Machine Learning Classification of Significant Tornadoes and Hail in the United States Using ERA5 Proximity Soundings

VITTORIO A. GENSINI,[a] CODY CONVERSE,[a] WALKER S. ASHLEY,[a] AND MATEUSZ TASZAREK[b]

[a] *Department of Geographic and Atmospheric Sciences, Northern Illinois University, DeKalb, Illinois*
[b] *Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma*

(Manuscript received 8 April 2021, in final form 23 September 2021)

ABSTRACT: Previous studies have identified environmental characteristics that skillfully discriminate between severe and significant-severe weather events, but they have largely been limited by sample size and/or population of predictor variables. Given the heightened societal impacts of significant-severe weather, this topic was revisited using over 150 000 ERA5 reanalysis-derived vertical profiles extracted at the grid point nearest—and just prior to—tornado and hail reports during the period 1996–2019. Profiles were quality controlled and used to calculate 84 variables. Several machine learning classification algorithms were trained, tested, and cross validated on these data to assess skill in predicting severe or significant-severe reports for tornadoes and hail. Random forest classification outperformed all tested methods as measured by cross-validated critical success index scores and area under the receiver operating characteristic curve values. In addition, random forest classification was found to be more reliable than other methods and exhibited negligible frequency bias. The top three most important random forest classification variables for tornadoes were wind speed at 500 hPa, wind speed at 850 hPa, and 0–500-m storm-relative helicity. For hail, storm-relative helicity in the 3–6 km and −10° to −30°C layers, along with 0–6-km bulk wind shear, were found to be most important. A game theoretic approach was used to help explain the output of the random forest classifiers and establish critical feature thresholds for operational nowcasting and forecasting. A use case of spatial applicability of the random forest model is also presented, demonstrating the potential utility for operational forecasting. Overall, this research supports a growing number of weather and climate studies finding admirable skill in random forest classification applications.

SIGNIFICANCE STATEMENT: A majority of losses due to tornadoes and hail are attributable to significant events [i.e., (E)F2+ tornadoes or ≥50-mm hail]. The decision of whether or not to issue a forecast for significant severe weather can be reduced to a binary classification problem, optimal for machine learning methodologies. Random forest classification algorithms are shown to be the most skillful for this application, and their continued implementation in operational nowcasting and forecasting may aid in better anticipation of significant tornado and hail events.

KEYWORDS: Tornadoes; Hail; Machine learning

1. Introduction

From 2011 to 2020, the United States tallied 77 severe storm events resulting in at least $1 billion in consumer price index (CPI) adjusted losses (NCEI 2021). The year 2020 recorded 13 of these events, easily surpassing the previous record of 9 in 2011. The frequency and magnitude of these costly severe storm events has dramatically increased since 1980 and can be primarily attributed to rapidly expanding developed land uses and subsequent increase in exposure to natural hazards (e.g., Hall and Ashley 2008; Changnon 2009; Bouwer 2011; Paulikas and Ashley 2011; Ashley et al. 2014; Rosencrants and Ashley 2015; Strader and Ashley 2015; Strader et al. 2015; Ashley and Strader 2016; Strader et al. 2017a,b; Strader and Ashley 2018; Strader et al. 2018; Childs et al. 2020; Ash et al. 2020). In addition to the expanding footprint of the human-built environment, climatological analyses of the spatial distributions of tornado and severe hail reports, and their favorable environment frequencies, reveal increasing trends in recent decades for both hazards across portions of the midwestern, southeastern, and northeastern United States. Decreasing frequency trends for significant tornado events have occurred across the central and southern Great Plains, whereas negative trends for severe hail are evident in the immediate lee of the Rocky Mountains in Colorado (Gensini and Brooks 2018; Moore 2018; Tang et al. 2019; Gensini et al. 2020; Taszarek et al. 2021a). Going forward, changes in both the human-built environment (i.e., exposure) and climatological probability (i.e., risk) of severe convective hazards are important components for assessing the potential for current and future economic loss, with changes in the human-built environment projected to play the biggest role in driving future disasters (Strader et al. 2017b).

Depending on their location and magnitude, severe convective storm (SCS) events can create vastly different economic impacts. The current U.S. definition of an SCS event includes hail with a maximum dimension of at least 25.4 mm, a tornado of any strength, or a thunderstorm-induced wind gust of at least 25.7 m s^{-1}. In an effort to distinguish SCS events more likely to produce significant economic losses, a *significant* SCS category

Denotes content that is immediately available upon publication as open access.

Corresponding author: Vittorio A. Gensini, vgensini@niu.edu

DOI: 10.1175/WAF-D-21-0056.1

was introduced (Hales 1988). Significant thresholds include hail exceeding a maximum dimension of 50.8 mm, a tornado rated (E)F2 or greater, and a thunderstorm-induced wind gust of at least 33.4 m s^{-1} (Hales 1988). A total of 34% (13%) of U.S. counties average at least 1 day with a significant hail (tornado) report per decade, with nearly all of these counties east of the Continental Divide (Brooks et al. 2003a; Allen and Tippett 2015). Though significant-severe is considerably less frequent than severe events (Gensini and Ashley 2011; Taszarek et al. 2020), a disproportionate share of losses stem from significant-severe hazards (Agee and Childs 2014; Grieser and Terenzi 2016).

This research was motivated by previous related works and a hypothesis that environmental variables commonly used to anticipate SCSs can be leveraged as skillful statistical discriminators between severe and significant-severe tornado and hail events. To test this hypothesis, all U.S. severe and significant-severe tornado and hail reports east of the Continental Divide from 1996 to 2019 were paired with 84 variables derived from ERA5 reanalysis vertical profiles (i.e., modeled proximity soundings) and passed to various machine learning classification algorithms for training, testing, and cross-validation. Machine learning approaches for severe convective weather diagnostic and prognostic applications have garnered significant attention in the last five years, and their value added is demonstrable (e.g., Gagne et al. 2017; Lagerquist et al. 2017; McGovern et al. 2017; Czernecki et al. 2019; Gagne et al. 2019; Mostajabi et al. 2019; Burke et al. 2020; Hill et al. 2020; Jergensen et al. 2020; Lagerquist et al. 2020; Loken et al. 2020; Sobash et al. 2020; Flora et al. 2021). Given that prediction of a severe versus significant severe weather event can be distilled to a binary classification problem, it warrants such approaches. Such machine learning approaches could greatly benefit SCS forecasters and nowcasters that desire to understand the probability of a particular event occurrence given the background convective environment.

2. Background

a. *Observed and model analysis proximity soundings*

Observed proximity sounding studies have been conducted since the 1940s in an effort to improve our understanding of the environments that favor SCS. Showalter and Fulks (1943), Fawbush and Miller (1954), and Beebe (1958) were the first formal studies that attempted to associate observed vertical thermodynamic and kinematic profiles with SCS events. The first definition of a proximity sounding is often credited to Darkow (1969). By the 1980s and early 1990s, diagnostic and prognostic fields from NWP were employed to investigate SCS vertical profiles (Weisman and Klemp 1982; Schaefer and Livingston 1988; Johns et al. 1990; Davies and Johns 1993). These works helped to conclude that for tornadoes, deep-layer wind shear, storm-relative helicity (SRH), and low static stability were fundamental ingredients for SCS occurrence.

The mid-to-late 1990s and 2000s produced a majority of SCS proximity sounding literature. Soundings associated with 92 mesocyclones illustrated a relationship between mesocyclone maintenance and a balance of SRH, deep-layer wind shear, and maximum moisture content in the vertical profile (Brooks et al. 1994). To construct a baseline climatology for variables associated with SCS forecasting (including significant severe events), all 0000 UTC soundings which possessed nonzero CAPE across the United States during 1992 were analyzed (Rasmussen and Blanchard 1998; Rasmussen 2003). Composite parameters such as the energy–helicity index (EHI) and the vorticity generation parameter showed more discriminatory skill depending on the application. An inherent weakness of utilizing *observational* proximity soundings is the lack of spatiotemporal coverage they provide, which may require broad, and potentially unrepresentative spatiotemporal criteria to capture a SCS event (Brooks et al. 1994; Craven and Brooks 2004; Potvin et al. 2010).

To increase sample size and refine the results of previous works, Rapid Update Cycle-2 (RUC-2) model proximity soundings were examined near supercell or discrete nonsupercell storms (Thompson et al. 2003). Larger values of deep-layer vertical wind shear, 0–1-km SRH, 0–1-km relative humidity, CAPE, and lower mixed-layer LCL heights discriminated well between significantly tornadic and nontornadic supercells (Thompson et al. 2003). The same RUC-2 dataset was used to extract wind profiles and associated kinematic diagnostics, indicating that significant tornado environments were frequently associated with larger ground-relative wind speeds, 0–1-km SRH, 0–1-km wind shear, and streamwise vorticity compared to weakly tornadic or nontornadic environments (Markowski et al. 2003).

Further expansion of this work yielded use of an effective inflow layer (effective SRH; ESRH) and effective bulk wind shear, both of which were found to better discriminate between significantly tornadic, weakly tornadic, and nontornadic supercells, as well as between supercell and nonsupercell convective modes (Thompson et al. 2007). Effective-layer calculations were applied to the supercell composite parameter (SCP) and significant tornado parameter (STP), improving their diagnostic discriminatory skill. Recent studies have indicated that further improvements in the skill of STP as a statistical discriminator can be achieved by using shallower layers for SRH integration bounds (e.g., 0–500 or 0–100 m; Coffer et al. 2019, 2020), and, generally, stronger ground-relative winds and more rightward-deviant storm motions contribute to more favorable conditions for tornadoes (Coniglio and Parker 2020). RUC-2 data were also used to stratify large hail events, where significant class overlap (severe and significant severe classes) was noted for thermodynamic variables (Johnson and Sugden 2014). However, improved skill was documented by using nontraditional environmental parameters that resulted in creation of the large hail parameter (LHP; Johnson and Sugden 2014).

In what are probably the two most similar studies to the research conducted herein, a multivariate logistic regression equation was shown to be more skillful in discriminating between tornadic and significantly tornadic environments versus just using effective-layer STP (Togstad et al. 2011), especially when incorporating dominant convective mode. Machine

learning was also used on a RUC-2 database of 1185 surface-modified vertical profiles, each classified as nonsupercell, nontornadic supercell, weak tornadic supercell, or significant tornadic supercell (Nowotarski and Jensen 2013). Vertical profiles were fed into a self-organizing maps (SOM) algorithm and trained to predict storm classification based on the associated kinematic and thermodynamic vertical profile diagnostics. In general, Nowotarski and Jensen (2013) revealed that simple kinematic diagnostics performed better than more complex kinematic and thermodynamic diagnostics. Ground-relative winds outperformed storm-relative winds, and the best performing SOM utilized 0–6-km ground-relative wind speed.

b. *Use of reanalyses*

Numerous studies have used the increased sample sizes offered by reanalysis data to extract proximity vertical profiles (e.g., Brooks et al. 2003b, 2007; Gensini and Ashley 2011; Gensini et al. 2014; Czernecki et al. 2019; King and Kennedy 2019; Taszarek et al. 2020, 2021b). Reanalyses have varying intervals, horizontal grid spacing, vertical resolution, and relative quality for SCS research (Gensini et al. 2014; King and Kennedy 2019; Taszarek et al. 2021b). Though reanalyses are not a perfect substitute for direct observations, they can be useful when observations are not available, whether in space or time.

Brooks et al. (2003b) was the first to implement the use of reanalysis data for SCS research. Examining NCEP–NCAR reanalysis (Kalnay et al. 1996) proximity soundings from 1997 to 1999, a linear discriminant analysis (LDA) was developed using 0–6-km shear and CAPE that best stratified between significant severe and nonsignificant severe/nonsevere environments. LDA also showed that the combination of mixed-layer LCL, 0–1-km shear, and station elevation discriminated between significantly tornadic and nontornadic environments. This led to further research examining SCS annual cycles (Brooks et al. 2007) and comparisons of this initial work to higher-resolution reanalyses (Gensini and Ashley 2011; Gensini et al. 2014; King and Kennedy 2019; Taszarek et al. 2020).

c. *Choice of reanalysis*

King and Kennedy (2019) incorporated a suite of reanalysis datasets to compare and contrast their classification ability and biases against the RUC-2 and results of Thompson et al. (2003, 2007) by testing many sounding-derived diagnostics (e.g., SRH, CAPE, SCP, and STP). Reanalysis datasets in the analysis included NARR (32-km grid, 29 vertical levels), ERA-Interim (80-km grid, 60 vertical levels), MERRA-2 (50-km grid, 72 vertical levels), JRA-55 (55-km grid, 60 vertical levels), 20CR (200-km grid, 28 vertical levels), and CFSR (38-km grid, 64 vertical levels). Kinematic variables were relatively consistent across all datasets, whereas thermodynamic diagnostics—especially those dependent on boundary layer moisture—showed the greatest bias. King and Kennedy (2019) suggest that these thermodynamic biases were not a by-product of spatiotemporal resolution differences; rather, they were primarily attributable to differences in parameter calculation methods and surface/boundary layer parameterization schemes. Additionally, fixed-layer calculations were more consistent across reanalysis datasets as compared to effective-layer.

The most recent global reanalysis product is the ECMWF ERA5, which possesses a horizontal grid spacing of 31 km, 137 vertical levels, and a 1-h output interval (Hersbach et al. 2020). These data, along with radar reflectivity, observed lightning data, and large hail reports, were used in a machine learning algorithm to improve the prediction of large hail events over Poland (Czernecki et al. 2019). A decision tree classification model was trained using various combinations of 35 predictors across ERA5 derived diagnostics and remote sensing data. ERA5-derived indices such as the hail size index (HSI) and significant hail parameter (SHIP) were shown to have skill in forecasting for large hail in Europe, but combining ERA5 data with observed radar and lightning data produced the best model performance.

ERA5's vertical resolution of 137 hybrid-sigma levels is far superior to any global reanalysis dataset currently available, which permits better depiction of rapidly changing profiles (e.g., sharp temperature inversions) that are vital to any vertically integrated calculations (e.g., CAPE, CIN). An analysis of over 3.7 million soundings from the United States and Europe illustrates that ERA5 is the most reliable available reanalysis dataset for SCS climatological research, with correlations to observed soundings of 0.8 for thermodynamic and 0.9 for kinematic parameters, respectively (Taszarek et al. 2021b). Thus, we use ERA5 reanalysis proximity vertical profiles to build on previous research, using large sample sizes, 84 diagnostic variables, and modern machine learning methods described in the next section to approach the issue of severe versus significant-severe tornado/hail report classification in a novel way.

3. Data and methods

a. *SCS reports*

Hail and tornado report data from 1996 to 2019 were obtained from the Storm Prediction Center public web page at https://www.spc.noaa.gov/wcm/ (Schaefer and Edwards 1999). Though reports are commonly used as ground-truth data for SCS climatological studies, it is important to explicitly mention their caveats herein. For tornadoes, reports do not consistently capture hazard magnitude, as they rely on a damage-based scale (Fujita 1971; Doswell III et al. 2009; McDonald et al. 2010; Edwards et al. 2013; Wurman et al. 2021). Tornadoes of the same pathlength, width, and intensity can produce vastly different societal impacts depending on their geographic location and underlying affected assets (Ashley et al. 2014; Ashley and Strader 2016; Strader and Ashley 2018; Strader et al. 2018). Hail reports at the surface are often subject to melting before they can be reported and have been shown to have a positive frequency bias toward higher population and road network densities (Tippett et al. 2015; Blair et al. 2017). Report collection involves measured and estimated hazard magnitudes, locations, and timing—also increasing uncertainty (Allen and Tippett 2015). While they do have significant societal impact, thunderstorm-induced severe wind reports were not assessed in this study due to their relatively poor reliability

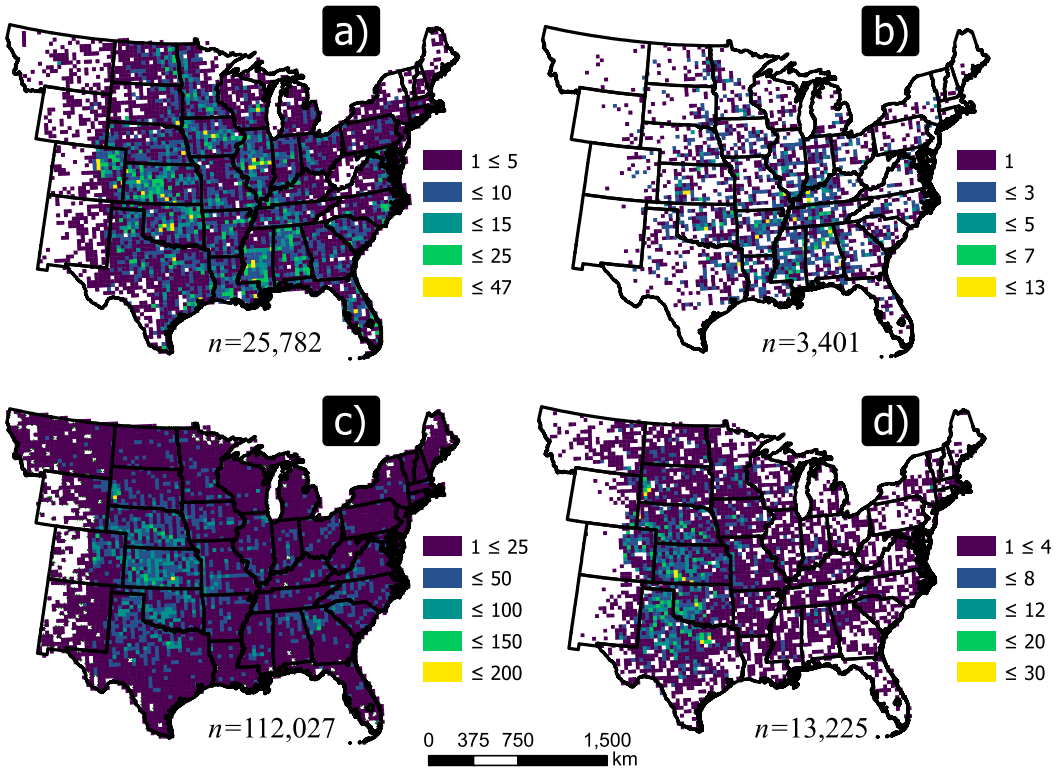

FIG. 1. Study domain and frequency of (a) (E)F0–1 tornadoes, (b) significant tornadoes, (c) severe hail, and (d) significant-severe hail assigned to the nearest ERA5 gridpoint. Sample sizes (n) for the study period (1996–2019) are also shown.

and quality compared to tornadoes and hail (Trapp et al. 2006; Smith et al. 2013; Edwards et al. 2018). It is important to emphasize that these are simply *reported* events, where a point-specific location and magnitude of the associated hazard are often approximated. Tornado reports from 1 January 1996 to 31 January 2006 utilized the Fujita scale, whereas all tornado reports from 1 February 2006 to 31 December 2019 used the enhanced Fujita scale (McDonald et al. 2010). This data discontinuity offers negligible implications for this research as the two scales differ little when binning nonsignificant and significant tornadoes. Hail reports contain maximum hailstone diameter, reported in 0.25 in. (6.35 mm) increments. Prior to 5 January 2010, the hail criterion for a severe thunderstorm was ≥0.75 in. (≈19 mm), which was thereafter increased to ≥1 in. (25.4 mm; NWS 2010). A severe hail size ≥ 25.4-mm threshold was implemented for the entire study to maintain physical consistency of the severe hail class.

b. *Study area, proximity definition, and calculation of variables*

The study area (Fig. 1) focuses on areas mostly east of the CONUS Continental Divide, covering the greatest climatological frequencies of SCSs (Brooks et al. 2003b, 2007; Gensini and Ashley 2011; Tippett et al. 2015; Gensini et al. 2020; Taszarek et al. 2020). For each tornado and hail report, a vertical profile (using all 137 ERA5 hybrid-sigma levels) of temperature, specific humidity, geopotential height, pressure, and (u, v) wind was extracted from the ERA5 grid cell nearest report location. To reduce soundings that were contaminated from the parent convective system, each vertical profile was extracted from the reanalysis interval 1 h prior to the hour floor of the associated report time. For example, a report time of 1835 UTC would yield a reanalysis sounding for 1700 UTC. Thus, all utilized model profiles were valid 60–119 min before report time. If multiple reports for the same hazard were recorded for the same ERA5 vertical profile, the highest report magnitude was assigned. To quality control for any issues related to boundary propagation and convective contamination, each sounding must have recorded nonzero MUCAPE, similar to previous research (Brooks et al. 2003b, 2007). In addition, profiles with average RH ≥ 90% in the 0.002–6-km layer were treated as contaminated and removed from the study (total of 192 and 645 for tornadoes and hail, respectively). Final sample sizes associated with each hazard class are as follows:

- Severe hail: 25.4 < 50.8 mm (1 < 2 in.): 112 027 profiles
- Significant-severe hail: ≥50.8 mm (≥2 in.): 13 225 profiles
- Tornado: (E)F0 or (E)F1: 25 782 profiles
- Significant tornado: ≥(E)F2: 3401 profiles

For each vertical profile, an assortment of 84 thermodynamic and kinematic variables (Table 1) were extracted or calculated using MetPy and SHARPpy (Unidata 2020; Blumberg et al. 2017) for use as predictors in each hazard class. Each profile

TABLE 1. Variables examined in this study.

Variable short name	Description	Units
Meltlvl	Height of 0°C T	m AGL
WetBulb0C	Height of 0°C T_w	m AGL
T_Sfc	2 m AGL T	°C
T_1km	1 km AGL T	°C
T_3km	3 km AGL T	°C
T_850 hPa	850-hPa T	°C
T_700 hPa	700-hPa T	°C
T_500 hPa	500-hPa T	°C
Theta_e_Sfc	2 m AGL θ_e	°C
Theta_e_1km	1 km AGL θ_e	°C
Theta_e_3km	3 km AGL θ_e	°C
Theta_e_850 hPa	850 hPa AGL θ_e	°C
Theta_e_700 hPa	700 hPa AGL θ_e	°C
Theta_e_500 hPa	500 hPa AGL θ_e	°C
Theta_e_Sfc1km	Avg θ_e between 2 m and 1 km AGL	°C
Theta_e_MDiff03km	Max difference in θ_e between 2 m and 3 km AGL	°C
Theta_Sfc	2 m AGL θ	°C
Theta_1km	1 km AGL θ	°C
Theta_3km	3 km AGL θ	°C
Theta_850 hPa	850-hPa θ	°C
Theta_700 hPa	700-hPa θ	°C
Theta_500 hPa	500-hPa θ	°C
MR1km	Mean w between 2 m and 1 km AGL	g kg^{-1}
MR2km	Mean w between 2 m and 2 km AGL	g kg^{-1}
MR3km	Mean w between 2 m and 3 km AGL	g kg^{-1}
MR4km	Mean w between 2 m and 4 km AGL	g kg^{-1}
MR5km	Mean w between 2 m and 5 km AGL	g kg^{-1}
MR6km	Mean w between 2 m and 6 km AGL	g kg^{-1}
MREL	Mean w of effective inflow layer	g kg^{-1}
RH_Sfc500 hPa	Mean RH between 2 m AGL and 500 hPa	%
RH_Sfc500m	Mean RH between 2 m and 500 m AGL	%
RH_Sfc1km	Mean RH between 2 m and 1 km AGL	%
RH_Sfc2km	Mean RH between 2 m and 2 km AGL	%
RH_Sfc3km	Mean RH between 2 m and 3 km AGL	%
RH_Sfc4km	Mean RH between 2 m and 4 km AGL	%
RH_Sfc5km	Mean RH between 2 m and 5 km AGL	%
RH_Sfc6km	Mean RH between 2 m and 6 km AGL	%
SRH05km	Storm-relative helicity between 10 and 500 m AGL using Bunker's storm motion	m^2 s^{-2}
SRH1km	Storm-relative helicity between 10 m and 1 km AGL using Bunker's storm motion	m^2 s^{-2}
SRH2km	Storm-relative helicity between 10 m and 2 km AGL using Bunker's storm motion	m^2 s^{-2}
SRH3km	Storm-relative helicity between 10 m and 3 km AGL using Bunker's storm motion	m^2 s^{-2}
SRH36km	Storm-relative helicity between 3 and 6 km AGL using Bunker's storm motion	m^2 s^{-2}
SRHEL	Storm-relative helicity in the effective inflow layer using Bunker's storm motion	m^2 s^{-2}
SRH020C	Storm-relative helicity in the 0° to −20°C layer using Bunker's storm motion	m^2 s^{-2}
SRH1030C	Storm-relative helicity in the −10° to −30°C layer using Bunker's storm motion	m^2 s^{-2}
Z020C	Thickness of 0° to −20°C layer	m AGL
Z1030C	Thickness of −10° to −30°C layer	m AGL
Shear05km	Bulk wind difference between 10 and 500 m AGL	kt
Shear1km	Bulk wind difference between 10 and 1 km AGL	kt
Shear2km	Bulk wind difference between 10 m and 2 km AGL	kt
Shear3km	Bulk wind difference between 10 m and 3 km AGL	kt
Shear6km	Bulk wind difference between 10 m and 6 km AGL	kt
Shear8km	Bulk wind difference between 10 m and 8 km AGL	kt
ShearEL	Bulk wind difference in the effective inflow layer	kt
WS_850 hPa	850-hPa wind speed	kt
WS_500 hPa	500-hPa wind speed	kt
WS_250 hPa	250-hPa wind speed	kt
Crit_Angle	Critical angle	°
LR700500	700–500-hPa lapse rate	°C km^{-1}
LR01km	2 m–1 km AGL lapse rate	°C km^{-1}
LR03km	2 m–3 km AGL lapse rate	°C km^{-1}
LR26km	2–6 km AGL lapse rate	°C km^{-1}
LCL_Sfc	Surface-based lifting condensation level (parcel from lowest model level)	m AGL
LCL_MU	Most-unstable lifting condensation level (most-unstable parcel in the profile)	m AGL
LCL_ML	Mixed-layer lifting condensation level (mean T and w of lowest 100 hPa)	m AGL
LCL_EL	Effective-layer lifting condensation level (mean T and w of effective layer)	m AGL
LFC_Sfc	Surface-based level of free convection (parcel from lowest model level)	m AGL
LFC_MU	Most-unstable level of free convection (most-unstable parcel in the profile)	m AGL

TABLE 1. (*Continued*)

Variable short name	Description	Units
LFC_ML	Mixed-layer level of free convection (mean T and w of lowest 100 hPa)	m AGL
LFC_EL	Effective-layer level of free convection (mean T and w of effective layer)	m AGL
CAPE_Sfc	Surface-based convective available potential energy (parcel from lowest model level)	J kg^{-1}
CAPE_MU	Most-unstable convective available potential energy (most-unstable parcel in the profile)	J kg^{-1}
CAPE_ML	Mixed-layer convective available potential energy (mean T and w of lowest 100 hPa)	J kg^{-1}
CAPE_EL	Effective-layer convective available potential energy (mean T and w of effective layer)	J kg^{-1}
CAPE_Sfc03	2 m–3 km AGL surface-based convective available potential energy (parcel from lowest model level)	J kg^{-1}
CAPE_MU03	2 m–3 km AGL most-unstable convective available potential energy (most-unstable parcel in lowest 3 km)	J kg^{-1}
CAPE_ML03	2 m–3 km AGL mixed-layer convective available potential energy (mean T and w of lowest 100 hPa)	J kg^{-1}
CAPE_EL03	2 m–3 km AGL Effective-layer convective available potential energy (mean T and w of effective layer)	J kg^{-1}
CIN_Sfc	Surface-based convective inhibition (parcel from lowest model level)	J kg^{-1}
CIN_MU	Most-unstable convective inhibition (most-unstable parcel in the profile)	J kg^{-1}
CIN_ML	Mixed-layer convective inhibition (mean T and w of lowest 100 hPa)	J kg^{-1}
CIN_EL	Effective-layer convective inhibition (mean T and w of effective layer)	J kg^{-1}
GRW_aEL	Difference of the effective layer and 3–6 km AGL average wind direction	°
SRW_aMid	Difference of the 10 m–1 km AGL and 3–6 km AGL layer average storm relative wind direction	°

was then assigned a label of 0 or 1 for a severe or significant-severe event, respectively.

c. *Machine learning classification*

Logistic regression, Gaussian naive Bayes, support vector machines, adaptive boosting, gradient boosting, decision trees, and random forests were all explored as potential classification models using Python's scikit-learn library (Pedregosa et al. 2011). Only logistic regression (Kleinbaum et al. 2002) and random forest classification (Breiman 2001) are discussed hereafter, as they were found to be the most skillful for this particular application. First, it should be noted that both tornado and hail reports have a class imbalance problem (i.e., there are approximately an order of magnitude more severe reports than significant-severe reports). Training any model designed to optimize accuracy on such data will simply predict the severe class only. This will lead to high model accuracy scores, but will also exhibit a critical success index (CSI) of 0. To overcome this, oversampling techniques are often performed on the minority class (in this case, significant-severe). The significant-severe class for both tornado and hail were oversampled using a borderline approach, where a support vector machine (SVM) is used to locate the decision boundary (Nguyen et al. 2011). The borderline area (i.e., decision boundary between classes) is approximated by the support vectors obtained after training a standard SVM classifier on the original training set. New instances will be randomly generated along the lines joining each minority class support vector with a number of its nearest neighbors using the interpolation. In addition to using an SVM, the technique attempts to select regions where there are fewer examples of the minority class and tries to extrapolate toward the class boundary. Samples from k nearest neighbors [set to 7; tested (3, 5, 7, 9, 15)] from the significant-severe class near the support vector boundary are the focus for generating synthetic samples to increase n for the minority class.

Training and evaluation of the classification models was performed using a leave-one-year-out-cross-validation approach. For instance, all models were trained on data from 1996 to 2018 and evaluated by predicting the outcome of unseen data from 2019. The process is then repeated for each year to create a diverse sample ($N = 24$) of model skill scores. Oversampling was only performed on the training data to prohibit synthetic overfitting. To ensure the model was not overfit to the hyperparameters, an 80%/20% random split of the training data (e.g., 1996–2018) was used for training and validation before applying to the testing data (e.g., 2019). We saw no differences in the optimal tuning parameters shown in Table 2, but the overall CSI of the models were marginally degraded (in some cases CSI values up to 0.02 points, likely due to the reduced amount of training data), but not statistically significant.

All data values were kept in their raw units and not scaled (scaling had negligible impacts and did not greatly improve skill for any of the tested models). Performance diagram variables (i.e., probability of detection and success ratio) were calculated for each model iteration using the deterministic 2 × 2 contingency table predictions through model.predict(). For random forest classification, the predicted class of an input sample is a vote by the trees in the forest, weighted by their probability estimates. That is, the predicted class is the one with highest mean probability estimate across the trees. For logistic regression, the probability threshold for classification was chosen by using the class with the highest probability (essentially a probability threshold of 0.5). Receiver operating characteristic (ROC) curves and attributes diagrams were created from the model predictions using model.predict_proba() to further assess model skill and reliability (Wilks 2011).

TABLE 2. Parameters tested during model tuning. Bold values indicate final model selections that produced the highest critical success index values.

Model parameter	RFC model (Tor)	RFC model (hail)
n_estimators	[50, 100, 200, **250**, 500, 750]	[50, 100, 200, **250**, 500, 750]
criterion	[**gini**, entropy]	[**gini**, entropy]
min_samples_split	[**2**, 3, 4, 5]	[**2**, 3, 4, 5]
min_samples_leaf	[1, 2, **4**, 6, 8, 10]	[1, 2, 4, 6, **8**, 10]
bootstrap	[**True**, False]	[**True**, False]
Model parameter	**LOGIT (Tor)**	**LOGIT (Hail)**
C	[0.1, 1, 5, **25**, 100]	[0.1, 1, **5**, 25, 100]
penalty	[l1, **l2**]	[l1, **l2**]
solver	[liblinear, **lbfgs**, saga]	[liblinear, **lbfgs**, saga]

Both random forest and logistic regression models were tuned using a "grid search" (Pedregosa et al. 2011) approach to achieve optimal CSI scores. For logistic regression, the only nondefault setting (scikit-learn v0.24.1) was the inverse of regularization strength parameter *C* (set to 25 for tornado and 5 for hail). For random forest classification, nondefault settings include the number of trees in the forest (set to 250 for both hazards) and the minimum number of samples required to be at a leaf node (set to 4 for tornado and 8 for hail). Other model parameters tested can be found in Table 2.

In the context of comparing different models, statistical significance was tested using 1000 bootstrapped (random resampling with replacement) samples of pairwise model scores (e.g., CSI). Models were deemed to be statistically significantly different if the bootstrapped p value was ≤ 0.05. Thus, the use of "statistically significant" hereafter should be interpreted as having at least 95% confidence that the compared models differ in skill.

d. Multicollinearity and variable importance

Some predictors in Table 1 exhibit multicolinearity (e.g., $R^2 \geq 0.7$ between variables). This does not have a significant consequence for the RFC model, as correlated values can be used as predictors at decision points (testing for mean reduction in gini impurity at each node) without a preference for a single variable. However, this can mask relative variable importance measurements in the RFC, making their interpretation more difficult. For example, perhaps 0–1-km SRH for tornado class forecasts is important, but assessing its value added relative to a model that also has 0–500-m SRH as a predictor reduces the relative importance of both variables given the colinearity. To address this, conditional permutation importance (Strobl et al. 2008) and Shapley additive explanations (SHAP) were used to evaluate the importance of each variable to the skillful model prediction of the severe report class.

Conditional permutation feature importance measures feature importance by observing how random reshuffling (thus preserving the distribution of the variable) of each predictor influences model performance. This is only possible with the RFC model, and not with logistic regression.

SHAP values follow a game theoretic approach to help explain the output of certain types of machine learning models (Shapley 2016; Lundberg et al. 2018, 2020). In this application, SHAP values allow for the evaluation of how each variable changes the log odds of the severe versus significant-severe class prediction. Useful thresholds can then be created at which each predictor begins to increase or decrease its probabilistic contribution to the forecast outcome.

Multicolinearity can be an important issue for logistic regression models. Thus, recursive feature elimination (RFE) was utilized. RFE starts with all features in the training dataset, ranks features by importance, discards the least important features, and refits the model (Guyon et al. 2002). This process is repeated until a specified number of features remains. The number of retained features was tested (3, 5, 10, 20, 40, 60, all), but the overall change in skill of various logistic regression models was negligible (skill tended to slowly increase with increasing n features before hitting an asymptote at $n \approx 15$). Further sensitivity tests conducted to see if model skill/feature importance changed by removing variables with high covariance (R^2 values ≥ 0.7) revealed negligible change in model skill and no change in variable importance rankings (unless the variable was specifically removed) as indicated by the training weights. RFE for logistic regression was set to find the 10 most important features for comparison to the top 10 variables for the RFC model. One potential issue with RFE is that one can obtain different variable importances by scaling and/or simply omitting variables. Thus, the authors are more confident in the ML interpretation methods for the RFC that uses the conditional permutation importance, and these variables can be further interpreted by examining the SHAP values.

4. Results

a. Model performance

1) TORNADOES

Random forest classification (RFC) exhibited the greatest skill for classification of a tornado (TOR) versus significant tornado (SIG TOR) event with a cross-validated mean area under the ROC curve (AUROC) value of 0.785 and a mean CSI value of 0.23. RFC classification skill was compared to three logistic regression models, one using logistic regression on all variables (hereafter LOGIT), another using Eq. (1) from Togstad et al. (2011) (hereafter T_{2011}), and another using effective-layer STP (STP_{EFF}), mirroring the logistic regression benchmark of Togstad et al. (2011) (Fig. 2a). The improvement noted by Togstad et al. (2011) in AUROC when comparing STP_{EFF} to T_{2011} was replicated herein (5.1%) using a much larger sample size (T_{2011} reported a 5.3% AUROC improvement), and represents a statistically significant increase in CSI and AUROC. Only minor improvement (not statistically significant) was noted over T_{2011} when using all variables in the LOGIT model. Thus, if using logistic regression for this classification task, the simpler (i.e., fewer variables) equation used in T_{2011} is likely an optimal choice. However, the RFC model provided statistically significant improvements in skill over all logistic regression models, with mean CSI increases found to be

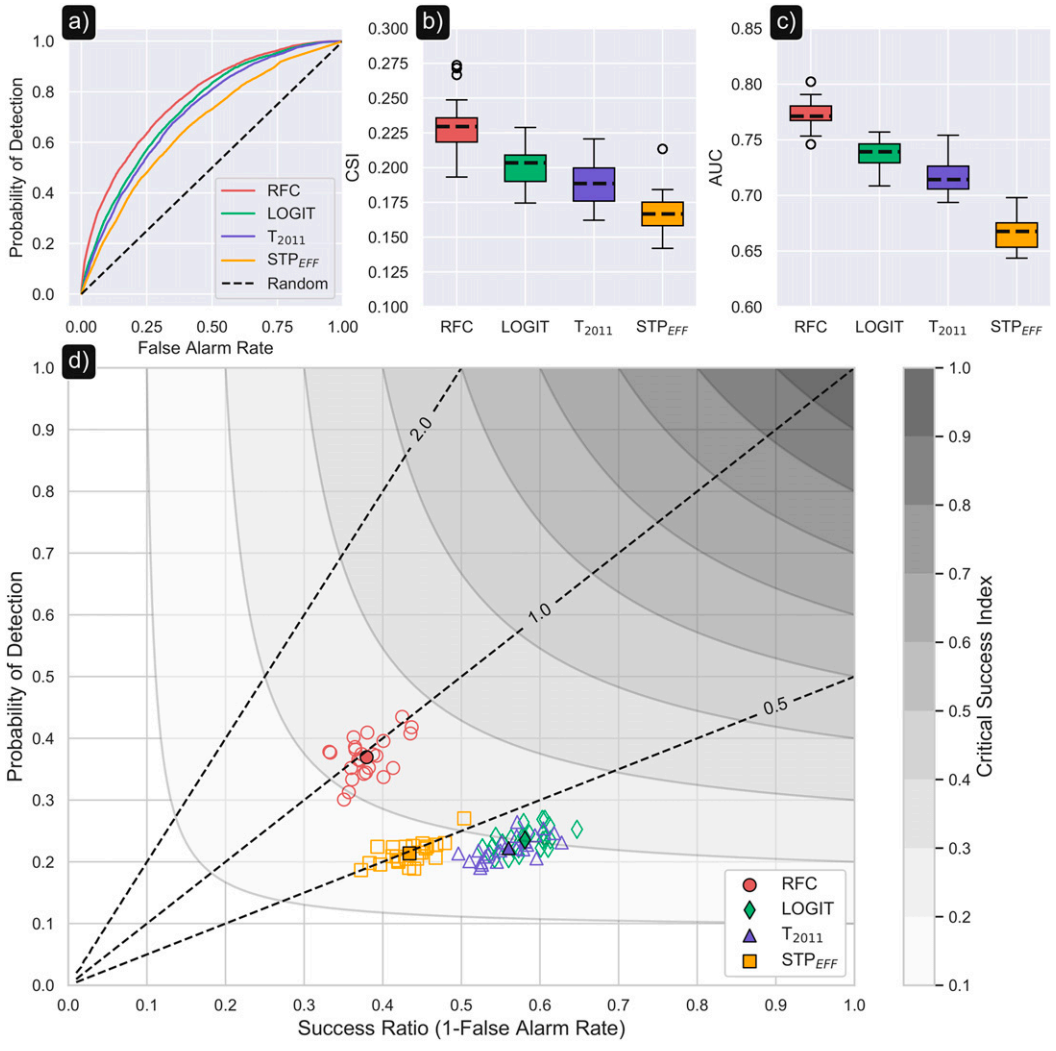

FIG. 2. (a) TOR vs SIG TOR classification mean receiver operating characteristic (ROC) curves for random forest classification (RFC), logistic regression using all variables (LOGIT), logistic regression using Eq. (1) from Togstad et al. (2011) (T_{2011}), and logistic regression using effective-layer STP (STP_{EFF}). (b) CSI and (c) AUROC value distributions (1000 random bootstrapped samples with replacement) from the leave-one-year-out cross validation are shown. Dashed lines on the box plots indicate the median value. Boxes display the interquartile range, and whiskers extend to the 10th and 90th percentiles (circles indicate outliers). (d) A performance diagram is shown comparing all models. Mean result from leave-one-year-out-cross validation ($n = 24$; 1996–2020) is noted by filled symbol and outlined in black, whereas all other years are noted as non-filled symbols. Black dashed lines on the performance diagram indicate frequency bias.

2.9%, 4.1%, and 6.3% over LOGIT, T_{2011}, and STP_{EFF}, respectively (Fig. 2b). RFC also exhibited a modest lift in the mean AUROC when compared to LOGIT (3.7%), T_{2011} (5.4%), and, especially, STP_{EFF} (10.5%; Fig. 2c). In fact, the interquartile range of RFC cross-validated CSI and AUROC scores were greater than—and had no overlap with—LOGIT, T_{2011}, or STP_{EFF}. To summarize classification skill of a TOR versus SIG TOR report, the ensemble of cross-validated results for all four models were plotted on a performance diagram (Fig. 2d). CSI increases as one moves from bottom left to top right of a performance diagram. Perhaps most notable here are the model frequency biases, noted as black dashed lines on the performance diagram. The mean RFC model has a frequency bias very close to 1, indicating that a SIG TOR is forecast by the RFC classifier exactly often as it is observed. All logistic regression techniques have a low frequency bias, predicting the SIG TOR class about half as often as they are observed in the report database.

An attributes diagram was also examined to assess the need for forecast calibration (Fig. 3). The one-to-one line (blue dashed) on an attributes diagram indicates perfect reliability. The solid blue vertical and horizontal lines represent climatological frequencies, and the black dotted line represents no skill (the line halfway between perfect reliability and the

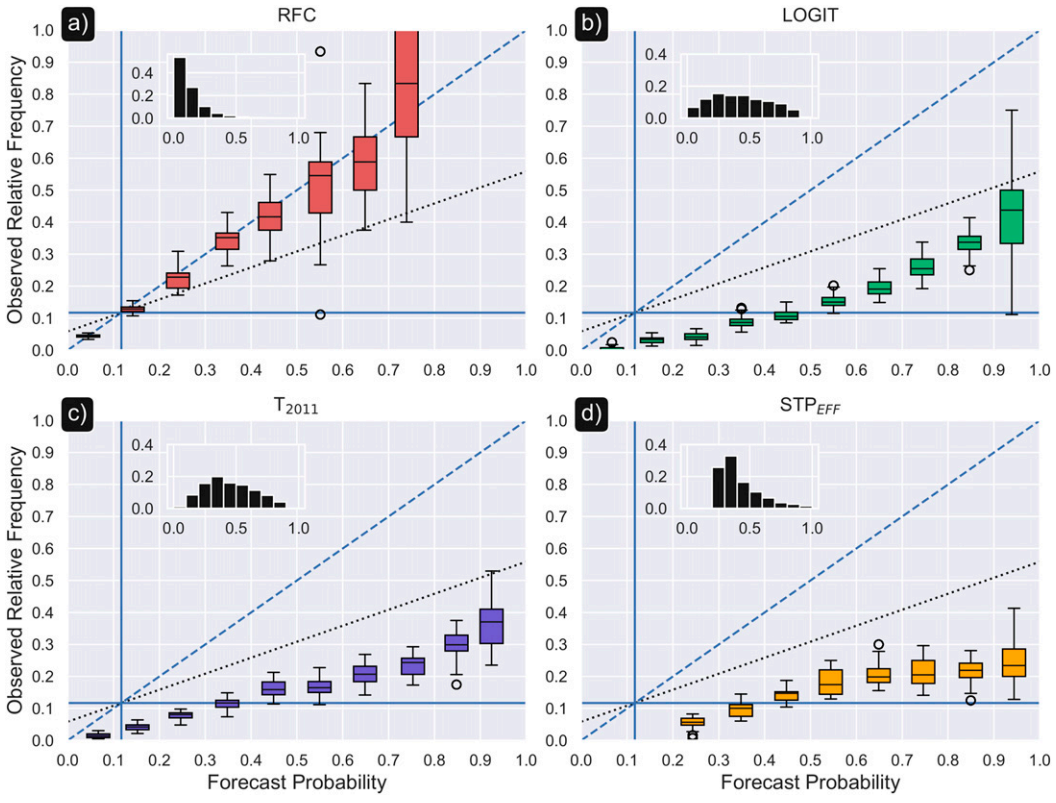

FIG. 3. Attributes diagram for (a) RFC, (b) LOGIT, (c) T_{2011}, and (d) STP_{EFF} to classify TOR vs SIG TOR reports. Solid lines on the box plots indicate the median value. Boxes display the interquartile range, and whiskers extend to the 10th and 90th percentiles (circles indicate outliers). Vertical and horizontal solid blue lines indicate climatology. Dashed blue line indicates perfect reliability. Dotted black line indicates no skill. Inset histograms indicate the respective forecast sharpness, with the *y* axis scaled to the fractional forecast frequency.

horizontal climatological frequency). To show positive forecast skill, points on the right side of the vertical climatological frequency line must lie above the no-skill line, and on the left side must be below the no-skill line. Spread among the cross-validated runs for RFC indicate good reliability, with median values falling along the one-to-one line of forecast probability and observed frequency, suggesting that the RFC classifier is skillful at all forecast probabilities (Fig. 3a). All logistic regression models showed tendencies to overforecast at forecast probabilities higher than climatology (Figs. 3b–d). Most of the classification skill associated with logistic regression models stems from lower-probability forecasts, similar to results shown in Togstad et al. (2011). Using T_{2011} (Fig. 3c), when a SIG TOR category has a forecast probability equal to 45% the actual chance of observing the event is closer to 20%. This is a nearly identical result to that presented in T_{2011}. Overall, this analysis indicates that the RFC model is well calibrated, whereas the logistic regression models would require calibration to become more reliable.

2) HAIL

Overarching results for severe hail (HAIL) versus significant-severe hail (SIG HAIL) are similar. The RFC model again exhibited the greatest skill, with a cross-validated mean AUROC value of 0.772 and a mean CSI value of 0.237. RFC was benchmarked against three logistic regression models comprised of all variables (LOGIT), the significant hail parameter (SHIP), and the large hail parameter (LHP; Johnson and Sugden 2014). RFC AUROC lift was marked (and statistically significant) over these benchmark models, recorded as 9.7%, 12.9%, and 15.7%, respectively (Figs. 4a,c). Logistic regression with SHIP performed slightly better (as interpreted by AUROC) than LHP, particularly at probability of detection values ≥ 0.5, but CSI distributions were nearly identical and not statistically significantly different. LOGIT performed statistically significantly better than SHIP and LHP, with AUROC lift at most values of probability of detection and false alarm rate (Fig. 4a). Interestingly, the lowest performing cross-validated RFC CSI (AUROC) score was still a 4% (3.6%) improvement over the best performing LOGIT score (Figs. 4b,c). RFC classification was found to have a mean frequency bias ≈ 1 and was clearly a superior performing model on the performance diagram (Fig. 4d). All logistic regression techniques exhibited frequency biases < 0.5 and were not able to reach deterministic probability of detection values exceeding 0.2. SHIP and LOGIT models performed better than LHP by reducing the false alarm rate.

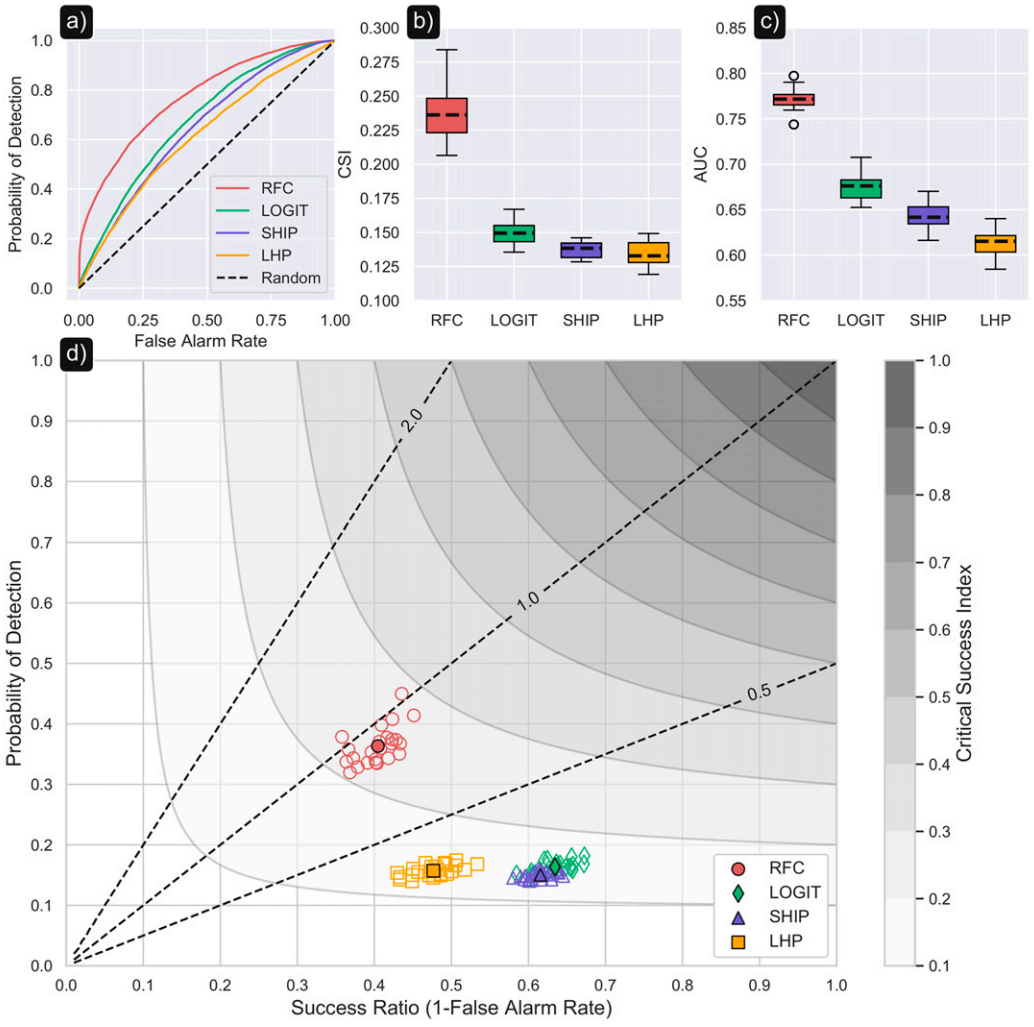

FIG. 4. As in Fig. 2, but for HAIL vs SIG HAIL classification using RFC, LOGIT, logistic regression using the significant hail parameter (SHIP), and logistic regression using the large hail parameter (LHP).

HAIL versus SIG HAIL model attributes diagrams indicated one notable difference when compared to tornado classification. The RFC classifier is less reliable at forecast probabilities between 0.3 and 0.9 (overforecasting bias), and garners most of its skill outside of that range (Fig. 5a). Recall the RFC TOR versus SIG TOR model was reliable at all forecast probabilities suggesting that the greatest signal for a HAIL versus SIG HAIL report classification forecast originates from relatively low or high forecast probabilities. Logistic regression methods were not as reliable and exhibit little-to-no skill at most forecast probabilities (Figs. 5b–d). We reiterate that all models may benefit from some degree of calibration to increase reliability.

b. Variable importance

1) TORNADO

The 500-hPa wind speed, 850-hPa wind speed, 0.01–3-km bulk wind shear, effective-layer SRH, and 2–6-km lapse rate all ranked in the top 10 most important variables for both LOGIT and RFC models (Table 3). Low-level SRH ranked in the top 5 for both models, but with different choices of integration bounds (10–500 m for RFC, and 0.01–1 km for LOGIT). A majority (7 of the top 10) of predictors for the RFC model were noted as being kinematic, which agrees with previous research examining the best environmental discriminators for tornadoes of various magnitudes (e.g., Brooks et al. 2003a; Rasmussen 2003; Markowski et al. 2003; Thompson et al. 2003; Nowotarski and Jensen 2013; Hampshire et al. 2018; Coffer et al. 2019; King and Kennedy 2019; Coffer et al. 2020). LOGIT had fewer kinematic variables (5) in the top 10. Results of relative importance through conditional permutation were not sensitive to the number of times permuting a variable or random shuffling.

2) HAIL

Variable importance for HAIL versus SIG HAIL was mixed between RFC and LOGIT models. Surface-based CAPE, 850-hPa θ_e, 2–6-km lapse rate, freezing level, 0.002–2-km average RH, and

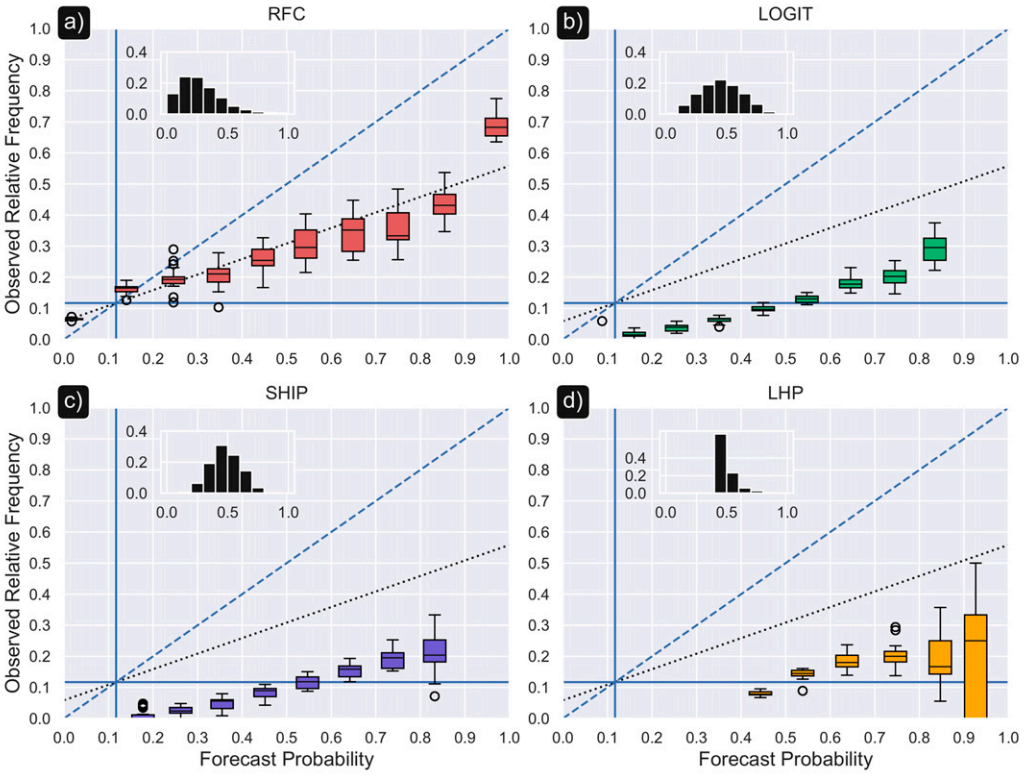

FIG. 5. As in Fig. 3, but for HAIL vs SIG HAIL classification using RFC, LOGIT, SHIP, and LHP.

mixed-layer LCL height were the thermodynamic variables with greatest importance from the LOGIT model. 0.01–3-km bulk shear was again in the top 10 for both models, along with 0.01–6-km bulk shear and 2–6-km lapse rate. LOGIT parameters were nearly evenly split between thermodynamic and kinematic variables, whereas 8 of the top 10 most important variables for SIG HAIL classification by the RFC model were kinematic in nature. Interestingly, the RFC model indicated that 3–6-km SRH and SRH in the −10° to −30°C layer were the two most important predictors. Details of the storm-relative hodograph have been hypothesized to play an important role in the hail growth (Kumjian and Lombardo 2020), and these results suggest that mid-to-upper-level SRH plays an important role in determining the likelihood of the environment supporting SIG HAIL.

3) EVALUATING RFC WITH SHAP

The SHAP value (Štrumbelj and Kononenko 2014) from Shapley game theory (Shapley 2016) for each predictor tries to identify the correct weight such that the sum of all Shapley values is the difference between the predictions and average

TABLE 3. TOR vs SIG TOR and HAIL vs SIG HAIL relative variable importance (1 highest; 10 lowest) for LOGIT and RFC models. FI indicates the feature importance value scaled from 0 to 1 using recursive feature elimination (LOGIT) and conditional permutation importance (RFC).

	TOR vs SIG TOR				HAIL vs SIG HAIL			
Rank	LOGIT	FI	RFC	FI	LOGIT	FI	RFC	FI
1	SRH1km	1	WS_500 hPa	1	T_1km	1	SRH36km	1
2	WS_500 hPa	0.764	WS_850 hPa	0.567	WS_500 hPa	0.583	SRH1030C	0.874
3	CAPE_ML	0.568	SRH05km	0.491	Shear3km	0.436	Shear6km	0.771
4	LCL_EL	0.507	CAPE_EL	0.365	Theta_e_850 hPa	0.342	Shear3km	0.753
5	CAPE_EL03	0.284	LR_26km	0.349	Theta_e_1km	0.337	Crit_Angle	0.377
6	SRH05km	0.232	CAPE_ML	0.316	Shear1km	0.319	WS_500 hPa	0.356
7	CIN_EL	0.162	Shear6km	0.306	Shear6km	0.294	GRW_aEL	0.309
8	RH_Sfc500m	0.117	ShearEL	0.303	RH_Sfc2km	0.284	LR_26km	0.306
9	Shear6km	0.116	Shear3km	0.291	T_850 hPa	0.239	Shear8km	0.24
10	Shear1km	0.113	SRHEL	0.284	Theta_3km	0.222	LR700500	0.162

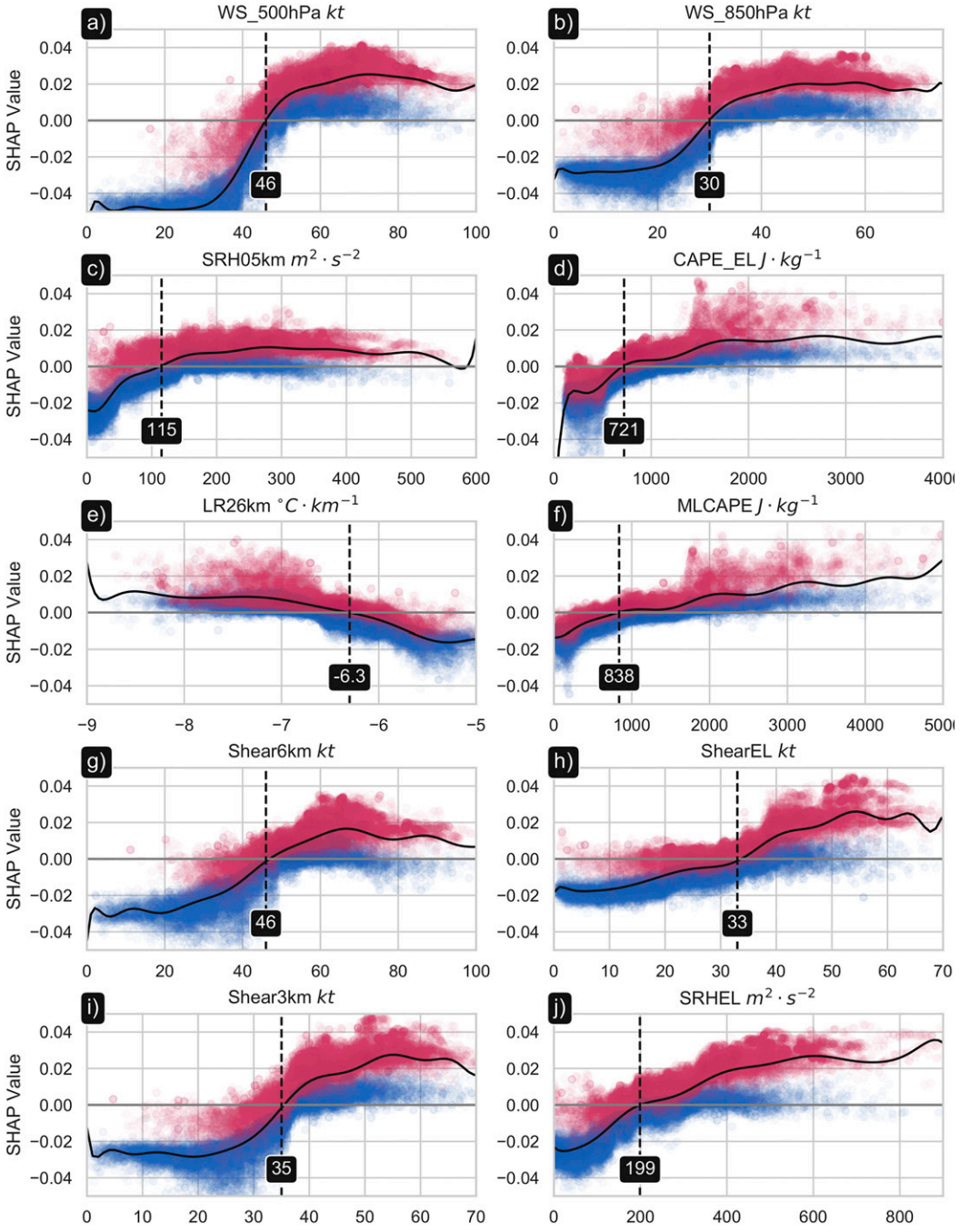

FIG. 6. Random forest classification SHAP values for the top-10 variables from Table 3 for TOR (blue dots) vs SIG TOR (red dots). The vertical black dashed line and associated x-axis variable value indicates the intersection of the SHAP = $f(x)$ polynomial (black solid line) with the 0 SHAP value.

value of the model. Essentially, Shapley values correspond to the contribution of each variable toward pushing the prediction away from the expected value. Shapley values consider all possible predictions for an instance using all inputs from the variable set. When fitting a polynomial to the distribution of $\text{SHAP}_x = f(x)$, one can examine the critical value(s) of x that solve $0 = f(x)$. These can be thought of as critical x values that demarcate a ± directional change toward contribution of a prediction. Figure 6 displays $\text{SHAP}_x = f(x)$ plots for the top ten variables from the RFC TOR versus SIG TOR model. As an example, the critical SHAP value for 500-hPa wind speed is 46 kt (25.7 m s^{-1}), essentially meaning that values below (above) 46 kt negatively (positively) contributed to the log odds of a SIG TOR (Fig. 6a). Unsurprisingly, best-fit polynomial

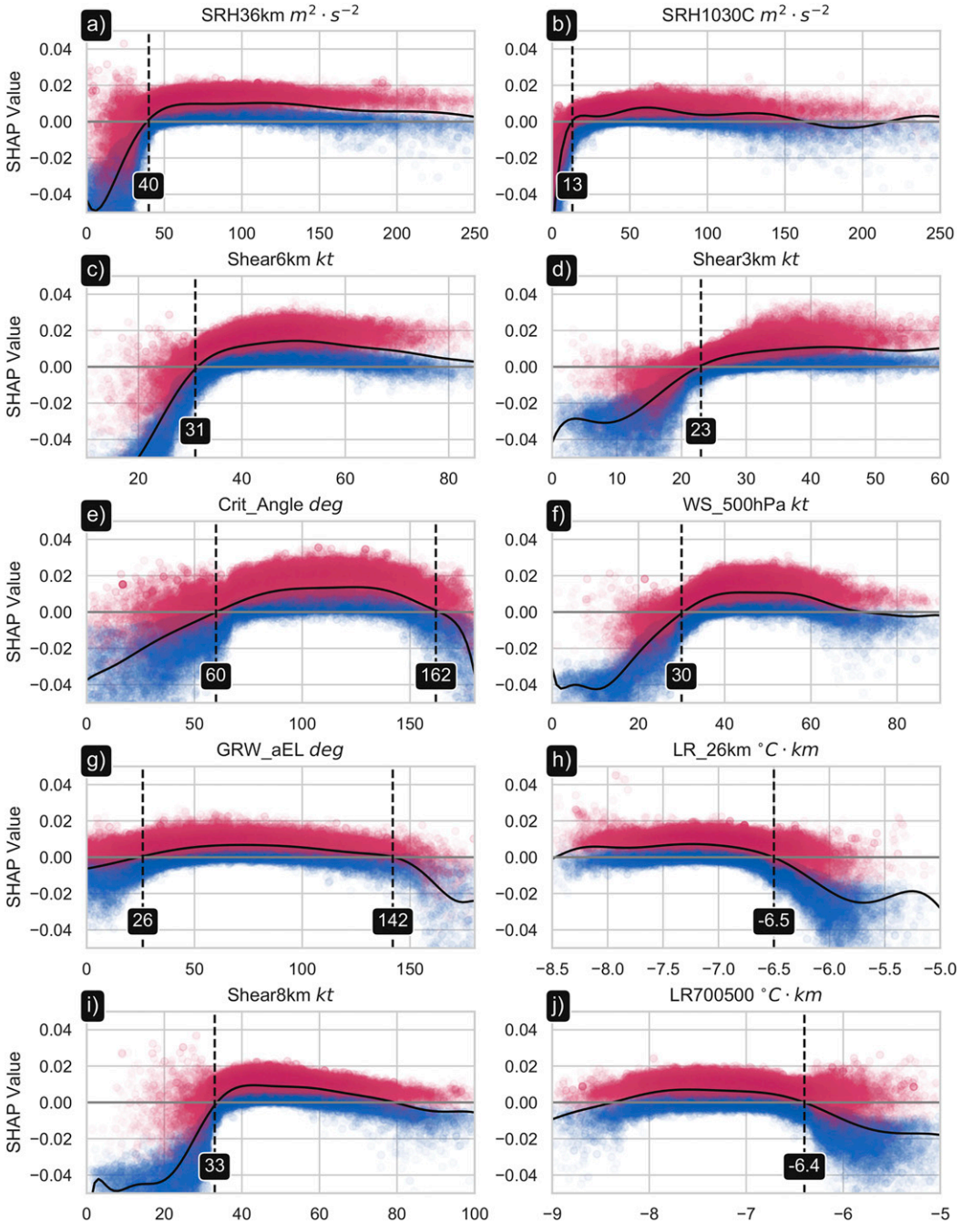

FIG. 7. As in Fig. 6, but for top-10 variables from Table 3 for HAIL (blue dots) vs SIG HAIL (red dots).

shape/structure of SHAP changes depending on the variable (e.g., cubic for Fig. 6b, parabolic for Fig. 7e).

Low-level, upper-level, and deep-layer winds/shear exhibit distributions that suggest the greatest change in $SHAP_x$ [i.e., $f'(x)$] happens within close proximity to the critical value and contributions toward the prediction of a SIG event do not change much for certain $f(x)$ values (Figs. 6a,b,g,h,i; 7c,d,f,i). This suggests that, for example, once wind speed at 500 hPa reaches $\approx$ 70 kt (36 m s^{-1}), no changes in the log odds of prediction occur (Fig. 6a). Other variables (e.g., MLCAPE and SRHEL for tornadoes) tend to scale $SHAP_x$ in an approximately linear manner (Figs. 6f,j), and some even substantially reintersect the critical $0 = f(x)$ value to again change the direction of contribution toward prediction (Figs. 7e,g). This is clear with critical angle for SIG HAIL, which shows a positive contribution to the outcome of a SIG HAIL event if the angle is between 60° and 162°. Physically, this would suggest a range of optimal angles between surface storm relative wind and the low-level shear

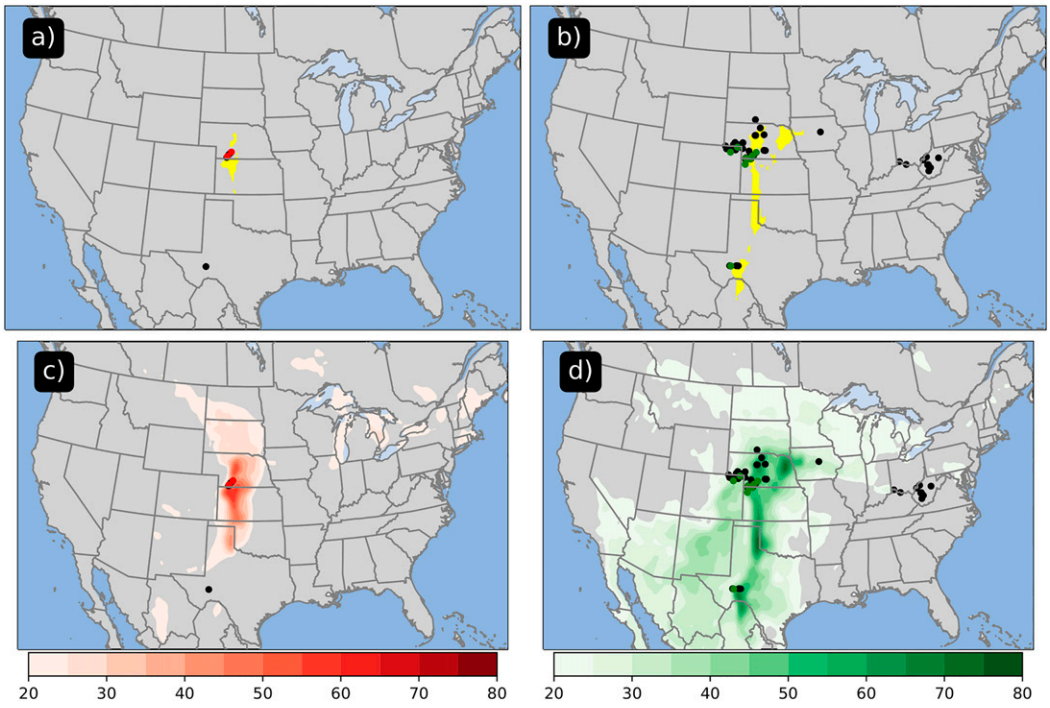

FIG. 8. Example use case for 2100 UTC 17 May 2019 using RFC models trained on data from 1996 to 2018 showing deterministic yes/no predictions (yes; yellow fill) for (a) SIG TOR and (b) SIG HAIL and the corresponding forecast probabilities for (c) SIG TOR and (d) SIG HAIL. Black dots indicate severe reports. Red and green dots indicate SIG TOR and SIG HAIL reports, respectively. All reports valid between 2100 UTC 17 May 2019 and 0000 UTC 18 May 2019.

vectors. This also relates to storm relative helicity which is important for mesocyclone formation and maintenance (Thompson et al. 2007; Esterheld and Giuliano 2008). While SIG events do occur on either side of the critical $SHAP_x$ value, these thresholds serve as potential discriminating values from an aggregate climatological approach when considering game theory. In short, they offer a simple, yet statistically meaningful, value that can be used by forecasters or normalization techniques in composite parameters. At the very least, SHAP value analysis permits the interpretation of an otherwise ''black box'' of decision-making in certain machine learning algorithms (McGovern et al. 2019).

c. *Example RFC use case: 17 May 2019*

The model training set from 1996 to 2018 was used to predict spatial locations favoring SIG TOR and SIG HAIL in a hindcast setting using ERA5 input from 2100 UTC 17 May 2019. This essentially resembles a quasi-operational setting in which a tool could ingest real-time environmental variables and provide a classification prediction. The prediction takes a matter of seconds once the variables are ingested, which is important for rapidly updating operational systems. The longest part of the process is training the RFC models, which could be done antecedently. The date 17 May 2019 exhibited broad southwesterly upper-level flow over the High Plains atop favorable boundary layer mixing ratio profiles and southerly winds east of a dryline positioned near a line extending from Imperial, Nebraska (KIML), to Dodge City, Kansas (KDDC), to Childress, Texas (KCDS), to Fort Stockton, Texas (KFST), at 2100 UTC (https://www.spc.noaa.gov/exper/archive/event.php?date=20190517).

RFC TOR versus SIG TOR models indicated that—should a severe weather report occur—the environment favored SIG TOR reports in portions of northwest Kansas and southwest Nebraska at 2100 UTC (Fig. 8). Two EF2 tornado reports (2250 and 2358 UTC) were recorded in the three hours following 2100 UTC in southwest Nebraska in the northwest quadrant of a surface θ_e axis. A tornado did occur near the Davis Mountains in Texas between 2100 and 0000 UTC, but it was not significant. The RFC SIG HAIL model highlighted a narrow zonal corridor just east of the surface dryline favoring SIG HAIL reports (Fig. 8b). A majority (13 of 15) of hail reports in far western Nebraska and northeast Colorado were not significant and were outside of the delineated SIG HAIL area. The environment further south into western Kansas and the eastern Texas Panhandle favored SIG HAIL, but convection had not initiated/matured there by 0000 UTC 18 May 2019. Additional knowledge about the probability of convective initiation would certainly be beneficial to reduce false alarm area. In addition, 12 hail reports were recorded in West Virginia–Oklahoma–Kentucky between 2100 and 0000 UTC, but none of them were significant as correctly indicated by the RFC model (Figs. 8b,d). While this is only one

example case, it demonstrates feasibility as a tool for operational applications.

5. Discussion and conclusions

Over 150 000 ERA5 reanalysis vertical profiles were extracted near severe weather reports in an attempt to investigate the ability of 84 atmospheric variables to stratify environments favorable for severe versus significant-severe tornadoes and hail. Key results of this study include:

- Random forest classification models were the most skillful in predicting whether or not a tornado or hail report would be classified as severe or significant-severe.
- Random forest classification models were more reliable and had less frequency bias compared to logistic regression.
- Conditional permutation importance indicated that kinematic variables generally showed greater discriminatory power for both tornadoes and hail.
- Shapley values from game theory were found to be useful for assessing individual contributions from each variable to the random forecast classification models.

Some study caveats are worthy of additional discussion. First, vertical profiles were uniformly assessed across the entire study domain and temporal record. Previous studies have examined SCS environments as a function of space and time, noting that a variety of parameters used in this study displayed geographic and/or seasonal relevance (e.g., Davies and Johns 1993; Johns et al. 1993; Brooks et al. 2003b; Gensini and Ashley 2011; Thompson et al. 2012; Sherburn and Parker 2014; Sherburn et al. 2016; Coffer et al. 2019; Gensini and Bravo de Guenni 2019; Taszarek et al. 2020). Seasonal and geographic variability was not examined in this study, however, as performing this for the current study design would lead to significant reduction in sample sizes associated with SIG events, both seasonally and geographically. Second, though the study temporal record was chosen based on previous literature, issues still likely remain in the quantification of SCS report magnitude (Tippett et al. 2015; Gensini et al. 2020). Such inconsistencies in the report data may also introduce additional variability in the skill of machine learning models. Finally, limitations do exist in the use of a reanalysis dataset as the baseline for the generation of a proximity sounding profile (Gensini et al. 2014; King and Kennedy 2019; Taszarek et al. 2021b).

There does not exist a "silver bullet" for the discrimination between severe and significant-severe events. Yet, results herein, and derivatives of this type of analysis, should aid in operational forecasting skill. Future work may specifically benefit by incorporating aspects related to convective mode using convection-allowing models (Smith et al. 2012; Thompson et al. 2012; Ashley et al. 2019; Sobash et al. 2020). Emerging techniques in data science—especially random forest algorithms—appear to be promising tools for certain diagnostic and prognostic applications in weather analysis and forecasting.

Acknowledgments. The authors wish to thank Dr. David Changnon for his feedback on early versions of this study. In addition, Dr. Aaron Hill and two anonymous reviewers also provided suggestions that greatly improved the manuscript. This research was partially funded by a Northern Illinois University Research and Artistry Award and the National Science Foundation (Award 2048770).

Data availability statement. Hail and tornado report data are open-source and can be obtained from https://www.spc.noaa.gov/wcm/. ERA5 data were downloaded from the European Centre for Medium-Range Weather Forecasts (ECMWF), Copernicus Climate Change Service (C3S) available at https://cds.climate.copernicus.eu/.

REFERENCES

Agee, E., and S. Childs, 2014: Adjustments in tornado counts, F-scale intensity, and path width for assessing significant tornado destruction. *J. Appl. Meteor. Climatol.*, **53**, 1494–1505, https://doi.org/10.1175/JAMC-D-13-0235.1.

Allen, J. T., and M. K. Tippett, 2015: The characteristics of United States hail reports: 1955–2014. *Electron. J. Severe Storms Meteor.*, **10** (3), https://ejssm.org/archives/2015/vol-10-3-2015/.

Ash, K. D., M. J. Egnoto, S. M. Strader, W. S. Ashley, D. B. Roueche, K. E. Klockow-McClain, D. Caplen, and M. Dickerson, 2020: Structural forces: Perception and vulnerability factors for tornado sheltering within mobile and manufactured housing in Alabama and Mississippi. *Wea. Climate Soc.*, **12**, 453–472, https://doi.org/10.1175/WCAS-D-19-0088.1.

Ashley, W. S., and S. M. Strader, 2016: Recipe for disaster: How the dynamic ingredients of risk and exposure are changing the tornado disaster landscape. *Bull. Amer. Meteor. Soc.*, **97**, 767–786, https://doi.org/10.1175/BAMS-D-15-00150.1.

——, S. Strader, T. Rosencrants, and A. J. Krmenec, 2014: Spatiotemporal changes in tornado hazard exposure: The case of the expanding bull's-eye effect in Chicago, Illinois. *Wea. Climate Soc.*, **6**, 175–193, https://doi.org/10.1175/WCAS-D-13-00047.1.

——, 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.

Beebe, R. G., 1958: Tornado proximity soundings. *Bull. Amer. Meteor. Soc.*, **39**, 195–201, https://doi.org/10.1175/1520-0477-39.4.195.

Blair, S. F., and Coauthors, 2017: High-resolution hail observations: Implications for NWS warning operations. *Wea. Forecasting*, **32**, 1101–1119, https://doi.org/10.1175/WAF-D-16-0203.1.

Blumberg, W. G., K. T. Halbert, T. A. Supinie, P. T. Marsh, R. L. Thompson, and J. A. Hart, 2017: SHARPpy: An open-source sounding analysis toolkit for the atmospheric sciences. *Bull. Amer. Meteor. Soc.*, **98**, 1625–1636, https://doi.org/10.1175/BAMS-D-15-00309.1.

Bouwer, L. M., 2011: Have disaster losses increased due to anthropogenic climate change? *Bull. Amer. Meteor. Soc.*, **92**, 39–46, https://doi.org/10.1175/2010BAMS3092.1.

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

Brooks, H. E., C. A. Doswell III, and J. Cooper, 1994: On the environments of tornadic and nontornadic mesocyclones. *Wea. Forecasting*, **9**, 606–618, https://doi.org/10.1175/1520-0434(1994)009<0606:OTEOTA>2.0.CO;2.

——, ——, and M. P. Kay, 2003a: Climatological estimates of local daily tornado probability for the United States.

Wea. Forecasting, **18**, 626–640, https://doi.org/10.1175/1520-0434(2003)018<0626:CEOLDT>2.0.CO;2.

——, J. W. Lee, and J. P. Craven, 2003b: The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. *Atmos. Res.*, **67**, 73–94, https://doi.org/10.1016/S0169-8095(03)00045-0.

——, A. R. Anderson, K. Riemann, I. Ebbers, and H. Flachs, 2007: Climatological aspects of convective parameters from the NCAR/NCEP reanalysis. *Atmos. Res.*, **83**, 294–305, https://doi.org/10.1016/j.atmosres.2005.08.005.

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

Changnon, S. A., 2009: Increasing major hail losses in the U.S. *Climatic Change*, **96**, 161–166, https://doi.org/10.1007/s10584-009-9597-z.

Childs, S. J., R. S. Schumacher, and S. M. Strader, 2020: Projecting end-of-century human exposure from tornadoes and severe hailstorms in eastern Colorado: Meteorological and population perspectives. *Wea. Climate Soc.*, **12**, 575–595, https://doi.org/10.1175/WCAS-D-19-0153.1.

Coffer, B. E., M. D. Parker, R. L. Thompson, B. T. Smith, and R. E. Jewell, 2019: Using near-ground storm relative helicity in supercell tornado forecasting. *Wea. Forecasting*, **34**, 1417–1435, https://doi.org/10.1175/WAF-D-19-0115.1.

——, M. Taszarek, and M. D. Parker, 2020: Near-ground wind profiles of tornadic and nontornadic environments in the United States and Europe from ERA5 reanalyses. *Wea. Forecasting*, **35**, 2621–2638, https://doi.org/10.1175/WAF-D-20-0153.1.

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.

Craven, J. P., and H. E. Brooks, 2004: Baseline climatology of sounding derived parameters associated with deep, moist convection. *Natl. Wea. Dig.*, **28**, 13–24.

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.

Darkow, G. L., 1969: An analysis of over sixty tornado proximity soundings. *Sixth Conf. on Severe Local Storms,* Chicago, IL, Amer. Meteor. Soc., 218–221.

Davies, J. M., and R. H. Johns, 1993: *Some Wind and Instability Parameters Associated with Strong and Violent Tornadoes: 1. Wind Shear and Helicity*. *Geophys. Monogr.*, Vol. 79, Amer. Geophy. Union, 573–582, https://doi.org/10.1029/GM079p0573.

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.

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

Esterheld, J. M., and D. J. Giuliano, 2008: Discriminating between tornadic and non-tornadic supercells: A new hodograph technique. *Electron. J. Severe Storms Meteor.*, **3** (2), https://ejssm.org/archives/2008/vol-3-2-2008/.

Fawbush, E. J., and R. C. Miller, 1954: The types of airmasses in which North American tornadoes form. *Bull. Amer. Meteor. Soc.*, **35**, 154–165, https://doi.org/10.1175/1520-0477-35.4.154.

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.

Fujita, T. T., 1971: Proposed characterization of tornadoes and hurricanes by area and intensity. SMRP Research Paper 91, 48 pp., https://swco-ir.tdl.org/bitstream/handle/10605/261875/ttu_fujita_000292.pdf.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Wea. Forecasting*, **32**, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

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

Gensini, V. A., and W. S. Ashley, 2011: Climatology of potentially severe convective environments from the North American Regional Reanalysis. *Electron. J. Severe Storms Meteor.*, **6** (8), https://ejssm.org/archives/2011/vol-6-8-2011/.

——, and H. E. Brooks, 2018: Spatial trends in United States tornado frequency. *npj Climate Atmos. Sci.*, **1**, 38, https://doi.org/10.1038/s41612-018-0048-2.

——, and L. Bravo de Guenni, 2019: Environmental covariate representation of seasonal us tornado frequency. *J. Appl. Meteor. Climatol.*, **58**, 1353–1367, https://doi.org/10.1175/JAMC-D-18-0305.1.

——, T. L. Mote, and H. E. Brooks, 2014: Severe-thunderstorm reanalysis environments and collocated radiosonde observations. *J. Appl. Meteor. Climatol.*, **53**, 742–751, https://doi.org/10.1175/JAMC-D-13-0263.1.

——, A. M. Haberlie, and P. T. Marsh, 2020: Practically perfect hindcasts of severe convective storms. *Bull. Amer. Meteor. Soc.*, **101**, E1259–E1278, https://doi.org/10.1175/BAMS-D-19-0321.1.

Grieser, J., and F. Terenzi, 2016: Modeling financial losses resulting from tornadoes in European countries. *Wea. Climate Soc.*, **8**, 313–326, https://doi.org/10.1175/WCAS-D-15-0036.1.

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.

Hales, J., 1988: Improving the watch/warning program through use of significant event data. Preprints, *15th Conf. on Severe Local Storms*, Baltimore, MD, Amer. Meteor. Soc., 165–168.

Hall, S. G., and W. S. Ashley, 2008: Effects of urban sprawl on the vulnerability to a significant tornado impact in northeastern Illinois. *Nat. Hazards Rev.*, **9**, 209–219, https://doi.org/10.1061/(ASCE)1527-6988(2008)9:4(209).

Hampshire, N. L., R. M. Mosier, T. M. Ryan, and D. E. Cavanaugh, 2018: Relationship of low-level instability and tornado damage rating based on observed soundings. *J. Oper. Meteor.*, **6** (1), 1–12, https://doi.org/10.15191/nwajom.2018.0601.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

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.

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

Johns, R. H., J. M. Davies, and P. W. Leftwitch, 1990: An examination of the relationship of 0–2-km agl positive wind shear to potential buoyant energy in strong and violent tornado situations. *16th Conf. on Severe Local Storms,* Kananaskis Park, AB, Canada, Amer. Meteor. Soc., 593–598.

——, ——, and P. W. Leftwich, 1993: *Some Wind and Instability Parameters Associated with Strong and Violent Tornadoes: 2. Variations in the Combinations of Wind and Instability Parameters*. *Geophys. Monogr*., Vol. 79, Amer. Geophys. Union, 583–590, https://doi.org/10.1029/GM079p0583.

Johnson, A. W., and K. E. Sugden, 2014: Evaluation of sounding-derived thermodynamic and wind-related parameters associated with large hail events. *Electron. J. Severe Storms Meteor.*, **9** (5), https://ejssm.org/archives/2014/vol-9-5-2014/.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

King, A. T., and A. D. Kennedy, 2019: North American supercell environments in atmospheric reanalyses and RUC-2. *J. Appl. Meteor. Climatol.*, **58**, 71–92, https://doi.org/10.1175/JAMC-D-18-0015.1.

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

Kumjian, M. R., and K. Lombardo, 2020: A hail growth trajectory model for exploring the environmental controls on hail size: Model physics and idealized tests. *J. Atmos. Sci.*, **77**, 2765–2791, https://doi.org/10.1175/JAS-D-20-0016.1.

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

——, ——, 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.

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.

Lundberg, S. M., and Coauthors, 2018: Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.*, **2**, 749–760, https://doi.org/10.1038/s41551-018-0304-0.

——, 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.

Markowski, P., C. Hannon, J. Frame, E. Lancaster, A. Pietrycha, R. Edwards, and R. L. Thompson, 2003: Characteristics of vertical wind profiles near supercells obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1262–1272, https://doi.org/10.1175/1520-0434(2003)018<1262:COVWPN>2.0.CO;2.

McDonald, J. R., K. C. Mehta, D. A. Smith, and J. A. Womble, 2010: The enhanced Fujita scale: Development and implementation. *Forensic Engineering 2009: Pathology of the Built Environment*, S.-E. Chen et al., Eds., ASCE, 719–728.

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

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

Moore, T. W., 2018: Annual and seasonal tornado trends in the contiguous United States and its regions. *Int. J. Climatol.*, **38**, 1582–1594, https://doi.org/10.1002/joc.5285.

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.

NCEI, 2021: U.S. Billion-Dollar Weather and Climate Disaster. NOAA/National Centers for Environmental Information, accessed March 2021, http://www.ncdc.noaa.gov/billions.

Nguyen, H. M., E. W. Cooper, and K. Kamei, 2011: Borderline over-sampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Paradig.*, **3**, 4–21, https://doi.org/10.1504/IJKESDP.2011.039875.

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.

NWS, 2010: National implementation of the use of 1-inch diameter hail criterion for severe thunderstorm warnings in the NWS. NWS, 2 pp., https://nws.weather.gov/products/PDD/OneInchHail_Oper_PDD.pdf.

Paulikas, M. J., and W. S. Ashley, 2011: Thunderstorm hazard vulnerability for the Atlanta, Georgia metropolitan region. *Nat. Hazards*, **58**, 1077–1092, https://doi.org/10.1007/s11069-010-9712-5.

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

Potvin, C. K., K. L. Elmore, and S. J. Weiss, 2010: Assessing the impacts of proximity sounding criteria on the climatology of significant tornado environments. *Wea. Forecasting*, **25**, 921–930, https://doi.org/10.1175/2010WAF2222368.1.

Rasmussen, E. N., 2003: Refined supercell and tornado forecast parameters. *Wea. Forecasting*, **18**, 530–535, https://doi.org/10.1175/1520-0434(2003)18<530:RSATFP>2.0.CO;2.

——, and D. O. Blanchard, 1998: A baseline climatology of sounding-derived supercell and tornado forecast parameters. *Wea. Forecasting*, **13**, 1148–1164, https://doi.org/10.1175/1520-0434(1998)013<1148:ABCOSD>2.0.CO;2.

Rosencrants, T. D., and W. S. Ashley, 2015: Spatiotemporal analysis of tornado exposure in five US metropolitan areas. *Nat. Hazards*, **78**, 121–140, https://doi.org/10.1007/s11069-015-1704-z.

Schaefer, J. T., and R. L. Livingston, 1988: The typical structure of tornado proximity soundings. *J. Geophys. Res.*, **93**, 5351–5364, https://doi.org/10.1029/JD093iD05p05351.

——, and R. Edwards, 1999: The SPC tornado/severe thunderstorm database. Preprints, *11th Conf. on Applied Climatology*, Dallas, TX, Amer. Meteor. Soc., 6.11, https://ams.confex.com/ams/older/99annual/abstracts/1360.htm.

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.

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

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

Showalter, A. K., and J. R. Fulks, 1943: Preliminary report on tornadoes. U.S. Weather Bureau, Washington,. DC, 162 pp.

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part I: Storm classification and climatology. *Wea. Forecasting*, **27**, 1114–1135, https://doi.org/10.1175/WAF-D-11-00115.1.

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

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

Strader, S. M., and W. S. Ashley, 2015: The expanding bull's-eye effect. *Weatherwise*, **68**, 23–29, https://doi.org/10.1080/00431672.2015.1067108.

——, and ——, 2018: Finescale assessment of mobile home tornado vulnerability in the central and southeast United States. *Wea. Climate Soc.*, **10**, 797–812, https://doi.org/10.1175/WCAS-D-18-0060.1.

——, W. Ashley, A. Irizarry, and S. Hall, 2015: A climatology of tornado intensity assessments. *Meteor. Appl.*, **22**, 513–524, https://doi.org/10.1002/met.1482.

——, W. S. Ashley, T. J. Pingel, and A. J. Krmenec, 2017a: Observed and projected changes in United States tornado exposure. *Wea. Climate Soc.*, **9**, 109–123, https://doi.org/10.1175/WCAS-D-16-0041.1.

——, ——, ——, and ——, 2017b: Projected 21st century changes in tornado exposure, risk, and disaster potential. *Climatic Change*, **141**, 301–313, https://doi.org/10.1007/s10584-017-1905-4.

——, ——, ——, and ——, 2018: How land use alters the tornado disaster landscape. *Appl. Geogr.*, **94**, 18–29, https://doi.org/10.1016/j.apgeog.2018.03.005.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, 2008: Conditional variable importance for random forests. *BMC Bioinfo.*, **9**, 307, https://doi.org/10.1186/1471-2105-9-307.

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

Tang, B. H., V. A. Gensini, and C. R. Homeyer, 2019: Trends in United States large hail environments and observations. *npj Climate Atmos. Sci.*, **2**, 45, https://doi.org/10.1038/s41612-019-0103-7.

Taszarek, M., J. T. Allen, T. Púčik, K. A. Hoogewind, and H. E. Brooks, 2020: Severe convective storms across Europe and the United States. Part 2: ERA5 environments associated with lightning, large hail, severe wind and tornadoes. *J. Climate*, **33**, 10 263–10 286, https://doi.org/10.1175/JCLI-D-20-0346.1.

——, ——, H. E. Brooks, N. Pilguj, and B. Czernecki, 2021a: Differing trends in United States and European severe thunderstorm environments in a warming climate. *Bull. Amer. Meteor. Soc.*, **102**, E296–E322, https://doi.org/10.1175/BAMS-D-20-0004.1.

——, N. Pilguj, J. T. Allen, V. Gensini, H. E. Brooks, and P. Szuster, 2021b: Comparison of convective parameters derived from ERA5 and MERRA2 with rawinsonde data over Europe and North America. *J. Climate*, **34**, 3211–3237, https://doi.org/10.1175/JCLI-D-20-0484.1.

Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close proximity soundings within supercell environments obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18**, 1243–1261, https://doi.org/10.1175/1520-0434(2003)018<1243:CPSWSE>2.0.CO;2.

——, 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.

Tippett, M. K., J. T. Allen, V. A. Gensini, and H. E. Brooks, 2015: Climate and hazardous convective weather. *Curr. Climate Change Rep.*, **1**, 60–73, https://doi.org/10.1007/s40641-015-0006-6.

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

Unidata, 2020: Metpy: A Python Package for Meteorological Data. Boulder, CO, UCAR/Unidata Program Center, accessed 10 June 2021, https://doi.org/10.5065/D6WW7G29.

Weisman, M. L., and J. B. Klemp, 1982: The dependence of numerically simulated convective storms on vertical wind shear and buoyancy. *Mon. Wea. Rev.*, **110**, 504–520, https://doi.org/10.1175/1520-0493(1982)110<0504:TDONSC>2.0.CO;2.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Wurman, J., K. Kosiba, T. White, and P. Robinson, 2021: Supercell tornadoes are much stronger and wider than damage-based ratings indicate. *Proc. Natl. Acad. Sci. USA*, **118**, e2021535118, https://doi.org/10.1073/pnas.2021535118.