A Random-Forest Model to Assess Predictor Importance and Nowcast Severe Storms Using High-Resolution Radar–GOES Satellite–Lightning Observations

JOHN R. MECIKALSKI,[a] THEA N. SANDMÆL,[b] ELISA M. MURILLO,[b] CAMERON R. HOMEYER,[b] KRISTOPHER M. BEDKA,[c] JASON M. APKE,[d] AND CHRIS P. JEWETT[e]

[a] *Atmospheric Science Department, University of Alabama in Huntsville, Huntsville, Alabama*
[b] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[c] *Science Directorate, NASA Langley Research Center, Hampton, Virginia*
[d] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, Colorado*
[e] *Earth Systems Science Center, Huntsville, Alabama*

(Manuscript received 23 August 2019, in final form 23 February 2021)

ABSTRACT: Few studies have assessed combined satellite, lightning, and radar databases to diagnose severe storm potential. The research goal here is to evaluate next-generation, 60-s update frequency geostationary satellite and lightning information with ground-based radar to isolate which variables, when used in concert, provide skillful discriminatory information for identifying severe (hail ≥ 2.5 cm in diameter, winds ≥ 25 m s^{-1}, and tornadoes) versus nonsevere storms. The focus of this study is predicting severe thunderstorm and tornado warnings. A total of 2004 storms in 2014–15 were objectively tracked with 49 potential predictor fields related to May, daytime Great Plains convective storms. All storms occurred when 1-min *Geostationary Operational Environmental Satellite* (*GOES*)-*14* "super rapid scan" data were available. The study used three importance methods to assess predictor importance related to severe warnings and used random forests to provide a model and skill evaluation measuring the ability to predict severe storms. Three predictor importance methods show that GOES mesoscale atmospheric-motion-vector-derived cloud-top divergence and above-anvil cirrus plume presence provide the most satellite-based discriminatory power for diagnosing severe warnings. Other important fields include Earth Networks Total Lightning flash density, GOES estimated cloud-top vorticity, and overshooting-top presence. Severe warning predictions are significantly improved at the 95% confidence level when a few important satellite and lightning fields are combined with radar fields, versus when only radar data are used in the random-forest model. This study provides a basis for including satellite and lightning fields within machine-learning models to help forecast severe weather.

KEYWORDS: Convective storms; Convective-scale processes; Radars/Radar observations; Satellite observations; Nowcasting; Probability forecasts/models/distribution

1. Introduction and background

Statistical models have proven to be valuable for making short-term forecasts of convective storms (Dixon and Wiener 1993; Wilson and Mueller 1993; Wilson et al. 1998; Mueller et al. 2003; Lin et al. 2012). With the advent of high-resolution satellite datasets (up to 0.5-km spacing per pixel, over 10 channels, available every ~1 min), National Weather Service (NWS) and other forecasters are challenged to integrate all information in a timely manner. Other high-resolution datasets include operational numerical weather prediction (NWP) models (~3–6-km grid spacing), ground- and space-based lightning networks (e.g., Krehbiel et al. 2000; Koshak et al. 2004; Goodman et al. 2013), and advanced Doppler-radar products (Zhang et al. 2011; Smith et al. 2016). The data-integration task frequently proves very challenging in an operational environment, often causing forecasters to rely on outdated products or methods without taking more current datasets into account, despite results from the latest research that frequently show considerable diagnostic value within individual products derived from the abovementioned datasets (e.g., Kumjian and Ryzhkov 2008; Schultz et al. 2015; Apke et al. 2018; Bedka et al. 2018). Forecasters have recently favored multi-data-source products, which presently operate using combined predictors that contain the most value for objectively identifying pending events, for example severe (or soon-to-be-severe) deep convection (Cintineo et al. 2014, 2020).

Present state-of-the-art methods that integrate a combination of weather datasets rely on raw and derived geostationary satellite parameters, gridded radar observations and derived products [e.g., the Multi-Radar Multi-Sensor (MRMS) product suite; Zhang et al. 2016], and NWP model fields. With respect to severe weather nowcasting (0–1 h forecasting), Probability of Severe Convection (ProbSevere; Cintineo et al. 2014, 2018, 2020), the MeteoSwiss Context and Scale Oriented Thunderstorm Satellite Predictors Development (COALITION; Nisi et al. 2014), and the European Organisation for the Exploitation of Meteorological Satellites (EUMETSAT) Nowcasting Satellite Application Facilities (NWCSAF) Rapidly Developing Thunderstorm (RDT; Autonès and Moisselin 2010; Gijben and de Coning 2017) systems are statistical models that perform such integration. Other models that have been developed to diagnose or forecast convective storms and related hazards use machine-learning methods for automated convective storm

Corresponding author: John R. Mecikalski, johnm@nsstc.uah.edu

DOI: 10.1175/MWR-D-19-0274.1

initiation nowcasting (Mecikalski et al. 2015), mesoscale convective complex identification (Ahijevych et al. 2016), automated overshooting top (OT) identification (Bedka and Khlopenkov 2016), hail forecasting (Gagne et al. 2017), ensemble forecasts (Gagne et al. 2014), and damaging straight-line wind diagnosis (Lagerquist et al. 2017). Machine-learning approaches have recently become popular for use with satellite remote sensing datasets (e.g., Kühnlein et al. 2014; Meyer et al. 2016; Beusch et al. 2018).

A guiding research question becomes: How useful are new research-based fields derived from 1-min resolution data for discriminating between severe and nonsevere convective storms, especially when used in a machine-learning nowcast model with many predictor fields? The specific science goals we address are the following: 1) Given use of only 5-min storm-centered radar products in a machine-learning approach, what is the skill for diagnosing severe weather? 2) Can the addition of 1-min resolution GOES satellite and lightning fields increase diagnostic or predictive skill when used in combination with radar data in a machine-learning model? 3) Which GOES-derived and lightning fields are most beneficial for severe weather diagnosis and nowcasting? The hypothesis guiding this study is that if 1-min-resolution satellite or lightning datasets provide useful information toward diagnosing and predicting severe weather, *then the addition of said fields with radar observations will make for a more skillful severe weather diagnostic and prediction system*.

To address the above question and goals, storm cells are tracked throughout their lifetime using gridded volumetric dual-polarization NEXRAD datasets (GridRad; Homeyer et al. 2017). GridRad analyses of storm microphysics and inferences of updraft and rotation intensity, *GOES-14* fields, Earth Networks Total Lightning Network (ENTLN; Rudlosky 2015) fields, and severe weather reports and human issued NWS warnings (severe thunderstorm or tornado) are linked to the storm tracks. Several methods are subsequently used to determine predictor importance beginning with 49 predictors. The most useful predictors are then used within a random-forest predictive model, and the accuracy of the model to identify storms with severe thunderstorm or tornado warnings, with and without satellite and lightning fields, is assessed. A first evaluation considered performing this study using severe weather reports, which is described further below.

It is well known that ground-based radar on its own provides a valuable dataset for diagnosing and predicting severe weather potential from convective storms (see Fabry 2015; chapter 7). However, prior to the present study, it is unclear how several new lightning- and satellite-derived products produced from 1-min resolution data can improve a machine-learning model's ability to automatically detect and predict severe storm occurrence. Until *GOES-14*, geostationary satellites had not routinely observed storms at time frequencies high enough (<5 min) to capture the often-rapid changes in cloud characteristics indicative of impending severe weather, limiting our ability to derive such new satellite fields. Hence, the novel aspects of the present study are the application of machine learning to identify severe convective storms using new nonoperational state-of-the-art radar, lightning and geostationary satellite-based fields together, as derived from 1-min resolution datasets, and determining which of these fields are most important to this forecast process. The desired outcome is to provide forecasters and developers of forecast systems increased understanding of the utility of new GOES-R-era experimental datasets designed for severe weather prediction.

2. Data

The following sections overview each dataset, and specifically how they were formed and processed into the main cell-track database as used for random-forest model development. Figure 1 shows a few of the datasets for a supercell storm near the Denver (Colorado) International Airport. When used individually, a forecaster (and similarly, a machine) can isolate the severe convection using strong divergence near a satellite-observed OT, which was associated with high lightning flash rates and strong reflectivity.

a. GridRad data and storm cell tracks

The main dataset used, as described in Sandmæl et al. (2019) and Murillo and Homeyer (2019), was a database of 49 radar, satellite, and lightning fields every minute along the cell tracks of convective storms across seven case study days listed in Table 1. The terms "storm" and "cell" are used interchangeably throughout this paper. This cell-track database consists of 1) *GOES-14* 1-min "super rapid scan" (SRS) observations and derived fields that were collected periodically from 2012 to 2015, yet cases from 2014 to 2015 were used for this study given the availability of additional satellite-based fields; 2) four-dimensional volumes of GridRad variables at 5-min resolution on each day processed; and 3) ENTLN lightning observations at 1-min resolution. The specific case study days were selected when severe weather occurred within the *GOES-14* SRS 1000 km × 1000 km Flex Mesoscale sectors, analogous to Mesoscale Domain Sectors collected by *GOES-16* and *-17* from the Advanced Baseline Imager (ABI).

NEXRAD Level II (i.e., volume) data used to derive the cell-track database were gathered from the National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Information (NCEI). All NEXRAD observations were obtained typically at 14 elevations per volume, at a range resolution of 250 m, and an azimuthal resolution of 0.5°, for the lowest three–four elevations, and 1.0° otherwise. The radar data were processed using the four-dimensional space–time merging methods described in Homeyer and Bowman (2017), which provide volumes of GridRad radar variables at 2-km horizontal, 1-km vertical, and 5-min temporal resolutions over the extent of the *GOES-14* SRS domains for all case study days. A total of 2004 cells were tracked for all study days. It is well recognized that the main database used is limited to May, daytime U.S. Great Plains severe weather events. Hence, the results to follow in terms of diagnosing and predicting severe storms may not generalize well to broader regions of the United States or to nighttime and more "pulse like" summertime severe weather events.

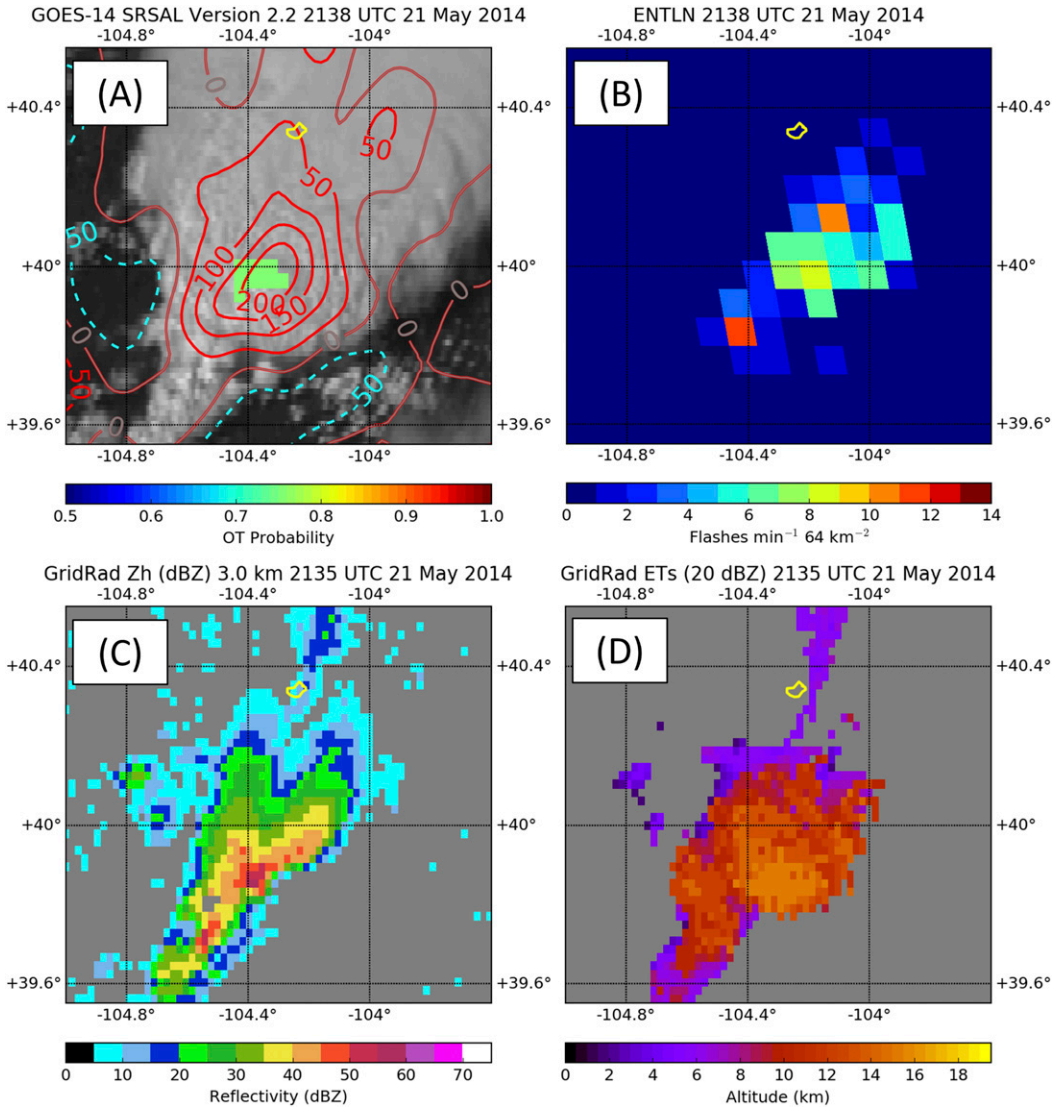

FIG. 1. Four of the main GridRad, ENTLN, and *GOES-14* datasets used in this study, specifically for 2138 UTC 21 May 2014 over central Colorado (this is just after a supercell overtook the Denver radar KFTG): (a) satellite-derived cloud-top divergence (CTD; s^{-1}) with overshooting tops (OTs; green filled pixels), (b) ENTLN flash extent density, (c) GridRad horizontal reflectivity Z_h at 3 km, and (d) GridRad 20-dBZ echo tops (ETs).

Cell tracking was performed using an automated echo-top algorithm introduced by Homeyer et al. (2017) and Bedka et al. (2018). Gaussian-smoothed local maxima in echo-top altitudes, where the horizontal reflectivity Z_H is at least 40 dBZ, are identified in each 5-min radar observation, and individual echo-top maxima are temporally pieced together if they lie within 12.5 km of each other over a period of 5 min. The 40-dBZ echo-top maxima are subsequently filtered by the convective echo classification output, as provided by the Storm Labeling in Three Dimensions (SL3D) algorithm (Starzec et al. 2017). The SL3D algorithm uses GridRad data to objectively classify radar echoes based on storm height, depth, and intensity for storm updraft classification. Tracked echo-top maxima are required to exceed 4 km above ground level altitude and occur for at least 15 min of radar analyses; hence storms with lifetimes < 15 min are not analyzed. Radar Z_H imagery of the objectively tracked storms were quality-controlled (as in Sandmæl et al. 2019), and then were used to extract maximum or minimum field values from each dataset within a 10-km radius of the storm location at 1-min intervals. Missing and unavailable radar data at resolutions > 5 min were interpolated linearly in time to the storm track locations at 1-min resolution. Once accomplished, the final cell-track database contained all variables listed in Table 2 every minute of a given storm's lifetime for all 2004 cells. Figure 2 shows tracks for the 2004 storms that were processed on the seven days examined (see Table 1). Severe Weather Data Inventory (SWDI; National Centers for Environmental Information 2017)

TABLE 1. List of case days used to form the cell-track database used in this study. The number of storm cells tracked, and the number of minutes all storm cells on a given day were tracked at 1-min time intervals, are also listed. Totals are listed in the bottom row.

Day	Cells per day	1-min cell-track times
10 May 2014	112	6977
11 May 2014	330	14 980
21 May 2014	54	2484
19 May 2015	329	18 004
24 May 2015	123	6208
25 May 2015	669	23 114
27 May 2015	387	20 463
Total	2004	92 216

severe weather reports from the NCEI and severe weather warnings (severe thunderstorm and tornado) were then linked to nearby storms in the database using the methodology described in Sandmæl et al. (2019).

The GridRad predictor variables (upper portion of Table 2) were selected to capture storm dynamical and microphysical characteristics and were derived using automated methods that could be or have previously been related to the occurrence of severe weather at the ground. The polarimetric GridRad fields afford us a quantitative ability to identify in-cloud precipitation fields as related to the main updraft and downdraft structures within severe convective storms (Klemp and Rotunno 1983; Rotunno 1993; Markowski and Richardson 2010). The more unique radar fields used in the analysis to follow include the ratio between the maximum differential reflectivity Z_{DR} and specific differential phase K_{DP}, radial velocity spectrum width (related to so-called Z_{DR} columns; Hall et al. 1980; Illingworth et al. 1987; Kumjian et al. 2014), correlation coefficient (CC) minimum in volumes where $Z_H > 45$ dBZ, hail differential radar reflectivity (H_{DR}; Depue et al. 2007), volume of hail $H_{DR} > 20$ dB, volume of hail identified by a hydrometeor classification algorithm (HCA; Park et al. 2009), maximum vertically integrated liquid (VIL) density (Amburn and Wolf 1997), maximum expected size of hail (MESH; Marzban and Witt 2001; Stumpf et al. 2004), area of MESH ≥ 1 in. (1 in. = 2.54 cm), radar-estimated radial divergence, and radar-estimated implied ascent (Kumjian and Lombardo 2017).

b. GOES-14 *super rapid scan data*

Beginning in summer 2012, *GOES-14* satellite operated periodically in "Super Rapid Scan Operations for GOES-R" mode to collect 1-min SRS imagery in preparation for GOES-R series ABI imagery (from 30 s to 1 min; Schmit et al. 2005, 2014, 2015). Use of actual *GOES-16/-17* data, along with Geostationary Lightning Mapper (GLM) lightning fields instead of ENTLN data, would extend the present study.

For this study, innovative new fields developed from *GOES-14* SRS data were demonstrated. These satellite fields include a probabilistic infrared- and NWP-based OT identification and cloud-top texture, based on spatial patterns in visible imagery (Bedka and Khlopenkov 2016), and mesoscale atmospheric motion vector (mAMV) derived cloud-top divergence (CTD) and cloud-top vertical vorticity (CTV; Apke et al. 2016, 2018). In addition, infrared temperature comparisons to the surrounding anvil and the tropopause, using the North American Regional Reanalysis (NARR; Mesinger et al. 2006) tropopause temperature, were also evaluated, which serve as another GOES metrics of updraft intensity since it relates to the depth of an OT and how far an OT may extend above the tropopause. CTD fields were found to be significantly larger for severe, deep convective storms compared to benign storms, based on a smaller sample of these data (Apke et al. 2018), and we would like to see whether CTD provides useful predictive information when used in concert with radar and lightning fields. Above-anvil cirrus plumes (AACP) were manually identified in the GOES SRS data using methods described in Bedka et al. (2018). These plumes and the associated "enhanced V" signatures have been found atop severe storms in many studies (Bedka et al. 2018, and references therein). All satellite fields were parallax corrected to spatially match the radar and lightning data; in this study, the 5-dBZ radar Z_H echo-top altitude is used to estimate cloud-top height and thus to correct for parallax.

Given the reliance on visible *GOES–14* data to create many of the above satellite fields, only daytime cases were used. Despite this requirement, >90% of the 2004 storm-cell database was available for analysis. Daytime over the U.S. Midwest is between 1300 UTC (local morning, 0700 or 0800 depending on daylight saving time) and 0100 UTC (local evening, 1900 or 2000).

In cases of weaker convective storms, some predictors were not defined during the lifetime of a radar-defined storm cell. For example, if the OT probability is zero (i.e., no OT was detected within infrared imagery), the "GOES brightness temperature (T_B) minimum–NARR tropopause temperature" (i.e., the depth of an OT relative to the local tropopause) field is set to zero or a small number to reflect a cloud-top T_B near the NARR tropopause temperature. OT probability is a logistic regression determination that an OT exists based on several satellite parameters including T_B gradients atop anvil clouds and the T_B with respect to the tropopause, with probabilities from 0% to 100% (see Bedka and Khlopenkov 2016). In the predictor importance analysis and random-forest training, all predictors were required to be available (not missing); if this was not done, a given predictor's relative importance within a random-forest model would be inaccurately described (Hapfelmeier and Ulm 2013), and similarly the random-forest model would not be trained in an accurate manner. Furthermore, in random-forest predictive modeling, missing "important" predictors will cause less accurate forecasts because, without them, the random-forest model's performance would be limited (Hapfelmeier and Ulm 2013). Thus, if the random-forest model developed in this study were run operationally, its overall predictive skill could be less than that documented here because of occasional missing predictors. The stated sizes of the training, validation and testing databases below have excluded the times with missing predictors.

c. *ENTLN lightning flash detection*

The ENTLN fields were processed into total lightning flash density, intracloud flash density, and cloud-to-ground flash

TABLE 2. List of 49 GridRad radar, ENTLN lightning, and *GOES-14* satellite predictor variables evaluated in this study. See the list of acronyms in appendix A for definitions.

Predictor	Predictor variable	Unit	Reference(s)
1–4	Z_H 10-, 20-, 30-, and 40-dBZ echo-top altitude	km	
5	Radar Z_H column max	dBZ	
6	Max Z_{DR}/K_{DP} column altitude	km	Kumjian et al. (2014); Homeyer and Kumjian (2015)
7	Spectrum width column max	m s^{-1}	Zrnić and Doviak (1975); Zrnić et al. (1985)
8	Spectrum width 1–3-km max	m s^{-1}	Zrnić and Doviak (1975); Zrnić et al. (1985)
9	CC min where $Z_H \geq 45$ dBZ	Unitless	Picca and Pyzhkov (2012)
10–11	Z_{DR} min and median where $Z_H \geq 45$ dBZ	dB	
12	H_{DR} column max	dB	Depue et al. (2007)
13	Volume of $H_{DR} \geq 20$ dB	km^3	
14	Area of SL3D-classified convection	km^2	Starzec et al. (2017)
15	Volume fraction of HCA-classified hail	Unitless	Park et al. (2009)
16	Volume of radar echo	km^3	
17	Area of HCA-classified hail at 3 km	km^3	
18	Depth of HCA-classified hail	km	
19	Max VIL density	g m^{-3}	Amburn and Wolf (1997)
20	Area of VIL density > 1.5 g m^{-3}	km^2	
21	Max expected size of hail maximum	cm	
22	Area of MESH > 2.5 cm	km^2	
23	GridRad vorticity column max	$\times 10^{-3}$ s^{-1}	
24–26	GridRad vorticity 1–3-km, 4–7-km, and 8+-km max	$\times 10^{-3}$ s^{-1}	
27	GridRad divergence column max	$\times 10^{-3}$ s^{-1}	
28	GridRad divergence 8+-km max	$\times 10^{-3}$ s^{-1}	
29	GridRad divergence 1–3-km min	$\times 10^{-3}$ s^{-1}	
30–31	GridRad median and max implied ascent	m s^{-1}	
32	GridRad implied ascent area	km^2	
33	ENTLN total lightning flash density	Count	
34	ENTLN intracloud lightning flash density	Count	
35	ENTLN cloud-to-ground lightning flash density	Count	
36	GOES overshooting top detection	Unitless	Bedka and Khlopenkov (2016)
37	GOES max overshooting top probability	Percent	Bedka and Khlopenkov (2016)
38	GOES overshooting top area	km^2	Bedka and Khlopenkov (2016)
39	GOES anvil detection	Unitless	Bedka and Khlopenkov (2016)
40	GOES IR T_B min–NARR tropopause temperature	K	
41	GOES IR T_B difference OT and anvil	K	
42	GOES max visible texture detection rating	Unitless	Bedka and Khlopenkov (2016)
43–44	GOES mAMV CTV max and mean	$\times 10^{-4}$ s^{-1}	Apke et al. (2016)
45–46	GOES mAMV CTD max and mean	$\times 10^{-4}$ s^{-1}	Apke et al. (2016)
47	GOES mAMV area where CTD $>15 \times 10^{-4}$ s^{-1}	km^2	Apke et al. (2016)
48	GOES mAMV wind magnitude of mean flow	m s^{-1}	Apke et al. (2016)
49	Above-anvil cirrus plume	Unitless	Bedka et al. (2018)

density, as described in Sandmæl et al. (2019). First, lightning sources close together in space and time (1 km and 100 ms) are grouped into flashes, and then they are binned into 0.08° × 0.08° longitude–latitude flash density grids (Goodman et al. 2013), which was done to mimic the spatial resolution of data provided by the *GOES-16* and *-17* GLM instrument. These 100-ms point data flashes were integrated to 1-min times within the ~10-km (0.08°) boxes. Second, the 1-min spatial maximum of the total ENTLN lightning flash density was extracted over a radius of 10 km along each storm track over time and subsequently used to populate the cell-track database. Rapid increases in lightning over 5–10-min time periods, the so-called lightning jump as described in Schultz et al. (2011, 2015), have been correlated to increases in severe weather. Although lightning jumps specifically are not included in the random-forest framework here, we acknowledge any work using ENTLN or GLM lightning jump fields within the machine-learning model framework are avenues for future research.

d. SWDI

The NCEI hosts the SWDI storm event database that contains the time, duration, location, magnitude, and source of all confirmed U.S. severe weather reports. Severe weather is the occurrence of severe winds (≥25 m s^{-1}), large hail (≥2.5 cm in diameter), or tornadoes. The SWDI-based reports were processed into the cell-track database using the method of Sandmæl et al. (2019). Although the NCEI SWDI database provides the most comprehensive account of historical severe weather events in the United States, well-established reporting biases (e.g., population density) can influence severe weather-storm report relationships (Doswell et al. 2005; Trapp et al. 2005, 2006; Verbout et al. 2006; Brotzge et al. 2011).

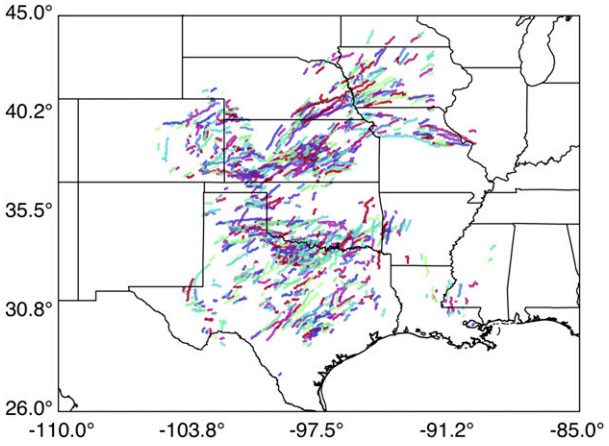

FIG. 2. Cell-track map for the 2004 storms that were used in this study, when storms lasted at least 30 min, for the days as listed in Table 1. The variation in line colors is to improve the visual interpretation of overlapping storm tracks.

Specifically, severe weather reports tend to be biased high in regions with higher population, simply because more people are likely to experience and report severe weather. In contrast, low biases in storm reports occur in regions with very low population.

Early efforts focused on use of severe weather reports as the ''predictand'' or response variable within the random forests' predictor importance and model training analysis. However, these analyses were made difficult when instantaneous point-based reports were used to represent the time continuum of severe weather events, for example, a single hail report at one time might be issued for a hailstorm that lasted for ~1 h. The choice was therefore to conduct our analysis using NWS severe thunderstorm and tornado warnings, representing an expert-generated dataset of the expected timing and locations of severe weather. Verification of random-forest model prediction skill is made against severe thunderstorm and tornado warnings, while nonstatistical comparison to NCEI SWDI reports is shown later to demonstrate utility of the predictions.

e. NWS severe thunderstorm and tornado warnings

All NWS warnings included issuance and expiration times, and polygon coordinates outlining the warned area, and were obtained from Iowa State University (Iowa Environmental Mesonet 2017). If a given storm track crossed a warning polygon, the warning was categorically (severe thunderstorm or tornado) documented at all 1-min valid times along a track in the cell-track database. Warnings as issued by the NWS are imperfect, while we felt that use of warnings was preferred to use of reports for the reasons already stated. It is important to note that forecasters use radar as a primary tool for severe weather warning operations (e.g., Fig. 1), so we expect the importance analysis to return relatively high importance for some common radar derived fields (e.g., Z_H and MESH). However, knowledge of the convection environment, local ground reports, and regional experience can also influence warning issuance, which sometimes deviates from radar field interpretations. It is here that additional datasets (e.g., satellites and lightning) could add value in diagnosing severe thunderstorms over radar analysis alone.

3. Method

Every minute in the database was classified as a ''severe'' (a warning was in effect) or ''non-severe'' (no warning was in effect) event. Thus, each minute of a storm lifetime is considered an ''event'' for random-forest training and variable importance analysis, which constitutes 92 216 separate events. The set of predictor variables used is shown in Table 2, and 523, 143, and 75 of the cells in the training, testing, and validation databases, respectively, had associated severe weather warnings, while 879, 259, and 125 of the cells, respectively, were unwarned (and assumed nonsevere) convective storm cells.

This study initially examined predictor importance and random-forest model development based on a single type of severe weather (e.g., hail only and wind only). In the end, it was found that random-forest model training was made very difficult because similar predictor values were often found for storm cells both with and without hail or wind reports in close time proximity. The net effect was very poor random-forest-based analyses with nearly no prediction skill. Specifically, when wind-only ±5-min time-padded SWDI reports were used, hit rates (HRs) were 0.178–0.259 and false alarm ratios (FARs) were from 0.625 to 0.655; the skill for forecasting hail-only reports was even lower with HRs of 0.036–0.071 and FARs of 0.792–0.816. When ±10-min time-padded reports were used for wind-only events, the above scores improved to 0.256 and 0.495 for HR and FAR, respectively; for hail-only events, the HR and FAR scores are 0.405 and 0.650, respectively. [All skill scores used in this paper are defined in appendix B]. The reason for these poor HR and FAR scores when time-padding reports were used is because of the artificial expansion of the hail and wind events into times when severe hail and winds were not occurring. Random-forest model training was thus not done in a manner that related predictor fields to severe weather reports in any meaningful way. In the case of more-long-lasting tornadoes, predictor importance analysis and random-forest forecasts for ±5-min time-padded reports yielded better results, yet HR and FAR scores were only 0.360 and 0.643, respectively. Because of these complications, the more-time-continuous predictand of severe warnings was used (typically lasting 20–40 min).

The MATLAB 9.5 (release R2018b) ''treebagger'' software package (MATLAB 2018a) was used to train the random-forest algorithm and make the predictions. In random forests, the choice was to use 100 trees since in early assessments the improvement in model performance [in terms of critical success index (CSI) skill] did not increase beyond ~90 trees.

There are many parameters that could be adjusted in random forests, such as the cost of classifying an example as class j if the true class is i (default value is 1 if $i \neq j$ and 0 if $i = j$), the number of predictor variables to loop through at each split node (default is all), and the minimum number of examples at a leaf node (default is 5). The default settings for each of these parameters were used on the basis of the documentation for

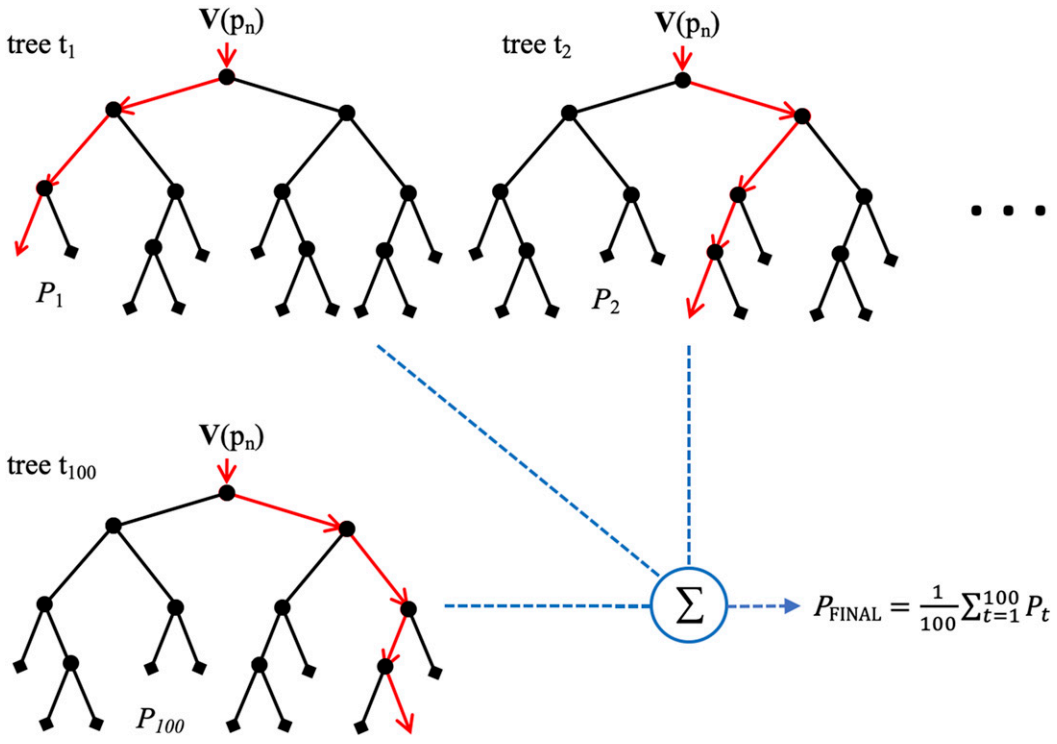

FIG. 3. Schematic diagram illustrating how random-forest decision trees are used to form a final prediction probability P_{FINAL}. Each decision tree $(t_1, t_2, \ldots, t_{100})$ produces a separate "vote" or outcome that is based on use of a vector of n predictors $V(p_n)$, which are in the form of probabilities P $(P_1, P_2, \ldots, P_{100})$ from 0% to 100%. Once all trees are formed, the 100 probabilities are summed and averaged to form P_{FINAL}.

MATLAB 9.5 R2018b. Random forests are an ensemble of individual decision trees that is used to predict a desired outcome (in this case, the occurrence of a severe weather warning), and uses "bagging/bootstrapping" and "subspace sampling" across the forest of trees. Bagging is done for each tree in the ensemble of decision trees as trained on a random "bootstrapped" partial sample of the training database, while keeping all predictor fields. Bagging/bootstrapping is thus resampling with replacement, drawing M examples from a dataset of M examples. This randomized training procedure guarantees that the decision trees are unique; individual decision trees often overfit with large biases, while a large set of diverse trees should have roughly offsetting biases such that the ensemble of trees would have an overall low bias. Each tree (100 in this case) in the forest determines a class prediction or "vote" on the outcome, and the vote counts can be calibrated into reliable probability forecasts, as shown schematically in Fig. 3.

Predictor importance was done using the training database given that it was larger than the validation database (as described below), which is expected to provide more statistically significant results. Within random-forest predictor importance is so-called Gini-based, meaning that the decision to split at each node of a decision tree (while it is being constructed) is based on a Gini impurity (GI) measure. GI is one way to evaluate the importance of a set of predictor variables as averaged over N_t trees in the forest ($T = 100$ trees in this study). The relationship is given (based on Louppe et al. 2013):

$$\text{GI} = \frac{1}{N_t}\sum_{T}\sum_{i=1}^{L} p_i(1 - p_i).$$

The parameter L is the number of possible class labels, which is set to 2 for this study, where p_1 is the percentage (or probability) of severe-warned events and p_2 is the percentage (or probability) of unwarned events. When using GI to measure impurity this is known as the Gini importance. GI is used in decision tree algorithms to both determine the optimal split from a root node as a tree is grown and to determine additional node splits. GI tells us the likelihood or probability of an incorrect classification of an event when using a given predictor. In random forests, the lower the GI for all predictors across all N_t trees is, the lower is the chance of an event being misclassified (Tan et al. 2005; chapter 4). For each predictor variable, the sum of the "Gini impurity decrease" is computed for all trees of the forest every time a given variable is chosen to split a decision tree node. The scale of predictor importance is irrelevant, while instead the relative importance magnitudes between predictors are most relevant. For a given predictor variable, the smaller the GI, the more that variable contributed to the decision to use it to split a growing decision tree and create a new node.

Three other predictor-ranking-importance methods were used: permutation feature selection (PFS; Breiman 2001) and sequential selection [forward (SFS) and backward (SBS)], with both implementations done in MATLAB 9.5 R2018b

(MATLAB 2018b,c). McGovern et al. (2019) provides an explanation of PFS, SFS, and SBS methods, which are briefly summarized here. In PFS, the importance of a predictor is measured by computing the increase in the model's prediction error after randomly permuting the predictor set, or specifically permuting predictors one variable at a time and measuring changes in model performance when a comparison is made to the unpermuted data. If after shuffling a predictor x^f the model error increases, a predictor is determined to be ''important,'' which implies that the model relied on the predictor for a prediction. In contrast, if after shuffling model error remains essentially unchanged, x^f is deemed ''unimportant'' because the model ignored x^f for the prediction. The main purpose of using PFS is to determine how model performance degrades when the statistical relationship between a given predictor and the predictand is purposely broken.

SFS is a greedy search algorithm, which means that it sequentially adds in an x^f that decreases the deviance $D(X_k + x^f)$ when combined with the predictors X_k that have already been selected. Deviance D is a generalization of the residual sum of squares. In SFS, sequential selection is incorporated into the model-training procedure, whereas the permutation test in PFS is applied to an already-trained model, which is the main difference between the two selection tests. SFS begins with a climatological model (for which forecast severe-weather probability is always the frequency over the training data), with a predictor being added per iteration provided that the change in D is more than the change expected from random chance, based on a chi-squared distribution with one degree of freedom (3.8415). SBS, in contrast, begins with a model containing all predictor variables; this is the model one would train by default if one were not concerned with predictor importance. SBS removes an x^f if the increase in D is less than 3.8415.

In this study, SFS yielded a smaller predictor set (below the 49 total predictors in Table 2), while the SBS solution included all 49 variables. SBS retained all predictors because the D increase was less than the D change caused by random chance, even if any predictor was removed. Retaining all predictors in SBS, while some were dismissed in SFS, suggests that some variables have predictive power only in combination, that is, only if certain other variables are included as well. Because the SBS results to follow are therefore no different than when all 49 variables are included in the random-forest model, SBS is not discussed further. Despite its use, a drawback of SFS is that it will not remove a predictor previously selected if an older predictor becomes obsolete because of redundancy (Rückstieß et al. 2011). Given use of PFS and SFS, in addition to the GI approach, as the predictor reductions yielded generally similar results regardless of the method (see below), the determination was made to not explore other predictor selection methods, such as multipass permutation or partial-dependence plots (see McGovern et al. 2019).

Related to correlated predictors, for GI it has been shown that when two predictors are correlated and deemed to be important, a duplicated predictor lowers the importance of the original predictor, and their importance will tend to be equivalent (Strobl et al. 2008). When two predictors are correlated, the PFS will rank neither as important (McGovern et al. 2019). For SFS, if a set of two or more variables jointly are important, but individually those variables are not important for the predictand, then they might not be selected, explaining why SFS may yield a smaller set of predictors than SBS.

Of the entire database, 70% of randomly selected complete storm tracks across the 2014 and 2015 database were used to assess predictor importance and to train the random-forest model, 10% of randomly selected complete storm tracks were used for a validation database (as noted above), and the remaining 20% of complete storm tracks were used to test the random-forest predictions, with no overlap in these three samples. Hence, time steps from a given storm were not split and are all in one of the three databases, which ensures independence at the storm level. Therefore, of the 92 216 total events, 64 551 events composed the training database, 9222 events composed the validation database, and 18 443 events made up the testing database. Storm tracks in both years compose all three databases. For the 49-predictor GI predictor importance evaluation, one variable reduction step was conducted. Predictors were manually removed during the GI evaluations if their importance was below the median importance value. For the PFS importance approach, when one importance value was extremely high (see results for details) variables were manually removed if they fell below the 25th-percentile importance value. The predictor reduction steps were done to identify a reduced set of more important predictors for use within a predictive random-forest model. Similar variable reduction approaches have been done in a wide range of studies (Evans and Cushman 2009; Rehfeldt et al. 2012; Hill et al. 2013). Following the predictor importance analysis using the 70% training database, the probability threshold (that which maximizes CSI) used to evaluate model performances was determined using the validation database, and the random-forest models for diagnosing severe warnings were evaluated with the 20% testing database.

Contingency-table values were computed to measure prediction skill (Wilks 2011, 260–275; see Table 3). A ''hit'' is a random-forest model forecast of severe warning conditions occurring along a given cell's track, whereas a ''miss'' is a forecast of no severe warning conditions (see Table 3 for definitions of the other two forecast categories). In this analysis, a random-forest probability of 48% was chosen to maximize the CSI on the testing data, as determined from the validation database. Using the validation database, the maximum CSI where the model biases were near 1.0 were determined using 500 randomly selected sets of size ''0.75 × validation database'' (0.75N, or 6917) with replacement for the three top-performing models (32Radar, 12PFSRadar, 17GIRadSatLight), as shown in Figs. 4–6 . The 0.75N size was used to avoid oversampling the validation database. Forecast probabilities ≥ 48% indicate a likelihood of severe weather, and those below 48% indicate nonsevere conditions.

We evaluated models trained on full predictor sets, reduced sets of important GridRad radar fields alone, and predictor sets composed of mixed important *GOES-14* SRS-derived and ENTLN lightning fields along with radar data (related to science question 2 in section 1). Six common scores were used: (i) CSI, (ii) FAR, (iii) HR (also known as probability of detection), (iv) Heidke skill score (HSS), (v) Peirce skill score

TABLE 3. Example contingency table, as used to develop the skill scores used in this study.

Event forecast	Event observed: Yes	Event observed: No	Equation
Yes	a	b	$a + b$
No	c	d	$c + d$
	$a + c$	$b + d$	$a + b + c + d = n$

(PSS) (also known as Hanssen–Kuiper skill score or true skill statistic; Manzato 2007), (vi) probability of false detection (POFD), and (vii) frequency bias. The area under the receiver operating characteristic (ROC) curve (AUC) was also computed to compare model performance with different predictor sets (Wilks 2011). For FAR, HR, and POFD the range of values is 0–1. The HSS ranges from $-\infty$ to 1, and the PSS (which is not a true skill score) ranges from -1 to 1. The bias ranges from 0 to ∞ in this study. Optimal values for each score are a FAR of 0, HR of 1, HSS of 1, PSS of 1, POFD of 0, bias of 1, and AUC of 1 (Wilks 2011).

4. Results

Prior to our analysis, an evaluation of SWDI-based severe weather in the storm cell-track database was done (see Table 4). From Table 4, severe weather of any type occurred $<$ 10% of the time within training and testing databases. No severe reports were associated with 84% of the training database, and hence this is not a severe weather-dominated database. The Table 4 data are different than the severe weather warning statistics presented above at the beginning of section 3.

a. Predictor importance evaluation

Figure 4 presents predictor importance results for all 49 predictors, using the GI, PFS and SFS importance methods. Again, predictor importance ranking values are unitless and represent a relative measure of importance (Tan et al. 2005; Louppe et al. 2013). For model 5SFSRadSat, the first five predictors chosen in the SFS selection method were used to form this predictor set, and for the other two models the predictors above the median importance (17GIRadSatLight) and 75% (10PFSRadSat) importance values were used to form these predictor sets. From Fig. 4, across all three predictor importance evaluations, the most consistently important fields are: 40-dBZ echo-top altitude, spectrum width 1–3-km maximum, volume of $H_{DR} > 20$ dB, CTD $> 15 \times 10^{-4}$ s^{-1} (hereinafter CTDA$_{15}$) and AACP presence. The area of VIL density > 1.5 g m^{-3} and GridRad divergence 8+-km maximum are also important, while the ENTLN total lightning flash density was found to be important only when GI is used. GOES mAMV CTV maximum was an important satellite field when

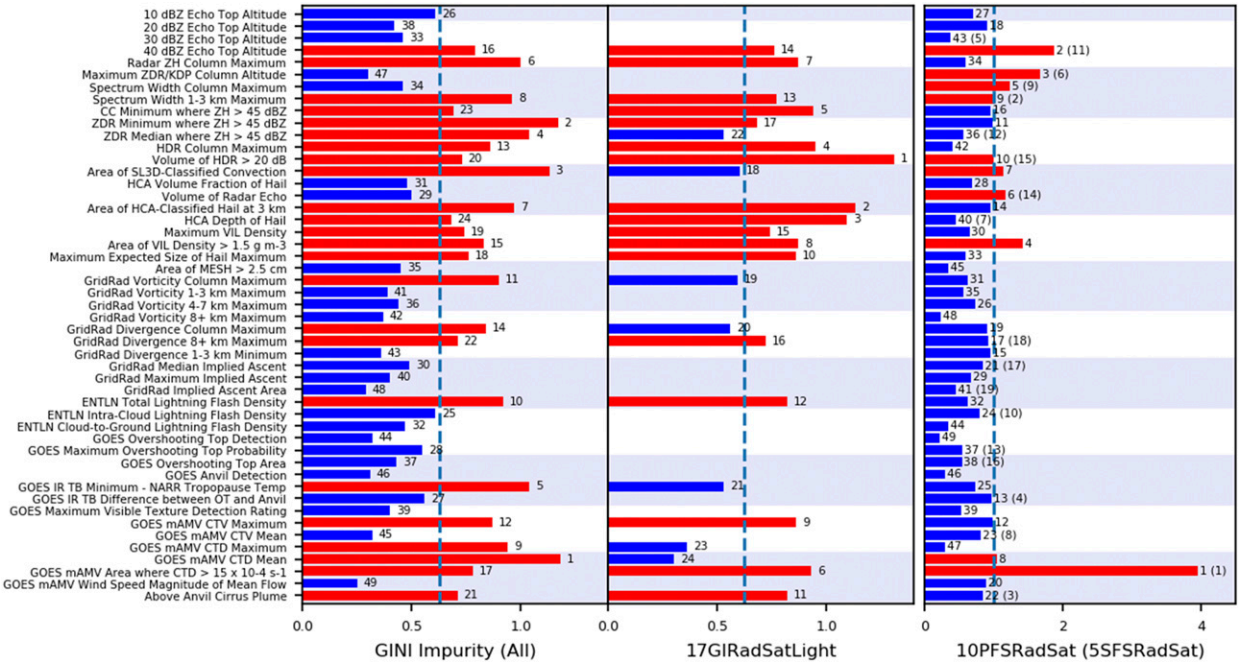

FIG. 4. Predictor importance (unitless) on the training data. (left) Gini impurity for all predictors. (center) Importance after one reduction step. Red and blue bars respectively denote where predictor importance is above and below the median for the given set. (right) Permutation feature selection (PFS) and sequential feature selection [forward (SFS), in parentheses] were used to identify important fields. (right) The red bars denote fields that are above the 75th percentile of 1.0. For model 5SFSRadSat, the first five predictors chosen in the SFS selection method were used to form this predictor set, and for the other two models the predictors above the median importance value (17GIRadSatLight) and 75% importance value (10PFSRadSat) were used to form these predictor sets. Numbers next to each bar denote the importance ranking per column, and parenthetical numbers for the SFS selections denote the order in which fields were included in that importance ranking method.

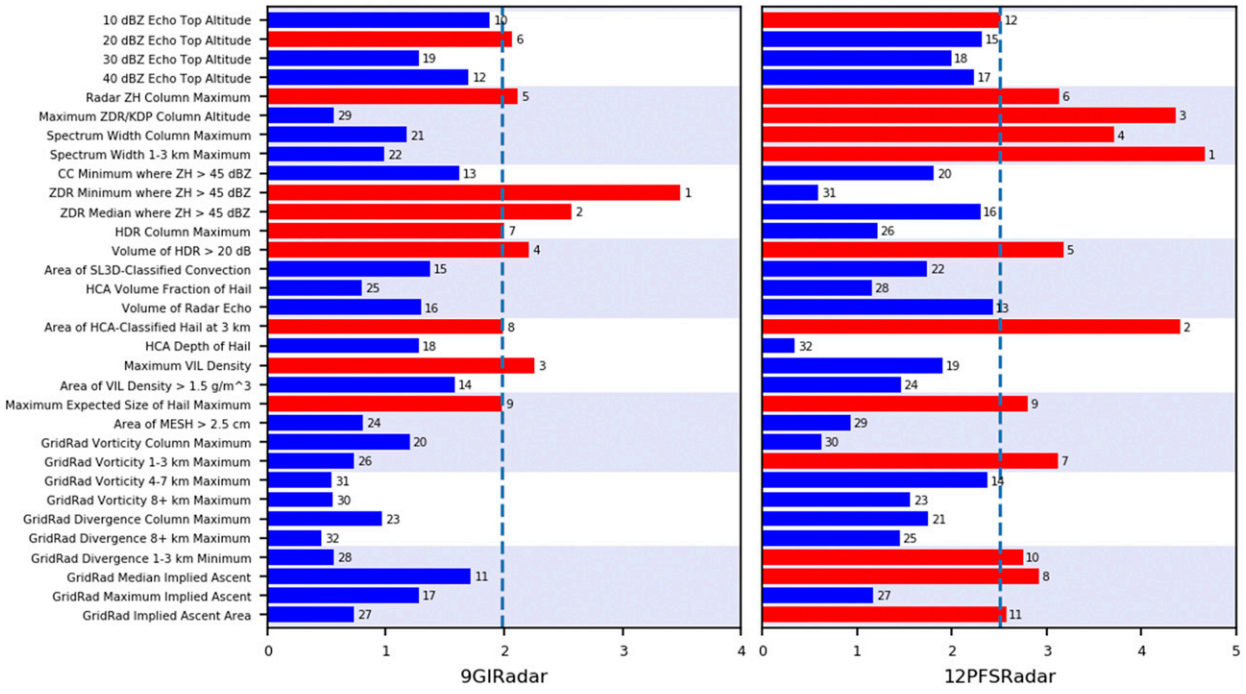

FIG. 5. As in Fig. 4, but with predictor importance (unitless) relative to severe thunderstorm and tornado warnings for the 32 radar predictors of the full predictor list is in Table 2. All fields in which the predictor importance is above the median values are denoted as red bars, which are the fields that compose the predictor sets for the 9-field model 9GIRadar and the 12-field model 12PFSRadar. As in Fig. 4, PFS refers to permutation feature selection. Numbers next to each bar denote the importance ranking per column.

GI was used. ENTLN total lightning flash density significantly helped to discriminate between warned and unwarned events yet was relatively less important when used along with 16 other radar and satellite fields (see model 17GIRadSatLight below) versus when used together with all other predictors.

The GridRad 40-dBZ echo top implies a deep convective updraft (enough to loft a high density of large detectable hydrometeors to high altitudes), while the $CTDA_{15}$ and AACP also imply the presence of a storm with a strong updraft, often a supercell according to the analysis of Bedka et al. (2018). The spectrum width 1–3-km maximum, $H_{DR} > 20$ dB, and VIL density >1.5 g m^{-3} are indicative of large hail, melting hail, and intense precipitation within a deep convective updraft. The ENTLN total lightning flash density is related to severe weather in a manner as described by Schultz et al. (2015, 2017), specifically to large amounts of mixed-phase precipitation particles within a storm's main updraft that generate high flash rates. Determining the top two–three fields by averaging the results of the three importance measures, $CTDA_{15}$, the AACP, and spectrum width 1–3-km maximum are most correlated with severe warnings. Since the 40-dBZ echo top was used in tracking, this may have led to its high importance in this predictor importance evaluation. The 40-dBZ echo top field also helps identify storms with strong updrafts that can loft significant hydrometeors (e.g., large raindrops and graupel) to high altitudes, enough to cause substantial in-cloud charge generation and large hail formation.

Predictor importance results for only radar fields (Fig. 5) were obtained using GI and PFS, while the SFS approach selected all 32 fields and hence provided no new information. Figure 5 shows that within the overall 32 GridRad fields the three most important fields are the volume of $H_{DR} > 20$ dB, area of HCA-classified hail at 3 km, and MESH. However, the predictor importance method used led to considerable differences in which fields were most important, with GI suggesting that Z_{DR} minimum where $Z_H \geq 45$ dBZ was the most important, compared to the spectrum width 1–3-km maximum when PFS was used. All fields are correlated to convective storms containing wide, strong updrafts supportive of long-term hail production, and large hail is most likely to occur in storms that have associated severe thunderstorm and tornado warnings. Figure 5 lists other fields of higher importance as found by PFS, all similarly related to the presence of hail and rotating updrafts in a storm (maximum Z_{DR}/K_{DP} column altitude, spectrum width column maximum, GridRad vorticity 1–3-km maximum, and GridRad median implied ascent).

When satellite and lightning fields are considered alone using the three importance methods (Fig. 6), $CTDA_{15}$, AACP, GOES mAMV CTD mean, and GOES T_B minimum–NARR tropopause temperature are the four most important fields. All four fields imply a tall OT and sustained updraft, and in the case of the AACP, a sufficient upper-tropospheric storm relative wind shear environment supportive of gravity wave breaking (Homeyer et al. 2017). Several of these important fields are consistent with the results discussed related to Fig. 4. The selection of these predictors confirms the prior studies by Apke et al. (2016, 2018) for CTD and CTV, and Bedka et al. (2018) for the AACP, as related to severe weather–producing

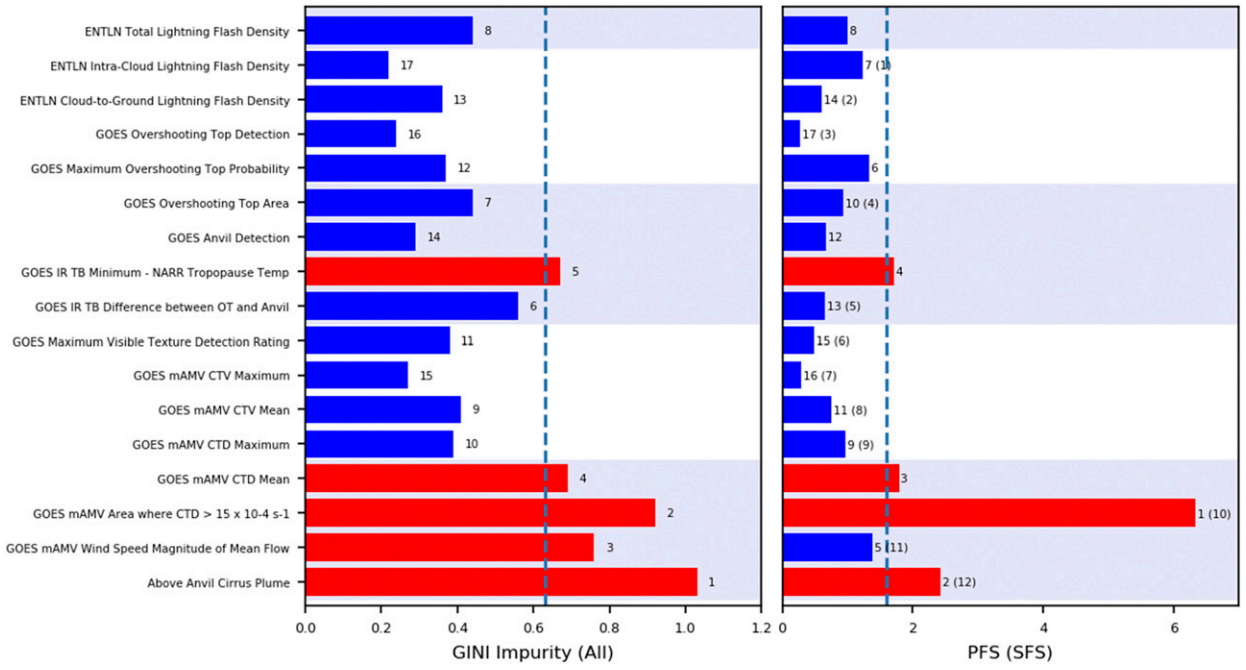

FIG. 6. As in Figs. 4 and 5, with predictor importance rankings (unitless) relative to severe thunderstorm and tornado warnings (the "predictand") for the 17 satellite and lightning predictors of the full predictor list is in Table 2. Importance values using (left) Gini impurity (GI) importance and (right) PFS and SFS were used to identify important fields. The red bars in the left column are above the median importance value, and the red bars in the right column denote fields that are above the 25th percentile. Numbers next to each bar denote the importance ranking per column, and parenthetical numbers for the SFS selections denote the order in which fields were included in that importance ranking method.

storms. Bedka et al. (2018) found that AACPs occur ~30 min in advance of severe weather at the ground, and some storms were found to continue to generate AACPs after producing severe weather and weakening, analogous to mesocyclones persisting in supercell storms after a tornado has ended. Hence, AACP presence could artificially extend a random-forest model warning prediction. In our database, 113 storms had AACPs, while 21 of these had associated tornado reports, 68 had associated large hail reports, 18 had associated high wind reports, and 43 were not associated with any severe weather. Also, 98 of the AACP storms occurred during a severe warning, while 15 AACP storms did not.

b. *Skill of different predictor sets*

Eight random-forest models composed of eight unique predictor combinations were evaluated. Figure 7 summarizes Figs. 4–6, listing the predictors used in five random-forest models: 9GIRadar, 17GIRadSatLight, 12PFSRadar, 10PFSRadSat, and 5SFSRadSat. Two additional models with no predictor reductions (all 49 fields—49AllPredictors, and all 32 radar-only fields—32Radar) were evaluated, along with another when the 32 radar fields were combined with two most important satellite fields ($CTDA_{15}$ and AACP; model 34RadSatLight).

Figure 8 shows ROC curves for the three random-forest models 32Radar, 10PFSRadSat and 17GIRadSatLight, with the 95% confidence intervals shaded per curve, which were formed using 500 randomly selected sets of size "0.75 × testing database" (0.75*N*, or 13 832) with replacement for the three models. The 0.75*N* size was used to avoid oversampling the testing database. Summarizing Fig. 8: 1) The three models are comparable, yet highest AUC of 0.78 is found for model 10PFSRadar. The 10PFSRadar random-forest model is composed of a mix of radar, lightning and satellite fields (CTD mean and $CTDA_{15}$; Fig. 7); 2) the second most skillful model is 17GIRadSatLight with an AUC = 0.75, which includes CTD maximum, $CTDA_{15}$, and AACP presence; and 3) use of only radar predictors slightly diminished skill, given the 32Radar model's AUC of 0.73. A main conclusion here is that satellite and lightning data in model 10PFSRadar add skill over the radar-only model; however, since the confidence intervals are largely overlapping, this skill increase may not be statistically significant (which is discussed further below).

TABLE 4. The percentage of times that large hail (>2.5 cm), strong winds (>25 m s^{-1}), and tornadoes occurred for each storm cell in the training, testing, and validation databases, as well as in the entire database. Percentages were compiled using the NCEI SWDI storm report dataset.

	Severe weather type			
	Hail	Wind	Tornadoes	No. of cells
Training	7.4%	5.8%	1.8%	1402
Testing	9.2%	2.2%	3.5%	402
Validation	6.5%	5.5%	1.5%	200
Entire	7.7%	5.0%	2.1%	2004

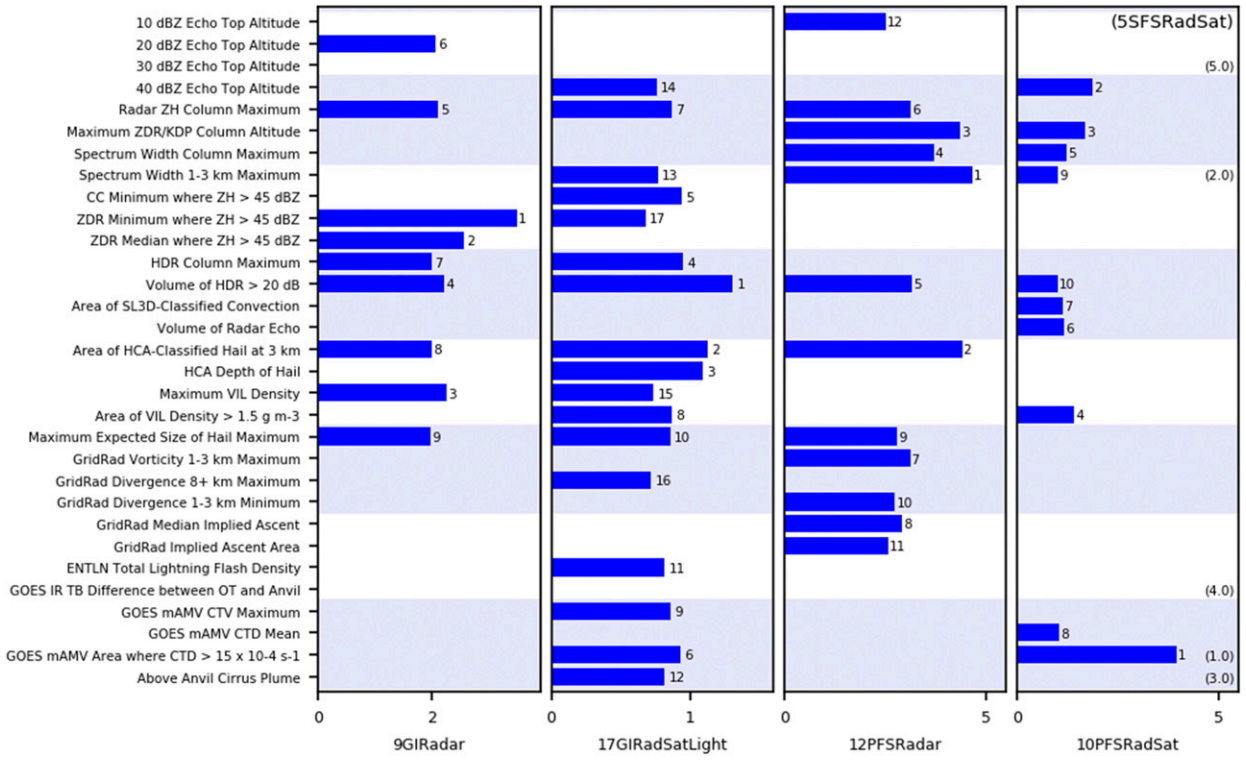

FIG. 7. Description of predictors used in the five random-forest predictor sets defined in Figs. 4 and 5. Shown are importance values (unitless) as determined using GI, PFS, and SFS. Numbers next to each bar denote the importance ranking per column.

As another way of assessing model skill, CSI and HSS scores are compared for the eight models (predictors in Figs. 4–7) in box-and-whisker plots in Figs. 9a and 9b. Like the ROC curves, the box-plot distributions are formed using 500 randomly selected sets of size 0.75N with replacement (13 832) for all eight models. For the analysis of Figs. 9a and 9b, a 48% random-forest probability was chosen as it maximized the CSI score for every random forest. Figures 9a and 9b exemplify further that when a combination of radar, lightning, and satellite fields is used in a random-forest predictive model the CSI scores are highest, whereas HSS scores for the 32Radar model are much higher than those from the 17GIRadSatLight model, showing that model 17GIRadSatLight provides better deterministic forecasts of severe and nonsevere conditions. With the HSS values being much higher for model 32Radar, it performs better relative to standard forecasts expected to verify based on chance. From Figs. 9a and 9b, the mean CSI and HSS values for model 17GIRadSatLight peaked at 0.747 and 0.307, respectively, and the 32Radar model produced the highest HSS of all models at 0.477 yet had a lower CSI of 0.64. Model 10PFSRadSat produced a mean CSI of 0.719 and a mean HSS of 0.27. The GI and PFS importance methods identified sets of predictors with the highest model predictive CSI skills. From a visual comparison of the 17GIRadSatLight and 10PFSRadSat models to the 32Radar model in Figs. 9a and 9b, it is less clear whether these two models are significantly better than 32Radar. When the CSI differences (32Radar–10PFSRadSat, 17GIRadSatLight–32Radar, and 17GIRadSatLight–10PFSRadSat) are computed using 500 bootstrapped CSI values per model, there is no overlap in the 17GIRadSatLight–32Radar and 32Radar–10PFSRadSat CSI-difference distributions. The nonoverlapping distributions show that the p values are < 0.05, demonstrating that the 32Radar model is significantly less skillful than the 17GIRadSatLight and 10PFSRadSat models (i.e., we reject the null hypothesis that the two model are equivalent). For the 17GIRadSatLight–10PFSRadSat CSI-difference distribution, the percentile of zero difference is 0.0762 (i.e., the p value), meaning that we cannot reject the null hypothesis since the distributions are not statistically different at the 95% confidence interval.

Several other conclusions can be drawn from Figs. 9a and 9b. First, when the $CTDA_{15}$ and AACP fields were included with all radar fields (model 34RadSatLight), CSI and HSS scores were substantially lower relative to model 32Radar yet were much higher than when all 49 variables were used. This behavior indicates that some of the 13 radar fields shared by the 17GIRadSatLight and 34RadSatLight models do not correlate well in space and time to severe warnings, which leads to inaccurate decision tree-based predictions. Specifically, since some of the important radar fields observe severe weather below cloud top (e.g., area of HCA-classified hail at 3 km) tens of minutes before satellite-derived fields (e.g., the AACP) can infer severe weather, a classification method using both types of fields together will not provide the best severe warning classification. Second, with respect to CSI, the 17GIRadSatLight model with mixed predictor sets was significantly more skillful than

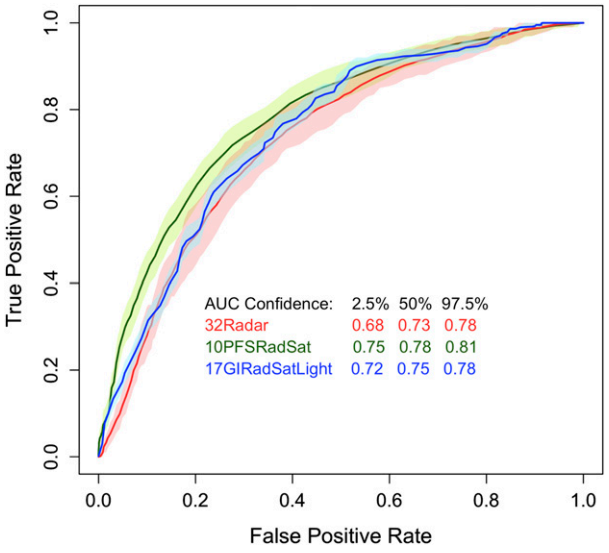

FIG. 8. Receiver operating characteristic (ROC) curves for three random-forest models: 32Radar, 10PFSRadSat, and 17GIRadSatLight. See Fig. 7 for the predictor sets. The 95% confidence interval per ROC curve is shaded, and the area under the ROC curve (AUC) for each model is shown. The 95% confidence interval was developed using 500 randomly selected datasets of size "0.75 × testing database" (13 832) for the three models. The 2.5%, 50%, and 97.5% confidence interval values of the AUC are listed per model.

models 9GIRadar, 12PFSRadar, 5SFSRadSat, yet not compared to model 10PFSRadSat. For the models with the highest CSI (17GIRadSatLight and 10PFSRadSat), model bias was near 1.0 with random-forest probability thresholds of 48%. The 48% probability is the threshold used to convert the forecasts from probabilistic to deterministic given that is was the probability that maximized CSI values across the vast majority of the 500 random-forest models used to make Fig. 9. The yellow plus sign and times sign denoted the highest CSI with the associated biases for models 17GIRadSatLight and 32Radar, respectively, as shown in Fig. 10.

Figure 10 is a performance diagram that compares the two models with the best CSI scores, 17GIRadSatLight (red line) and 32Radar (black line), with the 95% confidence intervals shaded for each model, as developed using the 500 randomly selected with replacement 0.75*N* (13 832) datasets for both models. The combined radar-satellite 17GIRadSatLight model's mean CSI peaks at 0.747 (with a bias of 0.94) as the mean POD increases to 0.72, which is higher than for model 32Radar (mean CSI of 0.636, bias of 1.0, and mean POD of 0.66), consistent with Fig. 9a.

In summary, Figs. 8–10 show that GI predictor importance resulted in a combination of 17 radar, lightning, and satellite fields, which provided the best overall forecast skills. Use of 32 radar-only fields also shows high skill, yet there is quantifiable added benefit of including satellite-based and lightning fields.

c. *Predictor field time series analysis*

Figures 11a–h present a time series analysis for eight select storms from the 2004 cell track database, specifically for storms in the testing database. The random-forest model used to develop predictions in Fig. 11 was 17GIRadSatLight. In Fig. 11, time series of only the top three most important radar fields are shown (as listed in Fig. 4), along with model probabilities and two satellite predictors (Figs. 4 and 6). The intent is to highlight

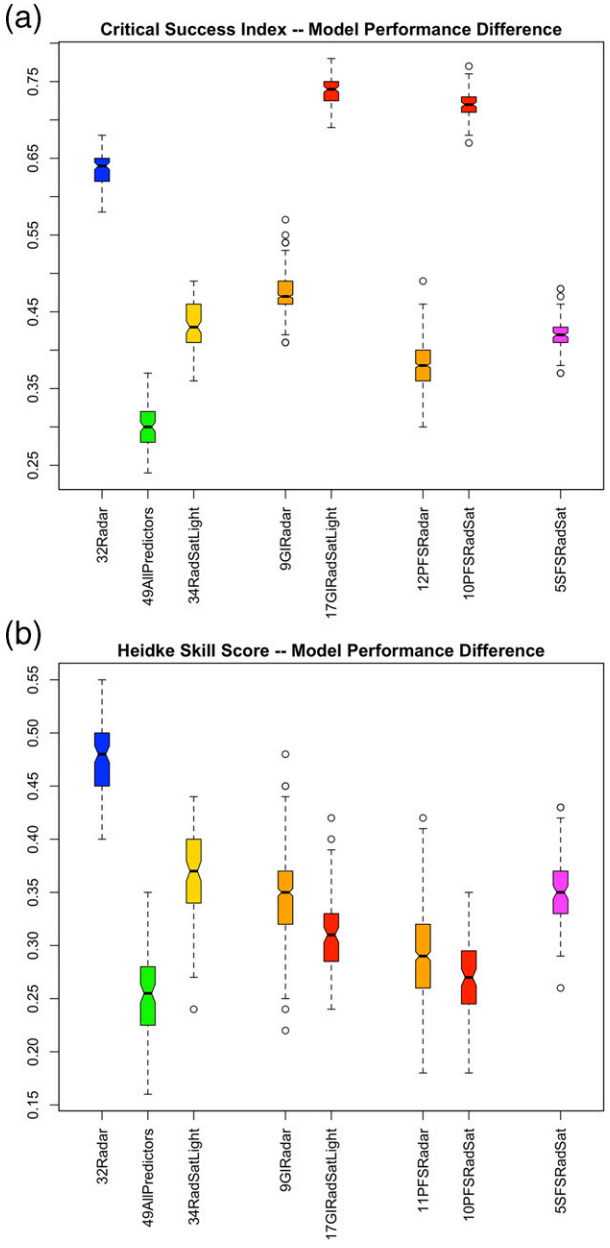

FIG. 9. Box-and-whiskers plots showing (a) critical success index and (b) Heidke skill score for all random-forest models using the predictor selections as listed in Fig. 7. The data for these boxplot curves were generated using 500 randomly selected datasets of size "0.75 × testing database" (13 832) for all eight models. The middle of the box is 50th percentile, edges of the box are 25th and 75th percentiles, the ends of whiskers are 2.5th and 97.5th percentiles, and dots are outliers (below 2.5th or above 97.5th). The 95% confidence interval therefore is between the 2.5th and 97.5th percentile for each case.

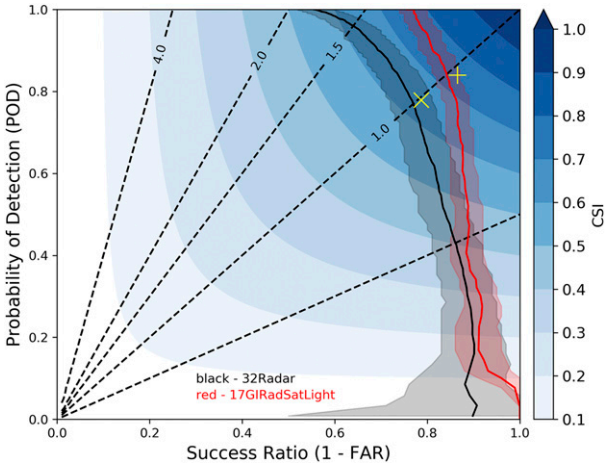

FIG. 10. Performance diagram that compares the success ratio [1 − false alarm ratio (FAR)] to the probability of detection (POD) for two models. The corresponding 95% confidence intervals are shaded as pink and gray, which were formed using 500 randomly selected datasets of size "0.75 × testing database" (13 832) for the two models. Dashed lines show frequency bias. The yellow plus sign and times sign represent the highest CSI and bias values for models 17GIRadSatLight and 32Radar, respectively, as related to Fig. 9a.

model performance relative to severe thunderstorm (black circles) and tornado (upside-down triangles) warning periods. A random-forest prediction value of 0.00 represents no diagnosed or predicted chance of severe weather, while a value of 1.00 represents a 100% chance for severe weather. Data in Figs. 11a–h are shown every 1 min, yet there are a few 1–3-min periods of missing data.

Highlights of Fig. 11 include: (i) there is a general tendency for the probabilities (ranging from 0 to 1) to increase prior to times of severe warnings, and also to fall toward the end of or after a severe warning period, (ii) there is a tendency for the radar predictor variable magnitudes (HDRg20, HCAarea and HDRcolMax as described in Fig. 11) to decrease toward the ending of a warning, and (iii) both of the $CTDA_{15}$ and AACP fields are typically well-defined when storms are particularly strong. Two of the more interesting time series analyses are in Figs. 11d and 11f. In Fig. 11d, increases in the forecast probabilities correlate to warning times, while the CTD maximizes during and surrounding the time of a tornado warning. The CTD trend is particularly pronounced in Fig. 11f, in which an AACP was also present, yet as the tornado warning time persists, the probability decreases along with magnitudes of all predictor fields, suggesting that the tornado warning period was too long or the model probabilities remained high for too short of a time. Figure 11g shows model prediction values in the ~0.60–0.62 range for the span of time severe thunderstorm and tornado warnings were in effect. Furthermore, Fig. 11h shows that a 20+-min "lead time" was given for this particular storm prior to when the actual warning was issued.

Figure 12 shows the radar fields as commonly seen in displays in NWS Forecast Offices for the storm in Fig. 11h, along with four important *GOES-14* satellite and GridRad radar predictors used in the 17GIRadSatLight model (Fig. 13). The radar fields shown in Fig. 12 are Z_H, radial velocity, Z_{DR}, and CC. Shown in Fig. 13 are *GOES–14* visible imagery with CTD, *GOES-14* 10.7-μm T_B, GridRad H_{DR} volume > 20 dB, and GridRad 20-dBZ echo tops. Given the fields shown in Fig. 12, it is apparent that hail exists within the high-reflectivity core, and broad rotation was present, although this storm was unwarned at the time (2050 UTC). In Fig. 13, this same storm (near Amarillo, Texas) possessed the radar predictors for ordinarily warned storms, and also attained forecast probabilities ≥90% for large portions of the tracked period as shown in Fig. 11h, from 2147 to 2208 UTC; that is, both satellite and GridRad variables identified this as a severe weather–producing storm prior to the warning issuance. This example storm highlights the use of multiplatform predictors in concert with a machine-learning system.

d. *Along-cell-track random-forest model performance*

Last, as a means of demonstrating how the random-forest 17GIRadSatLight model predictions perform along cell tracks relative to 32Radar model results, Figs. 14a and 14b provide a broad view of all cell tracks over the state of Kansas for the testing database for all 7 days, in which the contingency-table variables (Table 3) are shown as four different colors. The 0.48 threshold is used in Figs. 14a and 14b as this was the probability where the CSI peaked for the 17GIRadSatLight model, at 0.747. In Fig. 14a, the 32Radar model results are presented, while Fig. 14b shows the 17GIRadSatLight model results. The simple interpretation of Fig. 14 is that gray and green tracks are successful correct predictions, while blue and yellow are not. Red line segments in Fig. 14a are added to show locations where the 17GIRadSatLight model improved both the over and underpredictions of warning conditions relative to the 32Radar model, and vice versa for Fig. 14b where red lines show where model 32Radar showed improvements over model 17GIRadSatLight. When comparing the figures, model 17GIRadSatLight improved on 32Radar for 407 min of storm tracks (Fig. 14a), while model 32Radar improved on 17RadSatLight for 450 min of storm tracks (Fig. 14b), and thus the 32Radar model slightly outperformed the 17GIRadSatLight model in this case example centered over Kansas. Other notable differences between Figs. 14a and 14b are that the 17GIRadSatLight model produced a 7% increase in true positives (gray-colored lines), yet 45% more false positives (blue lines), while the 32Radar model had 11% more correct true negatives (green lines) and 17% more false negatives of severe weather (yellow lines). In Fig. 14a, the HR is 64%, the FAR is 25%, and the POFD is 33%, while in Fig. 14b the HR is 75%, the FAR is 17%, and the POFD is 45%.

5. Discussion and conclusions

The research goal here was to evaluate next-generation, 1-min-update-frequency geostationary satellite and lightning information with ground-based radar to isolate which variables, when used in concert, provide skillful discriminatory information in identifying severe versus nonsevere storms. To address the first science question stated in the introduction, the

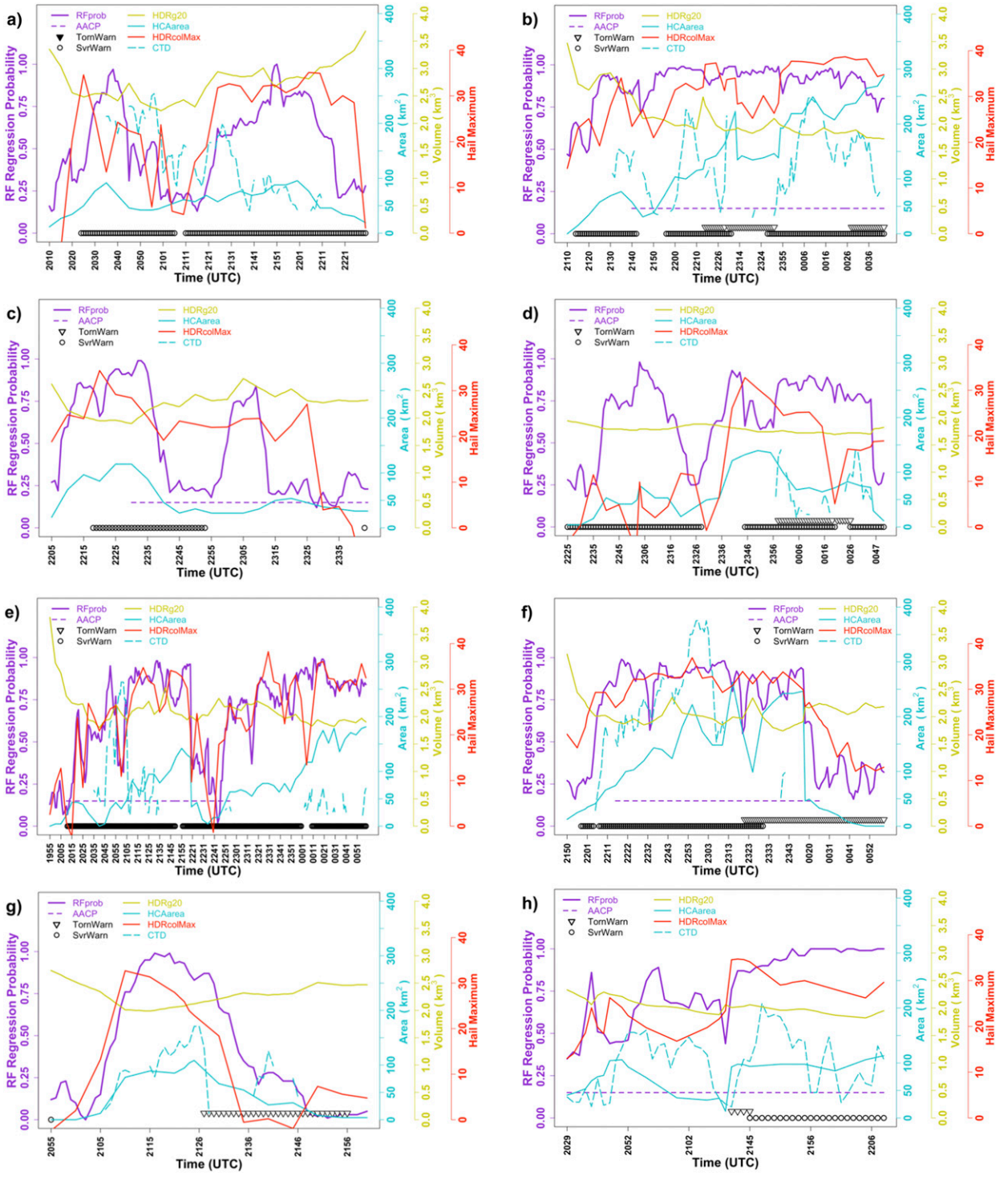

FIG. 11. Time series for select storm-track times, showing the six main predictors used in the random-forest model, along with probabilities, and severe storm (SvrWarn) and tornado (TornWarn) warning times, denoted by the black open circles and inverse filled triangles, respectively. The main predictors arevolume of $H_{DR} \geq 20$ dB (HDRg20), area of HCA-classified hail at 3 km (HCAarea), H_{DR} column maximum (HDRcolMax), above-anvil cirrus plume (AACP), and GOES mAMV area where CTD $>15 \times 10^{-4}\,s^{-1}$ (CTD). All predictors are listed in Figs. 4 and 6. See the text for acronym definitions.

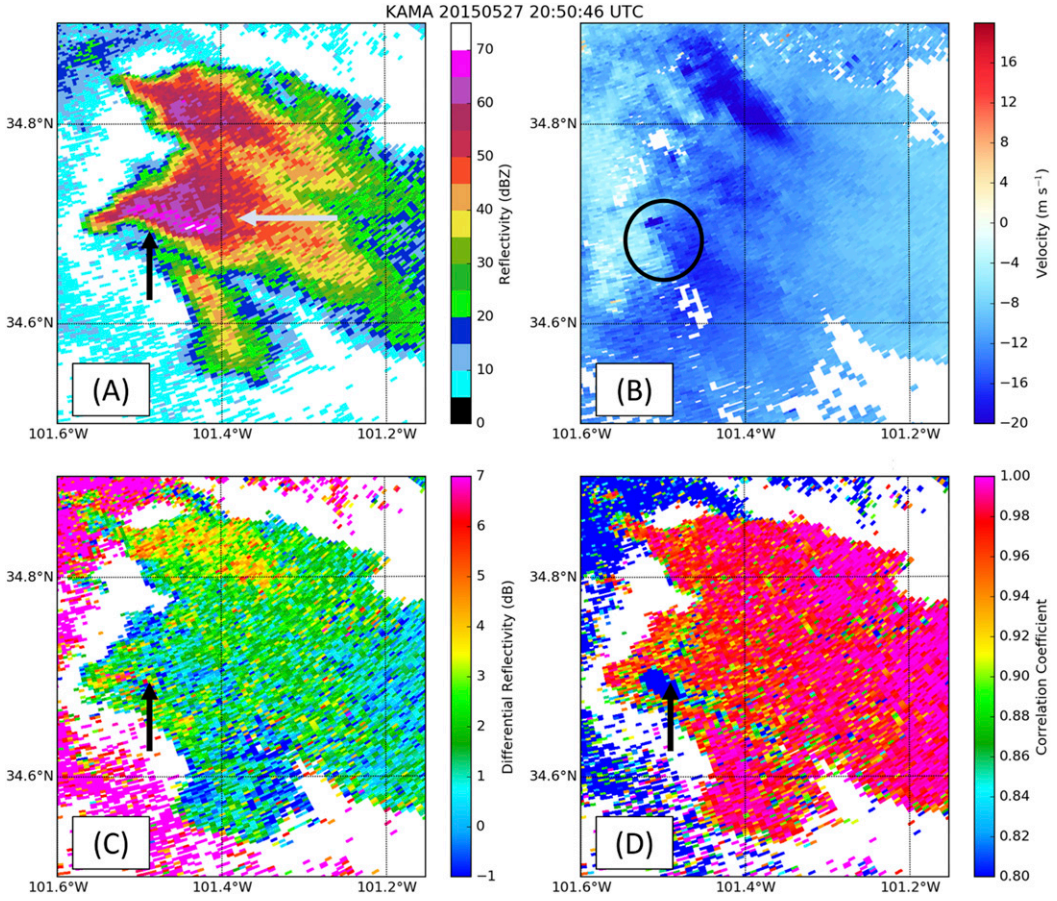

FIG. 12. For a convective storm at 2050:46 UTC 27 May 2015 (Fig. 11h), the Amarillo, radar (a) reflectivity (dBZ) with a white arrow highlighting the storm of interest, (b) radial velocity (m s^{-1}) with a circle highlighting the broad and finescale rotation, (c) differential reflectivity (dB), and (d) correlation coefficient covering a splitting supercell in north-central Texas. The black arrow highlights the location of a likely hail core. Radar plots were created using the Python Atmospheric Radiation Measurement Radar Toolkit (Py-ART; Helmus and Collis 2016). Two confirmed hail reports did occur with this storm, one at 2224 UTC of 2.25-in. hail, and another at 2243 UTC of 1.75-in. hail.

peak skill when predicting severe weather likelihood (e.g., whether warning conditions are occurring) when only 5-min storm-centered GridRad radar data are used was measured as a mean CSI score of 0.64 and mean HSS of 0.477 for model 32Radar (see Figs. 9a,b). The answers to our second and third study questions related to the value of satellite and lighting data, and which satellite and lightning fields are most beneficial, are that two models, one with 13 radar fields, 3 satellite fields, and 1 lightning field combined (model 17GIRadSatLight), and another with 8 radar and 2 satellite fields (model 10PFSRadSat), showed better severe weather diagnostic skill than models that used radar fields alone. Four satellite-based fields were found to be the most important of an initial 14 satellite fields using GI, SFS, and PFS importance methods, when used in concert with radar and ENTLN lightning data. These satellite fields are the $CTDA_{15}$, the AACP, GOES mAMV CTV maximum, and GOES mAMV CTD mean. Two other important satellite fields were the GOES T_B minimum–NARR tropopause temperature and GOES maximum OT probability. The use of GOES-derived fields adds 10%–30% predictive skill for severe convective storms, given CSI, HSS, and AUC (Figs. 8–10), which is supported by recent research.

Two additional points are worth noting: Although NWP model data are used to identify areas where severe storms are likely, no NWP-based fields were included in the analysis here. For storms in close proximity, even if high-spatial resolution convection-allowing NWP data were analyzed, a severe storm adjacent to a nonsevere storm could be assigned the same NWP fields in terms of kinematic and thermodynamic fields, hence not adding much to our current understanding. Also, given the use of 1-min-resolution satellite and lightning observations, it is recognized that the 1-min predictors for a given storm are highly correlated in both space and time (i.e., temporal autocorrelation as storm cell characteristics remain nearly constant over 5–10-min intervals). Therefore, over short segments of a storm cell's track, across-predictor relationships will likely be similar, which may have led to a degree of

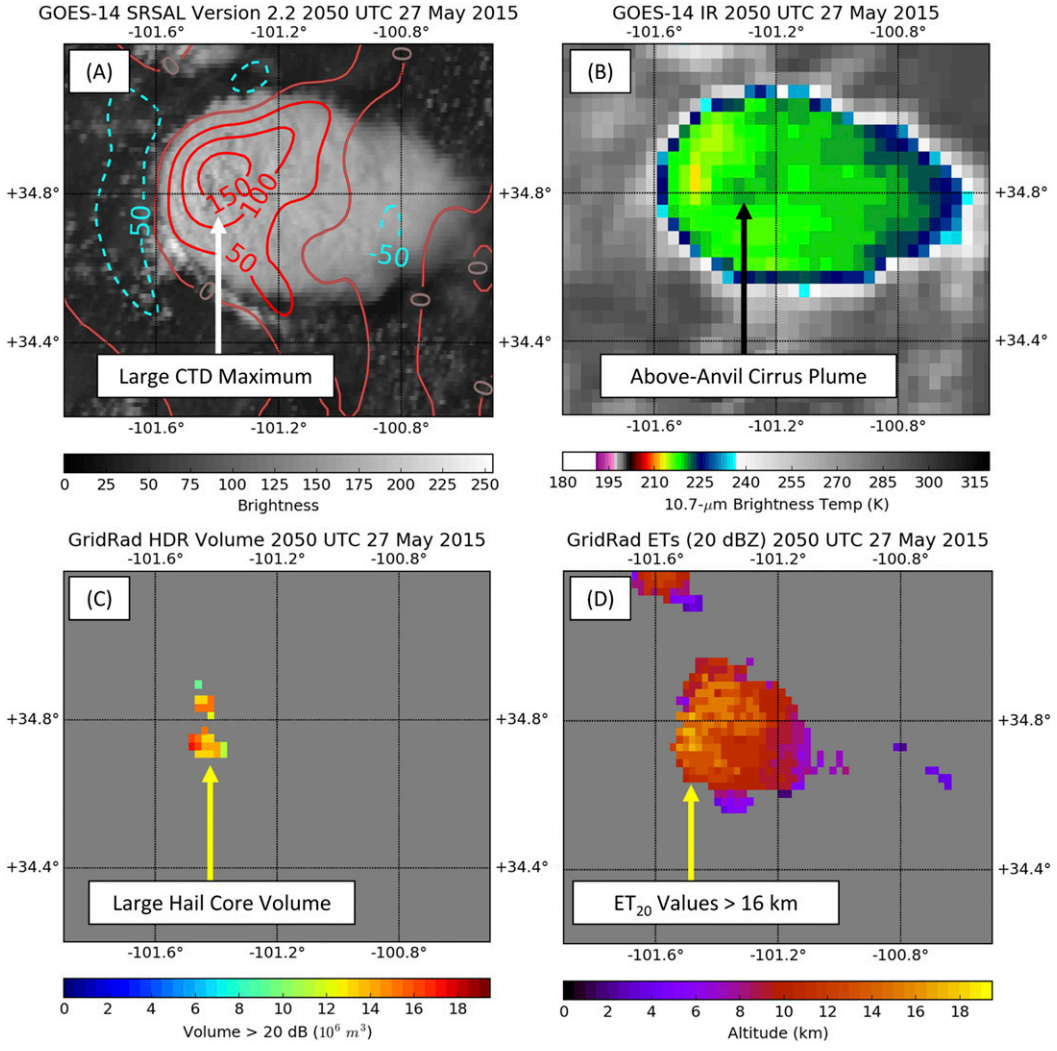

FIG. 13. Random-forest predictor datasets for the north-central Texas storm that is shown in Fig. 12, with (a) *GOES-14* visible imagery with positive cloud-top divergence (CTD) contoured in red and negative CTD contoured in blue dashes every $50 \times 10^{-5}\,s^{-1}$, (b) *GOES-14* 10.7-$\mu$m T_B (K), (c) GridRad-derived total hail differential reflectivity (H_{DR}) volume where $H_{DR} > 20$ dB (m^3), and (d) GridRad derived 20-dBZ echo tops (km).

overfitting and enhanced prediction accuracies. However, as many storm cells over long lifetime periods were examined in the training database, the confidence is high in our predictor field importance results given that model skills were similar between the three best-performing models 32Radar, 10PFSRadSat and 17GIRadSatLight.

As noted above, in a number of cases, especially related to the nonsevere convective storms tracked within the database, several satellite-derived variables did not reach key thresholds or did not occur within a given cell's track, and hence the most important fields $CTDA_{15}$, AACP, and CTV are not always well defined. Therefore, one key indicator of severe weather for a forecaster can be when the most important satellite fields become identifiable, or visible from the satellite in the absence of overlying higher clouds, and hence usable within a machine-learning approach. In general, most nonsevere storms never produce an AACP, have tall OTs (Bedka et al. 2018), or produce unobscured CTV and CTD fields of high magnitudes (Apke et al. 2018), as compared with their severe weather–producing counterparts. Although this study focused on severe weather warnings, a machine-learning approach could have instead emphasized tornado-only events, which is another option given the database used here (for that demonstration, see Sandmæl et al. 2019). Last, there is no intent within this study to suggest that a model such as random forests can consistently outperform a well-trained human forecaster or human expert, yet such a machine-learning approach can provide deterministic or probabilistic guidance within the severe weather forecast environment, much like the ProbSevere, RDT, COALITION or other models highlighted in the introduction.

The present study is limited especially by the seasonal and geographic sampling of the cases composing the cell-track

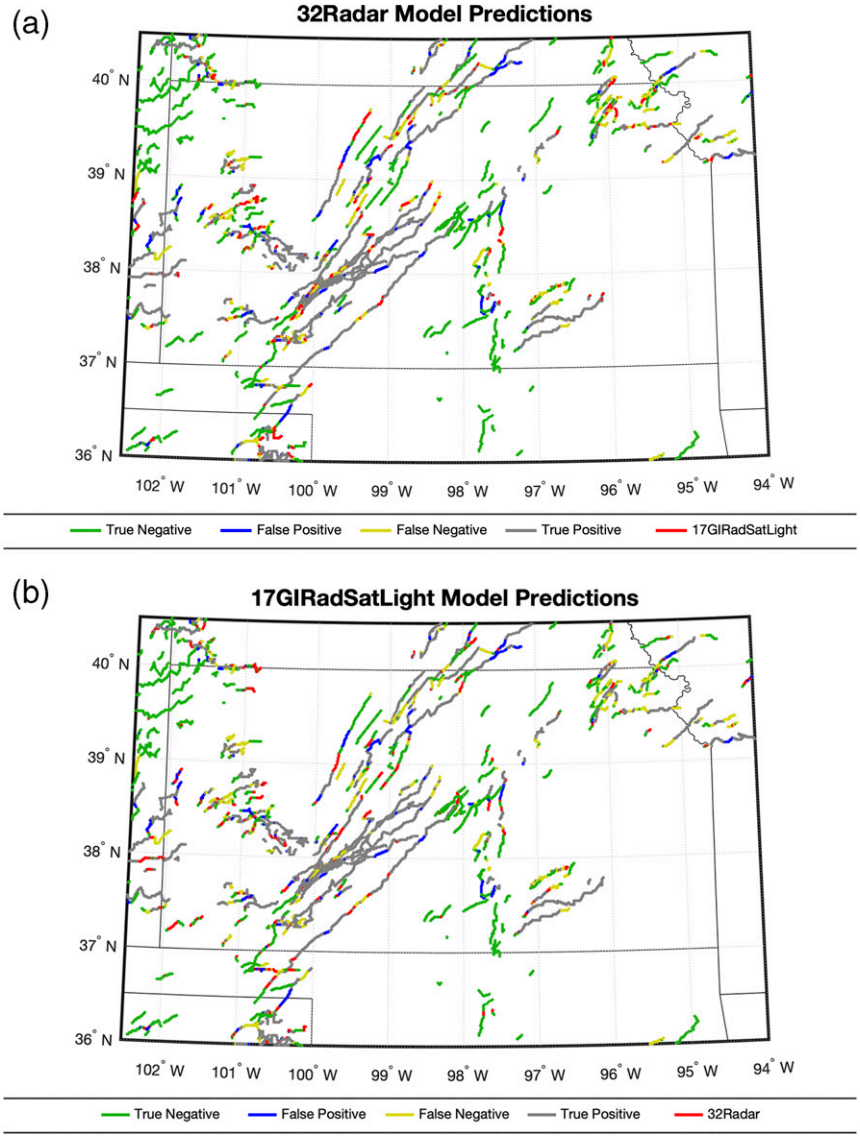

FIG. 14. Track-based performance of the (a) 32Radar model and (b) model 17GIRadSatLight on the testing dataset in and near Kansas. The tracks are plotted in various colors to signify whether a severe weather warning was in effect. The four colors represent the contingency-table entries (Table 3). The 48% criterion was used since it maximized the CSI as described in the text. In this figure, the red lines in (a) denote when model 17GISatRadLight improved the forecast from model 32Radar and the red lines in (b) denote when model 32Radar improved the forecast from model 17GISatRadLight.

database. Specifically, only daytime severe weather days in the month of May were considered, all within the central and southern Great Plains and *GOES–14* field of view (which has since been replaced by *GOES–16* and *–17*). Therefore, the random-forest model developed here would be biased to predict severe weather from more organized convective storm types, versus severe weather from pulse and quasi-linear convective line convection, which more often occurs during the summer months. A solution for future research would be to form a much more expansive cell-track database, containing *GOES-16* and *-17* satellite observed storms across the United States for a longer period during the convective weather season. The enhanced spatial resolution of the *GOES-16*/*-17* data would lead to a more robust random-forest model as a result. Also, performing this study using GLM lightning fields instead of ENTLN data represents an area of future work, in addition to using ENTLN or GLM lightning jump fields within the machine-learning model.

The hope from the present study is that future research will be focused toward making the key *GOES-14* and GridRad predictor fields into real-time day/night products based on use of *GOES-16* and *-17* data, and to improve the quality of current

probability-based nowcasting and diagnosis systems by using different and new predictor fields. The supporting of applied research that capitalizes on using observations from geostationary satellites with from 30-s to 2.5-min time resolution is a recommendation, as a means of maximizing value from extensive, SRS real-time datasets, which are very often only collected during severe weather episodes (e.g., Setvák and Müller 2013; Schmit et al. 2015; Setvák 2015). Another final aspect of this study is that the GridRad radar and GOES-derived fields are not yet available in real time, and hence new research and applied work must occur to transition important research-grade algorithms to run in an operational setting.

Acknowledgments. This research was supported by National Aeronautics and Space Administration (NASA) Langley Research Award NNX15AV82G. A portion of this research was also sponsored by National Science Foundation (NSF) research Grant AGS-1746119. The authors thank the University of Wisconsin Space Science and Engineering Center Data Center for providing the *GOES-14* satellite data analyzed in this study. The authors thank three anonymous reviewers for providing comments to earlier versions of this paper, which have significantly improved the paper's quality. Specifically, we are very grateful to one reviewer who spent a considerable amount of time helping to improve the presentation of the research results and paper figures.

Data availability statement. The main cell-track database used for this study is available from the lead author upon request. Following Sandmæl et al. (2019), NWS severe weather warnings were obtained from the archive maintained by Iowa State University (Iowa Environmental Mesonet 2017). All radar data used to form the cell-track database can be obtained from NCEI (www.ncdc.noaa.gov/data-access), and the *GOES-14* data can be obtained from the NOAA Comprehensive Large Array-data Stewardship System (CLASS; www.class.noaa.gov). As part of our Data Sharing Plan, the main cell-track database, which represents key databases used in this study and which contains all radar, satellite, and environmental data, is available upon request. The authors will provide guidance on interpreting the cell-track database and/or help others to construct a similar dataset using raw radar, satellite, and environmental datasets.

APPENDIX A

Acronym List

AACP	Above-anvil cirrus plume
CC	Correlation coefficient
CTD	Cloud-top divergence
$CTDA_{15}$	Cloud-top divergence area $> 15 \times 10^{-4}\,s^{-1}$
CTV	Cloud-top vorticity
GOES	Geostationary Operational Environmental Satellite
ENTLN	Earth Network Total Lightning Network
HCA	Hail classification algorithm
H_{DR}	Hail differential reflectivity
K_{DP}	Specific differential phase
mAMV	Mesoscale atmospheric motion vector
MESH	Maximum expected size of hail
NARR	North American Regional Reanalysis
NCEI	National Centers for Environmental Information
OT	Overshooting top
RF	Random forest
SL3D	Storm Labeling in Three Dimensions
SRS	Super rapid scan
SWDI	Severe Weather Data Inventory
VIL	Vertically integrated liquid
Z_H	Horizontal reflectivity
Z_{DR}	Differential reflectivity

APPENDIX B

Definition of Skill Scores

The following skill scores are used in the study. The definitions follow the contingency-table definitions in Table 3:

$$\text{hit rate (HR)} = \frac{a+d}{n},$$

$$\text{false positive rate} = \text{probability of false detection (POFD)} = \frac{b}{b+d},$$

$$\text{false alarm ratio (FAR)} = \frac{b}{a+b},$$

$$\text{critical success index (CSI)} = \frac{a}{a+b+c},$$

$$\text{Heidke skill score (HSS)} = \frac{2(ad-bc)}{(a+c)(c+d)+(a+b)(b+d)},$$

$$\text{Peirce skill score (PSS)} = \frac{ad-bc}{(a+c)(b+d)}, \quad \text{and}$$

$$\text{bias} = \frac{a+b}{a+c}.$$

REFERENCES

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

Amburn, S. A., and P. L. Wolf, 1997: VIL density as a hail indicator. *Wea. Forecasting*, **12**, 473–478, https://doi.org/10.1175/1520-0434(1997)012<0473:VDAAHI>2.0.CO;2.

Apke, J. M., J. R. Mecikalski, and C. P. Jewett, 2016: Analysis of mesoscale atmospheric flows above mature deep convection using Super Rapid Scan geostationary satellite data. *J. Appl. Meteor. Climatol.*, **55**, 1859–1887, https://doi.org/10.1175/JAMC-D-15-0253.1.

——, ——, K. M. Bedka, E. W. McCaul, C. R. Homeyer, and C. P. Jewett, 2018: Relationships between deep convection updraft characteristics and satellite-based super rapid scan mesoscale atmospheric motion vector-derived flow. *Mon. Wea. Rev.*, **146**, 3461–3480, https://doi.org/10.1175/MWR-D-18-0119.1.

Autonès, F., and J.-M. Moisselin, 2010: Algorithm theoretical basis document for "Rapid Development Thunderstorms" (RDT-PGE11 v2.2). Meteo-France Doc. SAF/NWC/CDOP/MFT/SCI/ATBD/11 (issue 2 Rev. 2), 65 pp., http://www.eumetrain.org/data/2/274/media/flash/SAF-NWC-CDOP-MFT-SCI-ATBD-11_v2.2.pdf.

Bedka, K. M., and K. Khlopenkov, 2016: A probabilistic multispectral pattern recognition method for detection of overshooting cloud tops using passive satellite imager observations. *J. Appl. Meteor. Climatol.*, **55**, 1983–2005, https://doi.org/10.1175/JAMC-D-15-0249.1.

——, E. M. Murillo, C. R. Homeyer, B. Scarino, and H. Mersiovsky, 2018: The above anvil cirrus plume: An important severe weather indicator in visible and infrared satellite imagery. *Wea. Forecasting*, **33**, 1159–1181, https://doi.org/10.1175/WAF-D-18-0040.1.

Beusch, L., L. Foresti, M. Gabella, and U. Hamann, 2018: Satellite-based rainfall retrieval: From generalized linear models to artificial neural networks. *Remote Sens.*, **10**, 939, https://doi.org/10.3390/rs10060939.

Breiman, L., 2001: Random forests. *Machine Learning*, R. E. Schapire, Ed., Vol. 45, Springer, 5–32.

Brotzge, J., S. Erickson, and H. Brooks, 2011: A 5-yr climatology of tornado false alarms. *Wea. Forecasting*, **26**, 534–544, https://doi.org/10.1175/WAF-D-10-05004.1.

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

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

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

Depue, T. K., P. C. Kennedy, and S. A. Rutledge, 2007: Performance of the hail differential reflectivity (H_{DR}) polarimetric radar hail indicator. *J. Appl. Meteor. Climatol.*, **46**, 1290–1301, https://doi.org/10.1175/JAM2529.1.

Dixon, M., and G. Wiener, 1993: TITAN: Thunderstorm Identification, Tracking, Analysis, and Nowcasting—A radar-based methodology. *J. Atmos. Oceanic Technol.*, **10**, 785–797, https://doi.org/10.1175/1520-0426(1993)010<0785:TTITAA>2.0.CO;2.

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

Evans, J. S., and S. A. Cushman, 2009: Gradient modeling of conifer species using random forests. *Landscape Ecol.*, **24**, 673–683, https://doi.org/10.1007/s10980-009-9341-0.

Fabry, F., 2015: *Radar Meteorology*. Cambridge University Press, 256 pp.

Gagne, D. J., II, A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1.

——, ——, S. 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.

Gijben, M., and E. de Coning, 2017: Using satellite and lightning data to track rapidly developing thunderstorms in data sparse regions. *Atmosphere*, **8**, 67, https://doi.org/10.3390/atmos8040067.

Goodman, S. J., and Coauthors, 2013: The GOES-R Geostationary Lightning Mapper (GLM). *Atmos. Res.*, **125–126**, 34–49, https://doi.org/10.1016/j.atmosres.2013.01.006.

Hall, M. P. M., S. M. Cherry, J. W. F. Goddard, and G. R. Kennedy, 1980: Rain drop sizes and rainfall rate measured by dual-polarization radar. *Nature*, **285**, 195–198, https://doi.org/10.1038/285195a0.

Hapfelmeier, A., and K. Ulm, 2013: Variable selection with random forests for missing data. University of Munich Department of Statistics Tech. Rep. 137, 13 pp., https://epub.ub.uni-muenchen.de/14344/1/TechnicalReport_LMU_10012013.pdf.

Helmus, J. J., and S. M. Collis, 2016: The Python ARM Radar Toolkit (Py-ART), a Library for working with weather radar data in the Python programming language. *J. Open Res. Software*, **4**, e25, https://doi.org/10.5334/jors.119.

Hill, R. A., C. P. Hawkins, and D. M. Carlisle, 2013: Predicting thermal reference conditions for USA streams and rivers. *Freshwater Sci.*, **32**, 39–55, https://doi.org/10.1899/12-009.1.

Homeyer, C. R., and M. R. Kumjian, 2015: Microphysical characteristics of overshooting convection from polarimetric radar observations. *J. Atmos. Sci.*, **72**, 870–891, https://doi.org/10.1175/JAS-D-13-0388.1.

——, and K. P. Bowman, 2017: Algorithm description document for version 3.1 of the Three-Dimensional Gridded NEXRAD WSR-88D Radar (GridRad) dataset. GridRad Tech. Rep., 23 pp., http://gridrad.org/pdf/GridRad-v3.1-Algorithm-Description.pdf.

——, J. D. McAuliffe, and K. M. Bedka, 2017: On the development of above-anvil cirrus plumes in extratropical convection. *J. Atmos. Sci.*, **74**, 1617–1633, https://doi.org/10.1175/JAS-D-16-0269.1.

Illingworth, A. J., J. W. F. Goddard, and S. M. Cherry, 1987: Polarization radar studies of precipitation development in convective storms. *Quart. J. Roy. Meteor. Soc.*, **113**, 469–489, https://doi.org/10.1002/qj.49711347604.

Iowa Environmental Mesonet, 2017: Archived NWS watch/warnings. Iowa State University, accessed 10 May 2017, http://mesonet.agron.iastate.edu/request/gis/watchwarn.phtml.

Klemp, J. B., and R. Rotunno, 1983: A study of the tornadic region within a supercell thunderstorm. *J. Atmos. Sci.*, **40**, 359–377, https://doi.org/10.1175/1520-0469(1983)040<0359:ASOTTR>2.0.CO;2.

Koshak, W. J., and Coauthors, 2004: North Alabama Lightning Mapping Array (LMA): VHF source retrieval algorithm and error analyses. *J. Atmos. Oceanic Technol.*, **21**, 543–558, https://doi.org/10.1175/1520-0426(2004)021<0543:NALMAL>2.0.CO;2.

Krehbiel, P. R., R. J. Thomas, W. Rison, T. Hamlin, J. Harlin, and M. Davis, 2000: GPS-based mapping system reveals lightning inside storms. *Eos, Trans. Amer. Geophys. Union*, **81**, 21–25, https://doi.org/10.1029/00EO00014.

Kühnlein, M., T. Appelhans, B. Thies, and T. Nauss, 2014: Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—A random forests–based approach applied to MSG SEVIRI. *Remote Sens. Environ.*, **141**, 129–143, https://doi.org/10.1016/j.rse.2013.10.026.

Kumjian, M. R., and A. V. Ryzhkov, 2008: Polarimetric signatures in supercell thunderstorms. *J. Appl. Meteor. Climatol.*, **47**, 1940–1961, https://doi.org/10.1175/2007JAMC1874.1.

——, and K. A. Lombardo, 2017: Insights into the evolving microphysical and kinematic structure of northeastern U.S. winter storms from dual-polarimetric Doppler radar. *Mon. Wea. Rev.*, **145**, 1033–1061, https://doi.org/10.1175/MWR-D-15-0451.1.

——, A. P. Khain, N. Benmoshe, E. Ilotoviz, A. V. Ryzhkov, and V. T. J. Phillips, 2014: The anatomy and physics of ZDR columns: Investigating a polarimetric radar signature with a spectral bin microphysical model. *J. Appl. Meteor. Climatol.*, **53**, 1820–1843, https://doi.org/10.1175/JAMC-D-13-0354.1.

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

Lin, P.-F., P.-L. Chang, B. J.-D. Jou, J. W. Wilson, and R. D. Roberts, 2012: Objective prediction of warm season afternoon thunderstorms in northern Taiwan using a fuzzy logic approach. *Wea. Forecasting*, **27**, 1178–1197, https://doi.org/10.1175/WAF-D-11-00105.1.

Louppe, G., L. Wehenkel, A. Sutera, and P. Geurts, 2013: Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, C. J. C. Burges et al., Eds., 431–439, https://papers.nips.cc/paper/2013/hash/e3796ae838835da0b6f6ea37bcf8bcb7-Abstract.html.

Manzato, A., 2007:A note on the maximum Peirce skill score. *Wea. Forecasting*, **22**, 1148–1154, https://doi.org/10.1175/WAF1041.1.

Markowski, P., and Y. Richardson, 2010: *Mesoscale Meteorology in Midlatitudes*. John Wiley and Sons, 430 pp.

Marzban, C., and A. Witt, 2001: A Bayesian neural network for severe-hail size prediction. *Wea. Forecasting*, **16**, 600–610, https://doi.org/10.1175/1520-0434(2001)016<0600:ABNNFS>2.0.CO;2.

MATLAB, 2018a: TreeBagger documentation. Accessed 4 November 2019, www.mathworks.com/help/stats/treebagger.html.

——, 2018b: Permuted predictor importance documentation. Accessed 4 November 2019, www.mathworks.com/help/stats/classificationbaggedensemble.oobpermutedpredictorimportance.html.

——, 2018c: Sequentialfs: Sequential feature selection using custom criterion documentation. Accessed 4 November 2019, www.mathworks.com/help/stats/sequentialfs.html.

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

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. Climate Appl. Meteor.*, **54**, 1039–1059, https://doi.org/10.1175/JAMC-D-14-0129.1.

Mesinger, F., and Coauthors, 2006: North American Regional Reanalysis. *Bull. Amer. Meteor. Soc.*, **87**, 343–360, https://doi.org/10.1175/BAMS-87-3-343.

Meyer, H., M. Kühnlein, T. Appelhans, and T. Nauss, 2016: Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals. *Atmos. Res.*, **169**, 424–433, https://doi.org/10.1016/j.atmosres.2015.09.021.

Mueller, C. K., T. Saxen, R. Roberts, J. Wilson, T. Betancourt, S. Dettling, N. Oien, and J. Yee, 2003: NCAR auto-nowcast system. *Wea. Forecasting*, **18**, 545–561, https://doi.org/10.1175/1520-0434(2003)018<0545:NAS>2.0.CO;2.

Murillo, E. M., and C. R. Homeyer, 2019: Severe hail fall and hailstorm detection using remote sensing observations. *J. Appl. Meteor. Climatol.*, **58**, 947–970, https://doi.org/10.1175/JAMC-D-18-0247.1.

National Centers for Environmental Information, 2017: NOAA's Severe Weather Data Inventory. Accessed 15 June 2017, https://www.ncdc.noaa.gov/severe-weather/severe-weather-data-inventory.

Nisi, L., P. Ambrosetta, and L. Clementi, 2014: Nowcasting severe convection in the Alpine region: The COALITION approach. *Quart. J. Roy. Meteor. Soc.*, **140**, 1684–1699, https://doi.org/10.1002/qj.2249.

Park, H. S., A. V. Ryzhkov, D. S. Zrnić, and K.-E. Kim, 2009: The hydrometeor classification algorithm for the polarimetric WSR-88D: Description and application to an MCS. *Wea. Forecasting*, **24**, 730–748, https://doi.org/10.1175/2008WAF2222205.1.

Picca, J., and A. Pyzhkov, 2012: A dual-wavelength polarimetric analysis of the 16 May 2010 Oklahoma City extreme hailstorm. *Mon. Wea. Rev.*, **140**, 1385–1403, https://doi.org/10.1175/MWR-D-11-00112.1.

Rehfeldt, G. E., N. L. Crookston, C. Sáenz-Romero, and E. M. Campbell, 2012: North American vegetation model for land-use planning in a changing climate: A solution to large classification problems. *Ecol. Appl.*, **22**, 119–141, https://doi.org/10.1890/11-0495.1.

Rotunno, R., 1993: Supercell thunderstorm modeling and theory. *The Tornado: Its Structure, Dynamics, Prediction, and Hazards, Geophys. Monogr.*, Vol. 79, Amer. Geophys. Union, 57–73.

Rückstieß, T., C. Osendorfer, and P. van der Smagt, 2011: Sequential feature selection for classification. *Advances in Artificial Intelligence: 24th Australasian Joint Conf.*, Perth, Australia, Springer, 132–141.

Rudlosky, S. D., 2015: Evaluating ENTLN performance relative to TRMM/LIS. *J. Oper. Meteor.*, **3**, 11–20, https://doi.org/10.15191/nwajom.2015.0302.

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.

Schmit, T. J., M. M. Gunshor, W. P. Menzel, J. Li, S. Bachmeier, and J. J. Gurka, 2005: Introducing the next-generation Advanced Baseline Imager (ABI) on GOES-R. *Bull. Amer. Meteor. Soc.*, **86**, 1079–1096, https://doi.org/10.1175/BAMS-86-8-1079.

——, and Coauthors, 2014: Geostationary Operational Environmental Satellite (GOES)-14 super rapid scan operations to prepare for GOES-R. *J. Appl. Remote Sens.*, **7**, 073462, https://doi.org/10.1117/1.JRS.7.073462.

——, and Coauthors, 2015: Rapid Refresh information of significant events: Preparing users for the next generation of geostationary operational satellites. *Bull. Amer. Meteor. Soc.*, **96**, 561–576, https://doi.org/10.1175/BAMS-D-13-00210.1.

Schultz, C. J., W. A. Petersen, and L. D. Carey, 2011: Lightning and severe weather: A comparison between total and cloud-to-ground lightning trends. *Wea. Forecasting*, **26**, 744–755, https://doi.org/10.1175/WAF-D-10-05026.1.

——, L. D. Carey, E. V. Schultz, and R. J. Blakeslee, 2015: Insights into the kinematic and microphysical processes that control lightning jumps. *Wea. Forecasting*, **30**, 1591–1621, https://doi.org/10.1175/WAF-D-14-00147.1.

——, ——, ——, and ——, 2017: Kinematic and microphysical significance of lightning jumps versus nonjump increases in total flash rate. *Wea. Forecasting*, **32**, 275–288, https://doi.org/10.1175/WAF-D-15-0175.1.

Setvák, M., 2015: Experimentální 2,5minutové snímání družicemi MSG. *Meteor. Zprávy*, **68**, 65–73.

——, and J. Müller, 2013: MSG-3 Super Rapid Scan study. EUM/STG-SWG/34/13/DOC/06 (internal EUMETSAT document).

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

Starzec, M., C. R. Homeyer, and G. L. Mullendore, 2017: Storm Labeling in Three Dimensions (SL3D): A volumetric radar echo and dual-polarization updraft classification algorithm. *Mon. Wea. Rev.*, **145**, 1127–1145, https://doi.org/10.1175/MWR-D-16-0089.1.

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.

Stumpf, G., T. M. Smith, and J. Hocker, 2004: New hail diagnostic parameters derived by integrating multiple radars and multiple sensors. *22nd Conf. on Severe Local Storms*, Hyannis, MA, Amer. Meteor. Soc., P7.8, https://ams.confex.com/ams/11aram22sls/techprogram/paper_81451.htm.

Tan, P.-N., M. Steinbach, A. Karpatne, and V. Jumar, 2005: *Introduction to Data Mining*. 2nd ed. Pearson/Addison Wesley, 792 pp.

Trapp, R. J., G. J. Stumpf, and K. L. Manross, 2005: A reassessment of the percentage of tornadic mesocyclones. *Wea. Forecasting*, **20**, 680–687, https://doi.org/10.1175/WAF864.1.

——, 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.

Verbout, S. M., H. E. Brooks, L. M. Leslie, and D. M. Schultz, 2006: Evolution of the U.S. tornado database: 1954–2004. *Wea. Forecasting*, **21**, 86–93, https://doi.org/10.1175/WAF910.1.

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

Wilson, J. W., and C. K. Mueller, 1993: Nowcasts of thunderstorm initiation and evolution. *Wea. Forecasting*, **8**, 113–131, https://doi.org/10.1175/1520-0434(1993)008<0113:NOTIAE>2.0.CO;2.

——, N. A. Crook, C. K. Mueller, J. Sun, and M. Dixon, 1998: Nowcasting thunderstorms: A status report. *Bull. Amer. Meteor. Soc.*, **79**, 2079–2099, https://doi.org/10.1175/1520-0477(1998)079<2079:NTASR>2.0.CO;2.

Zhang, J., and Coauthors, 2011: National Mosaic and Multi-Sensor QPE (NMQ) system: Description, results, and future plans. *Bull. Amer. Meteor. Soc.*, **92**, 1321–1338, https://doi.org/10.1175/2011BAMS-D-11-00047.1.

——, and Coauthors, 2016: Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bull. Amer. Meteor. Soc.*, **97**, 621–638, https://doi.org/10.1175/BAMS-D-14-00174.1.

Zrnić, D. S. and R. J. Doviak, 1975: Velocity spectra of vortices scanned with a pulsed-Doppler radar. *J. Appl. Meteor.*, **14**, 1531–1539, https://doi.org/10.1175/1520-0450(1975)014<1531:VSOVSW>2.0.CO;2.

——, D. W. Burgess, and L. D. Hennington, 1985: Doppler spectra and estimated wind speed of a violent tornado. *J. Climate Appl. Meteor.*, **24**, 1068–1081, https://doi.org/10.1175/1520-0450(1985)024<1068:DSAEWO>2.0.CO;2.