Forecasting Severe Weather with Random Forests

AARON J. HILL, GREGORY R. HERMAN,[a] AND RUSS S. SCHUMACHER

Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado

(Manuscript received 15 October 2019, in final form 10 March 2020)

ABSTRACT

Using nine years of historical forecasts spanning April 2003–April 2012 from NOAA's Second Generation Global Ensemble Forecast System Reforecast (GEFS/R) ensemble, random forest (RF) models are trained to make probabilistic predictions of severe weather across the contiguous United States (CONUS) at Days 1–3, with separate models for tornado, hail, and severe wind prediction at Day 1 in an analogous fashion to the Storm Prediction Center's (SPC's) convective outlooks. Separate models are also trained for the western, central, and eastern CONUS. Input predictors include fields associated with severe weather prediction, including CAPE, CIN, wind shear, and numerous other variables. Predictor inputs incorporate the simulated spatiotemporal evolution of these atmospheric fields throughout the forecast period in the vicinity of the forecast point. These trained RF models are applied to unseen inputs from April 2012 to December 2016, and their forecasts are evaluated alongside the equivalent SPC outlooks. The RFs objectively make statistical deductions about the relationships between various simulated atmospheric fields and observations of different severe weather phenomena that accord with the community's physical understandings about severe weather forecasting. Using these quantified flow-dependent relationships, the RF outlooks are found to produce calibrated probabilistic forecasts that slightly underperform SPC outlooks at Day 1, but significantly outperform their outlooks at Days 2 and 3. In all cases, a blend of the SPC and RF outlooks significantly outperforms the SPC outlooks alone, suggesting that use of RFs can improve operational severe weather forecasting throughout the Day 1–3 period.

1. Introduction

Severe weather, as defined in the United States, includes three distinct phenomena: 1) the presence of one or more tornadoes of any intensity, 2) the presence of 1 in. (2.54 cm) or larger hail, or 3) convectively induced wind gusts of at least 58 mph (93 km h^{-1}). Beyond these criteria, tornadoes of F2 or EF2 strength or greater, hail 2 in. (5.08 cm) or larger in diameter, or wind gusts of at least 74 mph (119 km h^{-1}), pose particularly elevated threats to life and property and are considered supplementarily in a "significant severe" weather class (Hales 1988; Edwards et al. 2015). Collectively, these hazards have inflicted more than 1100 fatalities and $36.4B in damages across the contiguous United States (CONUS) in 2010–18 (NWS 2018). While inherently dangerous and damaging phenomena, accurate severe weather forecasts can increase preparedness and help mitigate inclement weather losses.

The hazards associated with severe weather are further encumbered by the challenge in accurately forecasting the phenomena. Due to the very small spatial scales associated with severe weather, it is often exceedingly difficult to model dynamically with operational weather models. Production of large hail involves a plethora of very small-scale microphysical processes that are necessarily parameterized in numerical models. The microphysical simplifications involved to hasten production of operational model output, including bulk rather than bin schemes (e.g., Khain et al. 2015), single moment microphysics (e.g., Igel et al. 2015; Labriola et al. 2017), and in some cases, not having an explicit category for hail at all (e.g., Hong and Lim 2006), all make direct prediction of severe hail from operational dynamical model output a perilous task. Tornadoes are in some respect even more difficult to simulate; while numerical tornado simulations have been conducted in a research setting (e.g., Orf et al. 2017), they occur on much too small of spatial scales to be resolved by any operational model. In forecasting severe weather, it is therefore necessary to relate simulated environmental factors across various scales, from

[a] Current affiliation: Amazon.com.

Corresponding author: Aaron J. Hill, aaron.hill@atmos.colostate.edu

DOI: 10.1175/MWR-D-19-0344.1

storm scale up to the synoptic scale, to severe weather risk. This is routinely performed in the human severe weather forecast process (e.g., Johns and Doswell 1992; Doswell 2004; Doswell and Schultz 2006), but in terms of producing automated guidance, statistical in addition to dynamical approaches are necessary for this important forecast problem.

CONUS-wide operational severe weather forecasts are issued routinely by the Storm Prediction Center (SPC) for Days 1–8 via their convective outlooks (Edwards et al. 2015). Forecasts are issued for 24-h 1200–1200 UTC periods, and are given as probabilities of observing the corresponding severe weather phenomenon within 40 km of the forecast point during the period. For Day 1, SPC issues separate probabilistic outlooks for each of the three severe weather predictands; for Day 2 and beyond, they are treated collectively in a single outlook. In the forecast process, the forecaster draws from a discrete set of allowable probability isopleths, where applicable. For Day 1 hail and wind outlooks, and Day 2 and 3 outlooks, permitted isopleths are 5%, 15%, 30%, 45%, and 60%; Day 1 tornado outlooks include 2% and 10% probability contours as well. For Day 4 and beyond, only 15% and 30% contours are issued, and for significant severe risk, only a single 10% contour is drawn. For more information on SPC's forecasting process, including historical changes to severe weather and product definitions, see Hitchens and Brooks (2014), Edwards et al. (2015), or Herman et al. (2018).

A limited number of published studies have quantified the skill of these convective outlooks and examined their strengths and weaknesses. Hitchens and Brooks (2012) investigated the skill of Day 1 categorical outlooks, and this effort was expanded to include evaluation of Days 2 and 3—among other additions—in Hitchens and Brooks (2014). Early published efforts to verify SPC's convective outlooks probabilistically (e.g., Kay and Brooks 2000) have received renewed attention in Hitchens and Brooks (2017) and more formally in Herman et al. (2018). Collectively, these studies have demonstrated improving skill in short- to medium-range severe weather forecasts in association with improved numerical weather prediction (NWP), though advances have been stagnating somewhat in recent years. These studies have shown that forecast skill is highest at the shortest lead times and gets progressively lower with increasing lead time. In general, wind is the most skillfully predicted severe weather phenomenon with tornado outlooks exhibiting the lowest skill, but this is reversed for significant severe events. Additionally, skill was maximized over the Midwest and Great Plains, and lowest over the South and West. Outlooks are generally most skillful in the winter and spring, and least skillful in the late summer into early autumn. Furthermore, skill is high when at least moderate amounts of both CAPE and wind shear are present, but struggle when CAPE is limited and shear is large, or vice versa (e.g., Sherburn and Parker 2014). As noted above, SPC's convective outlooks are based on only a finite set of probability contours, producing discontinuous jumps in gridded probability fields. Herman et al. (2018) demonstrated that forecast skill is improved, albeit not uniformly, when probabilities are interpreted as interpolated between human-drawn probability contours. In these interpolated outlooks, hail and wind forecasts exhibit an overforecast bias, while tornado and Day 2 and 3 outlooks exhibit a slight underforecast bias. Moreover, their evaluation provides quantitative benchmarks for placing newly developed statistical guidance in the place of existing operational performance.

There have been numerous forays into statistical prediction of severe weather in existing literature. These include applications for statistical prediction of tornadoes (e.g., Marzban and Stumpf 1996; Alvarez 2014; Sobash et al. 2016a; Gallo et al. 2018; McGovern et al. 2019), hail (e.g., Marzban and Witt 2001; Brimelow et al. 2006; Adams-Selin and Ziegler 2016; Gagne et al. 2017; McGovern et al. 2019; Burke et al. 2020), wind (e.g., Marzban and Stumpf 1998; Lagerquist et al. 2017), and severe weather more broadly (e.g., Gagne et al. 2009; Sobash et al. 2011; Gagne et al. 2012; Sobash et al. 2016b). Many of these studies have applied machine learning (ML) to the prediction task; in general, ML techniques have demonstrated great promise in applications to high-impact weather prediction (e.g., McGovern et al. 2017, 2019). In addition to severe weather, ML has demonstrated success in forecasting heavy precipitation (e.g., Gagne et al. 2014; Herman and Schumacher 2018a,b; Whan and Schmeits 2018; Loken et al. 2019), cloud ceiling and visibility (e.g., Herman and Schumacher 2016; Verlinden and Bright 2017), and tropical cyclones (Loridan et al. 2017; Alessandrini et al. 2018; Wimmers et al. 2019). Furthermore, automated probabilistic guidance, including ML algorithms, have been identified as a priority area for integrating with the operational forecast pipeline (e.g., Rothfusz et al. 2014; Karstens et al. 2018). However, many past applications have focused on either much shorter time scales, such as nowcast settings (e.g., Marzban and Stumpf 1996; Lagerquist et al. 2017), or on much longer time scales (e.g., Tippett et al. 2012; Elsner and Widen 2014; Baggett et al. 2018), with less emphasis on the day-ahead time frame and very little model development in the medium-range (e.g., Alvarez 2014). Furthermore, many studies have operated over only a regional domain (e.g., Elsner and Widen 2014) and no study

to date has exactly replicated the operational predictands of SPC's convective outlooks, making it difficult to make one-to-one comparisons between ML study outcomes and operational performance.

One such ML algorithm that has demonstrated success in numerous previous high-impact weather forecasting applications (e.g., Williams et al. 2008; Gagne et al. 2009; McGovern et al. 2011; Williams 2014; Clark et al. 2015; Ahijevych et al. 2016; Elmore and Grams 2016; Herman and Schumacher 2016; Gagne et al. 2017; Herman 2018; Herman and Schumacher 2018b; Whan and Schmeits 2018) is the random forest (RF; Breiman 2001). This study seeks to apply RF methodology to the generation of calibrated probabilistic CONUS-wide forecasts of severe weather with predictands analogous to those of SPC convective outlooks in the hope that the guidance produced can be used to improve operational severe weather forecasting. Section 2 provides further background and describes the data sources used and methodologies employed to create and evaluate these forecasts. Section 3 describes the statistical importance of simulated fields determined by the trained models. Section 4 evaluates the RF forecasts produced and places the results in the context of existing operational forecasts. Section 5 concludes the paper with a synthesis of the findings and a discussion of their implications.

2. Data and methods

Herman and Schumacher (2018b) and its companion paper, Herman and Schumacher (2018a), extensively explored the utility of applying RFs and other machine learning algorithms toward postprocessing global ensemble output to forecast locally extreme precipitation events across the CONUS at Days 2–3. This study follows analogous methodology. A relevant summary of the methodology of Herman and Schumacher (2018a,b) necessary for proper understanding of the methods employed in this study is provided here, but for more detailed explanations of the mathematical underpinnings of RFs and the numerous sensitivity experiments performed therein, the reader is invited to consult those studies. For brevity, several of the RF model configuration choices selected in this study are motivated by the findings of Herman and Schumacher (2018b) rather than reperforming all the same sensitivity experiments for this forecast problem (e.g., input predictor variables). Informal replications of those sensitivity experiments with the severe weather predictands used in this study produced similar findings (not shown).

An RF (Breiman 2001) is an ensemble of unique, weakly correlated decision trees. A decision tree makes successive splits into branches, with each split based on the value of a single randomly selected input predictor at a branch node. The splitting predictor and the value associated with each branch is determined by the combination that best separates severe weather events from nonevents in the supplied model training data. This process is recursive and continues until a termination criterion is satisfied, either because all of the remaining training examples are "events" or "nonevents," or because there are too few remaining training examples to continue splitting. At this point, a "leaf" is produced, which makes a forecast according to the proportion of remaining training examples associated with each event class (e.g., tornado or no tornado). In real-time forecasting, new inputs are supplied and the tree is traversed from its root according to the input values until a leaf is reached, which becomes the real-time prediction of the tree. An RF produces numerous unique and diverse decision trees by considering different subsets of training data and input features (i.e., predictors) for each tree generation process, as well as randomizing the input features at each branch split in the tree. An RF's forecast is simply calculated as the mean probabilistic forecast issued by the trees within the forest (e.g., Breiman 2001).

a. *Designing the random forests*

RF predictor information comes from NOAA's Second Generation Global Ensemble Forecast System Reforecast (GEFS/R) dataset (Hamill et al. 2013). The GEFS/R is a global, convection-parameterized 11-member ensemble with T254L42 resolution—which corresponds to an effective horizontal grid spacing of ~55 km at 40° latitude—initialized once daily at 0000 UTC beginning in December 1984. Perturbations are applied only to the initial conditions, and are made using the ensemble transform with rescaling technique (Wei et al. 2008). The ensemble system used to generate these reforecasts is nearly static throughout its 30+ year period of coverage, though updates to the operational data-assimilation system over time have resulted in some changes in the bias characteristics of its forecasts over the period of record (Hamill 2017). Most surface (or column-integrated) fields are preserved on the native Gaussian grid (~0.5° spacing), while upper-level and some other fields are available only on a 1° × 1° grid. Based on findings from Herman and Schumacher (2018b), this study derives predictors from the GEFS/R ensemble median. Model training employs a 9-yr training period, using daily initializations from 12 April 2003 to 11 April 2012. Temporally, forecast fields are archived every 3 h out to 72 h past initialization, and are available every 6 h beyond that. Accordingly, the RFs trained in this study use 3-hourly predictors for Day 1 (forecast hours 12–36)

TABLE 1. Summary of dynamical model fields examined in this study, including the abbreviated symbol to which each variable is referred throughout the paper, an associated description, the predictor group with which the field is associated in the manuscript text, and the highest resolution for which the field can be obtained from the GEFS/R. Variable symbols with an asterisk are used only in the Day 1 models.

Symbol	Description	Grid	Calculated	Class
APCP	Precipitation accumulation in past (3) 6 h	Native Gaussian	Archived	None
CAPE	Surface-based convective available potential energy	Native Gaussian	Archived	Thermodynamic
CIN	Suface-based convective inhibition	Native Gaussian	Archived	Thermodynamic
MSLP	Mean sea level pressure	Native Gaussian	Archived	Kinematic
PWAT	Total precipitable water	Native Gaussian	Archived	Thermodynamic
Q2M	Specific humidity two meters above ground	Native Gaussian	Archived	Thermodynamic
RH2M*	Relative humidity two meters above ground	1° × 1°	Derived	Thermodynamic
SHR500	Bulk wind difference magnitude between 10 m and 500 hPa	1° × 1°	Derived	Kinematic
SHR850	Bulk wind difference magnitude between 10 m and 850 hPa	1° × 1°	Derived	Kinematic
SRH*	Storm relative helicity from surface to 850 hPa	1° × 1°	Derived	Kinematic
T2M	Air temperature two meters above ground	Native Gaussian	Archived	Thermodynamic
U10	Zonal component of 10-m wind	Native Gaussian	Archived	Kinematic
UV10	10-m wind speed	Native Gaussian	Derived	Kinematic
V10	Meridional component of 10-m wind	Native Gaussian	Archived	Kinematic
ZLCL*	Height of lifted condensation level	1° × 1°	Derived	Thermodynamic

and 2 (hours 36–60) forecasts, and 6-hourly temporal resolution for Day 3 (hours 60–84).

Several GEFS/R simulated atmospheric fields with known or postulated physical relationships with severe weather are used as RF predictors (Table 1), referred to interchangeably as ''features.'' These include surface-based CAPE and CIN, 10-m winds (U10, V10, UV10); surface temperature and specific humidity (T2M, Q2M), precipitable water (PWAT), accumulated precipitation (APCP), wind shear from the surface to 850 and to 500 hPa (SHR850, SHR500), and mean sea level pressure (MSLP). For Day 1, three additional predictors are supplied: surface relative humidity (RH2M), lifting condensation level height above ground (ZLCL), and surface to 850 hPa storm relative helicity (SRH), approximated following Ramsay and Doswell (2005) as described in the appendix. Some of these variables are archived natively by the GEFS/R, while others are derived based on stored fields that are available. The full list of fields, their class, whether they are natively archived or derived, and the grid from which they are sampled is included in Table 1. Descriptions of how derived variables are calculated is provided in the appendix. For each field, in addition to sampling the temporal variation of the fields throughout the forecast period as noted above, spatial variations in the simulated fields are included as inputs to the RF. Specifically, predictors are constructed in a forecast point-relative sense, with predictors up to three grid boxes (1.5° or 3°, depending on the predictor) displaced in any horizontal direction relative to the forecast point. Forecasts are made on the Gaussian grid; for predictors on the 1° grid, the nearest point to the Gaussian point is used as the central point on that grid. In addition to this suite of meteorological predictors, forecast point latitude, longitude, and the Julian day associated with the forecast are included as predictors.

SPC storm reports are used as observed severe weather events for training and used equivalently for verification, obtained from the SPC Severe Weather Database (SPC 2017a). A 40-km neighborhood radius is used to classify GEFS/R grid points as nonevents/events/significant events (encoded with a binary 0 or 1 for nonevents and events, respectively); a grid point is encoded as an event for the day if a report of the relevant severe weather type is reported during the 1200–1200 UTC valid period anywhere within 40 km of the grid point. While 40 km is somewhat comparable, though generally smaller, than the GEFS/R effective grid spacing, every report gets encoded as an event for at least one model grid point. SPC outlooks are verified in a similar manner, but on a higher-resolution grid. Sensitivity to grid choice was explored in Herman and Schumacher (2018b) and found not to play an important role in the verification statistics. Past studies have shown that there are significant changes in reporting trends in the SPC report record

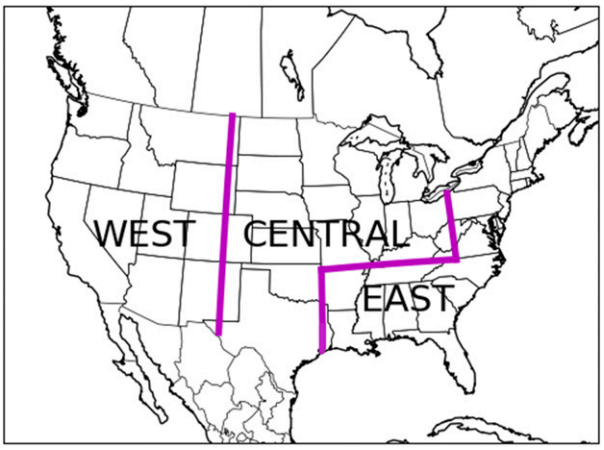

FIG. 1. Map depicting the training regions of the CONUS for the statistical models used in this study. The west region is bounded by 25°–49°N, 240°–255°E; the central region by 25°–36.5°N, 255°–265.4°E and 36.5°–49°N, 255°–279.5°E; and the east region by 25°–29°N, 277°–280.2°E; 29°–36.5°N, 265.4°–285°E; and 36.5°–49°N, 279.5°–294°E.

(Trapp et al. 2006; Verbout et al. 2006; Doswell 2007; Agee and Childs 2014), but the trends are negligible over this study's evaluation period, and the period does not overlap with any infrastructural changes (e.g., introduction of the WSR-88D network) that rapidly changed report counts over a short time period (e.g., Agee and Childs 2014). Thus, reporting trends are not a significant concern for this study. Additionally, while the SPC outlooks underwent a significant change in October 2014, when ''marginal'' and ''enhanced'' categories were added to outlooks on Days 1–3 (Jacks 2014), the changes only remapped categorical definitions to the unchanged probability contours.

Based on different diurnal and seasonal climatologies (e.g., Brooks et al. 2003; Nielsen et al. 2015; Krocak and Brooks 2018), and due to differing regimes and storm systems primarily responsible for severe weather across the CONUS (e.g., Smith et al. 2012), the country is partitioned into three regions: west, central, and east (Fig. 1). This study develops separate RFs for each of the three regions of the CONUS, with unique forests trained also for each of the five predictand lead-time combinations: 1) Tornado Day 1, 2) Hail Day 1, 3) Wind Day 1, 4) Severe Day 2, and 5) Severe Day 3. For the Day 1 models, the severity levels of the category are retained using a 3-category predictand (none, non-''significant'' severe, ''significant'' severe), while the severity levels are aggregated for longer lead times. Each of the 15 forests is trained using the 9-yr historical record, regardless of the existence of severe storm reports or SPC outlooks for a given day. As noted above, the focus of this study is on the model evaluation rather than on involved sensitivity experiments and parameter tuning. Models were trained and evaluated using Python's Scikit-Learn library (Pedregosa et al. 2011); deviations from defaults for this study were made based on a combination of performance considerations and computational constraints. The only parameters varied were the forest size and minimum number of training examples required to split an impure node in a decision tree. Informal tests were performed to assess parametric sensitivity (i.e., forest size and number of training examples) to the different hazard types, which determined relatively small sensitivity to forecast skill (not shown). For the interested reader, the final values used are furnished in Table 2.

b. Evaluation and analysis

Trained RFs are evaluated in two distinct ways. First, in section 3, the statistical relationships diagnosed by the RFs [feature importances (FIs)] are investigated to determine the insights gleaned about the forecast problem and assess whether the models are making predictions in ways consistent with our external understanding of the forecast problem. Due to the number and size of trees in a forest, it is not practical to investigate the complete structure of each tree in the forest; instead, FIs are used to capture the extent of use of predictor information in generating a final prediction. Though there are several ways that FIs can be quantified (e.g., Strobl et al. 2007, 2008; McGovern et al. 2019), this study uses the so-called ''Gini importance'' metric (e.g., Pedregosa et al. 2011; Herman and Schumacher 2018a; Whan and Schmeits 2018); we refer readers to McGovern et al. (2019) for a more complete assessment of various feature importance techniques, including the pros and cons of each.

TABLE 2. Parameter summary for the different RFs trained in the study. All RFs for a given region and lead time employ the same parameters, forest size and the minimum number of samples permitted to split an impure node. For more details, see Pedregosa et al. (2011) and Herman and Schumacher (2018b).

	West		Central		East	
Lead time	Forest size	Min No. of samples	Forest size	Min No. of samples	Forest size	Min No. of samples
Day 1	500	30	500	120	1000	120
Day 2	1000	30	1000	120	1000	120
Day 3	1000	30	1000	120	1000	120

A single FI is attributed to each input feature, and may be conceptualized as the number of node splits based on the given feature, weighted in proportion to the number of training examples encountering the split (Friedman 2001). The FIs are summed over each split in the tree for each tree in the forest, and normalized so that the sum of all FIs is unity. FIs thus range between zero and one, with larger values indicating that the associated predictor has more influence on the prediction values. In the extremes, an FI of zero means that the predictor has no influence on the prediction made by the RF, while a value of one indicates that the value of the associated predictor uniquely specifies the predictand. As noted above, input predictors to the RF vary in simulated forecast field, forecast time, and in space relative to the forecast point. In many cases, it is convenient to present FIs—which are calculated at every predictor point in time and space—summed or normalized over one or more of these dimensions to provide a summary aspect of which fields, times, and locations are being most and least used in generating predictions for different severe weather phenomena.

Second, in section 4, the probabilistic performance of the models is evaluated. The trained RFs are used to generate probabilistic convective outlooks over 4.5 yr of withheld model data spanning 12 April 2012–31 December 2016. Model skill is evaluated through the Brier skill score (BSS; Brier 1950), using an informed climatological reference, identical to official SPC severe climatologies (Kay and Brooks 2000; SPC 2017b) as described in Herman et al. (2018), while forecast calibration and resolution are assessed via reliability diagrams (Murphy and Winkler 1977; Bröcker and Smith 2007; Wilks 2011) and fractional coverage of severe weather reports (e.g., Erickson et al. 2019). While forecasts are evaluated in aggregate, they are also assessed both spatially and seasonally in order to assess the times and locations where the RFs perform most and least skillfully. To evaluate the RFs continuously across the CONUS, forecasts generated over the three geographic regions are stitched together using spatial smoothing with a sigmoid function at the regional borders so as to eliminate probability discontinuities. Additionally, following Herman et al. (2018), outlook skill is evaluated based on the large-scale environmental conditions associated with the forecast, as quantified based on CAPE and deep-layer bulk wind difference (hereafter referred to as shear) in the North American Regional Reanalysis (NARR; Mesinger et al. 2006). Findings are contextualized by comparing the RF performance against SPC convective outlooks for the same predictands issued with comparable lead times. Consistent with Herman et al. (2018), Day 1 SPC outlooks evaluated in this study come from the 1300 UTC forecast issuance, while Day 2 and 3 outlooks come from the 0100 CT (0600 or 0700 UTC) and 0230 CT (0730 or 0830 UTC) forecast issuances, respectively. Because the interpolated SPC probability grids verified more skillfully than the uninterpolated outlooks (Herman et al. 2018), the interpolated grids are used as the benchmark for comparison in this study. In most cases, the entire evaluation period is used for the comparison; due to data availability constraints, a slightly shorter 13 September 2012–31 December 2016 period is used for Day 2 and 3 verification, while 12 April 2012–31 December 2014 is used for the evaluation in the CAPE-versus-shear parameter space.

As a final evaluation of the operational utility of the ML-based forecast guidance provided by the trained RFs, a weighted blend of the SPC and RF-based convective outlooks is evaluated over the aforementioned periods at Days 1–3; the level of skill improvement, if any, quantifies the value added by the addition of the ML guidance to the operational forecast pipeline and human forecast process. The evaluation period is segmented into four quarters and weights are prescribed to the SPC and RF forecasts by using the average BSS of the two component outlooks based on the other three segments of the evaluation period that exclude the forecast being weighted. Weights are prescribed in this manner:

$$W_{\mathrm{SPC}} = \frac{\dfrac{1}{1-\mathrm{BSS}_{\mathrm{SPC}}}}{\dfrac{1}{1-\mathrm{BSS}_{\mathrm{SPC}}} + \dfrac{1}{1-\mathrm{BSS}_{\mathrm{RF}}}}; \quad W_{\mathrm{RF}} = 1 - W_{\mathrm{SPC}}. \tag{1}$$

In the event that one BSS is negative, the weight associated with that forecast is set to zero with the other set to one. In this way, if either forecast set has no climatology-relative skill on the portion of the evaluation period used to generate the weights, it does not contribute to the blended forecasts, while if either forecast set is perfect, it completely determines the blended forecast. Statistical significance of both the absolute climatology-relative skill and comparisons between forecast sets are assessed using bootstrapping whereby random samples of forecast days are sampled with replacement among the evaluation period to produce a realistic range of Brier and climatological Brier scores for each evaluated forecast set or forecast set comparison. Other uncertainty analysis follows the methods of Herman and Schumacher (2018b) and Herman et al. (2018); more details may be found there.

3. Results: Model internals

Predictive utility of different simulated atmospheric fields (Fig. 2) is found to vary somewhat by forecast

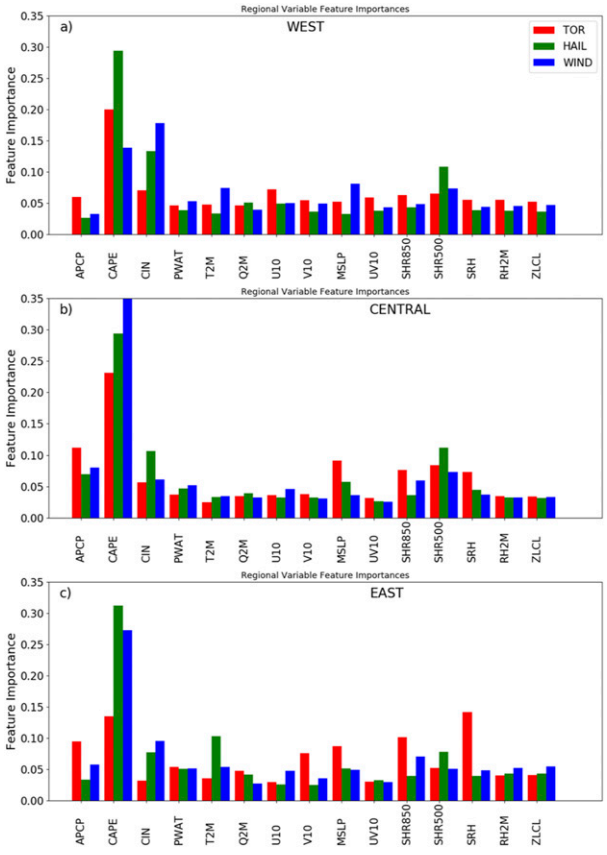

FIG. 2. FIs aggregated by atmospheric field for the Day 1 models in the (a) west, (b) central, and (c) east regions. Red bars correspond to FIs for the tornado predictive model, green bars to the hail predictive model, and blue bars to the wind predictive model for each region.

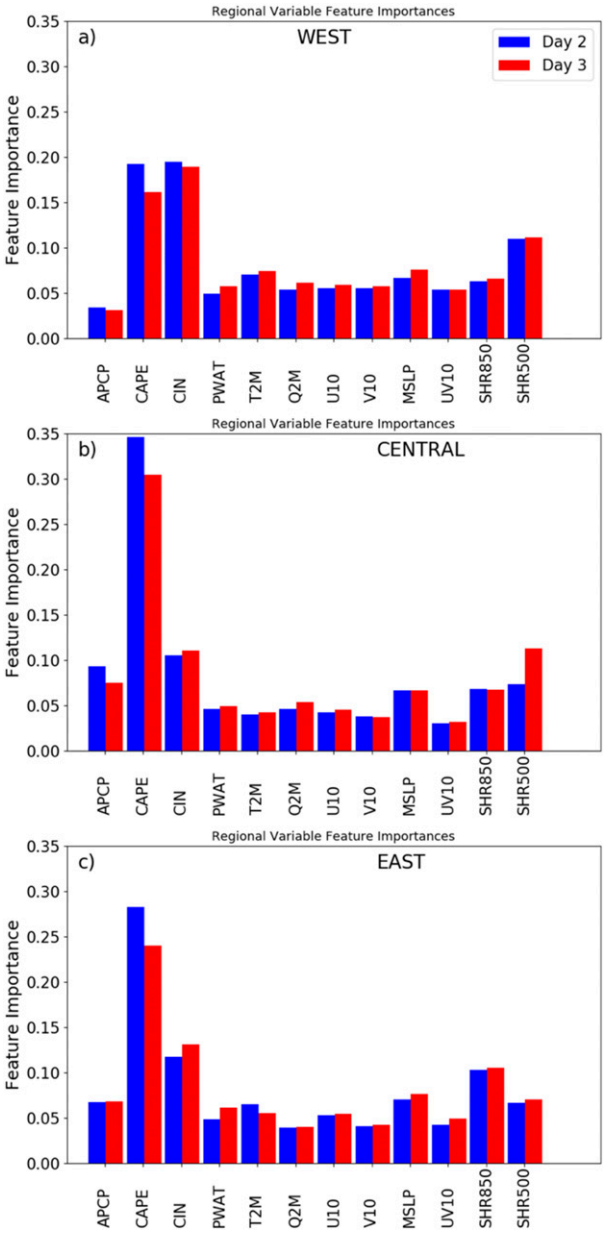

FIG. 3. As in Fig. 2, but for the Day 2 and 3 models. Day 2 and 3 FIs are indicated in blue and red bars, respectively.

region and severe predictand. Under almost all circumstances, CAPE is found to be the most predictive severe weather predictor by a fair margin, particularly for predicting hail and wind. CIN is generally identified as far less predictive, but still more so than other fields. The west is an exception, with CIN identified as quite predictive of hail and especially severe wind and actually having higher FIs than CAPE for wind (Fig. 2a). All fields contribute some to the output of each model, with a relatively balanced distribution outside of the more predictive fields. In addition to CAPE and CIN, SHR500 is found to be fairly predictive as well, and this is most evident for hail (Fig. 2). For tornadoes, shear over a shallower layer (SHR850) is found to be equally (e.g., Figs. 2a,b) or more (Fig. 2c) predictive than SHR500, and one of the more predictive variables overall. Other variables with high FIs for tornadoes include APCP, MSLP, and SRH. MSLP may be characterizing the synoptic environment and helping distinguish favorable from unfavorable environmental conditions for tornadoes. Additionally, SRH has often been noted as a predictive variable for determining tornado potential (e.g., Davies and Johns 1993; Thompson et al. 2007), and is found to be the most predictive field in the East (Fig. 2c). Overall, the RFs are largely following conventional wisdom about human forecasting of severe weather: CAPE and shear are some of the most important fields to consider; deep-layer bulk shear is important for hail and wind forecasts; bulk shear over a shallow layer and helicity (i.e., SRH) are good predictors for

tornado occurrence; and the kinematic environment plays a more significant role overall for tornadoes than for severe hail and wind. The RFs have simply learned these statistical relationships objectively and empirically based on analysis of many historical cases, and provide a quantitative assessment of variable importance.

In predicting any severe weather beyond Day 1 (Fig. 3), the trends largely follow the findings for hail and wind in their respective regions. Considering that the vast majority of severe observations are either hail or wind, that the FIs track those of hail and wind more closely than tornadoes is not surprising. CAPE and CIN are about equally predictive of severe weather at Days 2 and 3 in the west (Fig. 3a), with SHR500 the next most predictive. The relative ranking mostly holds for the Central and East regions (Figs. 3b,c), although CAPE is much more predictive than CIN, especially in the Central region. SHR850 becomes increasingly important with longitude, and is interestingly identified as more indicative of severe weather in the east region at these longer lead times. Importances are mostly similar between days, though CAPE importance tends to decline slightly from Day 2 to 3 (Fig. 3) and is distributed among the other fields. This is perhaps attributable to the noisy and highly sensitive nature of the CAPE field yielding less predictive utility with increasing forecast lead time and associated increasing uncertainty.

FIs may be further stratified along the time dimension (Fig. 4), revealing a clear diurnal peak in importance of model information throughout the forecast period, although in all cases the peak is much more uniformly distributed relative to the diurnal event climatology in the region. In the extreme, tornadoes in the west (Fig. 4a), there is little peak at all in FIs. In some cases, notably in the east (Figs. 4c,f,i), the importance peak is aligned with the climatological event maximum, while in other situations, it leads (e.g., Fig. 4h) or lags (e.g., Figs. 4d,e,g) the maximum. This result could be an initiation bias—particularly in the lagging cases—while it could also be attributable to the forecasted pre (or post) event environment being more predictive than the simulated evolution at event time. Breakdowns into thermodynamic and kinematic variables (Table 1) reveals that the thermodynamic variables are much more predictive of hail and wind than the kinematics, while the two classes are about equally predictive for tornadoes. Furthermore, while the thermodynamics have a sharp diurnal peak, the importance of the kinematic variables has little temporal dependence throughout the forecast period (Fig. 4). FI time series for Day 2 and 3 models (not shown) share similarities with their Day 1 counterparts, with importance peaks earliest in the east and latest in the west, and nearly constant predictive utility of simulated shear (i.e., sum of SHR500 and SHR850) across the forecast period (not shown).

In space (Fig. 5), RF FIs are typically highest near the forecast point and decrease with increasing distance from the point, but there are some notable anomalies. FIs are generally most spatially uniform for tornado prediction and have the sharpest peak in predicting severe hail; this is especially true in the west (cf. Figs. 4a,d). In the west, while FI maxima are collocated with the forecast point for tornadoes and wind, information to the east of the forecast point is more predictive of conditions at that point than the collocated simulated forecast values for hail and the medium-range forecasts. A variety of factors could be attributable to this observation, including a displacement or initiation bias in the model's placement of storms in the region, or the lopsided event climatology in the region, with most events occurring on the eastern fringes of the west region (see unfilled contours in Figs. 6 and 7). Additionally, the signal could be representative of poor forecast predictability over the Intermountain West due to orography or air mass interactions, whereas the synoptic environment over the Great Plains is better depicted and forecast by the GEFS/R, but more investigation is required to validate many of these dynamic hypotheses and is beyond the scope of this study. In the central region, FIs are highest from the forecast point south, with maxima southeast of the center point for every predictand except severe winds (Figs. 5b,e,k,n), which has an identified maximum in predictive utility southwest of the forecast point (Fig. 5h). The southern displacement in importance appears to become more pronounced with increasing forecast lead time, and is especially evident at Day 3 (Fig. 5n). FI maxima also become less pronounced with increasing forecast lead time (Figs. 5j–o), consistent with past studies (e.g., Herman and Schumacher 2018a). In the East, importances for all severe weather models maximize near the forecast point and extend to the south and west.

In summary, the RFs trained in this study appear to be making statistical deductions that are generally consistent with our current physical understanding of how these predictors–CAPE, CIN, SRH–may influence severe weather, and identify forecast fields to inspect that agree with conventional operational severe weather forecast practices (e.g., Johns and Doswell 1992). However, the RF provides an automated, objective, and quantitative synthesis of these many important factors that contribute to a skillful severe

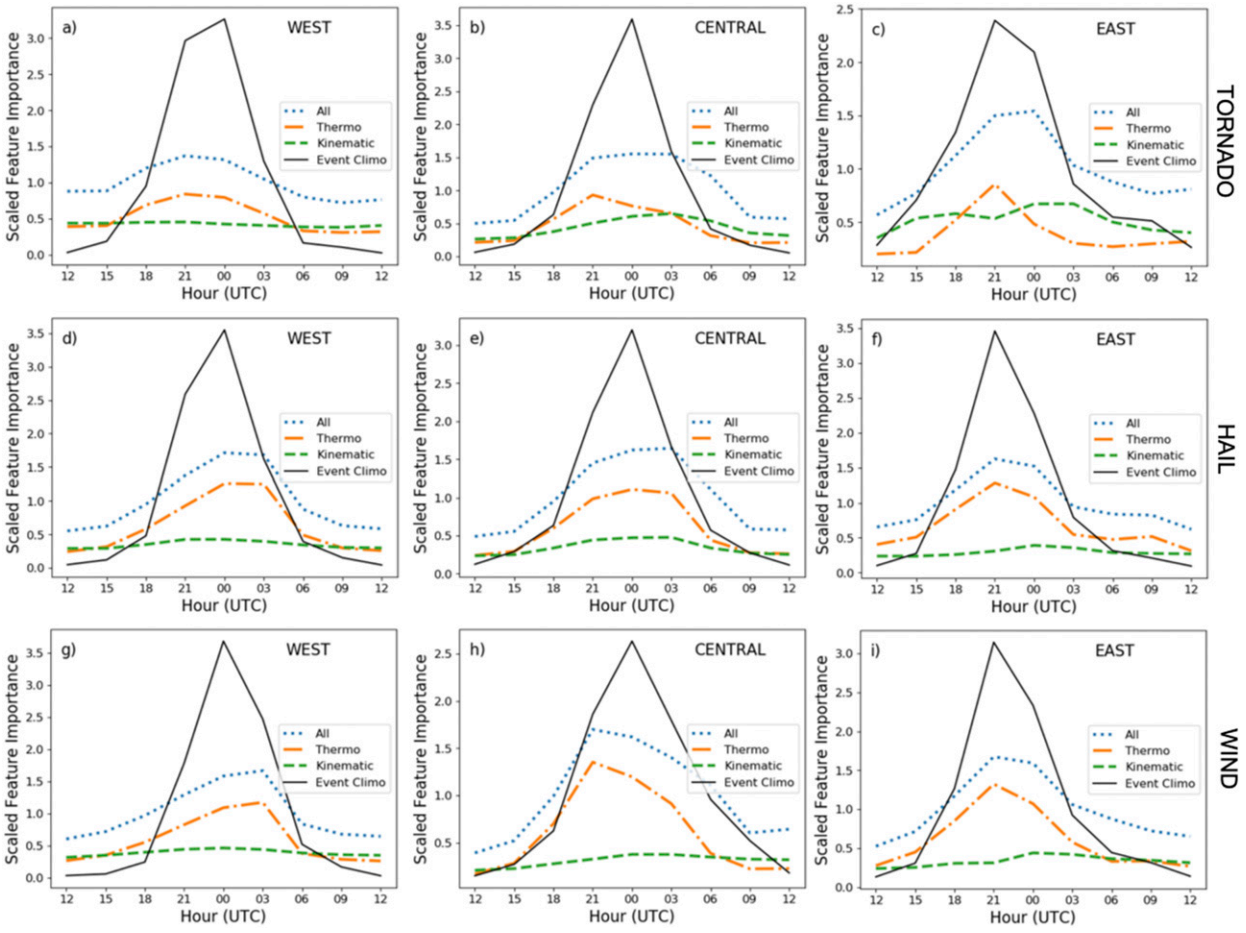

FIG. 4. Normalized FIs aggregated as a function of forecast hour for the Day 1 models. The top, middle, and bottom rows depict FIs for the tornado, hail, and wind models, respectively, while the left, center, and right columns, respectively, depict FIs for the west, central, and east regions. Severe phenomenon diurnal climatologies are depicted for each region in black. These and the total FIs, colored as indicated in the panel legend, are normalized so that the curve integrates to unity. FI time series broken down by thermodynamic and kinematic variables are also included, with lines as colored in the panel legend and using the variable partitioning depicted in Table 1.

weather forecast. The following section investigates the predictive performance of these models.

4. Results: Model performance

a. Model skill

The RFs show ability to skillfully predict all severe weather predictands (Fig. 6), though there are some differences in the details. Prediction of tornadoes (Fig. 6a) produced the most mixed verification results, with statistically significant positive skill over the central Great Plains, Mississippi Valley, Ohio River valley, and parts of the mid-Atlantic region and Florida Peninsula. However, BSSs are lower and in many cases less skillful than climatology—albeit not statistically significantly so—over the West, Northeast, upper Midwest, far northern and southern plains, and the Carolinas. These same general findings extend for significant tornadoes (Fig. 6b) but with lower skill overall, with CONUS-wide BSS decreasing from 0.029 for tornadoes to 0.013 for significant torrnadoes. The large area of extremely negative skill over the West is simply reflective of the fact that no significant tornadoes were observed over this region during the verification period, and the model had above climatological probabilities for some events. Due to the small or even nonexistent sample, the negative skill observed here is not statistically significant. Hail (Fig. 6c), wind (Fig. 6e), and the Day 2 and 3 (Figs. 6g,h) models all exhibit very similar spatial patterns of forecast skill, with near-uniform and statistically significant positive skill over much of the CONUS east of the Rocky Mountains. Somewhat degraded skill is seen over southern Texas, Florida, and pockets of the upper Midwest; these spatial variations are particularly pronounced in the hail verification (Fig. 6c). In the West, fewer of the results are found to be statistically significant due to the reduced event

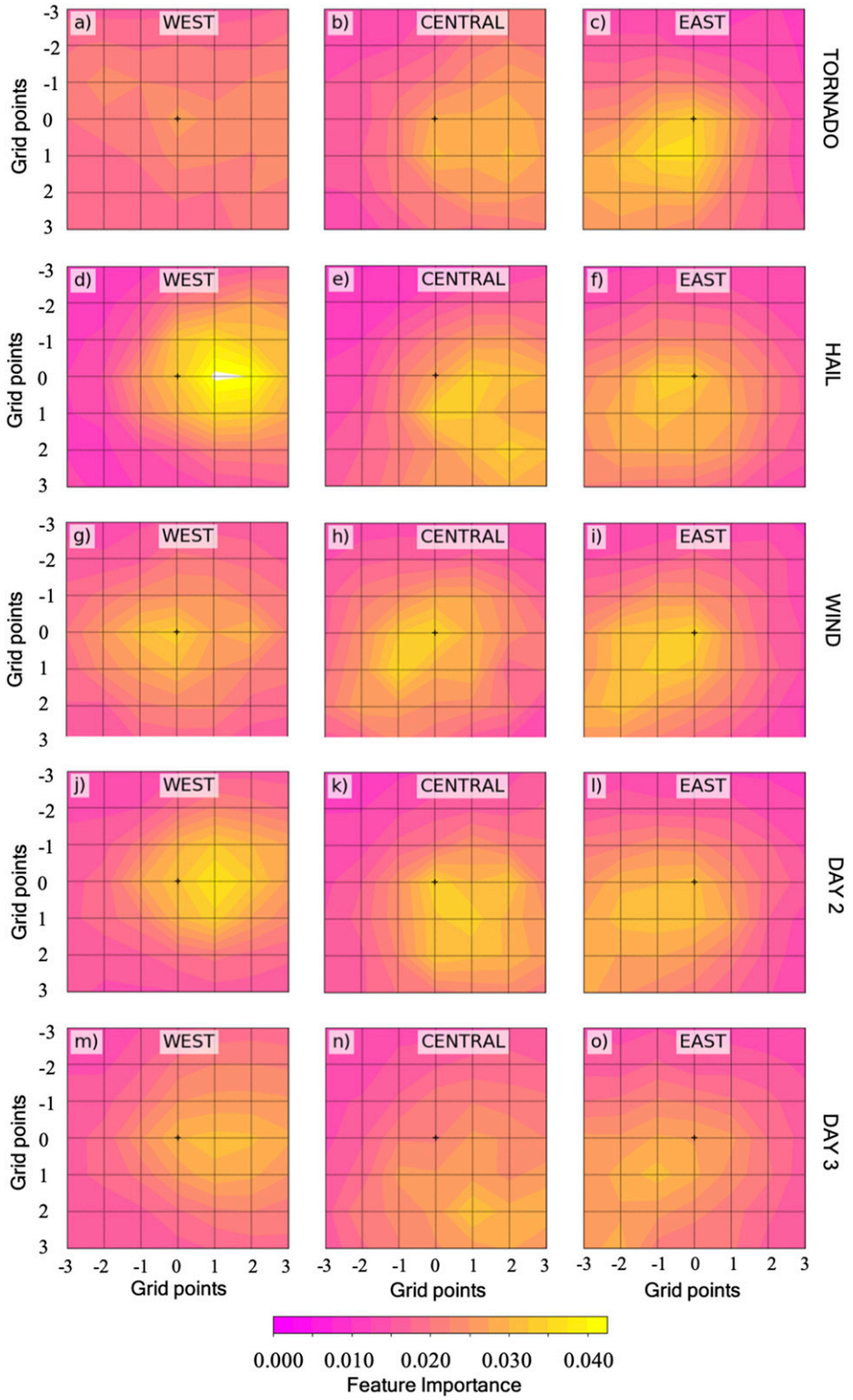

FIG. 5. FIs summed over all predictor variables and forecast times according to the corresponding predictor's position in gridpoint-relative space for the west, central, and east regions, respectively, in the left, center, and right columns. Tornado model FIs are depicted in the top row, followed by hail, wind, Day 2, and finally the Day 3 model on the bottom row. Yellows indicate high importance of information at the point, while magentas indicate lesser importance. The forecast point is shown with a black cross.

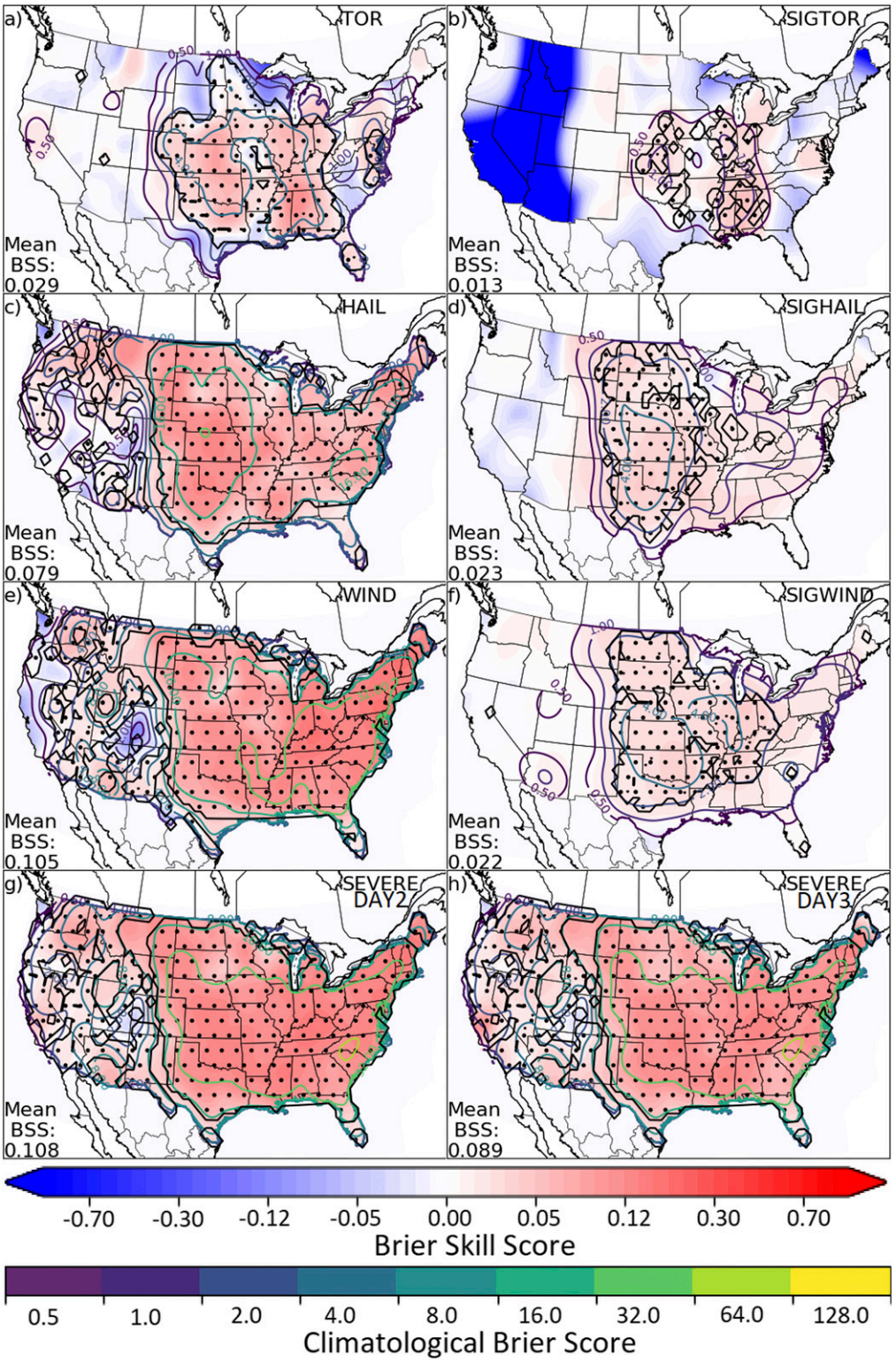

FIG. 6. Brier skill scores (filled contours) in space evaluated over the 12 Apr 2012–31 Dec 2016 verification period for each of the ML models trained in this study. (a)–(h) Performance of the tornado, significant tornado, hail, significant hail, severe wind, significant severe wind, Day 2, and Day 3 outlooks, respectively. Unfilled contours depict the Brier score of climatology at the point over the verification period; higher values indicate more common events. Stippling indicates areas where the sign of the skill score is statistically significant at 95% obtained from bootstrapping as described in the text.

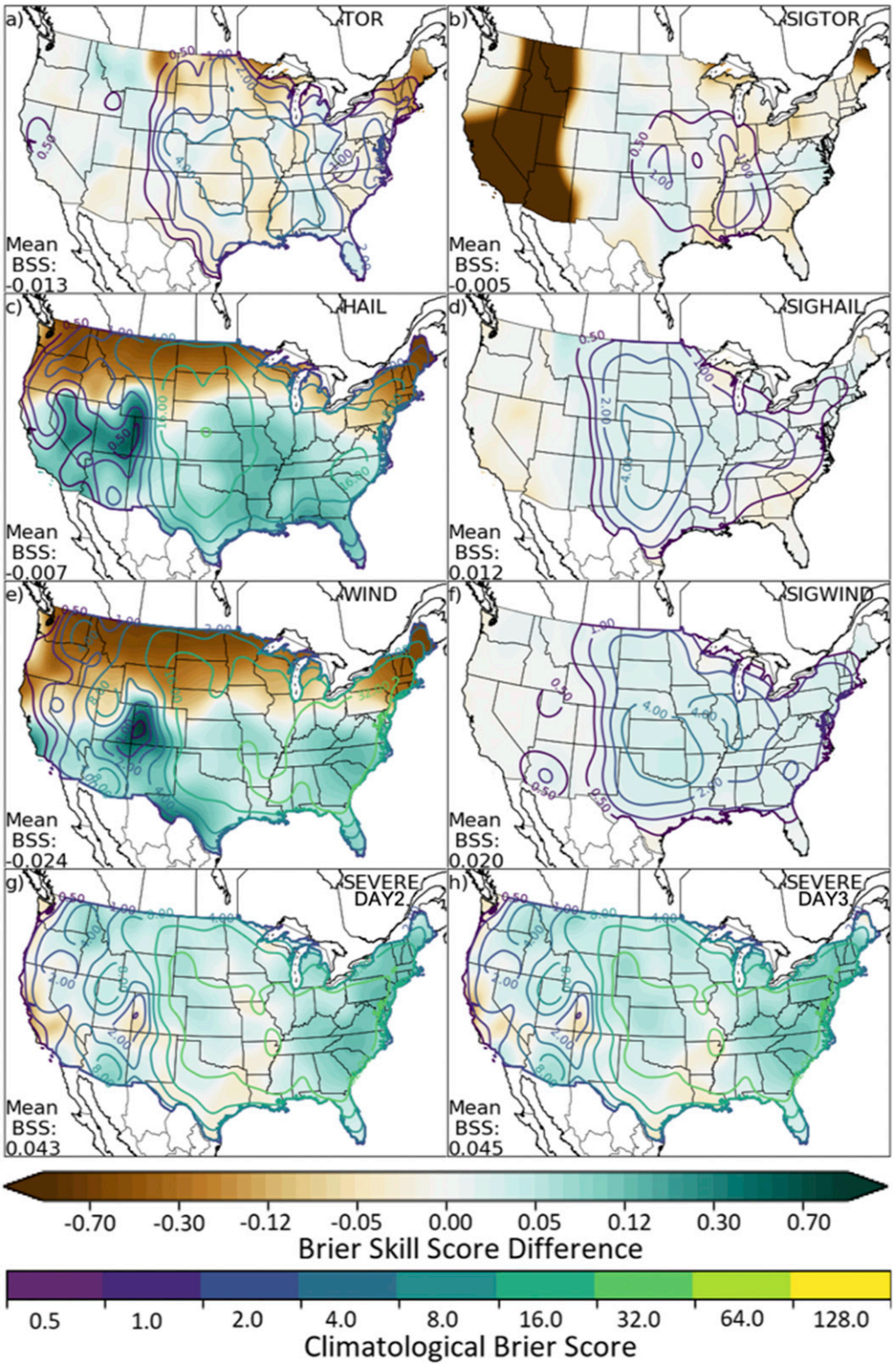

FIG. 7. As in Fig. 6, but it depicts the difference in BSS between ML outlooks and the analogous outlooks issued by SPC. Greens indicate ML forecasts outperform SPC; browns indicate the opposite. Due to data availability, a slightly shorter 13 Sep 2012–31 Dec 2016 period is used for the Day 2 and 3 outlook verification comparison.

frequency. Nevertheless, positive skill is still noted for these predictands over much of the West, with the exception of a pocket of southwestern Colorado and surroundings and the Pacific Coast. As with SPC convective outlooks (Herman et al. 2018), Day 1 forecast skill is highest for severe winds at 0.105, with hail in the middle at 0.079. Skill unsurprisingly decreases with increasing forecast lead time, and CONUS-wide BSSs of 0.108 and 0.089 are observed for Day 2 and Day 3 RF outlooks, respectively (Figs. 6g,h). Like with tornadoes, the spatial patterns are similar between hail and wind and their significant severe counterparts (Figs. 6d,f), except with lower skill magnitudes with CONUS-wide numbers of 0.023 and 0.022 for significant hail and wind. The highest (and statistically significant) skill exists over the central plains for these variables; positive but insignificant skill is observed in the East, and skill near climatology observed over much of the West.

Relative to SPC, the RF outlooks verify quite competitively (Fig. 7). On Day 1, where human forecasters have access to more skillful convection-allowing guidance and more updated observations and simulations, SPC outlooks are generally more skillful than the RF (negative BSS differences), with aggregate skill score differences of −0.007 for hail (Fig. 7c) increasing to −0.013 for tornadoes (Fig. 7a) and −0.024 for severe wind forecasts (Fig. 7e). However, the CONUS-wide summary gives an incomplete picture, as there are significant regional variations in skill differences. Unlike the RF outlooks, which exhibited fairly uniform skill in hail and wind across the eastern two-thirds of CONUS (Figs. 6c,e), SPC interpolated convective outlooks exhibited a strong latitudinal gradient in BSS, with higher skill to the north (Herman et al. 2018). This is reflected in the skill comparison, with SPC outlooks substantially outperforming the RF outlooks over far northern CONUS in predicting severe hail and wind (Figs. 7c,e). However, over the southern two-thirds of CONUS, the RF outlooks outperform the SPC outlooks in these fields. There is much more spatial inhomogeneity in the tornado outlooks (Fig. 7a). The magnitudes of the skill differences at a point are usually much smaller than in the hail and wind outlooks, but SPC outlooks still outperform the RF forecasts the most in the northern tier of states. The mixed spatial skill comparisons for tornadoes extend to verification of significant tornadoes (Fig. 7b) as well, but the comparison is much different for significant hail (Fig. 7d) and wind (Fig. 7f) events. Here, RF outlooks are actually found to exhibit higher probabilistic skill overall than the SPC outlooks, with skill differences of 0.012 and 0.020, respectively, for the significant severe hail and wind outlooks. The gains are largest over the central region.

For Day 2 and 3 periods (Figs. 7g,h), the RF outlooks exhibit higher probabilistic skill than the analogous SPC forecasts, with aggregate CONUS-wide skill differences of 0.043 and 0.045, respectively, for the Day 2 and 3 outlooks. RF outlooks demonstrate higher skill over almost all parts of CONUS, the primary exceptions being the Pacific Coast and western Colorado where the RFs had lower absolute skill (e.g., Fig. 6g), and over Louisiana, Arkansas, and eastern Texas. The biggest skill differences over SPC are in the east region domain, particularly the mid-Atlantic and southern New England. The general finding that the RF outlook skill becomes increasingly skillful relative to SPC outlooks with increasing forecast lead time is consistent with there being less information beyond global, convection-parameterized ensemble guidance on which to base a skillful forecast with increasing lead time, with the biggest jump between Days 1 and 2. Other factors (e.g., human-based) that may contribute to forecast skill differences at these longer lead times are presented in the summary section.

Except for hail, which exhibits a springtime maximum in skill (Fig. 8c), all RF outlooks exhibit a climatology-relative peak in skill during the cold-season (Figs. 8a,e,g). In fact, hail exhibits essentially an inverted seasonal cycle in forecast skill compared with the other variables, since hail outlooks verify worst in the winter and other variables verify worst in March. Tornadoes and wind also exhibit a skill minimum in late summer–early autumn, consistent with SPC outlooks (Herman et al. 2018). For all severe weather predictands, the severe and significant severe events have nearly identical seasonal cycles in forecast skill (Figs. 8a,c,e). Comparing against SPC, while there does not appear to be a clear seasonal or monthly signal in the skill difference for tornado outlooks (Fig. 8b), the primary advantage for SPC outlooks over the RF counterparts in hail and wind appears to come in the month of July, where SPC outlooks performed very well (Herman et al. 2018) and substantially outperform the RF outlooks.

In contrast, in the Day 2 and 3 comparison, RF outlooks outperform SPC by the most during the summer, maximizing in July (Fig. 8h). These differences are all consistent with the SPC being able to effectively harness the advantages of convection-allowing guidance for their Day 1 convective outlooks over the warm season, where the responsible physical processes are predominantly smaller-scale and more weakly forced than cold-season events. At Day 2 and 3, where convection-allowing guidance is largely unavailable, efforts to forecast severe weather are hampered by biased guidance that cannot come close to resolving

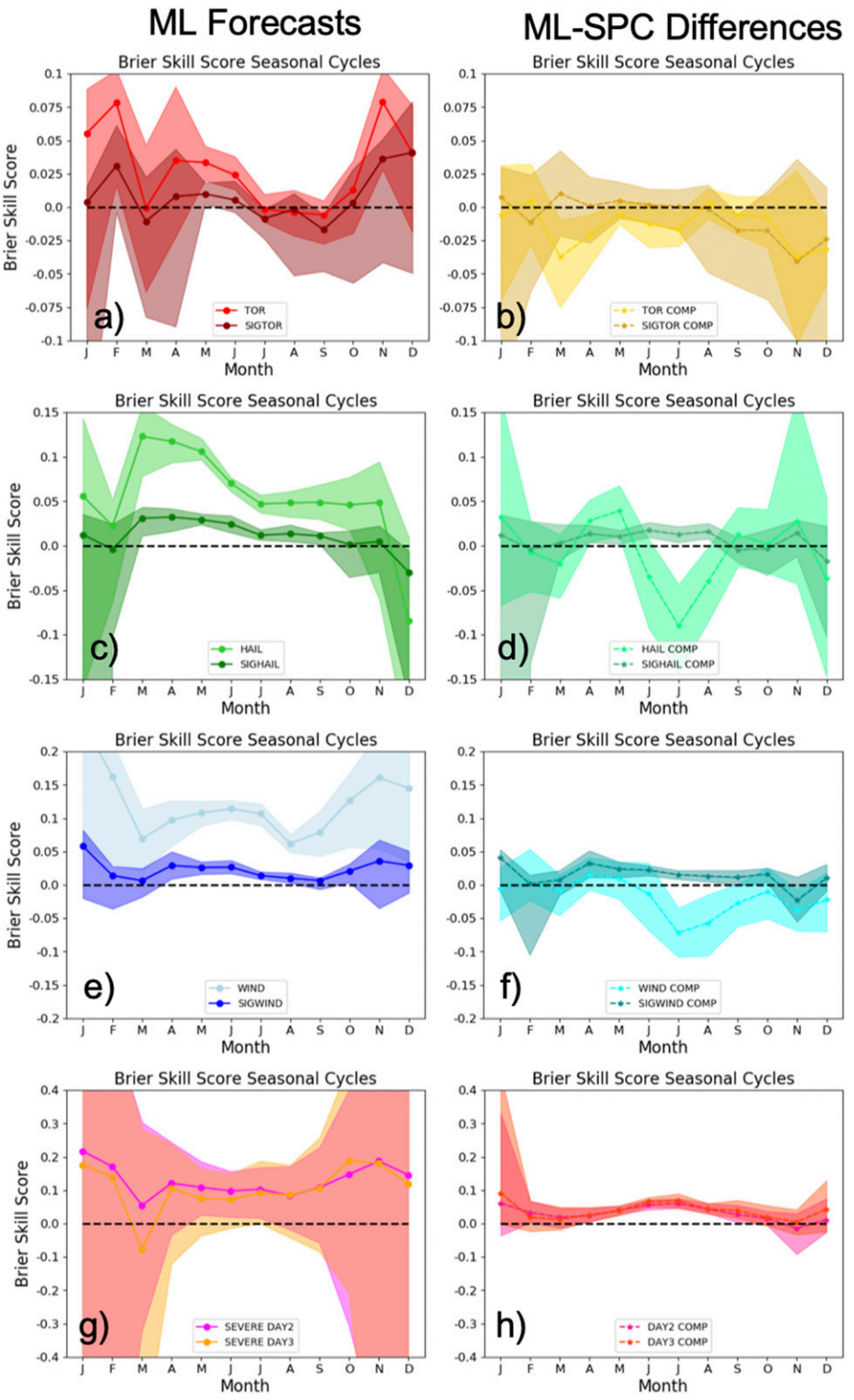

FIG. 8. (left) BSSs and (right) BSS differences by month between ML and SPC outlooks for (a),(b) tornado and significant tornado, (c),(d) hail and significant hail, (e),(f) wind and significant wind, and (g),(h) Day 2 and 3 outlooks. Lines are colored as indicated in the panel legend; shading about the line indicates 95% confidence bounds obtained by bootstrapping. Differences (COMP) are ML − SPC, positive numbers indicating ML outperforms SPC. Note that the *y* axis varies between rows.

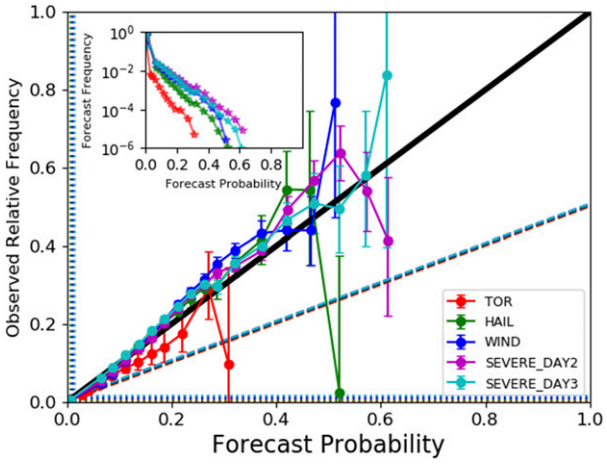

FIG. 9. Attribute diagrams for ML-based outlooks. Colored opaque lines with circular points indicate observed relative frequency as a function of forecast probability; the solid black line is the one-to-one line, indicating perfect reliability. Colors correspond to different severe predictands and lead times as indicated in the panel legend. Horizontal and vertical dotted lines denote the "no resolution" lines and correspond to the bulk climatological frequency of the given predictand. The tilted dashed lines depict the "no skill" line following the decomposition of the Brier score. Error bars correspond to 95% reliability confidence intervals using the method of Agresti and Coull (1998), and assuming nonoverlapping neighborhoods of the sample distribution of a binomial sample proportion (i.e., observed relative frequency) are independent. Sharpness diagram is inset with lines indicating the total proportion of forecasts falling in each forecast probability bin, using the logarithmic scale on the left hand side of the figure. Probability bins are delineated by 2.5%, 3.5%, 5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20%, 25%, and 30% thresholds for Day 1 tornado forecasts, and by 5.5%, 7.5%, 10%, 12.5%, 15%, 17.5%, 20%, 22.5%, 25%, 27.5%, 30%, 35%, 40%, 45%, 50%, 55%, and 60% for all other forecast sets.

the responsible physical processes. These biases are largest in the convectively active warm season; the RF outlooks, using years of historical data, are able to robustly identify and correct for many of these biases, leading to the largest improvements in skill when the model biases are largest and the least skillful external guidance is available to the human forecaster. In reality, both the forecasters and RF products partially correct for model biases, and both suffer from what they would be if their inputs were completely unbiased.

Reliability diagrams for the RF outlooks (Fig. 9) demonstrate quite calibrated forecasts along the spectrum of the probability distribution. A slight underconfidence bias is observed for most predictands, but otherwise calibration remains quite good until the highest probability bins, where sample size is very small. Maximum forecast probabilities get as high as approximately 30% for tornadoes, into the lower 50% range for hail and wind, and into the lower 60s for any severe event at Days 2 and 3. The main exception to calibration are the tornado forecasts, which are characterized by a slight overforecast bias. This may be attributable to large differences in the event frequency between the training sample, which featured many highly active tornadic years, and the test period, which was relatively quiet (Herman et al. 2018). Forecast resolution, which is insensitive to forecast calibration and assesses the forecast system's ability to discriminate observed frequencies, is largely positive and consistent across all predictands as the observed frequency of events increases proportionally with forecast probabilities. Noticeably, tornado forecasts tend to have the lowest resolution of all predictands.

b. *Blended forecast skill*

The weighted blend of SPC and RF outlooks described in section 2 unsurprisingly demonstrates forecast skill spatial characteristics of both the interpolated SPC (Herman et al. 2018) and RF outlooks (Fig. 10). Most prominently, the high skill in the northern states in the SPC outlooks is reintroduced to the blend in the hail and wind outlooks (cf. Figs. 6c,10c; Figs. 6e,10e). For predictands in which the skill difference is large between the two outlook sources, such as for significant wind (Fig. 10f) and the medium-range outlooks (Figs. 10g,h), the blended outlooks verify very similarly to the more skillful component, in part simply because the weights direct the blend heavily toward that component. Across the board, the SPC RF blend verifies as or more skillfully than the SPC outlooks alone—both in space (Fig. 11) and when aggregated across the CONUS (Fig. 12)—a testament to the utility of the RF guidance in improving operational severe weather forecasts. Even at Day 1, where SPC outlooks outperform the raw RF guidance (Fig. 12), the blended forecasts outperform both the raw SPC and raw RF outlooks. In the case of hail and wind, the margin of improvement is considerable, with BSS improvements of 0.061 and 0.053, respectively (Figs. 11c,e). At Day 2 and 3, while the blend is not able to improve skill over the RF outlooks (Fig. 12), that difference is already considerable when compared with the SPC outlooks at 0.044 and 0.048 (Figs. 11g,h). Consequently, the blended forecast exhibits much improved skill compared with the raw SPC outlooks for all eight forecast predictands evaluated (Fig. 12). Even more encouragingly, the skill improvements are seen across all regions of the CONUS (Fig. 11) with fairly uniform distribution. For hail, wind, and the Day 2–3 outlooks, the skill differences are statistically significant over all except for pockets of the western CONUS where the climatological event frequencies are insufficient to produce a robust sample. Hail outlooks are most improved over the Mississippi Valley region into the

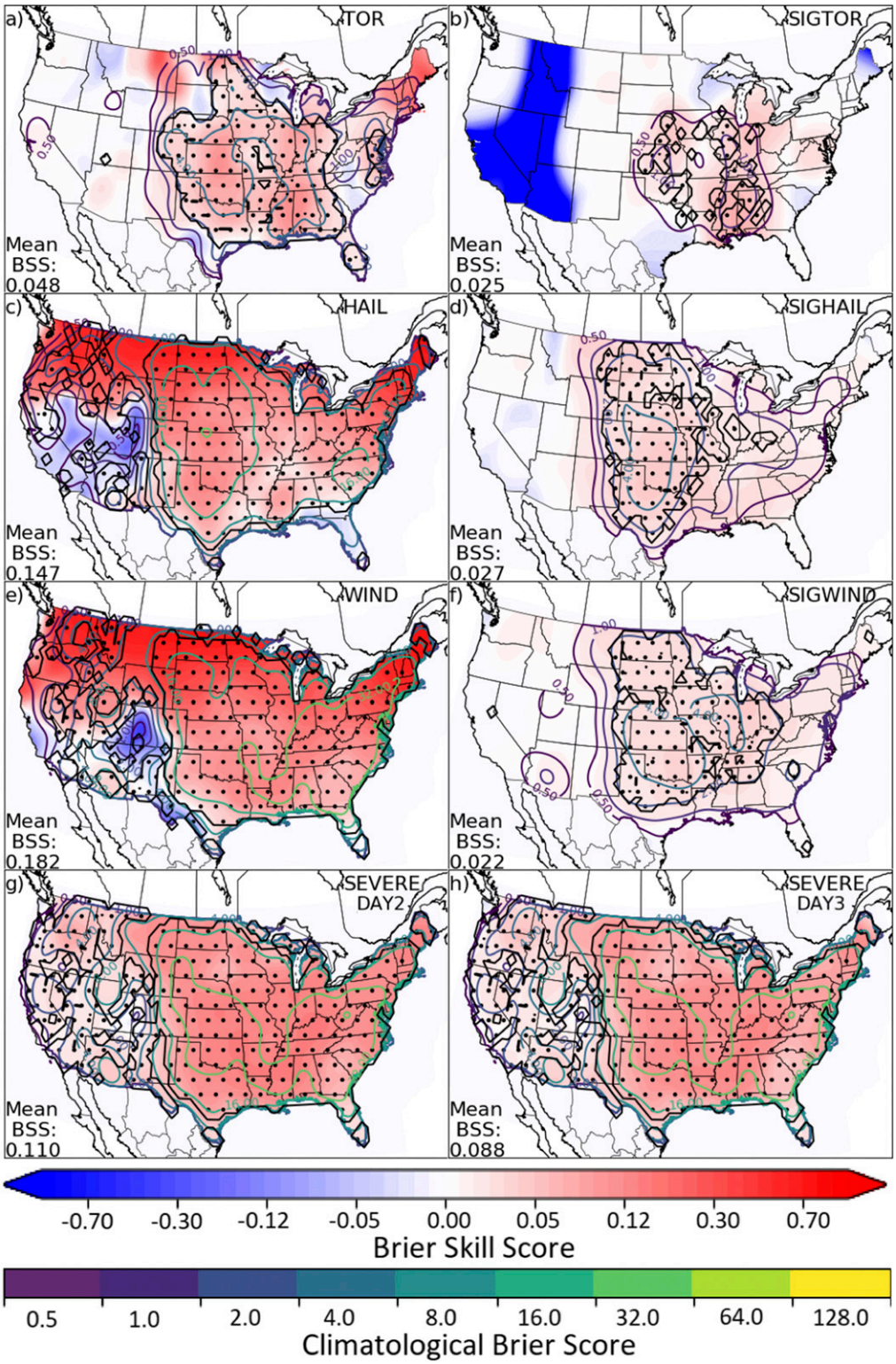

FIG. 10. As in Fig. 6, but for the weighted blend of SPC and ML outlooks.

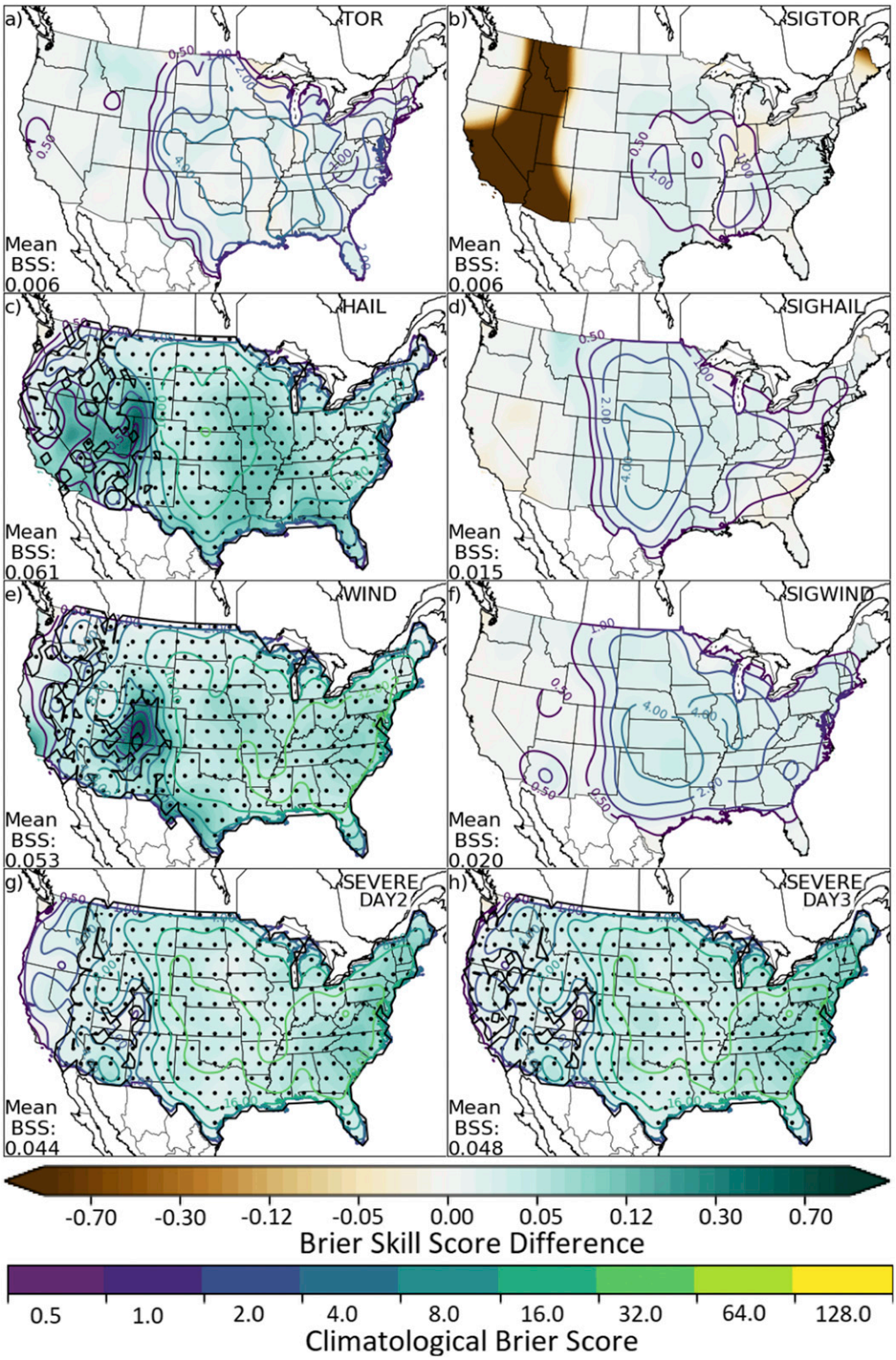

FIG. 11. As in Fig. 7, but skill for the weighted blend of SPC and ML outlooks against SPC outlooks.

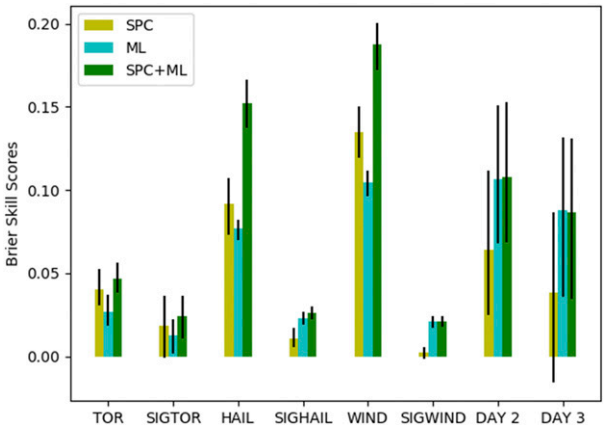

FIG. 12. CONUS-total BSS for each of the eight verified predictands for the SPC outlooks (yellow bars), ML forecasts (blue bars), and weighted average of the two (green bars). Error bars indicate 95% BSS confidence bounds obtained via bootstrapping.

Midwest, while wind outlooks are most improved over the southern plains, and the medium-range outlooks are most improved over the East Coast urban corridor.

c. *CAPE-shear space*

One additional instructive skill decomposition inspects forecast verification in the CAPE versus shear parameter space. The raw RF hail (Fig. 13d) and wind (Fig. 13g) forecasts exhibit high skill throughout much of the parameter space. Wind forecasts are skillful throughout essentially the entire space, with a skill minimum in the low CAPE, low shear corner of the parameter space. Hail exhibits a local BSS minimum in this region as well (Fig. 13d), but has primary skill minima in the high CAPE, low shear and especially the low CAPE, high shear corners of the parameter space. Tornado forecast verification results are more mixed. Like hail, forecast skill suffers in scenarios with ample supply of CAPE or shear, but little of the other (Fig. 13a). Skill is significantly positive when sufficient amounts of both ingredients are in place, but outlooks are not always skillful relative to climatology with less pronounced convective ingredients (Fig. 13a). The addition of the weighted average with SPC outlooks (Figs. 13b,e,h) improves outlook skill across the parameter space while leaving the character of the skill distribution much the same. Skill improvement is especially evident in low CAPE scenarios with low to moderate wind shear (e.g., Fig. 13e); skill improvement is minimal in the high CAPE, low shear and low CAPE, high shear corners of the parameter space, where SPC outlooks also struggle (Herman et al. 2018). In comparison to the raw SPC outlooks, the blend of the RF-based ML forecasts with the SPC outlooks yields skill improvements across the parameter space for hail (Fig. 13f) and wind (Fig. 13i) forecasts, and across much of the domain for tornadoes (Fig. 13c). The skill improvements are largest in the low shear end of the parameter space, especially with high CAPE. In general, where convective predictability is lowest in dynamical model guidance (i.e., low-shear) due to complex storm interactions and longevity dependence on storm morphology (e.g., Houston and Wilhelmson 2011), the combination of statistical guidance and forecaster knowledge leads to improved skill over the individual raw SPC and ML outlooks.

d. *Case study*

A brief case study example is provided in order to illustrate the real-time character of the ML model forecasts. The outlooks valid 1200 UTC 9 May–1200 UTC 10 May 2016 (Fig. 14) are chosen for evaluation, a period in the middle of a moderate-severity multiday outbreak, which spread from the Colorado plains to the Mississippi Valley. SPC's Day 1 tornado outlook (Fig. 14c) highlighted the southern domain reasonably well, with a 10% risk contour, but was generally too far southeast with many tornadoes occurring on the edge of the 2% probability contour, and most of the northern cluster was missed entirely. SPC forecasters identified hail (Fig. 14f) as the primary risk of the day, with a 30% risk contour in addition to a significant hail contour over eastern Oklahoma, western Arkansas, and far northeastern Texas. Their wind outlook (Fig. 14i) had essentially an identical outline to the severe hail one, except topping out with approximately 15% event probabilities and no significant wind contour.

In many respects, the ML Day-1 outlooks were improved, relative to the SPC outlooks. The tornado outlook (Fig. 14a) both indicates higher risk, with a maximum tornado probability over 15%; displaces the maximum to the northwest where more events were observed; and extends the probabilities farther north to at least indicate some appreciable risk in the northern cluster, albeit still lower than in the southern region. The hail (Fig. 14d) and wind (Fig. 14g) outlooks are more distinct, with higher hail probabilities to the north and west over Oklahoma, Kansas, and Nebraska and lower probabilities to the east; these changes again better collocate the high event probabilities with the observations. Compared with hail, wind probabilities maximize to the southeast over eastern Oklahoma and Arkansas. The RF models forecasts also had better spatial placement in the medium range, even indicating the two primary risk areas at Day 2 (Fig. 15a), and encompassing the western severe weather observations when the operational outlook (Fig. 15c) did not. This difference was further magnified at Day 3 when only a 15% severe probability was indicated and many

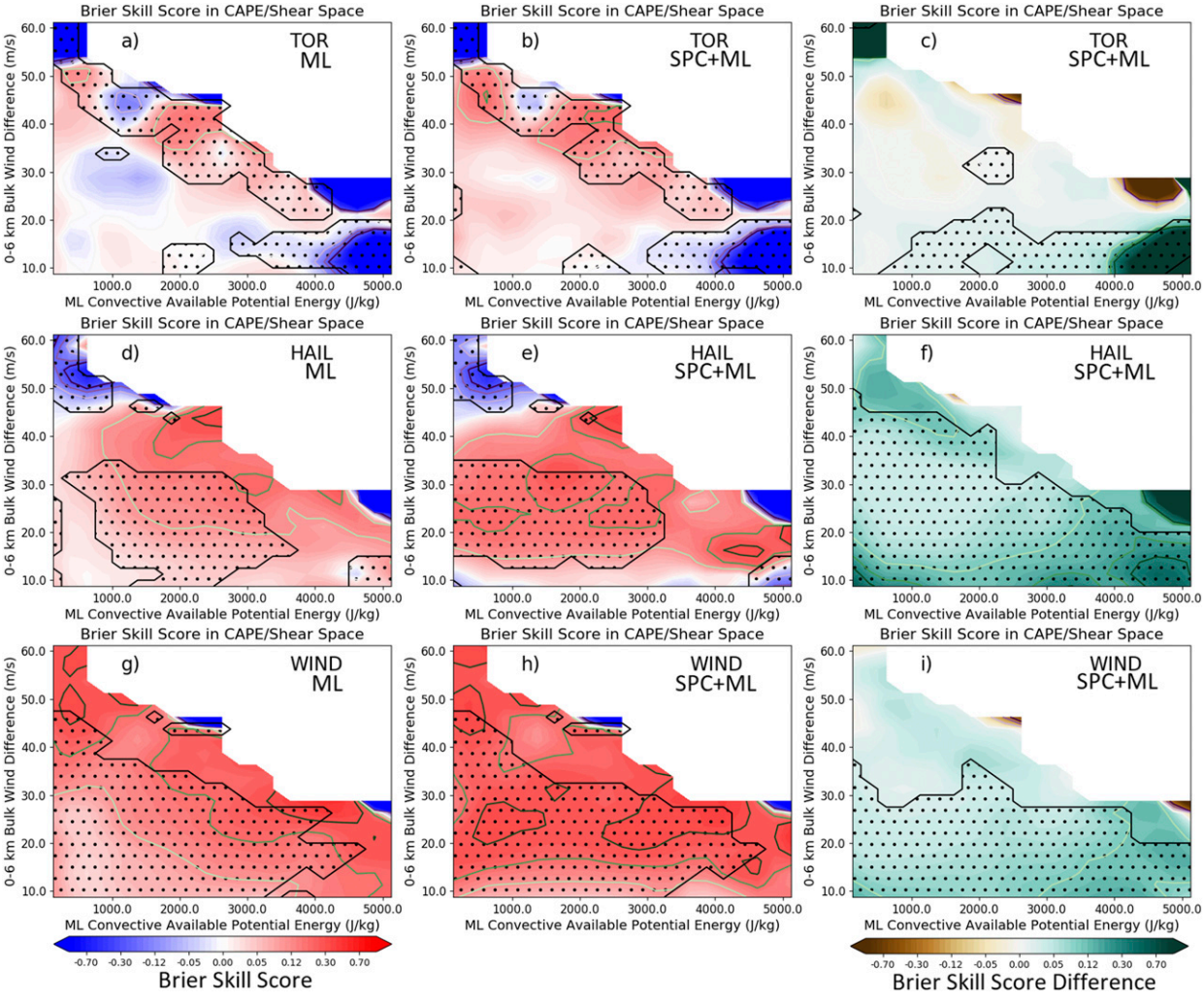

FIG. 13. BSS evaluation broken by CAPE vs shear parameter space for (a)–(c) tornado, (d)–(f) hail, and (g)–(i) wind outlooks as partitioned in Herman et al. (2018) and described in the manuscript text. Unfilled contours replicate the filled contours at the −0.3, −0.2, −0.1, 0.1, 0.2, and 0.3 levels and are included for quantitative clarity. The left column depicts verification of the ML forecasts, the center column to the evaluation of the weighted blend of SPC and ML outlooks, and the right column presents the skill score difference between the blend and the raw interpolated SPC outlooks, with greens indicating an improvement over the SPC outlooks and browns representing loss of skill. Stippling indicates regions where the sign of the BSS or BSS difference is statistically significant with $\alpha = 0.05$ based on bootstrap resampling.

severe weather reports over the central plains were not encompassed by the 5% marginal contour in the operational outlook (Fig. 15f), while nearly every observation was encompassed by a 2% contour at Day 3 in the ML outlook (Fig. 15d) and severe probabilities maximized over 30%. On the other hand, it should be noted that the ML model misidentified an area of enhanced probabilities in eastern Virginia and North Carolina at all lead times that did not receive severe weather reports. The blended model forecast (Figs. 14b,e,h and 15b,d) took many of the successful aspects of the ML and SPC outlooks and refined the area of severe threat, eroding lower probabilities at the edge of the outlooks (e.g., Fig. 14b) and minimizing erroneous probabilities in the mid-Atlantic region (e.g., Fig. 15e). While not all cases demonstrate this degree of success, this case study exemplifies many of the benefits consistently demonstrated by machine learning: relative spatial placement of risks, approximate risk magnitudes, and rarely missing observed events entirely.

e. *Spatial characteristics*

Finally, the spatial coverage of severe reports in each outlook probability threshold is considered across many cases through fractional coverage statistics. The

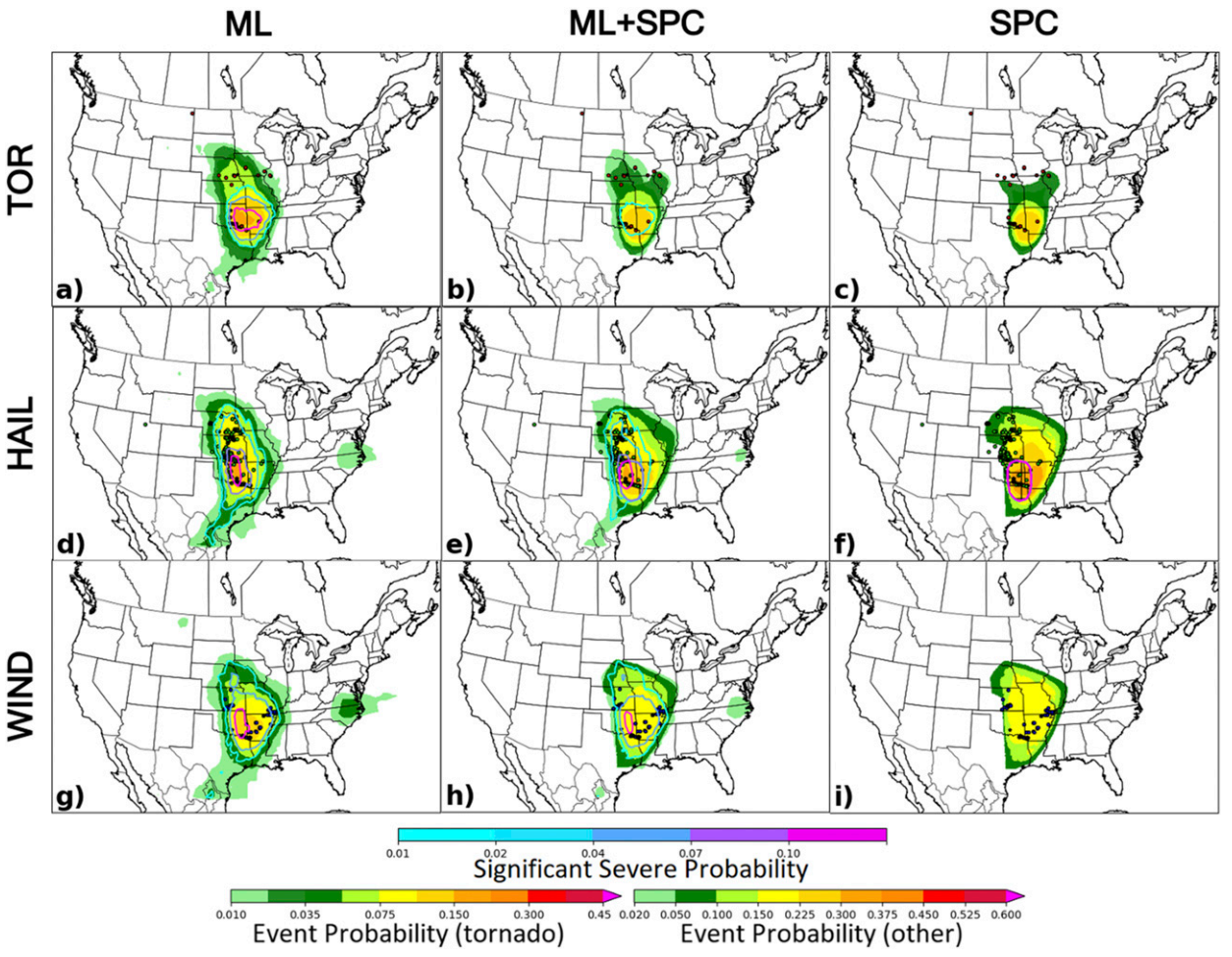

FIG. 14. Outlooks from the (left) RF models, (middle) blended ML+SPC model, and (right) raw interpolated SPC contours valid for the 24-h period ending 1200 UTC 10 May 2016. Filled contours depict severe probabilities as indicated by the corresponding colorbar on figure bottom; unfilled contours indicate significant severe probabilities for the corresponding phenomenon as applicable. Day 1 (a)–(c) tornado, (d)–(f) hail, and (g)–(i) wind outlooks are shown. Severe weather reports for the period are shown with red, green, and blue circles for tornadoes, hail, and wind. Darker colored stars indicate significant severe reports for the color-corresponding phenomenon.

SPC-defined outlook probabilities for each hazard type are compared against corresponding severe storm reports to determine the average fractional coverage of observations in each category (e.g., Erickson et al. 2019). For example, it should be expected that severe hail reports (with a 40-km neighborhood) encompass greater than or equal to 15% of the spatial area in a 15% forecast probability contour on average, not to exceed the next probability contour (e.g., 30%). To facilitate this analysis, each outlook probability threshold (for each hazard type) from 1 January 2013 to 31 December 2016 is compared to severe reports, and overlapping coverage of probability contours with severe reports is aggregated for the RF model outlooks (ML), blend of RF and SPC outlooks (BLEND), and SPC outlooks (Fig. 16). For Day 1 hazards, the SPC outlooks are well-calibrated for tornadoes (Fig. 16a), and poorly calibrated for hail and severe wind at probabilities above 5% (Figs. 16b,c); hail and wind outlooks are consistently too large or frequent (below the black line in Fig. 16) with low fractional coverage compared to the probabilistic category. In comparison, the ML and BLEND models are comparable to SPC outlooks for tornadoes, but better calibrated for severe hail and wind. In contrast to the Day 1 outlooks, nearly all outlooks generated for Days 2 and 3 are too small and infrequent (above the black line in Fig. 16), particularly at the higher outlook probability thresholds (Figs. 16d,e); the fractional coverage of severe weather reports is on average greater than the probabilistic forecast. The SPC, ML, and BLEND outlooks are fairly comparable at the 5% and 15% thresholds for Day 2 and 3 aggregated severe potential, and all perform poorly at the higher thresholds, where the outlooks are routinely too small and overconfident. It is

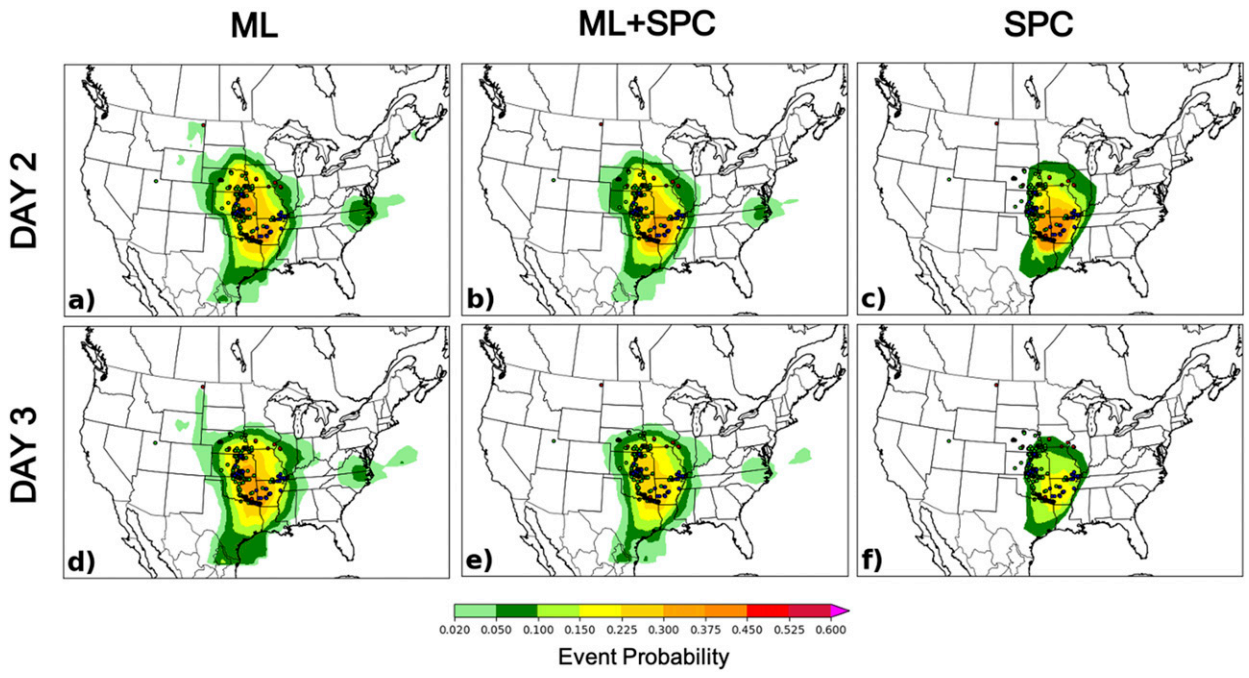

FIG. 15. As in Fig. 14, but the (a)–(c) Day 2 and (d)–(f) Day 3 outlooks are presented, issued previously for the same valid 24-h period ending 1200 UTC 10 May 2016.

notable that the statistical models do not forecast the highest probability thresholds often (Figs. 16b–e). Generally, well-calibrated guidance routinely comes from the BLEND outlooks for all Day-1 predictands, and low probability threshold Day-2 outlooks, illustrating the value of incorporating statistical forecast guidance with the human forecast process.

5. Summary and conclusions

RFs have been trained to generate probabilistic predictions of severe weather for Days 1–3 across the CONUS with analogous predictands to SPC's convective outlooks, with tornado, hail, and wind treated separately at Day 1 and collectively for Days 2–3. Distinct RFs were trained for the western, central, and eastern CONUS as partitioned in Fig. 1. Inputs to the RFs came from the GEFS/R ensemble median of 12 different atmospheric fields: APCP, CAPE, CIN, PWAT, U10, V10, UV10, T2M, Q2M, SHR850, and SHR500. For the Day 1 models, three additional predictors were used: RH2M, ZLCL, and SRH. The spatiotemporal evolution of each of these fields in the vicinity of the forecast point throughout the forecast period was included in the predictor set to provide a comprehensive assessment of the simulated environmental conditions for each severe weather forecast. Each of the fifteen RFs—three regions and five predictands—was trained on nine years of forecasts spanning 12 April 2003–11 April 2012. The identified relationships between simulated model variables and observed severe weather during that period were assessed using RF FIs. The trained RFs were then run over an extended withheld test period spanning 12 April 2012–31 December 2016 and the performance of these forecasts assessed, both in isolation with a climatological reference and relative to SPC convective outlooks issued during the same period.

The statistical relationships identified by the RFs bear considerable correspondence with known physical relationships between atmospheric variables and severe weather, lending credence to the veracity of the model solutions. For example, CAPE, CIN, and wind shear—some of the most commonly used variables to characterize severe weather environments (e.g., Johns and Doswell 1992)—are consistently identified as the most predictive variables for forecasting severe weather. More nuanced identifications are made as well, including more emphasis on kinematics in tornado prediction compared with hail and wind, and additionally, wind difference over a shallower vertical layer being more predictive for tornadoes than for hail and wind. Even spatiotemporal relationships that are identified accord with physical intuition of advective properties, such as enhanced importance to the south and southwest for variables in the central and east regions, respectively, particularly for longer lead times.

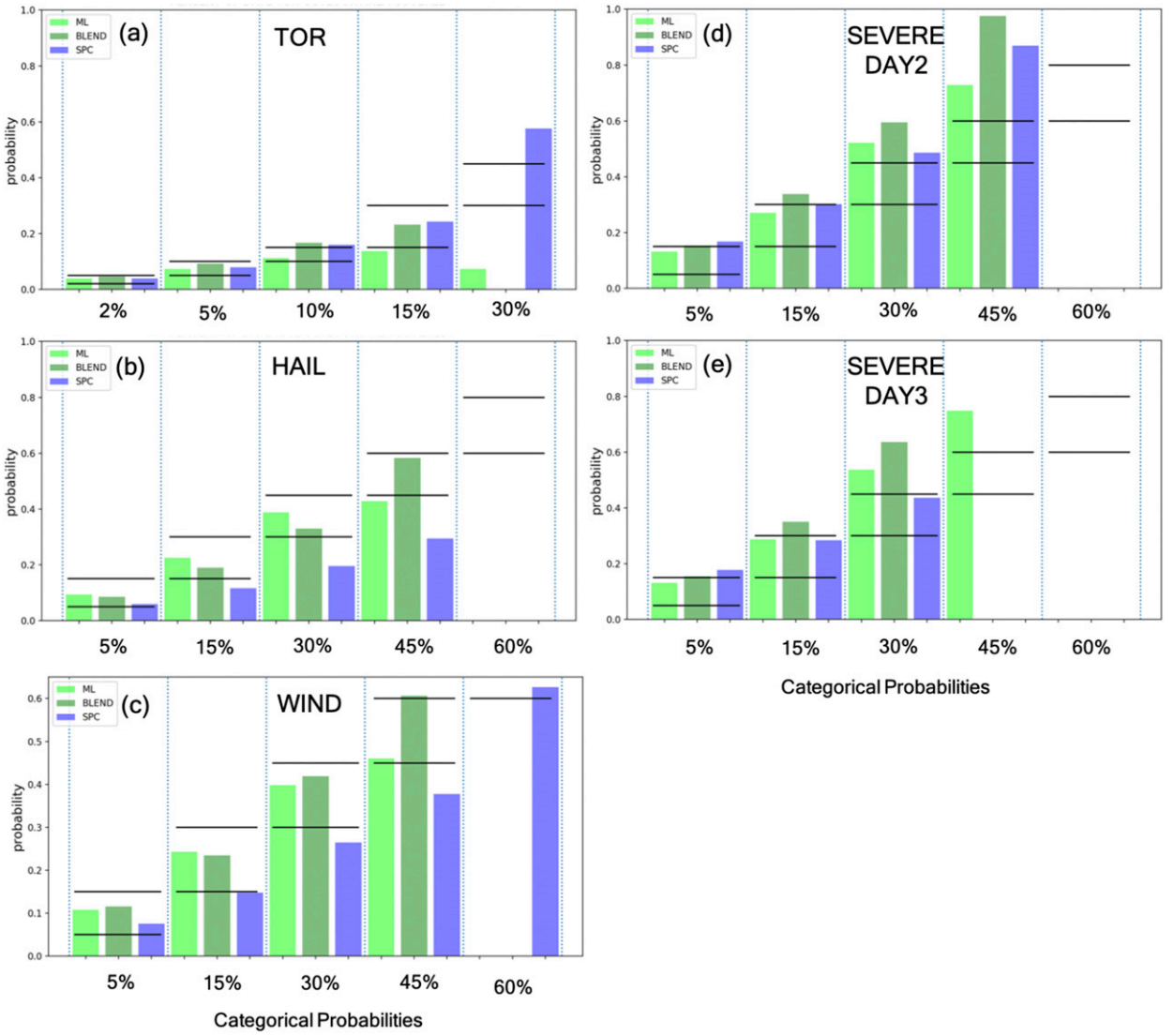

FIG. 16. Average coverage probability of severe weather reports for Day 1 (a) tornado, (b) hail, and (c) wind outlooks and (d) Day 2 and (e) Day 3 outlooks from the RF model (ML), blend of ML and SPC outlooks (BLEND) as discussed in the text, and SPC forecasts (SPC) calculated from 2013 to 2016. For each predictand, outlooks are broken down by respective categorical probability thresholds as defined by SPC. Horizontal black lines correspond to the expected range of area coverage for each probability threshold, with calibrated forecast products lying between the lines for a particular threshold.

In terms of aggregate performance, the outlooks demonstrate impressive probabilistic forecast skill, significantly outperforming equivalent SPC outlooks at Days 2 and 3 as well as for significant severe events at Day 1, while underperforming SPC outlooks somewhat in the standard categories at Day 1. However, a weighted blend of the two outlooks significantly outperformed the SPC outlooks for all phenomena and lead times, with the blend also significantly outperforming the raw ML-based outlooks at Day 1. The largest improvements came for hail and wind, with less gain seen in the tornado outlooks. Spatially, the skill gains of the blend were nearly uniform, although the most gain was generally seen in the Mississippi Valley at Day 1 and the East for Days 2 and 3 with the most variability in the West owing to the low climatological frequency and small sample size. Seasonally, the largest gains at Day 1 tended to occur during the winter and spring, with the largest medium-range gains seen in the summer. Additionally, the largest forecast skill improvements generally came when wind shear was relatively low, but across the spectrum of environmental CAPE. The area coverage of observations by the ML models and SPC outlooks were also evaluated. At Day 1, the ML and

blended models were effectively calibrated for all hazard types and all but the highest probability contours when few RF forecast contours are issued. SPC forecast outlooks were calibrated best for Day 1 tornado events, and typically oversized for hail and wind. All three forecast products were generally well calibrated for Day 2 and 3 aggregated severe threats at low probability thresholds, and overconfident (i.e., undersized) at higher probability contours.

Some limitations of this analysis should be noted. Principally, due to a combination of logistical and practical constraints, SPC outlooks are inherently limited in their probability contours, and the human forecaster cannot issue probabilities across the entire probability spectrum like ML-models can. Some of this is partly overcome here by interpolating between SPC probability contours, which Herman et al. (2018) demonstrated to yield higher probabilistic skill compared with the uninterpolated outlooks. However, some limitations remain. In particular, probabilities much above the highest risk contour, 60%, cannot be produced even with interpolation. More significantly, risk contours below the lowest risk contour—2% for tornadoes and 5% for everything else—cannot be produced at all without imposing additional assumptions about probabilities in the vicinity of but outside risk contours. Instead, all forecast probabilities outside the lowest risk contour are assumed to be zero. The ML-based outlooks frequently forecast event probabilities above 0 but below 2% or 5%, which contribute to mixed skill in areas where severe reports are infrequent (e.g., significant tornadoes in the West region), but substantial relative-skill improvements—skill relative to SPC outlooks—for nonsignificant Day-1 severe events (not shown). This effect is further exacerbated for significant severe events. Here, SPC only issues a 10% risk contour, and can thus only issue 0 or 0.1 event probabilities. Forecasts above 10% do occur, but are quite rare in the ML-based outlooks, and the majority of the skill reaped in its outlooks occur from its above-climatological event probabilities that are nevertheless below 10% (not shown). Additional limitations exist due to the prescribed, static set of input predictor variables, which were selected based on considerations from previous work (e.g., Herman and Schumacher 2018b). By choosing a static set of input predictors that are believed to be important for severe weather forecasting, the RF model is not necessarily learning anything new about the forecast problem, a major benefit of RFs that can be exploited to gain insight into a particular forecast problem. It may be beneficial to tailor individual hazard RF models to different sets of predictors and evaluate other sophisticated importance measures, which the authors feel is a worthwhile avenue for future research, but beyond the scope of this work.

Notwithstanding these limitations, the results of this study demonstrate great promise for the application of machine learning to operational severe weather forecasting, particularly in the medium range. Moreover, when combined with the outcomes of other studies (e.g., Herman and Schumacher 2016, 2018b), the favorable comparison with operational benchmarks across a wide range of applications suggests utility in analogous methods as a statistical postprocessing tool across the broader domain of high-impact weather prediction (e.g., McGovern et al. 2017). The approach taken here is fairly simple, and based on relatively unskillful dynamical guidance compared with the current state of operational dynamical NWP. Future work that investigates use of more sophisticated preprocessing; additional physically relevant predictors; use of additional data sources, including observations, convection-allowing guidance, and other dynamical ensembles; and more detailed and individualized treatments of the different severe weather predictands (e.g., Gagne et al. 2017) into a single synthesized machine learning–based probabilistic forecast model may yield considerable additional skill compared to what has been demonstrated here. Additionally, the blended RF+SPC model methodology could be easily adjusted for real-time evaluation, specifically using rolling weights (e.g., the last 90 days of forecasts) or fixed weights to generate forecasts. Nevertheless, even this straightforward implementation has illustrated considerable potential benefit for using machine learning in operational severe weather forecasting, and further research in this domain is certainly warranted.

Acknowledgments. The authors thank Erik Nielsen and Stacey Hitchcock for illuminating discussions and presentational suggestions, in addition to Erik's assistance with SPC forecast gridding. Roger Edwards provided considerable insight into SPC outlook details and practices, and the authors had several engaging discussions about machine learning model development and severe weather applications with David John Gagne. Both of these conversations greatly improved the quality of this study. The authors also thank three anonymous reviewers who provided helpful and constructive suggestions to improve our paper. Computational resources were generously afforded by the National Center for Atmospheric Research Computational Information Systems Laboratory. Funding for this research was supported by NOAA Awards NA16OAR4590238 and NA18OAR4590378 and NSF Grant ACI-1450089.

APPENDIX

Derived Variables

a. Relative humidity

Relative humidity is calculated as a function of specific humidity q, temperature T, and pressure P, all of which are natively archived. The surface pressure is assumed to be negligibly different from the air pressure two meters above ground. The variables are related through Clausius–Clapeyron, as employed in Bolton (1980) and elsewhere:

$$\mathrm{RH} = \frac{0.263 \times P \times q}{e^{\frac{17.67(T-T_0)}{T-29.65}}} \tag{A1}$$

where temperature is in K and pressure is in Pa, and a reference temperature T_0 of 273.15 K is used. RH is calculated on the 1° grid, since surface pressure is only archived on this grid.

b. Lifting condensation level height

An exact formula for the LCL height as a function of temperature, pressure, and relative humidity was described in Romps (2017), and that formulation is employed here. Relative humidity is not natively archived and is supplied to this formulation as calculated in the previous subsection.

c. Wind shear

SHR850 and SHR500—bulk wind differences between two vertical levels—are calculated straightforwardly:

$$\mathrm{SHR850} = \sqrt{(U_{850} - U_{10\mathrm{m}})^2 + (V_{850} - V_{10\mathrm{m}})^2}, \tag{A2}$$

$$\mathrm{SHR500} = \sqrt{(U_{500} - U_{10\mathrm{m}})^2 + (V_{500} - V_{10\mathrm{m}})^2}. \tag{A3}$$

Winds were used on the 1° grid for both levels.

d. Storm relative helicity

Limited information is available from which to calculate SRH, but given its demonstrated importance in severe environments (e.g., Kuchera and Parker 2006; Parker 2014), the forecast information is used to generate as accurate of SRH estimates as possible. Low-level vertical winds on pressure levels are provided at only 1000, 925, 850, and 700 hPa—quite insufficient for use in an SRH calculation. In height, winds are provided at only 10 and 80 m above ground level—again, insufficient. Hybrid levels provide some resolution in the low levels, with winds archived on the 0.996, 0.987, 0.977, and 0.965 sigma levels; geopotential heights are provided for these levels as well. Thus, for calculating SRH from the surface to 850 hPa, five layers are used: 1) 10 m–0.996σ, 2) 0.996σ–0.987σ, 3) 0.987σ–0.977σ, 4) 0.977σ–0.965σ, and 5) 0.965σ–850 hPa. Storm motion is estimated as 75% and 30° to the right of the mean wind, a common heuristic employed in Ramsay and Doswell (2005) and others. The mean wind is estimated as the average of the wind at 850, 500, and 200 hPa:

$$\overline{U} = \frac{U_{850} + U_{500} + U_{200}}{3}; \quad \overline{V} = \frac{V_{850} + V_{500} + V_{200}}{3}. \tag{A4}$$

Accordingly,

$$\mathrm{SRH} = \sum_{l=1}^{5} \max(0, \mathrm{SRH}_l), \tag{A5}$$

where

$$\mathrm{SRH}_l = (Z_l - Z_{l-1})\left[(\overline{V}_l - V_{\mathrm{st}})\frac{U_l - U_{l-1}}{Z_l - Z_{l-1}} - (\overline{U}_l - U_{\mathrm{st}})\frac{V_l - V_{l-1}}{Z_l - Z_{l-1}}\right], \tag{A6}$$

with

$$\overline{U}_l = \frac{U_l + U_{l-1}}{2}; \quad \overline{V}_l = \frac{V_l + V_{l-1}}{2}, \tag{A7}$$

and

$$U_{\mathrm{st}} = \sqrt{0.75} \times [\overline{U}\cos(-30°) - \overline{V}\sin(-30°)], \tag{A8}$$

$$V_{\mathrm{st}} = \sqrt{0.75} \times [\overline{U}\sin(-30°) + \overline{V}\cos(-30°)]. \tag{A9}$$

REFERENCES

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

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

Agresti, A., and B. A. Coull, 1998: Approximate is better than "exact" for interval estimation of binomial proportions. *Amer. Stat.*, **52**, 119–126, https://doi.org/10.1080/00031305.1998.10480550.

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

Alessandrini, S., L. D. Monache, C. M. Rozoff, and W. E. Lewis, 2018: Probabilistic prediction of tropical cyclone intensity with an analog ensemble. *Mon. Wea. Rev.*, **146**, 1723–1744, https://doi.org/10.1175/MWR-D-17-0314.1.

Alvarez, F. M., 2014: Statistical calibration of extended-range probabilistic tornado forecasts with a reforecast dataset. Ph.D. thesis, Saint Louis University, 210 pp.

Baggett, C. F., K. M. Nardi, S. J. Childs, S. N. Zito, E. A. Barnes, and E. D. Maloney, 2018: Skillful subseasonal forecasts of weekly tornado and hail activity using the Madden–Julian Oscillation. *J. Geophys. Res. Atmos.*, **123**, 12 661–12 675, https://doi.org/10.1029/2018JD029059.

Bolton, D., 1980: The computation of equivalent potential temperature. *Mon. Wea. Rev.*, **108**, 1046–1053, https://doi.org/10.1175/1520-0493(1980)108<1046:TCOEPT>2.0.CO;2.

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

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

Brimelow, J. C., G. W. Reuter, R. Goodson, and T. W. Krauss, 2006: Spatial forecasts of maximum hail size using prognostic model soundings and HAILCAST. *Wea. Forecasting*, **21**, 206–219, https://doi.org/10.1175/WAF915.1.

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

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

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

Clark, A. J., A. MacKenzie, A. McGovern, V. Lakshmanan, and R. Brown, 2015: An automated, multiparameter dryline identification algorithm. *Wea. Forecasting*, **30**, 1781–1794, https://doi.org/10.1175/WAF-D-15-0070.1.

Davies, J. M., and R. H. Johns, 1993: Some wind and instability parameters associated with strong and violent tornadoes: 1. Wind shear and helicity. *The Tornado: Its Structure, Dynamics, Prediction, and Hazards*, *Geophys. Monogr.*, Vol. 79, Amer. Geophys. Union, 573–582.

Doswell, C. A., III, 2004: Weather forecasting by humans: Heuristics and decision making. *Wea. Forecasting*, **19**, 1115–1126, https://doi.org/10.1175/WAF-821.1.

——, 2007: Small sample size and data quality issues illustrated using tornado occurrence data. *Electron. J. Severe Storms Meteor.*, **2** (5), 1–16.

——, and D. M. Schultz, 2006: On the use of indices and parameters in forecasting severe storms. *Electron. J. Severe Storms Meteor.*, **1** (3), http://www.ejssm.org/ojs/index.php/ejssm/article/viewArticle/11.

Edwards, R., G. W. Carbin, and S. F. Corfidi, 2015: Overview of the Storm Prediction Center. *13th History Symp.*, Phoenix, AZ, Amer. Meteor. Soc., 1.1, https://ams.confex.com/ams/95Annual/webprogram/Paper266329.html.

Elmore, K. L., and H. Grams, 2016: Using mPING data to generate random forests for precipitation type forecasts. *14th Conf. on Artificial and Computational Intelligence and Its Applications to the Environmental Sciences*, New Orleans, LA, Amer. Meteor. Soc., 4.2, https://ams.confex.com/ams/96Annual/webprogram/Paper289684.html.

Elsner, J. B., and H. M. Widen, 2014: Predicting spring tornado activity in the central Great Plains by 1 March. *Mon. Wea. Rev.*, **142**, 259–267, https://doi.org/10.1175/MWR-D-13-00014.1.

Erickson, M., J. Kastman, B. Albright, S. Perfater, J. Nelson, R. Schumacher, and G. Herman, 2019: Verification results from the 2017 HMT–WPC flash flood and intense rainfall experiment. *J. Appl. Meteor. Climatol.*, **58**, 2591–2604, https://doi.org/10.1175/JAMC-D-19-0097.1.

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

Gagne, D. J., A. McGovern, and J. Brotzge, 2009: Classification of convective areas using decision trees. *J. Atmos. Oceanic Technol.*, **26**, 1341–1353, https://doi.org/10.1175/2008JTECHA1205.1.

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

——, ——, 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.

Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2018: Blended probabilistic tornado forecasts: Combining climatological frequencies with NSSL–WRF ensemble forecasts. *Wea. Forecasting*, **33**, 443–460, https://doi.org/10.1175/WAF-D-17-0132.1.

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

Hamill, T. M., 2017: Changes in the systematic errors of global reforecasts due to an evolving data assimilation system. *Mon. Wea. Rev.*, **145**, 2479–2485, https://doi.org/10.1175/MWR-D-17-0067.1.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

Herman, G. R., 2018: Statistical calibration of extended-range probabilistic tornado forecasts with a reforecast dataset. Ph.D. thesis, Colorado State University, 210 pp.

——, and R. S. Schumacher, 2016: Using reforecasts to improve forecasting of fog and visibility for aviation. *Wea. Forecasting*, **31**, 467–482, https://doi.org/10.1175/WAF-D-15-0108.1.

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

——, and ——, 2018b: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, https://doi.org/10.1175/MWR-D-17-0250.1.

——, E. R. Nielsen, and R. S. Schumacher, 2018: Probabilistic verification of Storm Prediction Center convective outlooks. *Wea. Forecasting*, **33**, 161–184, https://doi.org/10.1175/WAF-D-17-0104.1.

Hitchens, N. M., and H. E. Brooks, 2012: Evaluation of the Storm Prediction Center's day 1 convective outlooks. *Wea. Forecasting*, **27**, 1580–1585, https://doi.org/10.1175/WAF-D-12-00061.1.

——, and ——, 2014: Evaluation of the Storm Prediction Center's convective outlooks from day 3 through day 1. *Wea. Forecasting*, **29**, 1134–1142, https://doi.org/10.1175/WAF-D-13-00132.1.

——, and ——, 2017: Determining criteria for missed events to evaluate significant severe convective outlooks. *Wea. Forecasting*, **32**, 1321–1328, https://doi.org/10.1175/WAF-D-16-0170.1.

Hong, S.-Y., and J.-O. J. Lim, 2006: The WRF single-moment 6-class microphysics scheme (WSM6). *J. Korean Meteor. Soc.*, **42**, 129–151.

Houston, A. L., and R. B. Wilhelmson, 2011: The dependence of storm longevity on the pattern of deep convection initiation in a low-shear environment. *Mon. Wea. Rev.*, **139**, 3125–3138, https://doi.org/10.1175/MWR-D-10-05036.1.

Igel, A. L., M. R. Igel, and S. C. van den Heever, 2015: Make it a double? Sobering results from simulations using single-moment microphysics schemes. *J. Atmos. Sci.*, **72**, 910–925, https://doi.org/10.1175/JAS-D-14-0107.1.

Jacks, E., 2014: Service change notice 14-42. National Weather Service, Fire and Public Weather Services Branch, NOUS41 KWBC 071749, 3 pp., https://www.weather.gov/media/notification/pdfs/scn14-42day1-3outlooks.pdf.

Johns, R. H., and C. A. Doswell III, 1992: Severe local storms forecasting. *Wea. Forecasting*, **7**, 588–612, https://doi.org/10.1175/1520-0434(1992)007<0588:SLSF>2.0.CO;2.

Karstens, C. D., and Coauthors, 2018: Development of a human–machine mix for forecasting severe convective events. *Wea. Forecasting*, **33**, 715–737, https://doi.org/10.1175/WAF-D-17-0188.1.

Kay, M. P., and H. E. Brooks, 2000: Verification of probabilistic severe storm forecasts at the SPC. Preprints, *20th Conf. on Severe Local Storms,* Orlando, FL, Amer. Meteor. Soc., 9.3, https://ams.confex.com/ams/Sept2000/techprogram/paper_15921.htm.

Khain, A., and Coauthors, 2015: Representation of microphysical processes in cloud-resolving models: Spectral (bin) microphysics versus bulk parameterization. *Rev. Geophys.*, **53**, 247–322, https://doi.org/10.1002/2014RG000468.

Krocak, M. J., and H. E. Brooks, 2018: Climatological estimates of hourly tornado probability for the United States. *Wea. Forecasting*, **33**, 59–69, https://doi.org/10.1175/WAF-D-17-0123.1.

Kuchera, E. L., and M. D. Parker, 2006: Severe convective wind environments. *Wea. Forecasting*, **21**, 595–612, https://doi.org/10.1175/WAF931.1.

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

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

Loken, E., A. Clark, A. McGovern, M. Flora, and K. Knopfmeier, 2019: Postprocessing next-day ensemble probabilistic precipitation forecasts using random forests. *Wea. Forecasting*, **34**, 2017–2044, https://doi.org/10.1175/WAF-D-19-0109.1.

Loridan, T., R. P. Crompton, and E. Dubossarsky, 2017: A machine learning approach to modeling tropical cyclone wind field uncertainty. *Mon. Wea. Rev.*, **145**, 3203–3221, https://doi.org/10.1175/MWR-D-16-0429.1.

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

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

——, 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.

McGovern, A., D. J. Gagne, N. Troutman, R. A. Brown, J. Basara, and J. K. Williams, 2011: Using spatiotemporal relational random forests to improve our understanding of severe weather processes. *Stat. Anal. Data Mining*, **4**, 407–429, https://doi.org/10.1002/sam.10128.

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

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

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.

Murphy, A. H., and R. L. Winkler, 1977: Reliability of subjective probability forecasts of precipitation and temperature. *Appl. Stat.*, **26**, 41–47, https://doi.org/10.2307/2346866.

Nielsen, E. R., G. R. Herman, R. C. Tournay, J. M. Peters, and R. S. Schumacher, 2015: Double impact: When both tornadoes and flash floods threaten the same place at the same time. *Wea. Forecasting*, **30**, 1673–1693, https://doi.org/10.1175/WAF-D-15-0084.1.

NWS, 2018: Summary of natural hazard statistics for 2018 in the United States. NOAA/National Weather Service, Office of Climate, Weather, and Water Services, 3 pp., https://www.weather.gov/media/hazstat/sum18.pdf.

Orf, L., R. Wilhelmson, B. Lee, C. Finley, and A. Houston, 2017: Evolution of a long-track violent tornado within a simulated supercell. *Bull. Amer. Meteor. Soc.*, **98**, 45–68, https://doi.org/10.1175/BAMS-D-15-00073.1.

Parker, M. D., 2014: Composite VORTEX2 supercell environments from near-storm soundings. *Mon. Wea. Rev.*, **142**, 508–529, https://doi.org/10.1175/MWR-D-13-00167.1.

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

Ramsay, H. A., and C. A. Doswell III, 2005: A sensitivity study of hodograph-based methods for estimating supercell motion. *Wea. Forecasting*, **20**, 954–970, https://doi.org/10.1175/WAF889.1.

Romps, D. M., 2017: Exact expression for the lifting condensation level. *J. Atmos. Sci.*, **74**, 3891–3900, https://doi.org/10.1175/JAS-D-17-0102.1.

Rothfusz, L., C. Karstens, and D. Hilderbrand, 2014: Next-generation severe weather forecasting and communication. *Eos, Trans. Amer. Geophys. Union*, **95**, 325–326, https://doi.org/10.1002/2014EO360001.

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

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes for significant severe thunderstorms in the contiguous United States. Part I: Storm

classification and climatology. *Wea. Forecasting*, **27**, 1114–1135, https://doi.org/10.1175/WAF-D-11-00115.1.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, https://doi.org/10.1175/WAF-D-10-05046.1.

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

——, C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016b: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31**, 255–271, https://doi.org/10.1175/WAF-D-15-0138.1.

SPC, 2017a: SVRGIS (updated: 15 May 2017). Storm Prediction Center, accessed 1 June 2017, http://www.spc.noaa.gov/gis/svrgis/.

——, 2017b: Severe weather climatology (1982–2011). Storm Prediction Center, accessed 1 June 2017, http://www.spc.noaa.gov/new/SVRclimo/climo.php?parm=anySvr.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn, 2007: Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinf.*, **8**, 25, https://doi.org/10.1186/1471-2105-8-25.

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

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

Tippett, M. K., A. H. Sobel, and S. J. Camargo, 2012: Association of U.S. tornado occurrence with monthly environmental parameters. *Geophys. Res. Lett.*, **39**, L02801, https://doi.org/10.1029/2011GL050368.

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

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

Verlinden, K. L., and D. R. Bright, 2017: Using the second-generation GEFS reforecasts to predict ceiling, visibility, and aviation flight category. *Wea. Forecasting*, **32**, 1765–1780, https://doi.org/10.1175/WAF-D-16-0211.1.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus*, **60A**, 62–79, https://doi.org/10.1111/j.1600-0870.2007.00273.x.

Whan, K., and M. Schmeits, 2018: Comparing area-probability forecasts of (extreme) local precipitation using parametric and machine learning statistical post-processing methods. *Mon. Wea. Rev.*, **146**, 3651–3673, https://doi.org/10.1175/MWR-D-17-0290.1.

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

Williams, J., 2014: Using random forests to diagnose aviation turbulence. *Mach. Learn.*, **95**, 51–70, https://doi.org/10.1007/s10994-013-5346-7.

——, R. Sharman, J. Craig, and G. Blackburn, 2008: Remote detection and diagnosis of thunderstorm turbulence. *Proc. SPIE*, **7088**, 708804, https://doi.org/10.1117/12.795570.

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