Forecasting Severe Weather with Random Forests

Gregory R. Herman*

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

Corresponding author address: Gregory R. Herman, Department of Atmospheric Science, Colorado State University, 1371 Campus Delivery, Fort Collins, CO 80523.

E-mail: gherman@atmos.colostate.edu

ABSTRACT

Using nine years of historical forecasts spanning April 2003–April 2012 from NOAA's Second Generation 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 Storm Prediction Center's (SPC's) convective outlooks. Separate models are also trained for western, central, and eastern CONUS. Input predictors include many fields—both native in the model archive and externally derived—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–December 2016 and evaluated alongside the equivalent SPC outlooks. The RFs objectively make statistical deductions about the relationship between various simulated atmospheric fields and observances 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 outlooks and RF outlooks significantly outperforms the SPC outlooks alone, suggesting that use of the method can improve operational severe weather forecasting throughout the Day 1–3 period.

1. Introduction

Severe weather is comprised of 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 this, 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 this decade alone (NWS 2018), making severe weather decidedly one of society's great hazards. 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 which 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), 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 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 III 2004; Doswell III 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). In these products, forecasts are issued for 24-hour 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. An additional categorical risk outlook is provided for Days 1–3, defined based on the probabilistic outlook values. 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 signifcant 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; e.g. Hitchens and Brooks 2012, 2014), though advances have been stagnating somewhat in recent years (Herman et al. 2018). Forecast skill is highest at the shortest lead times and gets progressively lower with increasing lead time (e.g. Hitchens and Brooks 2014; Herman et al. 2018). 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 (Hitchens and Brooks 2017; Herman et al. 2018). Additionally, skill was found to be maximum over the Midwest and Great Plains, and lowest over the South and West (Herman et al. 2018). Outlooks are generally most skillful in the winter and spring, and least successful in the late summer into early autumn (Herman et al. 2018). Furthermore, skill is high when at least moderate amounts of both CAPE and wind shear are present, but struggle in scenarios with large amount of one convective ingredient are present in the absence of the other (e.g. Sherburn and Parker 2014; Herman et al. 2018). 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 confining 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, the evaluation of Herman et al. (2018) 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), hail (e.g. Marzban and Witt 2001;

Brimelow et al. 2006; Adams-Selin and Ziegler 2016; Gagne et al. 2017), 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). In addition to severe weather, ML has demonstrated success in forecasting heavy precipitation (e.g. Gagne et al. 2014; Herman and Schumacher 2018b,a; Whan and Schmeits 2018), 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). 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 timescales, such as nowcast settings (e.g. Marzban and Stumpf 1996; Lagerquist et al. 2017), or on much longer timescales (e.g. Tippett et al. 2012; Elsner and Widen 2014; Baggett et al. 2018), with lesser 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. McGovern et al. 2011; Ahijevych et al. 2016; Herman and Schumacher 2016; Gagne et al. 2017; 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 investigates the RF-derived severe weather forecasting insights gleaned from 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 towards post-processing global ensemble output to forecast locally extreme precipitation events across CONUS at Days 2–3. This study follows analogous methodology. A relevant summary of the methodology of Herman and Schumacher (2018b,a) 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 as applied here and the numerous sensitivity experiments performed therein, the reader is invited to consult those studies. For the sake of brevity, several of the model configuration choices selected in this study are motivated by the findings of Herman and Schumacher (2018b) rather than reperforming all the same experiments for this forecast problem. Informal replications of those 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 input predictor. The splitting predictor and the value associated with each branch is determined by the combination that best separates severe weather events from non-events in the supplied model

training data. This process then continues for progressively smaller branched subsets based on only the training data that satifies the previous branching conditions. This process continues until a termination criterion is satisfied, either because all of the remaining training examples are either all events or all non-events, 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. In real-time forecasting, the 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 decision trees by considering different subsets of training data and input features for each tree generation process. An RF's forecast is simply calculated as the mean probabilistic forecast issued by the trees within the forest (e.g. Breiman 2001; Herman and Schumacher 2018a).

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 $\sim$55 km at 40° latitude—initialized once daily at 0000 UTC back to December 1984. Perturbations are applied only to the initial conditions, and are made using the ensemble transform with rescaling technique (Wei et al. 2008). The ensemble system used to generate these reforecasts is nearly static throughout its 30+ year period of coverage, though updates to the operational data assimilation system over time have resulted in some changes in the bias characteristics of its forecasts over the period of record (Hamill 2017). Most surface (or column-integrated) fields are preserved on the native Gaussian grid ($\sim 0.5^\circ$ spacing), while upper-level and some other fields are available only on a $1^\circ \times 1^\circ$ grid. Based on findings from Herman and Schumacher (2018b), this study derives predictors from the GEFS/R ensemble median. Model training employs a 9-year training period, using daily initializations from 12 April 2003–11 April 2012.

Temporally, forecast fields are archived every three hours out to 72 hours past initialization, and are available every six hours beyond that. Accordingly, the RFs trained in this study use 3-hourly predictors for Day 1 and 2 forecasts, and 6-hourly temporal resolution for Day 3.

Several different GEFS/R simulated atmospheric fields with known or postulated physical relationships with severe weather are used as RF predictors (Table 1). These include surface-based CAPE and CIN, 10-meter 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 500 hPa (MSHR, DSHR), 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 approximate storm relative helicity (SRH). 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 as well.

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 re-

sponsible for severe weather across CONUS (e.g. Smith et al. 2012), the country is partitioned into three regions as shown in Figure 1. This study develops separate RFs for each of the three regions of 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 a nine year historical record spanning 12 April 2003–11 April 2012. As noted above, the focus of this study is on the model evaluation rather than on involved sensitivity experiments and parameter tuning. Models were 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 B and minimum number of training examples required to split an impure node in a decision tree, Z. For the interested reader, the final values used are furnished in Table 2.

Trained RFs are evaluated in two distinct ways. First, in Section 3, the statistical relationships diagnosed by the RFs 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, summary statistics are used to capture the extent of use of different aspects of supplied forecast information in generating a final prediction. In particular, this is done by means of feature importances (FIs). Though there are several ways that FIs can be quantified (e.g. Strobl et al. 2007, 2008), this study uses the so-called "Gini importance" metric for consistency with prior ML research in the community (e.g. Pedregosa et al. 2011; Herman and Schumacher 2018a; Whan and Schmeits 2018). A single FI is

attributed to each input feature, and may be conceptualized as the number of splits based on the given feature, weighted in proportion to the number of training examples encountering the split (Friedman 2001). This is summed over each split in the tree for each tree in the forest, and then 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 associated simulated forecast field, forecast time, and in space relative to the forecast point. In many cases, it is convenient to present importances summed 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 years 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 as described in Herman et al. (2018), while forecast calibration is assessed via reliability diagrams (Murphy and Winkler 1977; Bröcker and Smith 2007; Wilks 2011). 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. 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 performance here against SPC convective outlooks for the same predictands is-

sues with comparable lead times. Consistent with Herman et al. (2018), Day 1 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 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 same period; the level of skill improvement, if any, quantifies the value added by the addition of the ML guidance to the operational forecast pipeline. Weights are supplied based on the BSS of the two component outlooks using three temporally-contiguous quarters of the evaluation period that excludes the forecast being weighted, based on the following formula:

$$W_{SPC} = \frac{\frac{1}{1-BSS_{SPC}}}{\frac{1}{1-BSS_{SPC}} + \frac{1}{1-BSS_{RF}}}; W_{RF} = 1 - W_{SPC} \tag{1}$$

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 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, with CIN 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, DSHR is found to be fairly predictive as well, and this is most evident for hail (Fig. 2). For tornadoes, shear over a shallower layer in MSHR is found to be equally (e.g. Fig. 2b) or more (Fig. 2c) predictive than DSHR, and one of the more predictive variables overall. Other variables with high RF FIs for tornadoes include APCP, MSLP, and SRH. The high FI attributed to model APCP in predicting tornadoes may be surprising, but heavy precipitation is often found to be associated with low-level rotation (e.g. Smith et al. 2001; Hitchens and Brooks 2013; Nielsen and Schumacher 2018). MSLP serves to characterize the synoptic environment and help distinguish favorable from unfavorable environmental conditions for tornadoes. 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, shear should be

considered over a deeper layer for hail and wind to ascertain supercell potential and over a shallow layer and in conjunction with helicity for tornado prediction in order to ascertain potential for low-level rotation, and the kinematics play a more significant role overall for tornadoes than for severe hail and wind. The RFs have simply learned these facts objectively and empirically based on analysis of many historical cases, and have provided a quantitative assessment of their findings.

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 DSHR the next most predictive. The relative ranking mostly holds for the Central and East regions (Fig. 3b,c), although CAPE is much more predictive than CIN, especially in the Central region. MSHR 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.

FI time series (Fig. 4) reveal 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 some cases, notably in the East (Fig. 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. Fig 4d,e,g) it. In some cases, this 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).

RF FI time series for Day 2 and 3 models (Fig. 5) again share similarities with their Day 1 counterparts. Like with the Day 1 models, importance peaks come earliest in the East (Fig. 5c) and latest in the West (Fig. 5a), ranging from 2100–0300 UTC. Interestingly, there is a shift in peak importance between Day 2 and 3 models towards earlier times, especially pronounced in the West and Central regions (Fig. 5a,b). This may simply be attributable to the degradation in temporal resolution between the two models, but it is possible that there is some lead time dependence on the diurnal climatology and biases in the GEFS/R. As was seen for kinematic variables overall in Day 1 (Fig. 4), the predictive utility of simulated shear is nearly constant at all times throughout the forecast period for both forecast lead times (Fig. 5). CAPE and CIN both have more pronounced diurnal signatures, but they are different from one another (e.g. Fig. 5a). CAPE FIs peak in association with the maximum in climatological event time frequency, while CIN has a primary peak after this and, in many circumstances (e.g. Fig 5a,c), a secondary peak before it. The secondary peak is perhaps the more intuitive of the two; the environmental CIN in the pre-event environment determines how much of a cap storms must overcome, and the potential for instability to build or storms to be prevented from initiating entirely. The primary peak may speak to the degree of stabilization associated with cold pool strength, instability release, anvil shading, and other factors as portrayed in the convection-parameterized GEFS/R, and the severe weather potential associated with these factors. However, more investigation into the causes of this peak may prove fruitful.

In space (Fig. 6), RF FIs are typically highest near the forecast point and decrease with increasing distance from the point, but there are some notably 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. Fig. 4a,d). In the West, while FI importance 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 with the primary storm ingredients just to the east over the Great Plains. Especially because this appears prominently in the hail signature but not in other fields, the interface between the simulated fields over the Great Plains and events over the far eastern Intermountain West appears a likely source, with more usefully predicted values over the Great Plains, but more investigation is required to validate that hypothesis. In the Central region, FIs are highest from the forecast point south, with downstream maxima for every predictand except severe winds, which has an identified maximum in predictive utility just upstream of the forecast point (Fig. 6h). The southern displacement in importance appears to become more pronounced with increasing forecast lead time, and is especially evident at Day 3 (Fig. 6n). FI maxima also become less pronounced with increasing forecast lead time (Fig. 6j–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.

The so-called ring plots of Figures 7–9 provide a more complete representation of the models' diagnoses and how the summary statistics of Figures 2, 4, and especially 6 were obtained. In the West (Fig. 7), the most predictive fields, CAPE and CIN (Fig. 2) are seen clearly for all

three predictands. In general the importance maxima for these fields occur near the forecast point, though CAPE FI maxima are displaced farther north relative to the forecast point in predicting tornadoes compared with hail and wind (Fig. 7a). For CIN (Fig. 7d), importances maximze downstream of the forecast point, particularly for wind. DSHR is predictive for both hail and wind (Fig. 2a, 7m), but is maximized on the upstream side of the forecast point for wind and downstream side for hail. A different moisture variable is found to be most predictive of for each severe weather predictand: APCP, Q2M, and PWAT for tornadoes, hail, and wind, respectively (Fig. 7h,e,g). In all cases, the spatial maximum in importance is found displaced to the north of the forecast point, likely associated with biases in the GEFS/R's positioning of precipitation systems (e.g Herman and Schumacher 2018a), also seen in other models with paramterized convection (e.g. Clark et al. 2010).

The RF FI maxima and spatial placement thereof displays some similarities and some differences between the West (Fig. 7) and Central (Fig. 8) regions. In the Central region, CAPE FI (Fig. 8a) are still of course paramount for all predictands (per Fig. 2), but unlike in the West region, the maxima are found to the south of the forecast point. This southern displacement is even more pronounced in CIN (Fig. 8d), particularly for forecasting severe hail. In the moisture variables, there is a shift from the West to Central region, with APCP becoming the preferred moisture variable for each predictand. Interestingly, APCP FI importance is consistently maximized late in the period to the northeast of the forecast point, perhaps noting with its late and eastward-displaced elevated FIs that many tornadoes occur during the afternoon hours with discrete supercell activity and during the upscale growth phase leading up to vigorous evening mesoscale convective systems which are common during the warm-season in this region (e.g. Nielsen et al. 2015). The northern displacement is again consistent with the documented displacement bias in the positioning of convective systems in convection-parameterized models such as the GEFS/R (e.g. Wang et al.

2009). DSHR's FI maxima (Fig. 8m) are again centered near the forecast point, although MSHR (Fig. 8j) and SRH (Fig. 8l), which are particularly predictive for tornadoes, have maximum predictive utility to the southeast of the forecast point. Finally, MSLP is also found to a useful severe weather predictor (Fig. 8i), and one observes its importances track from west to east across the forecast point domain throughout the forecast period.

The East region FIs (Fig. 9) display very similar spatial patterns in CAPE (Fig. 9a) and CIN (Fig. 9d) as seen in the Central region. In both cases, it appears that these thermodynamic indicators forecasted in the source region of moisture and instability are more predictive than at the point itself, particularly for CIN. A similar pattern is also seen in APCP (Fig. 9h), including the northward displacement. However, the late maximum in the northeast corner is entirely removed, as nocturnal mesoscale convective systems are not climatologically frequent over much of this region, and the synoptic conditions associated with tornadoes are often different between the regions (e.g. Smith et al. 2012). Shear is again most important nearly collocated with the forecast point (Fig. 9j,m) with MSHR (Fig. 9j)—especially late in the period—being more predictive for tornadoes and wind, while DSHR (Fig. 9m) is the dominant shear variable for predicting hail. In predicting tornadoes, meridional winds (Fig. 9n) to the south of the forecast point and MSLP (Fig. 9i) upstream of the forecast point are found to be good discriminators of tornado events and non-tornado events, speaking to both the degree of advection of convective ingredients from the south and the level of synoptic-scale forcing for ascent advecting into the region. SRH (Fig. 9l) is found to be predictive of tornadoes throughout the period, with FI maxima generally tracking west to east to the immediate south of the forecast point during the period. One other major difference between the East region and other regions is the importance of nighttime T2M (Fig. 9b) in predicting hail in the East; the exact reasoning for this identification is not obvious.

In summary, the RFs trained in this study appear to be making statistical deductions that are in strong agreement with our physical understanding of severe weather processes, and identify values to inspect—such as CAPE and shear near the forecast point and APCP to its north, and inspecting DSHR for hail but MSHR for tornadoes—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 weather forecast, in addition to identifying some factors, such as the southward CIN FI maxima displacement, that may be less well-documented but still contribute to a skillful forecast. The following section investigates the predictive performance of these models.

4. Results: Model Performance

The RFs show ability to skillfully predict all severe weather predictands (Fig. 10), though there are some differences in the details. Prediction of tornadoes (Fig. 10a) 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 Floridian 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. 10b) but with lower skill overall, with CONUS-wide skill 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 non-existent sample, the negative skill observed here is not statistically significant. Hail (Fig. 10c), wind (Fig. 10e), and the Day 2 and 3 (Fig. 10g,h) models all exhibit very similar spatial

patterns of forecast skill, with near uniform and statistically significant positive skill over much of 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. 10c). In the West, fewer of the results are found to be statistically significant due to the reduced event frequency. Nevertheless, positive skill is still noted for these predictands over much of the West, with the exceptions 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 (Fig. 10g,h). Like with tornadoes, the spatial patterns are similar between hail and wind and their significant severe counterparts (Fig. 10d,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 is seen 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 (Fig. 11), the RF outlooks verify quite competitively. 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, with aggregate skill score differences of 0.007 for hail (Fig. 11c) increasing to 0.013 for tornadoes (Fig. 11a) and 0.024 for severe wind forecasts (Fig. 11e). 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 (Fig. 10c,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 compari-

son, with SPC outlooks substantially outperforming the RF outlooks over far northern CONUS in predicting severe hail and wind (Fig. 11c,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. 11a). 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. 11b) as well, but the comparison is much different for significant hail (Fig. 11d) and wind (Fig. 11f) 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 outlooks (Fig. 11g,h), the RF outlooks exhibit substantially higher probabilistic skill than the analogous SPC convective outlooks, 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. 10g), and over Louisiana and Arkansas. 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.

Except for hail (Fig. 12b), which exhibits a springtime maximum in skill, all RF outlooks exhibit a climatology-relative peak in skill during the cold-season (Fig. 12a,c,d). 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 outloks (Herman et al. 2018). For all severe weather predictands, the severe and significant severe events have nearly identical seasonal cycles in forecast skill (Fig. 12). Comparing against SPC, while there does not appear to be a clear seasonal or monthly signal in the skill difference for tornado outlooks (Fig. 12a), 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 Day 3 comparison, RF outlooks outperform SPC by the most during the summer, maximizing in July. 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, SPC outlooks suffer from biased guidance that cannot come close to resolving 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.

Reliability diagrams for the RF outlooks (Fig. 13) 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 at Days 2 and 3. The main exception to calibration is 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 mostly relatively quiet (Herman et al. 2018).

The weighted blend of SPC and RF outlooks described in Section 2 (Fig. 14) unsurprisingly demonstrates forecast skill spatial characteristics of both the interpolated SPC (Herman et al. 2018) and RF (Fig. 10) outlooks. 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. Fig. 10c,14c; 10e,14e). For predictands in which the skill difference is large between the two outlook sources, such as for significant wind (Fig. 14f) and the medium-range outlooks (Fig. 14g,h), the blended outlooks verify very similarly to the more skillful component, in part simply because the weights direct the blend heavily towards that component. Across the board, the SPC RF blend verifies as or more skillfully than the SPC outlooks alone—both in space (Fig. 15) and when aggregated across CONUS (Fig. 16)—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. 16), 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 (Fig. 15c,e). At Day 2 and 3, while the blend is not able to improve skill over the RF outlooks (Fig. 16), that difference is already considerable when compared with the SPC outlooks at 0.044 and 0.048 (Fig. 15g,h). Consequently, the blended forecast exhibits much improved skill compared with the raw SPC outlooks for all eight forecast predictands evaluated (Fig. 16). Even more encouragingly, the skill improvements are seen across all regions of CONUS (Fig. 15) with fairly uniform distribution. For hail, wind, and the medium-range outlooks, the skill differences are statistically significant over all except for pockets of western CONUS where the climatological event frequencies are insufficient to produce a robust sample. Hail outlooks ae most improved over

the Mississippi Valley region into the Midwest, while wind outlooks are most improved over the southern Plains, and the medium-range outlooks most improved over the East Coast urban corridor.

One additional instructive skill decomposition inspects forecast verification in the CAPE vs. shear parameter space. The raw RF hail (Fig. 17d) and wind (Fig. 17g) 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 (Fig. 17g) exhibits a local BSS minimum in this region as well, 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 (Fig. 17a) verification results are more mixed. Like hail, forecast skill suffers in scenarios with ample supply of CAPE or shear, but little of the other. 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, as evidenced by the interior pockets of blue in Figure 17a. The addition of the weighted average with SPC outlooks (Fig. 17b,e,h) improve 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. 17e); 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. 17f) and wind (Fig. 17i) forecasts, and across much of the domain for tornadoes (Fig. 17c). The skill improvements are largest in the low shear end of the parameter space, especially with high CAPE. Moderate to high wind shear is a necessary ingredient for supercell activity, processes which can be much better resolved by CAMs than parameterized guidance like the GEFS/R. Benefit of employing these RF outlooks can likely be maximized on the low shear end of

the parameter space because the benefits from the statistical learning are more offset by an inferior representation of the underlying dynamics in the GEFS/R in high wind shear scenarios.

Finally, a brief case study example is provided in order to illustrate the real-time character of the ML model forecasts. Across many cases, the spatial character of the ML-based outlooks are often very similar to those produced by SPC. This is seen for the outlooks valid 1200 UTC May 9 2016–1200 UTC May 10 2016 (Fig. 18), a period in the middle of a moderate-severity multi-day outbreak which spread from the Colorado Plains out to Appalachia. This 24-hour period, while not the most intense outbreak of the evaluation period, garnered a considerable number of reports for each severe weather phenomenon in different areas, including significant severe observations for each. Tornadoes (Fig. 18a,b) occurred primarily in two groups. One cluster centered about southern and southeastern Oklahoma, with scattered reports up into central Oklahoma and south and east into Arkansas and far northeastern Texas. The second cluster was more broadly spread out from southern Nebraska and northern Kansas east across Iowa and Missouri into western Illinois. Both had at least one significant tornado embedded. Hail observations (Fig. 18c,d) were more focused in a north-south oriented region extending from the Oklahoma-Texas border into far northern and northeastern Nebraska, with significant observations seen throughout this region. Wind observations (Fig. 18e,f), in contrast, were observed only in two regions: a tightly clustered region in south central Kansas, and a broader region from the Texas/Oklahoma/Arkansas triple point extending northeast across Arkansas into southeastern Missouri. SPC's Day 1 tornado outlook (Fig. 18b) 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. They identified hail (Fig. 18d) 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 had

essentially an identical outline to the severe hail one, except topping out with approximately 15% event probabilities and no significant wind contour.

The ML Day 1 outlooks did several desirable changes compared with the SPC outlooks. The tornado outlook (Fig. 18a) 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. 18c) and wind (Fig. 18e) 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 models also had better spatial placement in the medium-range, even indicating the two primary risk areas at Day 2 (Fig. 18g), and encompassing the western severe weather observations when the operational outlook (Fig. 18h) did not. This was further magnified at Day 3 when only a 15% severe probability was indicated and many severe weather over the Central Plains were not encompassed by the 5% marginal contour in the operational outlook (Fig. 18j), while nearly every observation was encompassed by a marginal contour at Day 3 in the ML outlook (Fig. 18i) and severe probabilities maximized over 30%. 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.

5. Summary and Conclusions

RFs have been trained to generate probabilistic predictions of severe weather for Days 1–3 across CONUS with analogous predictands to SPC's convective outlooks, with tornado, hail, and wind

599 treated separately at Day 1 and collectively for Days 2–3. Distinct RFs were trained for western,
600 central, and eastern CONUS as partitioned in Figure 1. Inputs to the RFs came from the GEFS/R
601 ensemble median of 12 different atmospheric fields: APCP, CAPE, CIN, PWAT, U10, V10, UV10,
602 T2M, Q2M, MSHR, and DSHR. For the Day 1 models, three additional predictors were used:
603 RH2M, ZLCL, and SRH. The spatiotemporal evolution of each of these fields in the vicinity of
604 the forecast point—up to 1.5° away in any direction for some fields and up to 3° away in others,
605 depending on the grid resolution (see Table 1)—throughout the forecast period was included in the
606 predictor set to provide a comprehensive assessment of the simulated environmental conditions for
607 each severe weather forecast. 3-hourly temporal resolution is used for Day 1 and 2 models, and
608 6-hourly resolution was used for Day 3. Each of the fifteen RFs—three regions, five predictands—
609 was trained on nine years of forecasts spanning 12 April 2003–11 April 2012. The identified
610 relationships between simulated model variables and observed severe weather during that period
611 were assessed using RF FIs. The trained RFs were then run over an extended withheld test period
612 spanning 12 April 2012–31 December 2016 and the performance of these forecasts assessed, both
613 in isolation with a climatological reference and relative to SPC convective outlooks issued during
614 the same period.

615 The statistical relationships identified by the RFs bear considerable correspondence with known
616 physical relationships between atmospheric variables and severe weather, lending credence to the
617 veracity of the model solutions. For example, CAPE, CIN, and wind shear—some of the most
618 commonly used variables to characterize severe weather environments (e.g. Johns and Doswell
619 1992)—are consistently identified as the most predictive variables for forecasting severe weather.
620 More nuanced identifications are made as well, including more emphasis on kinematics in tor-
621 nado prediction compared with hail and wind, and additionally, wind difference over a shallower
622 vertical layer being more predictive for tornadoes than for hail and wind. Even spatiotemporal re-

lationships that are identified accord with physical intuition, such as meridional wind to the south of the forecast point speaking to the degree of temperature and moisture advection into the region, and upstream pressure transitioning to be over and eventually past the forecast point during the forecast period. Previously identified dynamical model biases (e.g. Wang et al. 2009; Herman and Schumacher 2018a) also emerge objectively from the analysis, including the northward displacement bias of convective systems in the GEFS/R and other convection-parameterized models.

The trained models produce real-time forecasts on unseen inputs that exhibit similar spatial and quantitative character to their human-produced counterparts. In general, they produce somewhat larger regions of marginal risk equivalence and fewer incidences of moderate and high risk-equivalent outlooks. This behavior can be largely attributed to the ML-based outlooks being informed by less total real-time information—a single ensemble rather than many different models coupled with observations—and lower-resolution output than is available to the human forecaster, leading to lower confidence and higher uncertainty. Nevertheless, ML outlooks do produce across the gamut of risk categories for all lead times, and the differences in real-time forecast guidance are typically merely quantitative, rather than highlighting completely different risk areas when compared with SPC outlooks.

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 statistically 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 spatial uniform, although the most gain was generally seen in the Mississippi Valley at Day 1 and the East

for Day 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. Finally, the largest forecast skill improvements generally came when wind shear was relatively low, but across the spectrum of environmental CAPE.

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 so 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%, and can gain considerable probabilistic skill simply by virtue of having higher resolution in this domain of the probability space. 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%.

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 post-processing 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 pre-processing; 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. 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 author greatly thanks advisor Russ Schumacher for guidance, support, and encouragement throughout this study, and on many fruitful discussions and penetrating insights related to this work and on previous foundational studies. The author also thanks 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 author had several engaging discussions about machine learning model development and severe weather applications with David John Gagne. Both of these greatly improved the quality of this study. Computational resources were generously afforded by the National Center for Atmospheric Research Computational Information Systems

Laboratory. Funding for this research was supported by NOAA Award NA16OAR4590238 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:

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

for temperature in K and pressure in Pa, where 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

SHEAR850 and SHEAR500—bulk wind differences between two vertical levels—are calculated straightforwardly:

$$SHEAR850 = \sqrt{(U_{850} - U_{10m})^2 + (V_{850} - V_{10m})^2} \tag{A2}$$

$$SHEAR500 = \sqrt{(U_{500} - U_{10m})^2 + (V_{500} - V_{10m})^2} \tag{A3}$$

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 meters 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) 10m–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:

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

Accordingly:

$$SRH850 = \sum_{l=1}^{5} \max(0, SRH_l) \tag{A5}$$

where

$$SRH_l = (Z_l - Z_{l-1}) \left((\overline{V_l} - V_{st}) \frac{U_l - U_{l-1}}{Z_l - Z_{l-1}} - (\overline{U_l} - U_{st}) \frac{V_l - V_{l-1}}{Z_l - Z_{l-1}} \right) \tag{A6}$$

with

$$\overline{U_l} = \frac{U_l + U_{l-1}}{2}; \overline{V_l} = \frac{V_l + V_{l-1}}{2} \tag{A7}$$

and

$$U_{st} = \sqrt{0.75} * (\overline{U}\cos(-30^\circ) - \overline{V}\sin(-30^\circ)) \quad \text{(A8)}$$

$$V_{st} = \sqrt{0.75} * (\overline{U}\sin(-30^\circ) + \overline{V}\cos(-30^\circ)) \quad \text{(A9)}$$

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.

Agresti, A., and B. A. Coull, 1998: Approximate is better than exact for interval estimation of binomial proportions. *The American Statistician*, **52**, 119–126, doi: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.

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.*, **in press**.

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 five week forecasts of tornado and hail activity. *J. Geophys. Res.*, **submitted**.

Bolton, D., 1980: The computation of equivalent potential temperature. *Mon. Wea. Rev.*, **108**, 1046–1053.

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

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, doi: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, doi:10.1175/WAF915.1.

Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, doi: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 (4)**, 626–640.

Clark, A. J., W. A. Gallus Jr, and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509.

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*, 573–582.

Doswell III, C. A., 2004: Weather forecasting by humansheuristics and decision making. *Wea. Forecasting*, **19**, 1115–1126.

Doswell III, C. A., and D. M. Schultz, 2006: On the use of indices and parameters in forecasting severe storms. *Electronic J. Severe Storms Meteor.*, **1**.

Edwards, R., G. W. Carbin, and S. F. Corfidi, 2015: Overview of the Storm Prediction Center. *Preprints, 13th History Symp., Phoenix, AZ, Amer. Meteor. Soc., 1.1*, [Available online at http://www.spc.noaa.gov/publications/edwards/spc-over.pdf].

771 Elsner, J. B., and H. M. Widen, 2014: Predicting spring tornado activity in the central Great Plains
772 by 1 March. *Mon. Wea. Rev.*, **142**, 259–267.

773 Friedman, J. H., 2001: Greedy function approximation: a gradient boosting machine. *Ann. Stat.*,
774 1189–1232, doi:10.1214/aos/1013203451.

775 Gagne, D. J., A. McGovern, J. B. Basara, and R. A. Brown, 2012: Tornadic supercell environments
776 analyzed using surface and reanalysis data: a spatiotemporal relational data-mining approach.
777 *J. Appl. Meteor. Climatol.*, **51**, 2203–2217.

778 Gagne, D. J., A. McGovern, and J. Brotzge, 2009: Classification of convective areas using decision
779 trees. *J. Atmos. Oceanic Technol.*, **26**, 1341–1353.

780 Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-
781 based probabilistic hail forecasting with machine learning applied to convection-allowing en-
782 sembles. *Wea. Forecasting*, **32**, 1819–1840.

783 Gagne, D. J., A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale
784 ensemble probabilistic quantitative precipitation forecasts. *Wea. Forecasting*, **29**, 1024–1043.

785 Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2018: Blended
786 probabilistic tornado forecasts: Combining climatological frequencies with NSSL–WRF en-
787 semble forecasts. *Wea. Forecasting*, **33**, 443–460.

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

790 Hamill, T. M., 2017: Changes in the systematic errors of global reforecasts due to an evolving data
791 assimilation system. *Mon. Wea. Rev.*, **145**, 2479–2485.

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

795 Herman, G. R., E. R. Nielsen, and R. S. Schumacher, 2018: Probabilistic verification of Storm
796 Prediction Center convective outlooks. *Wea. Forecasting*, **33**, 161–184.

797 Herman, G. R., and R. S. Schumacher, 2016: Using reforecasts to improve forecasting of fog and
798 visibility for aviation. *Wea. Forecasting*, **31**, 467–482.

799 Herman, G. R., and R. S. Schumacher, 2018a: Dendrology in numerical weather prediction: What
800 random forests and logistic regression tell us about forecasting extreme precipitation. *Mon. Wea.*
801 *Rev.*, **in press**.

802 Herman, G. R., and R. S. Schumacher, 2018b: Money doesn't grow on trees, but forecasts do:
803 Forecasting extreme precipitation with random forests. *Mon. Wea. Rev.*, **146**, 1571–1600, doi:
804 10.1175/MWR-D-17-0250.1.

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

807 Hitchens, N. M., and H. E. Brooks, 2013: Preliminary investigation of the contribution of super-
808 cell thunderstorms to the climatology of heavy and extreme precipitation in the United States.
809 *Atmospheric research*, **123**, 206–210.

810 Hitchens, N. M., and H. E. Brooks, 2014: Evaluation of the Storm Prediction Center's con-
811 vective outlooks from day 3 through day 1. *Wea. Forecasting*, **29**, 1134–1142, doi:10.1175/
812 WAF-D-13-00132.1.

Hitchens, N. M., and H. E. Brooks, 2017: Determining criteria for missed events to evaluate significant severe convective outlooks. *Wea. Forecasting*, **32**, 1321–1328, doi: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 (2)**, 129–151.

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.

Johns, R. H., and C. A. Doswell, III, 1992: Severe local storms forecasting. *Wea. Forecasting*, **7**, 588–612.

Karstens, C. D., and Coauthors, 2018: Development of a human-machine mix for forecasting severe convective events. *Wea. Forecasting*, **in press**.

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.

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

Krocak, M. J., and H. E. Brooks, 2018: Climatological estimates of hourly tornado probability for the United States. *Wea. Forecasting*, **33**, 59–69.

Kuchera, E. L., and M. D. Parker, 2006: Severe convective wind environments. *Wea. Forecasting*, **21**, 595–612.

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.

834 Loridan, T., R. P. Crompton, and E. Dubossarsky, 2017: A machine learning approach to modeling
835 tropical cyclone wind field uncertainty. *Mon. Wea. Rev.*, **145**, 3203–3221.

836 Marzban, C., and G. J. Stumpf, 1996: A neural network for tornado prediction based on Doppler
837 radar-derived attributes. *J. Appl. Meteor.*, **35**, 617–626.

838 Marzban, C., and G. J. Stumpf, 1998: A neural network for damaging wind prediction. *Wea.*
839 *Forecasting*, **13**, 151–163.

840 Marzban, C., and A. Witt, 2001: A Bayesian neural network for severe-hail size prediction. *Wea.*
841 *Forecasting*, **16**, 600–610.

842 McGovern, A., K. L. Elmore, D. J. Gagne, S. E. Haupt, C. D. Karstens, R. Lagerquist, T. Smith,
843 and J. K. Williams, 2017: Using artificial intelligence to improve real-time decision making for
844 high-impact weather. *Bull. Amer. Meteor. Soc.*, **98**, 2073–2090, doi:10.1175/BAMS-D-16-0123.
845 1.

846 McGovern, A., D. John Gagne, N. Troutman, R. A. Brown, J. Basara, and J. K. Williams, 2011:
847 Using spatiotemporal relational random forests to improve our understanding of severe weather
848 processes. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **4**, 407–429.

849 Mesinger, F., and Coauthors, 2006: North American regional reanalysis. *Bull. Amer. Meteor. Soc.*,
850 **87**, 343–360, doi:10.1175/BAMS-87-3-343.

851 Murphy, A. H., and R. L. Winkler, 1977: Reliability of subjective probability forecasts of precipi-
852 tation and temperature. *Appl. Statistics*, **26**, 41–47, doi:10.2307/2346866.

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

Nielsen, E. R., and R. S. Schumacher, 2018: Dynamical insights into extreme short-term precipitation associated with supercells and mesovortices. *J. Atmos. Sci.*, Submitted, doi:10.1175/JAS-D-18-0385.1.

NWS, 2018: Summary of natural hazard statistics in the United States. National Weather Service, Office of Climate, Weather, and Water Services, [Available online at http://www.nws.noaa.gov/om/hazstats.shtml].

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.

Parker, M. D., 2014: Composite VORTEX2 supercell environments from near-storm soundings. *Monthly Weather Review*, **142**, 508–529.

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.

Romps, D. M., 2017: Exact expression for the lifting condensation level. *J. Atmos. Sci.*, **74**, 3891–3900.

Rothfusz, L., C. Karstens, and D. Hilderbrand, 2014: Forecasting a continuum of environmental threats: Exploring next-generation forecasting of high impact weather. *Eos, Trans. Amer. Geophys. Union*, **95**, 325–326.

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, doi:10.1175/WAF-D-13-00041.1.

878 Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective modes
879 for significant severe thunderstorms in the contiguous United States. Part I: Storm classification
880 and climatology. *Wea. Forecasting*, **27**, 1114–1135.

881 Smith, J. A., M. L. Baeck, Y. Zhang, and C. A. Doswell III, 2001: Extreme rainfall and flooding
882 from supercell thunderstorms. *J. Hydrometeor.*, **2**, 469–489.

883 Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011:
884 Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme
885 phenomena in convection-allowing model forecasts. *Wea. Forecasting*, **26**, 714–728, doi:10.
886 1175/WAF-D-10-05046.1.

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

890 Sobash, R. A., C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016b: Severe
891 weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecast-*
892 *ing*, **31**, 255–271, doi:10.1175/WAF-D-15-0138.1.

893 Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, 2008: Conditional variable
894 importance for random forests. *BMC bioinformatics*, **9**, 307.

895 Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn, 2007: Bias in random forest variable
896 importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, **8**, 25.

897 Thompson, R. L., C. M. Mead, and R. Edwards, 2007: Effective storm-relative helicity and bulk
898 shear in supercell thunderstorm environments. *Wea. Forecasting*, **22**, 102–115.

Tippett, M. K., A. H. Sobel, and S. J. Camargo, 2012: Association of US tornado occurrence with monthly environmental parameters. *Geophys. Res. Lett.*, **39**, L02 801.

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.

Wang, S.-Y., T.-C. Chen, and S. E. Taylor, 2009: Evaluations of NAM forecasts on midtropospheric perturbation-induced convective storms over the US northern plains. *Wea. Forecasting*, **24**, 1309–1333.

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 A*, **60**, 62–79.

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.*, **submitted**.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. Academic press, 676 pp.

LIST OF TABLES

Symbol	Description	Grid	Calculated	Class
APCP	Precipitation accumulation in past (3) 6 hours	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	$1^\circ \times 1^\circ$	Derived	Thermodynamic
SHEAR500	Bulk wind difference mangitude between 10 meters and 500 hPa	$1^\circ \times 1^\circ$	Derived	Kinematic
SHEAR850	Bulk wind difference mangitude between 10 meters and 850 hPa	$1^\circ \times 1^\circ$	Derived	Kinematic
SRH850*	Storm relative helicity from surface to 850 hPa	$1^\circ \times 1^\circ$	Derived	Kinematic
T2M	Air temperature two meters above ground	Native Gaussian	Archived	Thermodynamic
U10	Zonal-component of 10-meter wind	Native Gaussian	Archived	Kinematic
UV10	10 meter wind speed	Native Gaussian	Derived	Kinematic
V10	Meridional-component of 10-meter wind	Native Gaussian	Archived	Kinematic
ZLCL*	Height of Lifted Condensation Level	$1^\circ \times 1^\circ$	Derived	Thermodynamic

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.

Lead Time	West	Central	East
Day 1	(500,30)	(500,120)	(1000,120)
Day 2	(1000,30)	(1000,120)	(1000,120)
Day 3	(1000,30)	(1000,120)	(1000,120)

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. In each data cell, the first number corresponds to the forest size, B, while the second number corresponds to the Z parameter, the minim number of samples permitted to split an impure node. For more details, see Pedregosa et al. (2011) and Herman and Schumacher (2018b).

LIST OF FIGURES

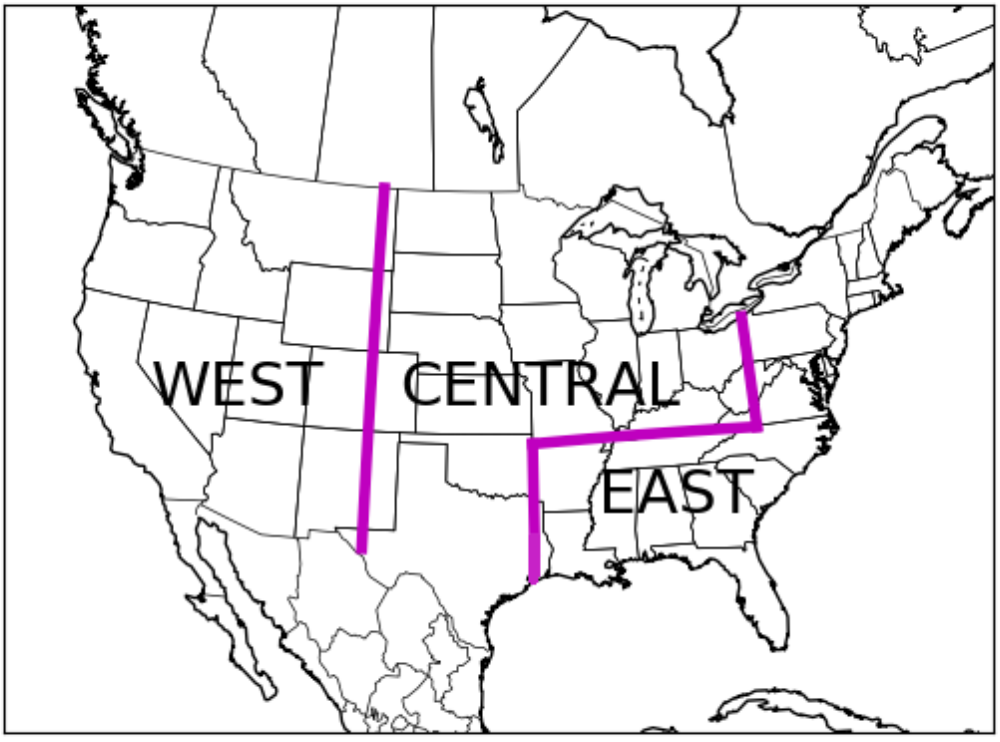


FIG. 1. Map depicting the training regions of CONUS for the statistical models used in this study.

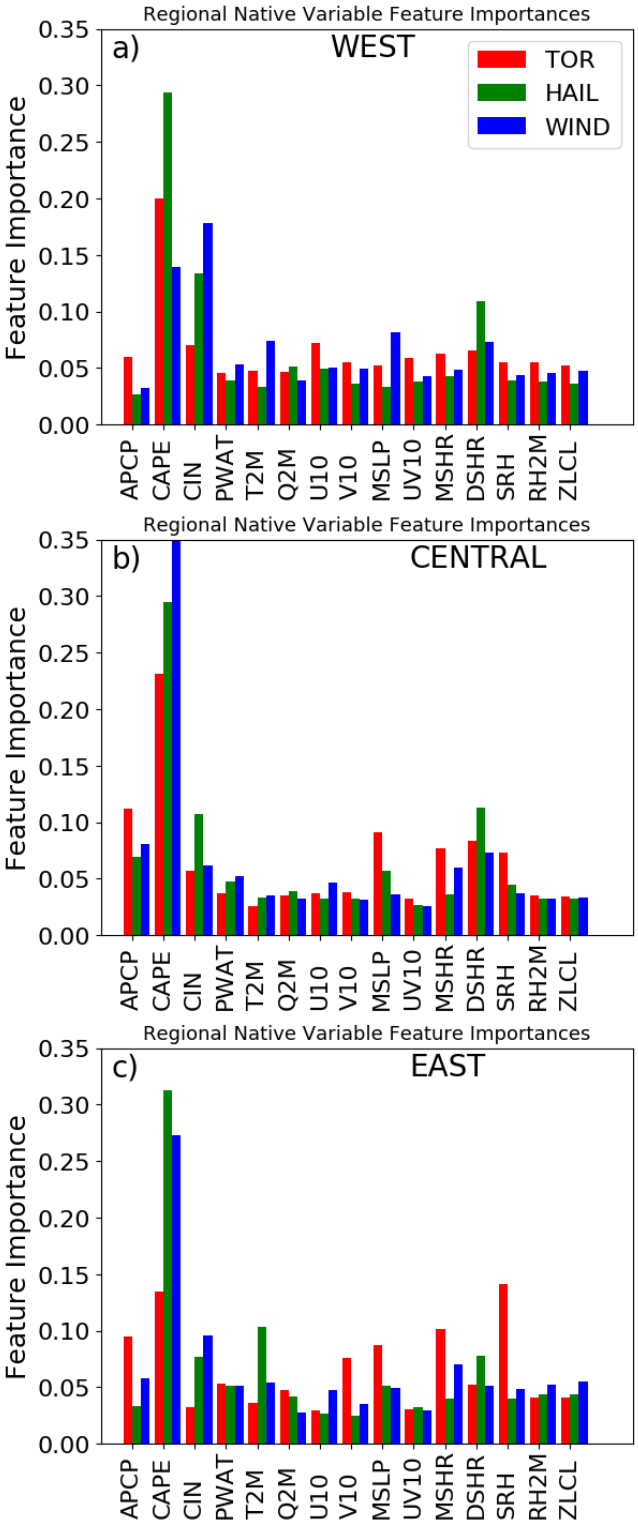
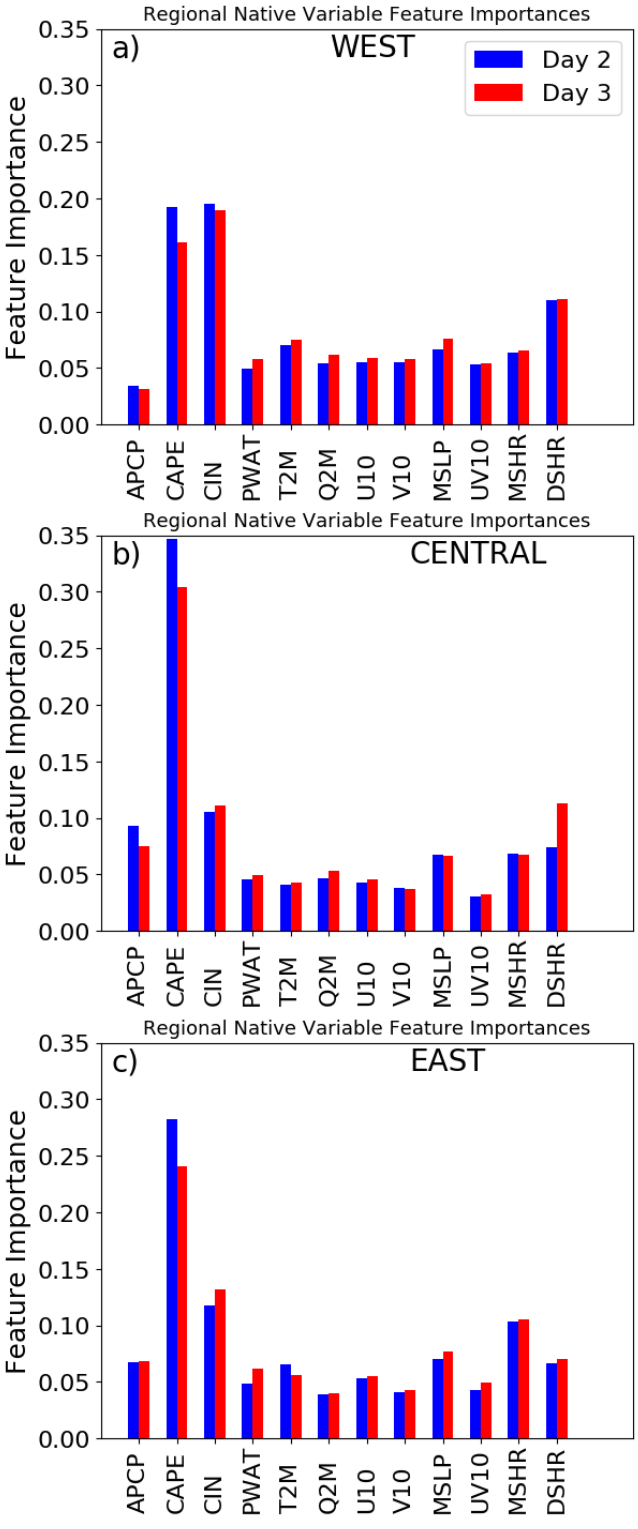


FIG. 2. FIs aggregated by atmospheric field for the Day 1 models in the WEST, CENTRAL, and EAST regions in panels (a)–(c), respectively. Red bars correspond to FIs for the tornado predictive model, green bars to the hail predictive model, and blue bars to the wind predictive model for each region.



1032 FIG. 3. Same as Figure 2, but for the Day 2 and 3 models. Day 2 and 3 FIs are indicated in red and blue bars,
1033 respectively.

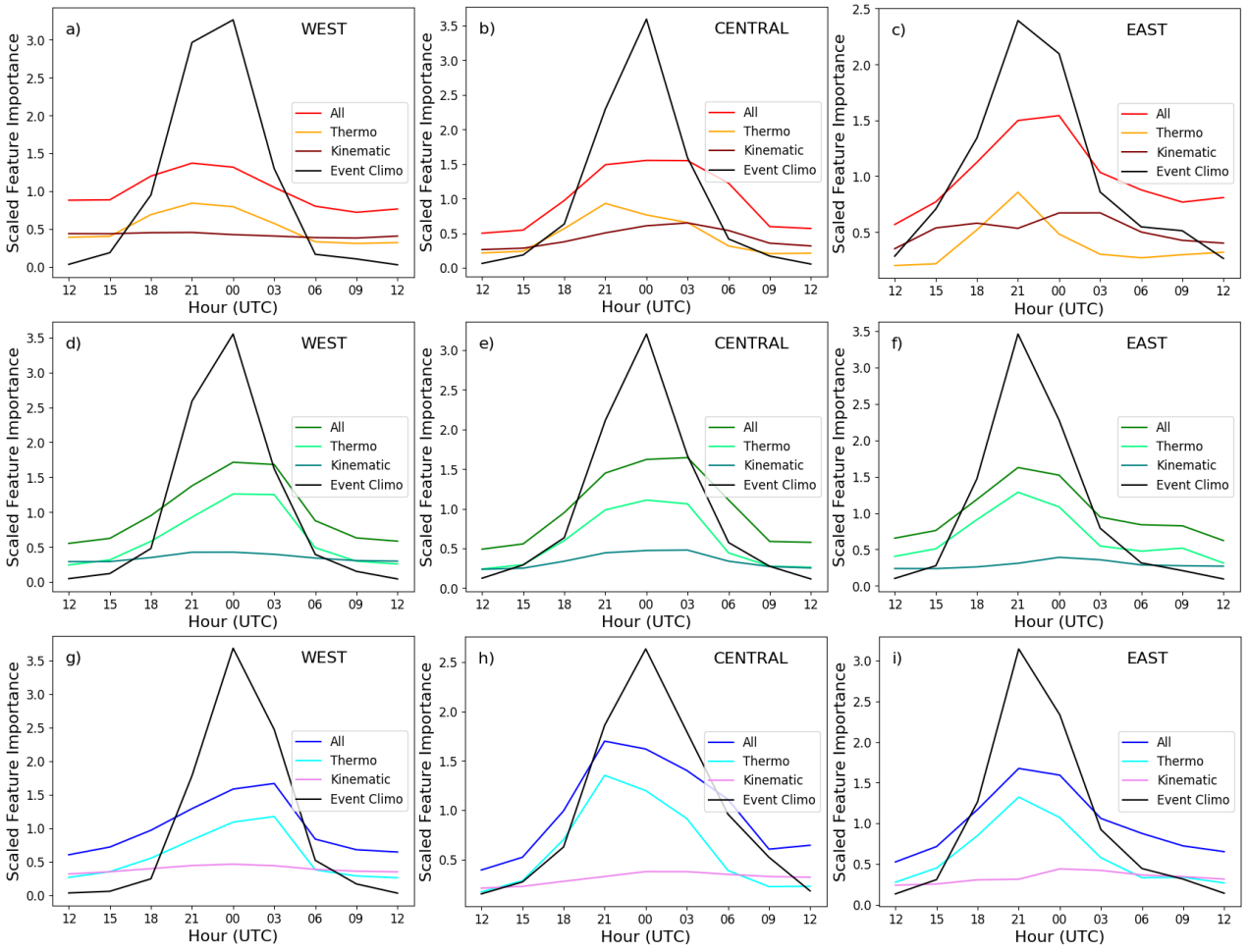


FIG. 4. Normalized FIs aggregated as a function of forecast hour for the Day 1 models. The top, middle, and bottom rows depict FIs for the tornado, hail, and wind models, respectively, while the left, center, and right columns respectively depict FIs for the WEST, CENTRAL, and EAST regions. Severe phenomenon 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.

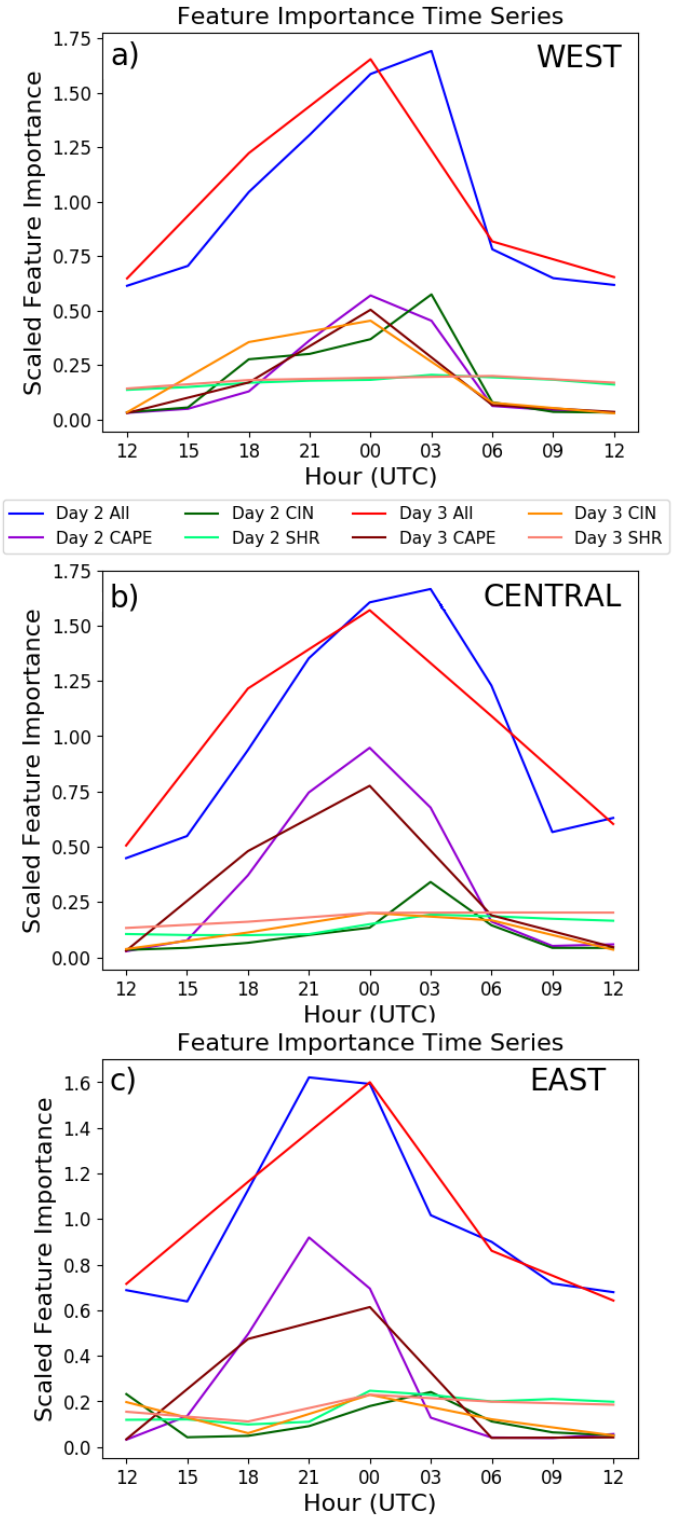


FIG. 5. Similar to Figure 4, except for the Day 2 and 3 models, which are combined onto single panels for the (a) WEST, (b) CENTRAL, and (c) EAST regions. FI time series of CAPE, CIN, shear, and all variables combined are shown for each forecast region, colored as indicated in the panel legend.

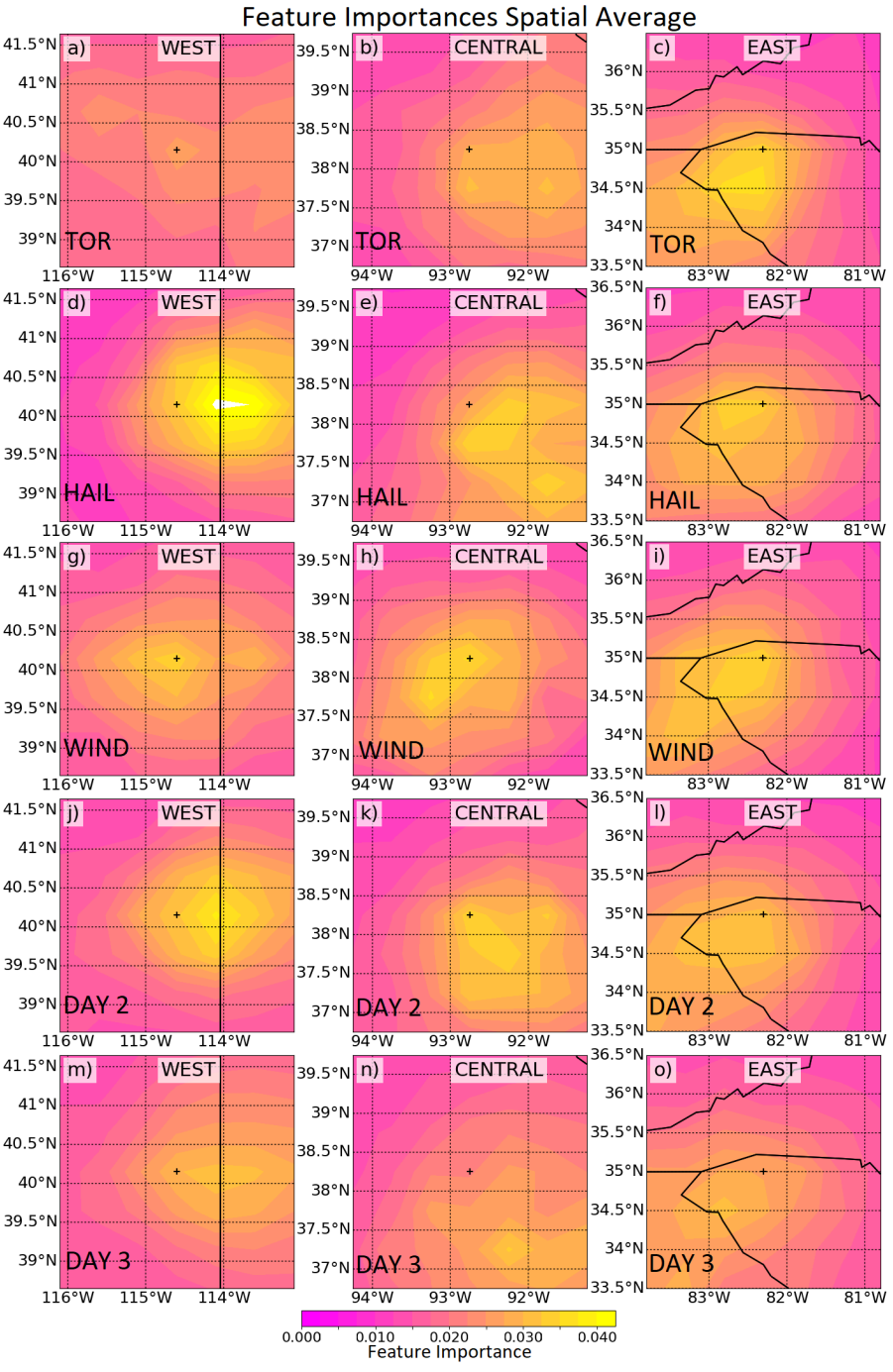


FIG. 6. FIs summed according to the corresponding predictor's position in point-relative space for the WEST, CENTRAL, and EAST regions respectively in the left, center, and right columns. Tornado model FIs are depicted in the top row, followed by hail, wind, Day 2, and finally the Day 3 model on the 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; latitude and longitude are presented using the region centroid, and are shown merely to provide improved sense of spatial scale.

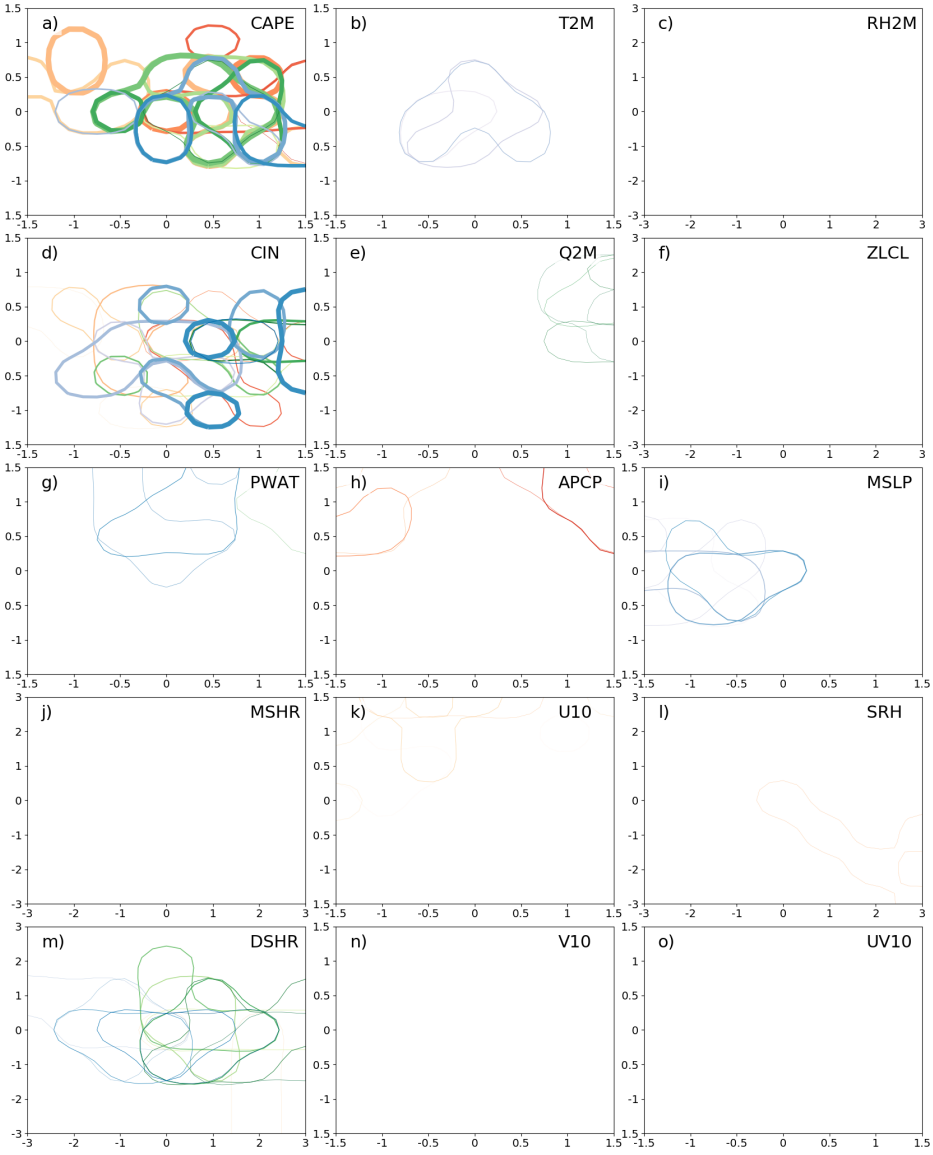


FIG. 7. Feature importances by space and atmospheric field for the Day 1 tornado, hail, and wind models in the WEST region. Rings enclose regions where the FI for the variable and time exceeds 1.5 standard deviations above the spatial mean FI for that variable and time. Ring colors vary according to the predictand of the model, with oranges and reds corresponding to FIs associated with predicting tornadoes, greens to predicting hail, and blues to predicting wind. Within these, colors darken and transition from orange (tornado), green-yellow (hail), and purple-blue (wind) to solid red, green, and blue with time throughout the forecast period, from the front-end 1200 UTC (forecast hour 12) to the back-end 1200 UTC (forecast hour 36). Line thickness is determined by the FI threshold associated with the ring, with thicker lines indicating higher FI and rings associated with below average thresholds (based on the +1.5 standard deviation exceedance given the predictand, predictor field, and time) are excluded entirely. Panels (a)–(o) correspond respectively to FIs for the CAPE, T2M, RH2M, CIN, Q2M, ZLCL, PWAT, APCP, MSLP, MSHR, U10, SRH, DSHR, V10, and UV10 fields.

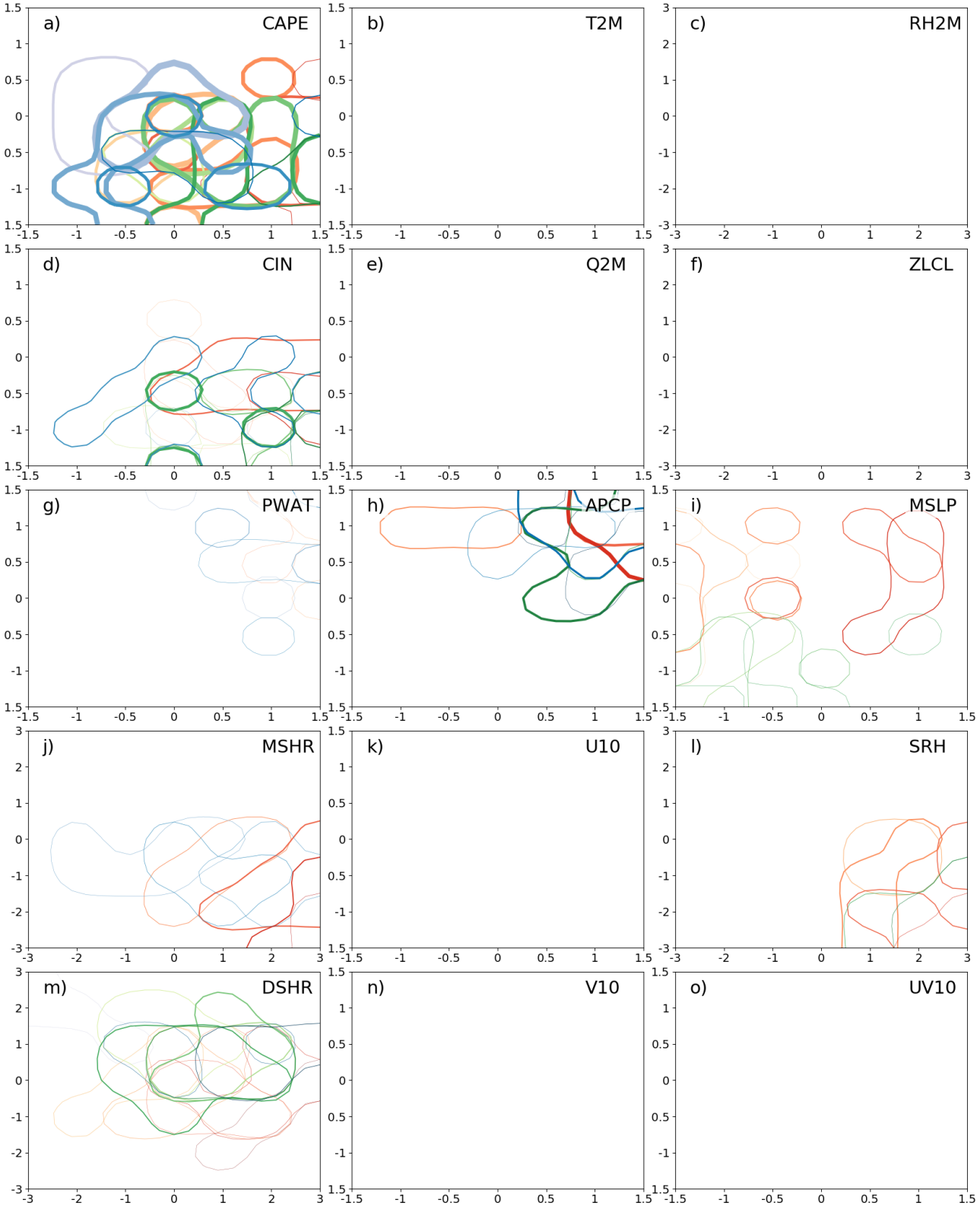


FIG. 8. Same as Figure 7, but for the CENTRAL region.

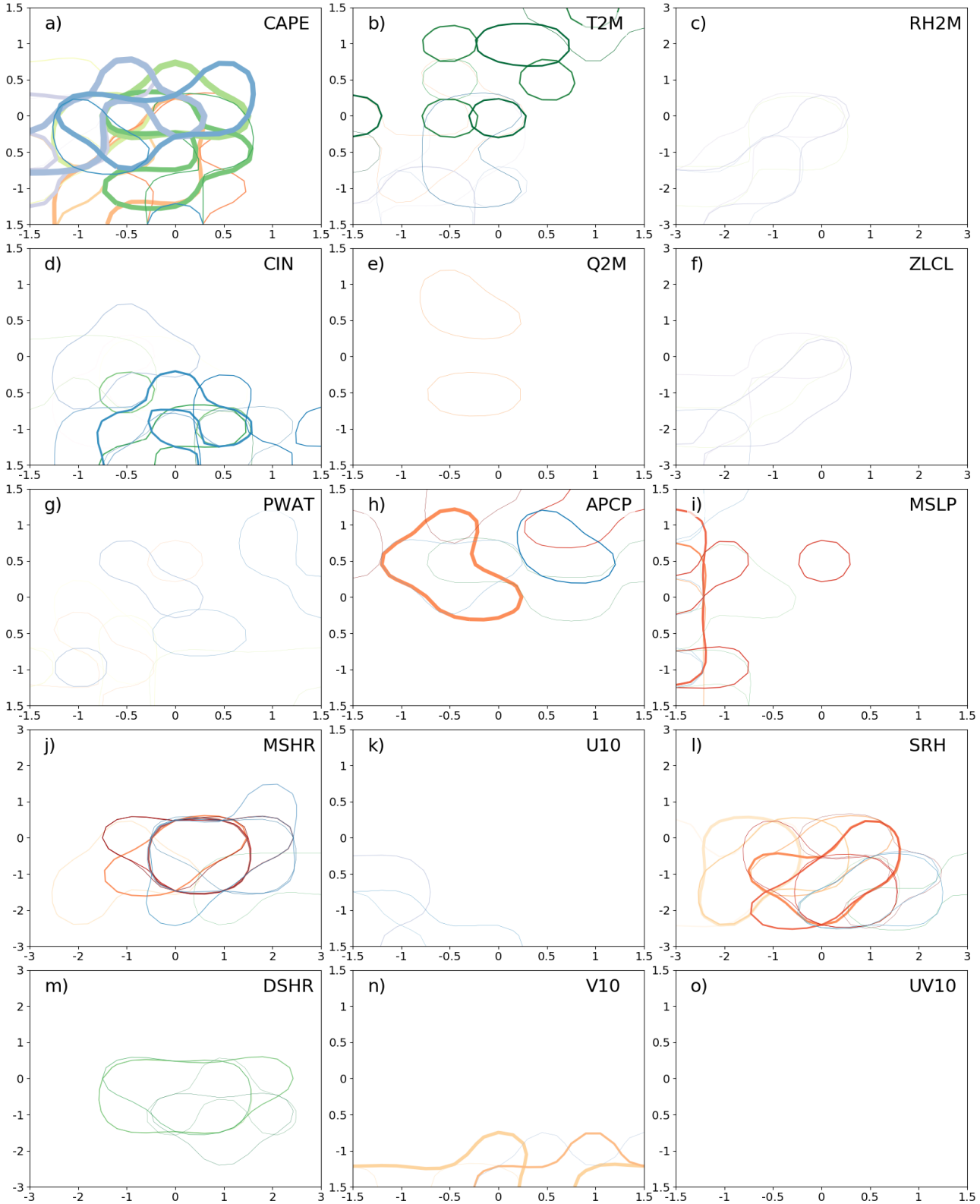


FIG. 9. Same as Figure 7, but for the EAST region.

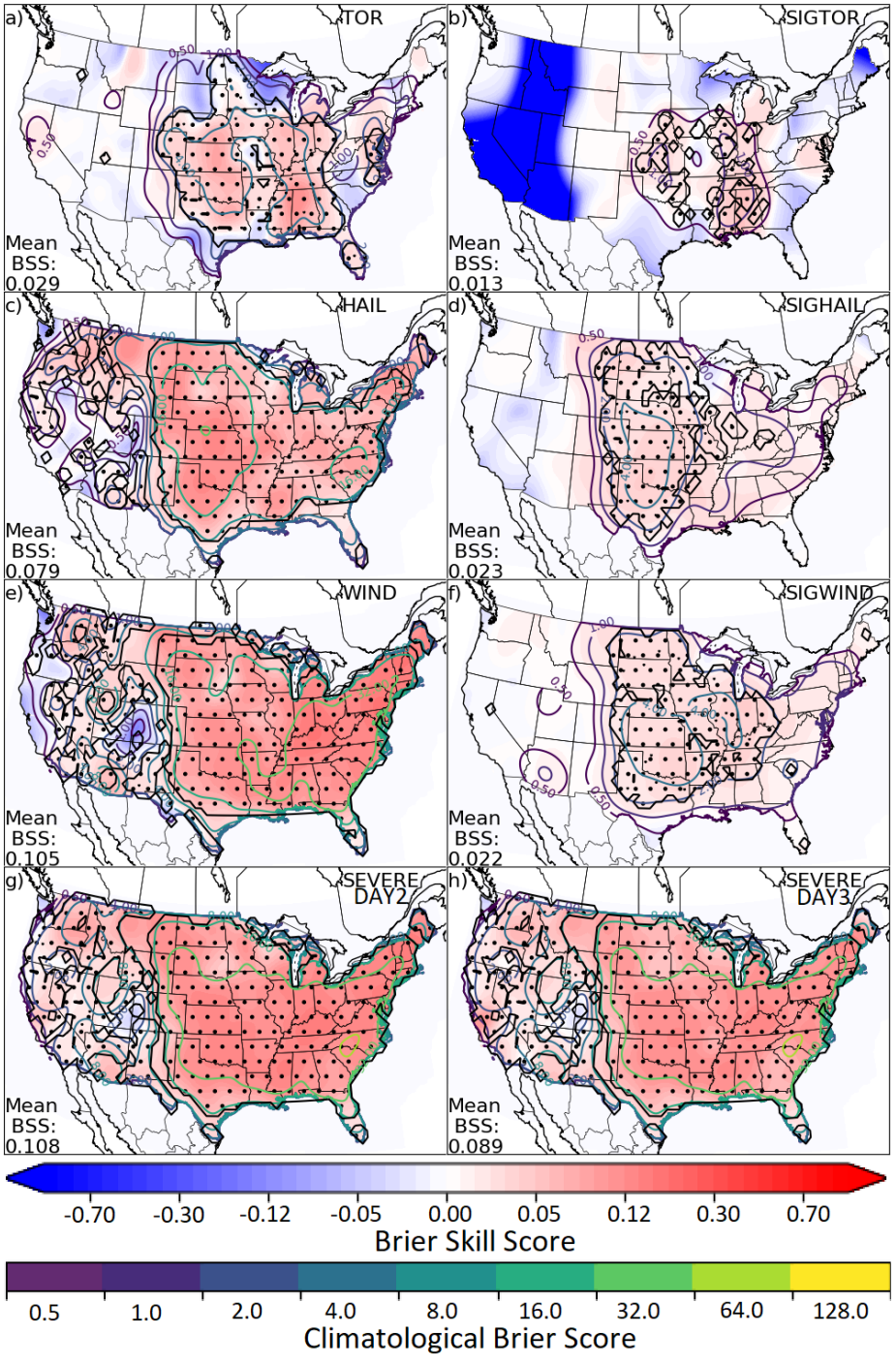


FIG. 10. Brier skill scores (filled contours) in space evaluated over the 12 April 2012–31 December 2016 verification period for each of the ML models trained in this study. Panels (a)–(h) correspond respectively to the performance of the tornado, significant tornado, hail, significant hail, severe wind, significant severe wind, Day 2, and Day 3 outlooks. 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.

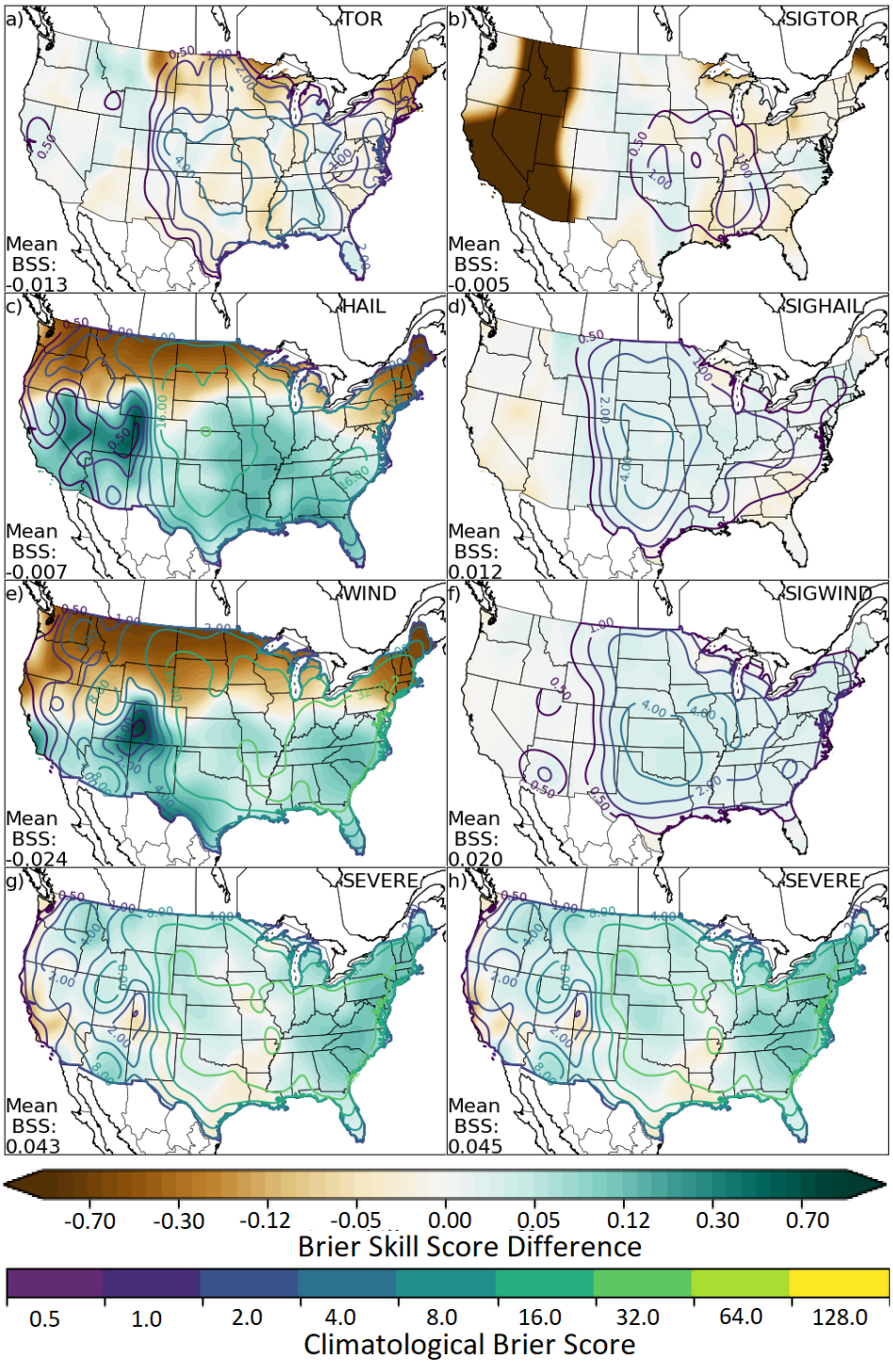


FIG. 11. Same as Figure 10, except depicts the difference in BSS between ML outlooks and the analogous outlooks issued by SPC. Greens indicate ML forecasts outperform SPC; browns suggest the opposite. Due to data availability, a slightly shorter 13 September 2012–31 December 2016 period is used for the Day 2 and 3 outlook verification comparison.

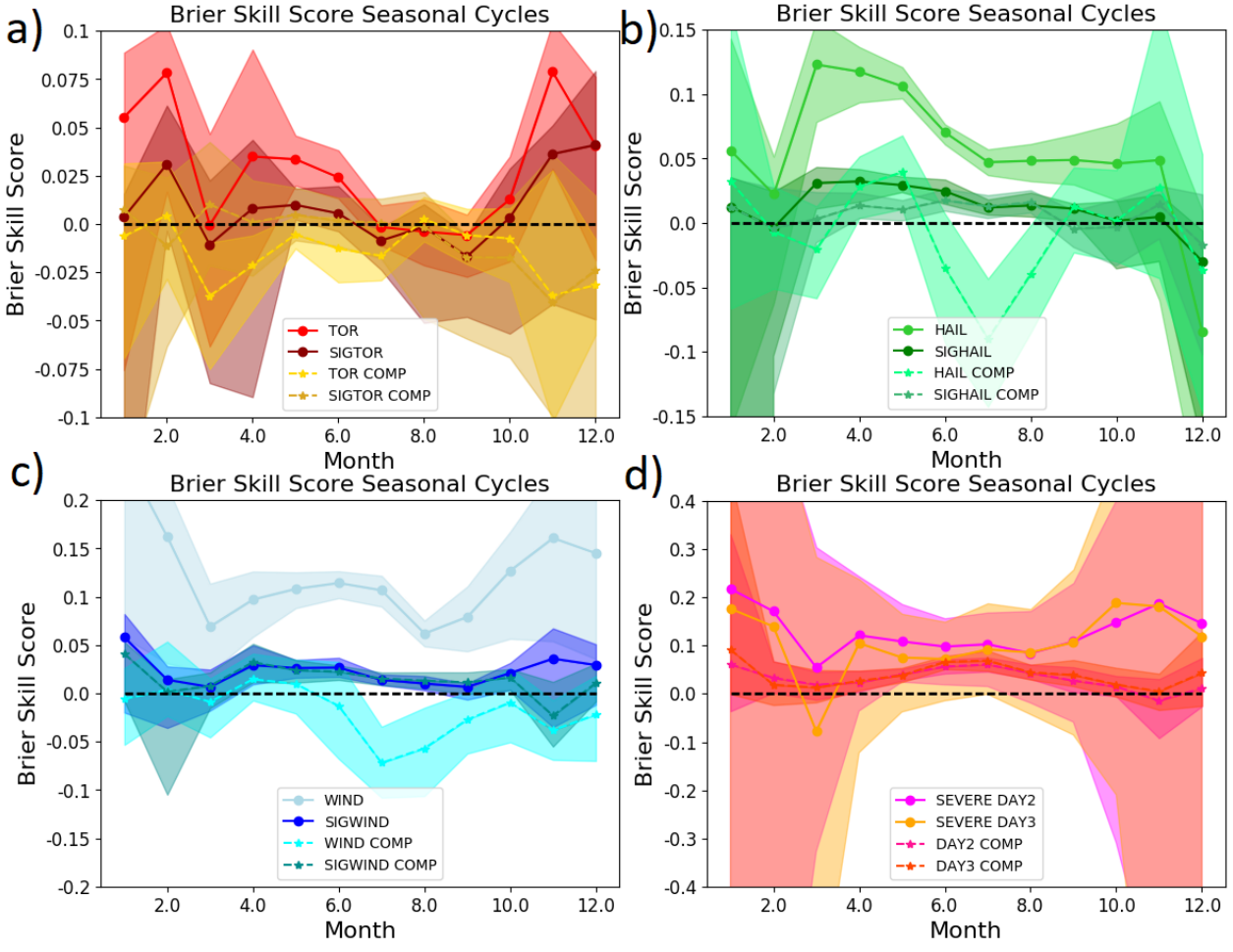


FIG. 12. BSSs by month and comparison between ML and SPC outlooks for (a) tornado and significant tornado, (b) hail and significant hail, (c) wind and significant wind, and (d) Day 2 and 3 outlooks. Lines are colored as indicated in the panel legend; shading about the line indicates 95% confidence bounds obtained by bootstrapping. Differences are ML-SPC, positive numbers indicating ML outperforms SPC. Note that the y-axis varies between panels.

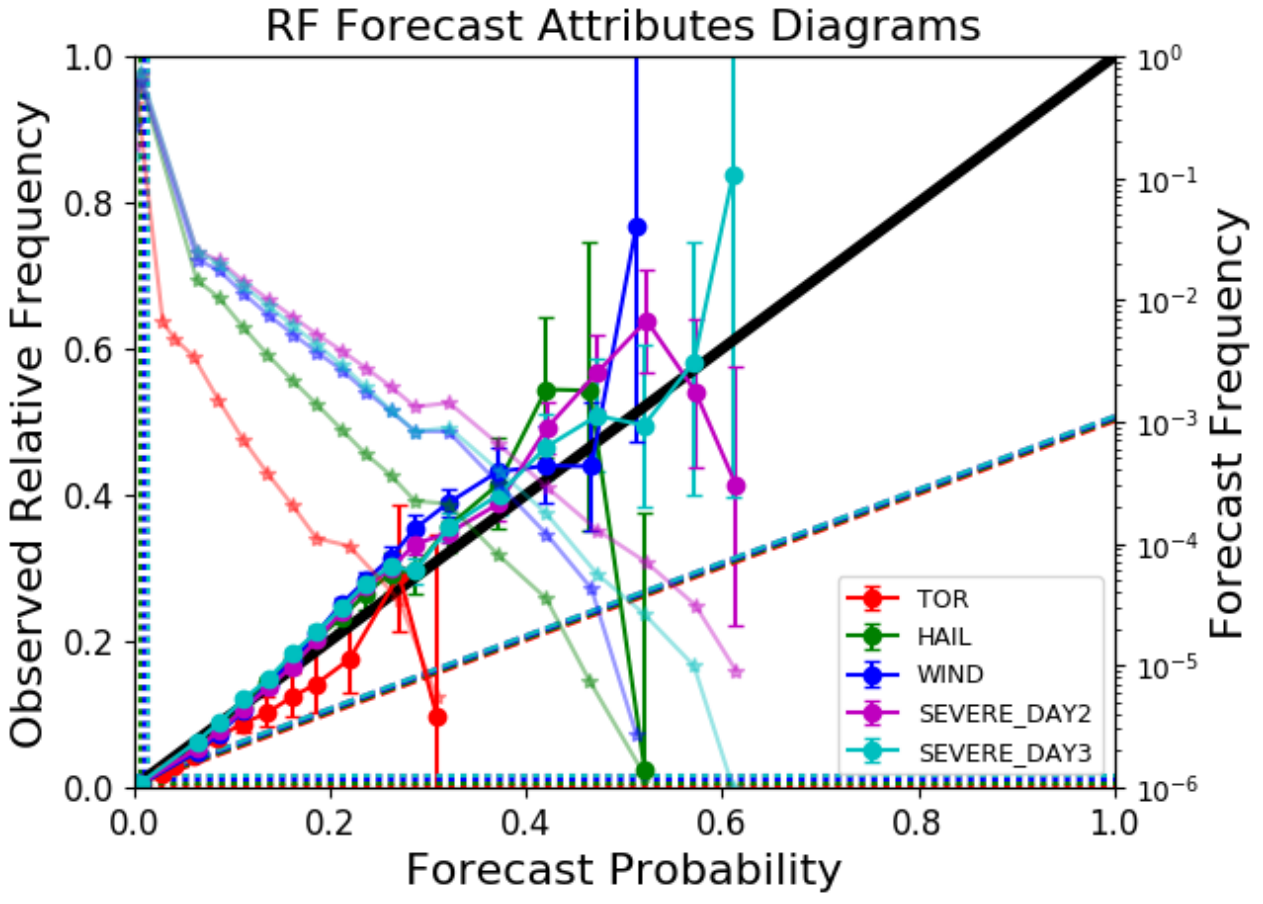


FIG. 13. Attributes diagrams for ML-based outlooks. Colored opaque lines with circular points indicate observed relative frequency as a function of forecast probability; the solid black line is the one-to-one line, indicating perfect reliability. Colors correspond to different severe predictands and lead times as indicated in the panel legend. Semi-transparent lines indicate the total proportion of forecasts falling in each forecast probability bin, using the logarithmic scale on the right 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. 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), where non-overlapping neighborhoods are assumed to be independent.

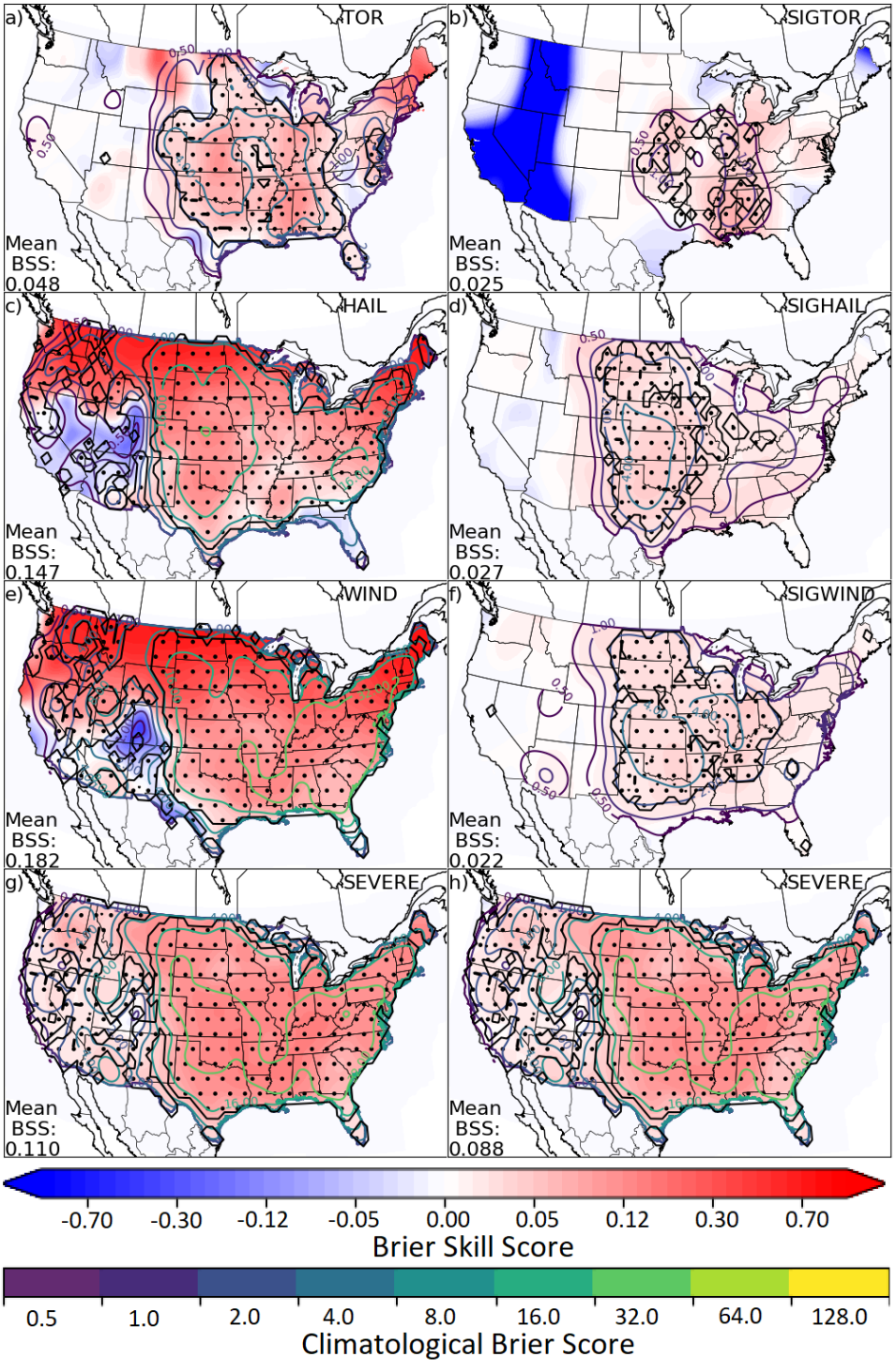


FIG. 14. Same as Figure 10, except for the weighted blend of SPC and ML outlooks.

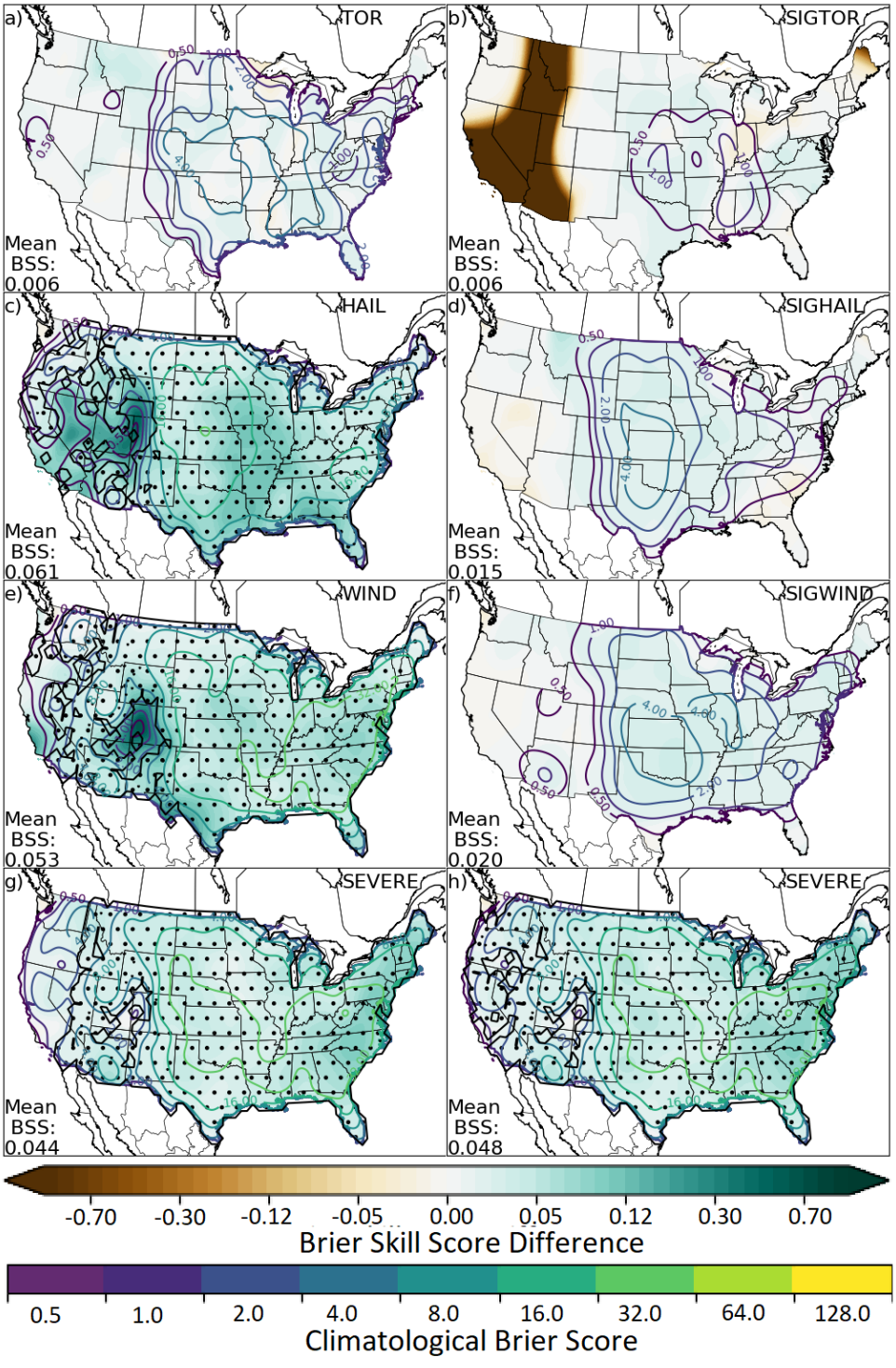


FIG. 15. Same as Figure 11, except for the weighted blend of SPC and ML outlooks.

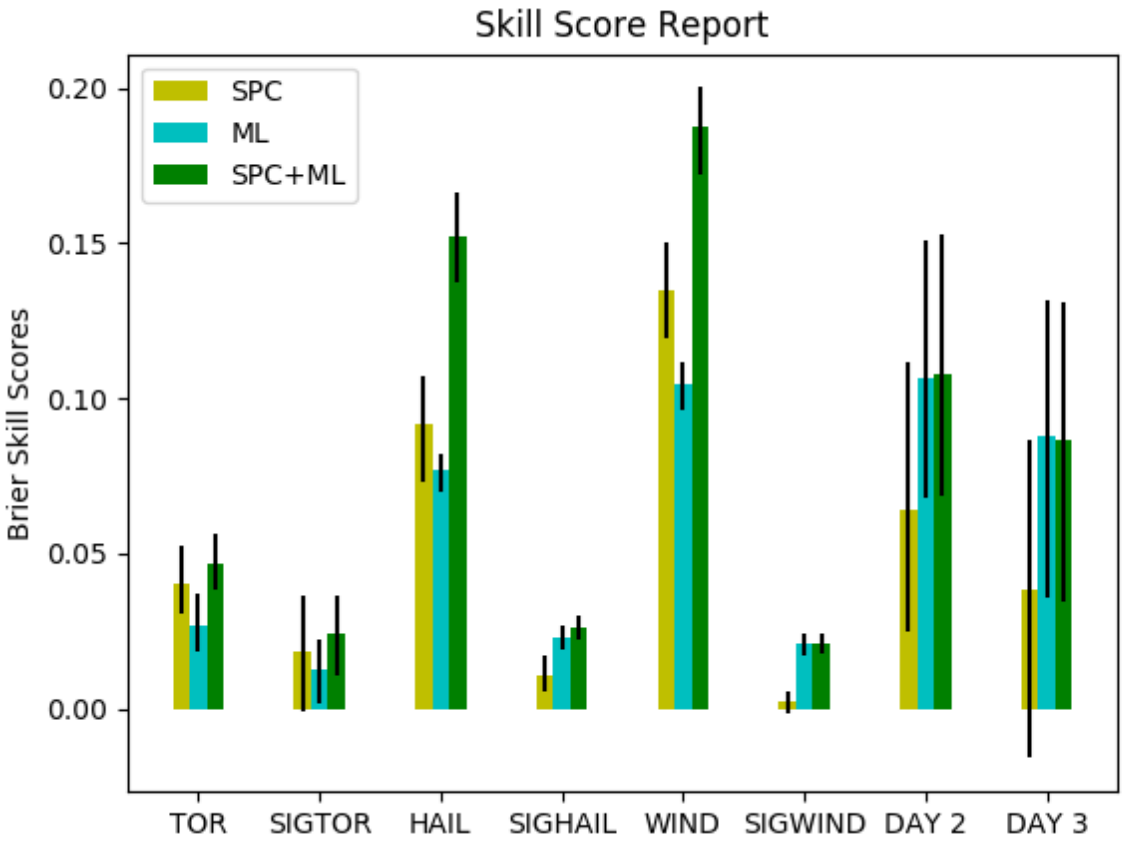
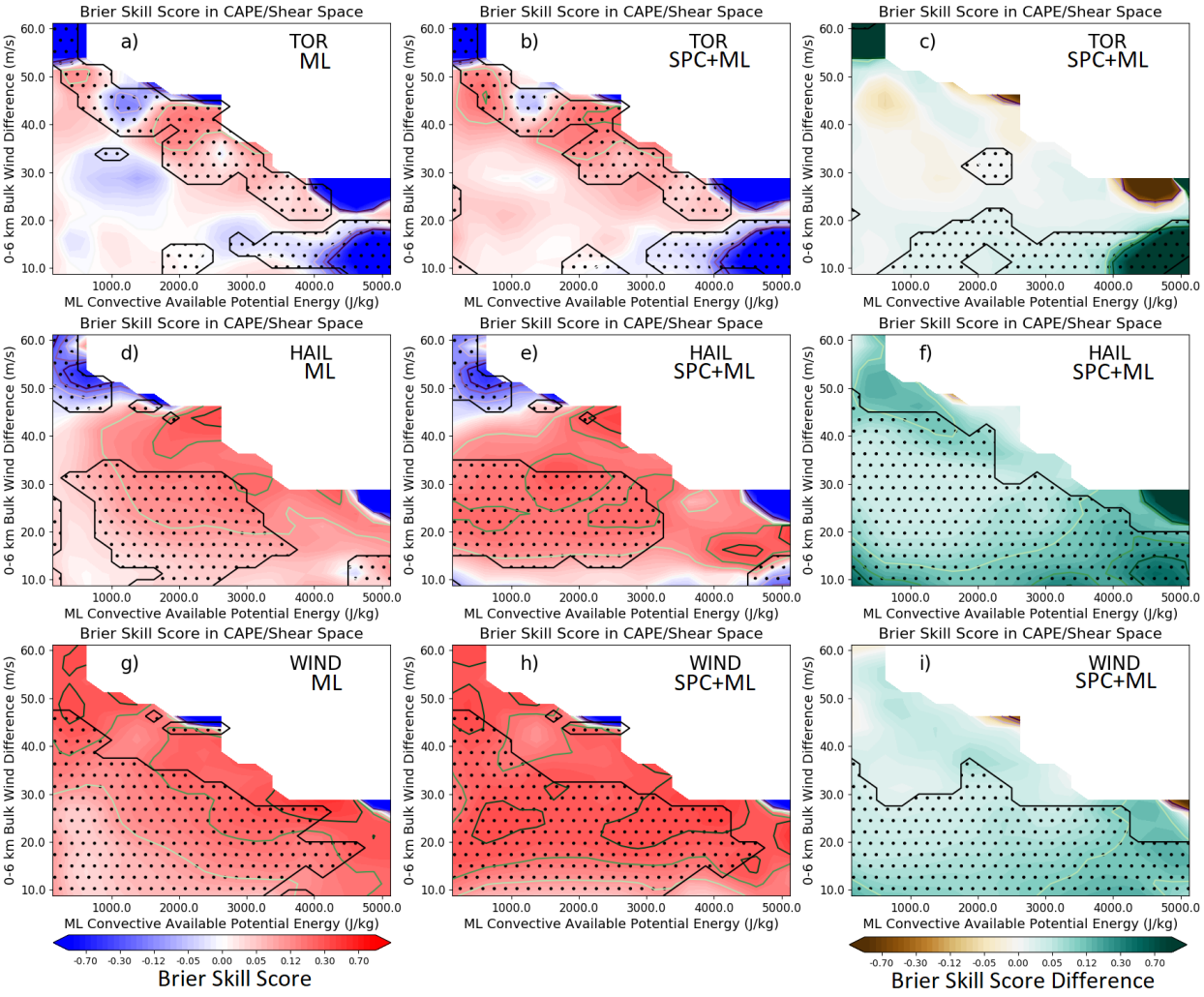


FIG. 16. CONUS-total BSS for each of the eight verified predictands for the SPC outlooks (yellow bars), ML forecasts (blue bars), and weighted average of the two (green bars). Error bars indicate 95% BSS confidence bounds obtained via bootstrapping.



1091 FIG. 17. BSS evaluation broken by CAPE versus shear parameter space for tornado, hail, and wind outlooks
1092 in panels (a)–(c), (d)–(f), and (g)–(i) as partitioned in Herman et al. (2018) and described in the manuscript text.
1093 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
1094 quantitative clarity. The left column depicts verification of the ML forecasts, the center column to the evaluation
1095 of the weighted blend of SPC and ML outlooks, and the right column presents the skill score difference between
1096 the blend and the raw interpolated SPC outlooks, with greens indicating an improvement over the SPC outlooks
1097 and browns representing loss of skill. Stippling indicates regions where the sign of the BSS or BSS difference
1098 is statistically significant with α=0.05 based on bootstrap resampling.

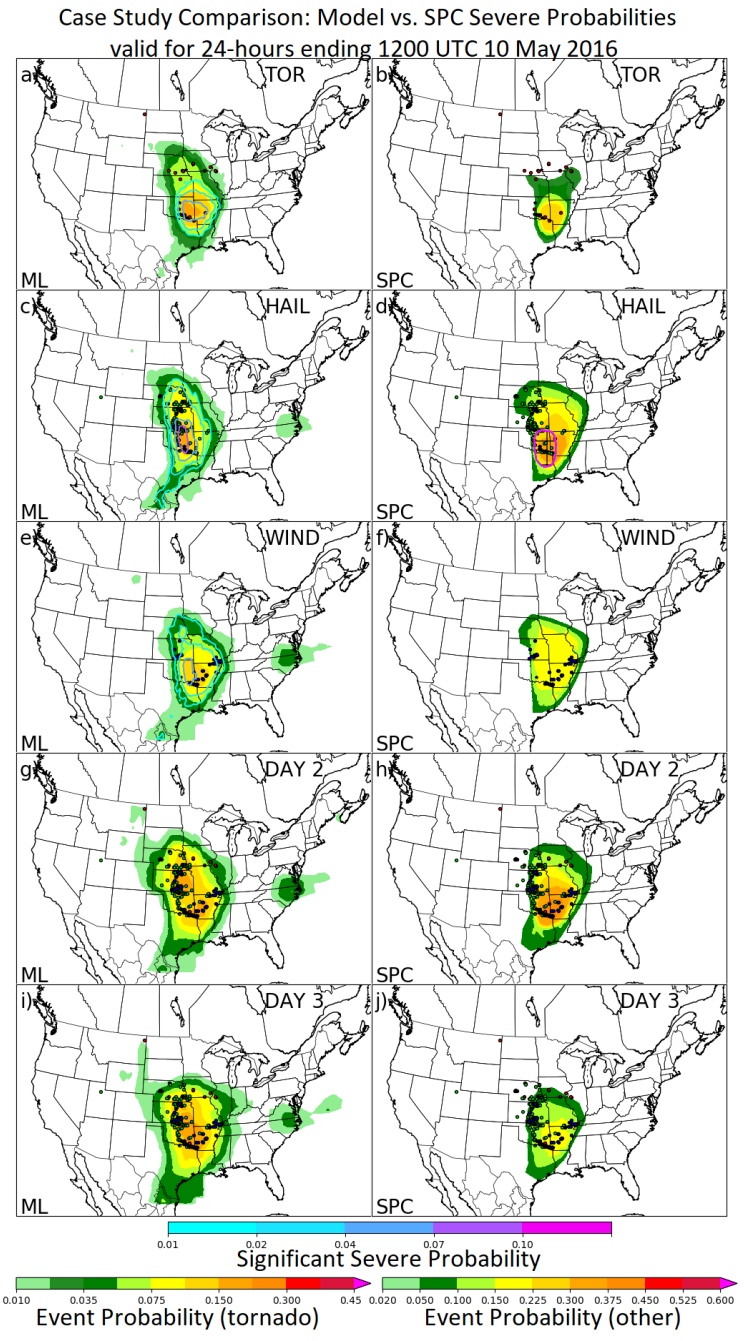


FIG. 18. Outlooks from the ML models and interpolated SPC contours valid for the 24-hour period ending 1200 UTC 10 May 2016 in the left and right columns, respectively. Filled contours depict severe probabilities as indicated by the corresponding colorbar on figure bottom; unfilled contours indicate significant severe probabilities for the corresponding phenomenon as applicable. Panels (a)–(b), (c)–(d), and (e)–(f) depict respectively Day 1 tornado, hail, and wind outlooks, while panels (g)–(h) and (i)–(j) show Day 2 and Day 3 outlooks issued previously for the same valid period. 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.