data sources (e.g., observed soundings) to see if they contribute additional skill to the classifiers' predictions or fill holes of the currently used predictors. Exploring other pretornadic radar metrics would allow for more confidence in the conclusion that pretornadic mesocyclone width is the most important radar metric for anticipating tornado damage intensity. Exploring and developing the best automated methods for extracting the running means of the radar metrics and the spatial means of the environmental parameters for ingestion into models would contribute to the continueddevelopment of this ML application for future operational usefulness. Additionally, the training and/or testing of the models from this study using a radar dataset created from automated methods would also reveal any potential impacts of automated methods on model performance, feature importance, and transitional values of predictors.

Overall, the results demonstrate a skilled binary prediction of tornado damage intensity, conditioned upon tornadogenesis, and the potential for these ML applications to develop into a helpful resource in an operational setting to better protect life and property by providing information about potential tornado damage intensity before tornadoes form. Additionally, this study adds to the growing number of studies demonstrating the skill and utility of ML in weather and climate studies and highlights the need for ML applications for tornadoes to focus on predictions of damage intensity in addition to genesis and detection.

*Acknowledgments.* We thank Roger Edwards and two anonymous peer reviewers for their helpful suggestions and feedback that thoroughly improved and refined the manuscript. This work was supported in part by NOAA Award NA17OAR4590195.

*Data availability statement.* The pretornadic radar dataset, as well as the near-storm environmental dataset, is available upon request and will be added to the University of Illinois Data Bank. GitHub provides a sample of the ML pipeline used in this study at https://github.com/michaelsessa/Machine_Learning_tor_int.

## REFERENCES

Adrianto, I., T. B. Trafalis, and V. Lakshmanan, 2009: Support vector machines for spatiotemporal tornado prediction. *Int. J. Gen. Syst.*, **38**, 759–776, https://doi.org/10.1080/03081070601068629.

Ahijevych, D., J. O. Pinto, J. K. Williams, and M. Steiner, 2016: Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. *Wea. Forecasting*, **31**, 581–599, https://doi.org/10.1175/WAF-D-15-0113.1.

Anderson-Frey, A. K., and H. Brooks, 2019: Tornado fatalities: An environmental perspective. *Wea. Forecasting*, **34**, 1999–2015, https://doi.org/10.1175/WAF-D-19-0119.1.

Ashley, W. S., 2007: Spatial and temporal analysis of tornado fatalities in the United States: 1880–2005. *Wea. Forecasting*, **22**, 1214–1228, https://doi.org/10.1175/2007WAF2007004.1.

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

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

Bothwell, P. D., J. A. Hart, and R. L. Thompson, 2002: An integrated three-dimensional objective analysis scheme in use at the Storm Prediction Center. Preprints, *21st Conf. on Severe Local Storms*, San Antonio, TX, Amer. Meteor. Soc., JP3.1, https://ams.confex.com/ams/SLS_WAF_NWP/techprogram/paper_47482.htm.

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

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

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

Casteel, M. A., 2016: Communicating increased risk: An empirical investigation of the National Weather Service's impact-based warnings. *Wea. Climate Soc.*, **8**, 219–232, https://doi.org/10.1175/WCAS-D-15-0044.1.

Chase, R. J., D. R. Harrison, A. Burke, G. M. Lackmann, and A. McGovern, 2022: A Machine learning tutorial for operational meteorology. Part I: Traditional machine learning. *Wea. Forecasting*, **37**, 1509–1529, https://doi.org/10.1175/WAF-D-22-0070.1.

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, 785–794, https://doi.org/10.1145/2939672.2939785.

Chicco, D., and G. Jurman, 2020: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, **21**, 6, https://doi.org/10.1186/s12864-019-6413-7.

Cintineo, J. L., 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.

——, ——, ——, A. Wimmers, J. Brunner, and W. Bellon, 2020a: A deep-learning model for automated detection of intense midlatitude convection using geostationary satellite images. *Wea. Forecasting*, **35**, 2567–2588, https://doi.org/10.1175/WAF-D-20-0028.1.

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

Coffer, B., M. Kubacki, Y. Wen, T. Zhang, C. A. Barajas, and M. K. Gobbert, 2021: Machine learning with feature importance

analysis for tornado prediction from environmental sounding data. *Proc. Appl. Math. Mech.*, **20**, e202000112, https://doi.org/10.1002/pamm.202000112.

Cohen, J., 1960: A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, **20**, 213–220, https://doi.org/10.1177/001316446002000104.

Coniglio, M. C., and M. D. Parker, 2020: Insights into supercells and their environments from three decades of targeted radiosonde observations. *Mon. Wea. Rev.*, **148**, 4893–4915, https://doi.org/10.1175/MWR-D-20-0105.1.

Crowe, C. C., C. J. Schultz, M. Kumjian, L. D. Carey, and W. A. Petersen, 2012: Use of dual-polarization signatures in diagnosing tornadic potential. *Electron. J. Oper. Meteor.*, **13**, 57–78.

Czernecki, B., M. Taszarek, M. Marosz, M. Półrolniczak, L. Kolendowicz, A. Wyszogrodzki, and J. Szturc, 2019: Application of machine learning to large hail prediction - The importance of radar reflectivity, lightning occurrence and convective parameters derived from ERA5. *Atmos. Res.*, **227**, 249–262, https://doi.org/10.1016/j.atmosres.2019.05.010.

Doswell, C. A., III, and D. W. Burgess, 1988: On some issues of United States tornado climatology. *Mon. Wea. Rev.*, **116**, 495–501, https://doi.org/10.1175/1520-0493(1988)116<0495:OSIOUS>2.0.CO;2.

——, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576–585, https://doi.org/10.1175/1520-0434(1990)005<0576:OSMOSI>2.0.CO;2.

Drosdowsky, W., and L. E. Chambers, 2001: Near-global sea surface temperature anomalies as predictors of Australian seasonal rainfall. *J. Climate*, **14**, 1677–1687, https://doi.org/10.1175/1520-0442(2001)014<1677:NACNGS>2.0.CO;2.

Edwards, R., J. G. LaDue, J. T. Ferree, K. Scharfenberg, C. Maier, and W. L. Coulbourne, 2013: Tornado intensity estimation: Past, present, and future. *Bull. Amer. Meteor. Soc.*, **94**, 641–653, https://doi.org/10.1175/BAMS-D-11-00006.1.

Esterheld, J. M., and D. J. Giuliano, 2021: Discriminating between tornadic and non-tornadic supercells: A new hodograph technique. *Electron. J. Severe Storms Meteor.*, **3** (2), https://doi.org/10.55599/ejssm.v3i2.15.

Flora, M. L., C. K. Potvin, P. S. Skinner, S. Handler, and A. McGovern, 2021: Using machine learning to generate storm-scale probabilistic guidance of severe weather hazards in the warn-on-forecast system. *Mon. Wea. Rev.*, **149**, 1535–1557, https://doi.org/10.1175/MWR-D-20-0194.1.

French, M. M., and D. M. Kingfield, 2021: Tornado formation and intensity prediction using polarimetric radar estimates of updraft area. *Wea. Forecasting*, **36**, 2211–2231, https://doi.org/10.1175/WAF-D-21-0087.1.

Friedman, J. H., 2001: Greedy function approximation: A gradient boosting machine. *Ann. Statist.*, **29**, 1189–1232, https://doi.org/10.1214/aos/1013203451.

Gagne, D. J., II, A. McGovern, J. B. Basara, and R. A. Brown, 2012: Tornadic supercell environments analyzed using surface and reanalysis data: A spatiotemporal relational data-mining approach. *J. Appl. Meteor. Climatol.*, **51**, 2203–2217, https://doi.org/10.1175/JAMC-D-11-060.1.

——, 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.

Gensini, V. A., C. Converse, W. S. Ashley, and M. Taszarek, 2021: Machine learning classification of significant tornadoes and hail in the United States using ERA5 proximity soundings. *Wea. Forecasting*, **36**, 2143–2160, https://doi.org/10.1175/WAF-D-21-0056.1.

Gibbs, J. G., and B. R. Bowers, 2019: Techniques and thresholds of significance for using WSR-88D velocity data to anticipate significant tornadoes. *J. Oper. Meteor.*, **7**, 117–137, https://doi.org/10.15191/nwajom.2019.0709.

Gron, A., 2017a: The machine learning landscape. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* 1st ed. O'Reilly Media, Inc., 3–31.

——, 2017b: Support vector machines. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems.* 1st ed. O'Reilly Media, Inc., 147–167.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik, 2002: Gene selection for cancer classification using support vector machines. *Mach. Learn.*, **46**, 389–422, https://doi.org/10.1023/A:1012487302797.

Haddad, S., R. E. Killick, M. D. Palmer, M. J. Webb, R. Prudden, F. Capponi, and S. V. Adams, 2022: Improved infilling of missing metadata from expendable bathythermographs (XBTs) using multiple machine learning methods. *J. Atmos. Oceanic Technol.*, **39**, 1367–1385, https://doi.org/10.1175/JTECH-D-21-0117.1.

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.

Homeyer, C. R., T. N. Sandmæl, C. K. Potvin, and A. M. Murphy, 2020: Distinguishing characteristics of tornadic and nontornadic supercell storms from composite mean analyses of radar observations. *Mon. Wea. Rev.*, **148**, 5015–5040, https://doi.org/10.1175/MWR-D-20-0136.1.

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

Jergensen, G. E., A. McGovern, R. Lagerquist, and T. Smith, 2019: Classifying convective storms using machine learning. *Wea. Forecasting*, **35**, 537–559, https://doi.org/10.1175/WAF-D-19-0170.1.

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.

Krstajic, D., L. J. Buturovic, D. E. Leahy, and S. Thomas, 2014: Cross-validation pitfalls when selecting and assessing regression and classification models. *J. Cheminform.*, **6**, 10, https://doi.org/10.1186/1758-2946-6-10.

Kleinbaum, D. G., K. Dietz, M. Gail, M. Klein, and M. Klein, 2002: *Logistic Regression.* Springer, 513 pp.

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.

——, ——, C. R. Homeyer, D. J. Gagne II, 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.

Leinonen, J., U. Hamann, U. Germann, and J. R. Mecikalski, 2022: Nowcasting thunderstorm hazards using machine learning: The impact of data sources on performance. *Nat. Hazards*

*Earth Syst. Sci.*, **22**, 577–597, https://doi.org/10.5194/nhess-22-577-2022.

Loken, E. D., A. J. Clark, 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.

——, ——, and A. McGovern, 2022: Comparing and interpreting differently designed random forests for next-day severe weather hazard prediction. *Wea. Forecasting*, **37**, 871–899, https://doi.org/10.1175/WAF-D-21-0138.1.

Lundberg, S. M., and S. I. Lee, 2017: A unified approach to interpreting model predictions. arXiv, 1705.07874v2, https://doi.org/10.48550/arXiv.1705.07874.

——, and Coauthors, 2020: From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.*, **2**, 56–67, https://doi.org/10.1038/s42256-019-0138-9.

Manzato, A., 2013: Hail in northeast Italy: A neural network ensemble forecast using sounding-derived indices. *Wea. Forecasting*, **28**, 3–28, https://doi.org/10.1175/WAF-D-12-00034.1.

Marion, G. R., R. J. Trapp, and S. W. Nesbitt, 2019: Using overshooting top area to discriminate potential for large, intense tornadoes. *Geophys. Res. Lett.*, **46**, 12520–12526, https://doi.org/10.1029/2019GL084099.

Martens, B., W. Waegeman, W. A. Dorigo, N. E. C. Verhoest, and D. G. Miralles, 2018: Terrestrial evaporation response to modes of climate variability. *npj Climate Atmos. Sci.*, **1**, 43, https://doi.org/10.1038/s41612-018-0053-5.

Marzban, C., and G. J. Stumpf, 1996: A neural network for tornado prediction based on Doppler radar-derived attributes. *J. Appl. Meteor.*, **35**, 617–626, https://doi.org/10.1175/1520-0450(1996)035<0617:ANNFTP>2.0.CO;2.

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

Matthews, B. W., 1975: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.*, **405**, 442–451, https://doi.org/10.1016/0005-2795(75)90109-9.

May, R. M., S. C. Arms, P. Marsh, E. Bruning, J. R. Leeman, K. Goebbert, J. E. Thielen, and Z. Bruick, 2022: MetPy: A Python package for meteorological data. Version 1.3.1, Unidata, https://doi.org/10.5065/D6WW7G29.

McGovern, A., D. J. Gagne II, J. K. Williams, R. A. Brown, and J. B. Basara, 2014: Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning. *Mach. Learn.*, **95**, 27–50, https://doi.org/10.1007/s10994-013-5343-x.

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

——, I. Ebert-Uphoff, D. Gagne, and A. Bostrom, 2022: Why we need to focus on developing ethical, responsible, and trustworthy artificial intelligence approaches for environmental science. *Environ. Data Sci.*, **1**, E6, https://doi.org/10.1017/eds.2022.5.

McGuire, M. P., and T. W. Moore, 2022: Prediction of tornado days in the United States with deep convolutional neural networks. *Comput. Geosci.*, **159**, 104990, https://doi.org/10.1016/j.cageo.2021.104990.

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.

Miller, D. E., Z. Wang, R. J. Trapp, and D. S. Harnos, 2020: Hybrid prediction of weekly tornado activity out to week 3: Utilizing weather regimes. *Geophys. Res. Lett.*, **47**, e2020GL087253, https://doi.org/10.1029/2020GL087253.

Mitchell, T., 1997a: Decision tree learning. *Machine Learning*, McGraw-Hill, 52–78.

——, 1997b: Bayesian learning. *Machine Learning*, McGraw-Hill, 154–199.

Mostajabi, A., D. L. Finney, M. Rubinstein, and F. Rachidi, 2019: Nowcasting lightning occurrence from commonly available meteorological parameters using machine learning techniques. *npj Climate Atmos. Sci.*, **2**, 41, https://doi.org/10.1038/s41612-019-0098-0.

Newman, J. F., V. Lakshmanan, P. L. Heinselman, M. B. Richman, and T. M. Smith, 2013: Range-correcting azimuthal shear in Doppler radar data. *Wea. Forecasting*, **28**, 194–211, https://doi.org/10.1175/WAF-D-11-00154.1.

Nowotarski, C. J., and A. A. Jensen, 2013: Classifying proximity soundings with self-organizing maps toward improving supercell and tornado forecasting. *Wea. Forecasting*, **28**, 783–801, https://doi.org/10.1175/WAF-D-12-00125.1.

Ojala, M., and G. C. Garriga, 2010: Permutation tests for studying classifier performance. *J. Mach. Learn Res.*, **11**, 1833–1863.

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

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

Saito, T., and M. Rehmsmeier, 2015: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, **10**, e0118432, https://doi.org/10.1371/journal.pone.0118432.

Sandmæl, T. N., C. R. Homeyer, K. M. Bedka, J. M. Apke, J. R. Mecikalski, and K. Khlopenkov, 2019: Evaluating the ability of remote sensing observations to identify significantly severe and potentially tornadic storms. *J. Appl. Meteor. Climatol.*, **58**, 2569–2590, https://doi.org/10.1175/jamc-d-18-0241.1.

——, and Coauthors, 2023: The tornado probability algorithm: A probabilistic machine learning tornadic circulation detection algorithm. *Wea. Forecasting*, **38**, 445–466, https://doi.org/10.1175/WAF-D-22-0123.1.

Schlef, K. E., H. Moradkhani, and U. Lall, 2019: Atmospheric circulation patterns associated with extreme United States floods identified via machine learning. *Sci. Rep.*, **9**, 7171, https://doi.org/10.1038/s41598-019-43496-w.

Sessa, M. F., and R. J. Trapp, 2020: Observed relationship between tornado intensity and pretornadic mesocyclone characteristics. *Wea. Forecasting*, **35**, 1243–1261, https://doi.org/10.1175/WAF-D-19-0099.1.

——, and ——, 2023: Environmental and radar-derived predictors of tornado intensity within ongoing convective storms. *J. Oper. Meteor.*, **11**, 49–71, https://doi.org/10.15191/nwajom.2023.1105.

Shapley, L. S., 2016: A value for n-person games. *Contributions to the Theory of Games (AM-28)*, Vol. II, L. S. Shapley, Ed., Princeton University Press, 307–318.

Shield, S. A., and A. L. Houston, 2022: Diagnosing supercell environments: A machine learning approach. *Wea. Forecasting*, **37**, 771–785, https://doi.org/10.1175/WAF-D-21-0098.1.

Simmons, K., and D. Sutter, 2011: *Economic and Societal Impacts of Tornadoes.* Amer. Meteor. Soc., 282 pp.

Smith, B. T., R. L. Thompson, D. A. Speheger, A. R. Dean, C. D. Karstens, and A. K. Anderson-Frey, 2020a: WSR-88D tornado intensity estimates. Part I: Real-time probabilities of peak tornado wind speeds. *Wea. Forecasting*, **35**, 2479–2492, https://doi.org/10.1175/WAF-D-20-0010.1.

——, ——, ——, ——, ——, and ——, 2020b: WSR-88D tornado intensity estimates. Part II: Real-time applications to tornado warning time scales. *Wea. Forecasting*, **35**, 2493–2506, https://doi.org/10.1175/WAF-D-20-0011.1.

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

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

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.

Štrumbelj, E., and I. Kononenko, 2014: Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.*, **41**, 647–665, https://doi.org/10.1007/s10115-013-0679-x.

Stumpf, G. J., A. Witt, E. D. Mitchell, P. L. Spencer, J. T. Johnson, M. D. Eilts, K. W. Thomas, and D. W. Burgess, 1998: The national severe storms laboratory mesocyclone detection algorithm for the WSR-88D. *Wea. Forecasting*, **13**, 304–326, https://doi.org/10.1175/1520-0434(1998)013<0304:TNSSLM>2.0.CO;2.

Thompson, R. L., R. Edwards, and J. A. Hart, 2002: Evaluation and interpretation of the supercell composite and significant tornado parameters at the Storm Prediction Center. Preprints, *21st Conf. on Severe Local Storms*, San Antonio, TX, Amer. Meteor. Soc., J3.2, https://ams.confex.com/ams/SLS_WAF_NWP/techprogram/paper_46942.htm.

——, ——, ——, K. L. Elmore, and P. M. Markowski, 2003: Close proximity soundings within supercell environments obtained from the rapid update cycle. *Wea. Forecasting*, **18**, 1243–1261, https://doi.org/10.1175/1520-0434(2003)018<1243:CPSWSE>2.0.CO;2.

——, C. M. Mead, and R. Edwards, 2007: Effective storm-relative helicity and bulk shear in supercell thunderstorm environments. *Wea. Forecasting*, **22**, 102–115, https://doi.org/10.1175/WAF969.1.

——, B. T. Smith, J. S. Grams, A. R. Dean, and C. Broyles, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part II: Supercell and QLCS tornado environments. *Wea. Forecasting*, **27**, 1136–1154, https://doi.org/10.1175/WAF-D-11-00116.1.

——, and Coauthors, 2017: Tornado damage rating probabilities derived from WSR-88D data. *Wea. Forecasting*, **32**, 1509–1528, https://doi.org/10.1175/WAF-D-17-0004.1.

Togstad, W. E., J. M. Davies, S. J. Corfidi, D. R. Bright, and A. R. Dean, 2011: Conditional probability estimation for significant tornadoes based on Rapid Update Cycle (RUC) profiles. *Wea. Forecasting*, **26**, 729–743, https://doi.org/10.1175/2011WAF2222440.1.

Trapp, R. J., G. R. Marion, and S. W. Nesbitt, 2017: The regulation of tornado intensity by updraft width. *J. Atmos. Sci.*, **74**, 4199–4211, https://doi.org/10.1175/JAS-D-16-0331.1.

Van Den Broeke, M. S., 2020: A Preliminary polarimetric radar comparison of pretornadic and nontornadic supercell storms. *Mon. Wea. Rev*, **148**, 1567–1584, https://doi.org/10.1175/MWR-D-19-0296.1.

Warning Decision Training Division, 2021: Warning content: Impact-based warnings. Accessed 19 May 2021, https://training.weather.gov/wdtd/courses/rac/warnings/warn-content/presentation_html5.html.

Wilks, D. S., 2019: *Statistical Methods in the Atmospheric Sciences.* 4th ed. Elsevier, 676 pp.

Wind Science and Engineering Center, 2006: A recommendation for an enhanced Fujita scale (EF-scale), revision 2. Texas Tech University Tech. Rep., 95 pp., https://www.depts.ttu.edu/nwi/pubs/fscale/efscale.pdf.

Zhou, K., Y. Zheng, W. Dong, and T. Wang, 2020: A deep learning network for cloud-to-ground lightning nowcasting with multisource data. *J. Atmos. Ocean. Technol.*, **37**, 927–942, https://doi.org/10.1175/JTECH-D-19-0146.1.