Comparing and Interpreting Differently Designed Random Forests for Next-Day Severe Weather Hazard Prediction

ERIC D. LOKEN,[a,b,c] ADAM J. CLARK,[c,b] AND AMY MCGOVERN[c]

[a] *Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma*
[b] *NOAA/OAR/National Severe Storms Laboratory, Norman, Oklahoma*
[c] *School of Meteorology, University of Oklahoma, Norman, Oklahoma*

(Manuscript received 19 August 2021, in final form 18 March 2022)

ABSTRACT: Recent research has shown that random forests (RFs) can create skillful probabilistic severe weather hazard forecasts from numerical weather prediction (NWP) ensemble data. However, it remains unclear how RFs use NWP data and how predictors should be generated from NWP ensembles. This paper compares two methods for creating RFs for next-day severe weather prediction using simulated forecast data from the convection-allowing High-Resolution Ensemble Forecast System, version 2.1 (HREFv2.1). The first method uses predictors from individual ensemble members (IM) at the point of prediction, while the second uses ensemble mean (EM) predictors at multiple spatial points. IM and EM RFs are trained with all predictors as well as predictor subsets, and the Python module tree interpreter (TI) is used to assess RF variable importance and the relationships learned by the RFs. Results show that EM RFs have better objective skill compared to similarly configured IM RFs for all hazards, presumably because EM predictors contain less noise. In both IM and EM RFs, storm variables are found to be most important, followed by index and environment variables. Interestingly, RFs created from storm *and* index variables tend to produce forecasts with greater or equal skill than those from the all-predictor RFs. TI analysis shows that the RFs emphasize different predictors for different hazards in a way that makes physical sense. Further, TI shows that RFs create calibrated hazard probabilities based on complex, multivariate relationships that go well beyond thresholding 2–5-km updraft helicity.

KEYWORDS: Ensembles; Forecasting; Artificial intelligence; Decision trees; Machine learning; Model interpretation and visualization

1. Introduction

Random forests (RFs; Breiman 2001) are appealing for numerical weather prediction (NWP) postprocessing because they can handle raw (i.e., nonnormalized) predictors, tend to produce reliable probabilistic predictions (Breiman 2001), are computationally efficient, and require little tuning compared to other machine learning (ML) methods. Moreover, RFs have recently demonstrated substantial skill in postprocessing precipitation (e.g., Gagne et al. 2014; Herman and Schumacher 2018; Loken et al. 2019) and severe weather (e.g., Gagne et al. 2017; Burke et al. 2020; Loken et al. 2020; Hill et al. 2020) forecasts from NWP ensembles. Indeed, Hill et al. (2020) found their 2- and 3-day lead-time RF-based severe weather forecasts had higher Brier skill scores (BSSs) than corresponding Storm Prediction Center (SPC) forecasts, while Loken et al. (2020) found that RFs using convection-allowing ensemble (CAE) predictors frequently produced more skillful day 1 hazard forecasts than those from the SPC. RFs have also performed well in testbed settings (e.g., Clark et al. 2021; Schumacher et al. 2021) and are now even considered in real-time operations (e.g., Schumacher et al. 2021).

Given the recent forecasting successes of RFs, it is natural and important to ask the following questions: How do RFs use simulated ensemble data to create skillful forecasts? What relationships does the RF learn between ensemble forecast variables and observed high-impact weather? Are current preprocessing techniques optimal? This study seeks to address these questions by comparing differently configured, severe-weather-predicting RFs and interpreting their output using the Python-based tree interpreter module (TI; Saabas 2016).

Uncovering the relationships learned by an ML algorithm is important because doing so can confirm the ML model is working as intended, build trust with product users, and provide new insights into underlying weather prediction tools (e.g., ensembles, satellites). There are multiple interpretability methods to help users better understand RF models. Single- (Breiman 2001) and multipass (Lakshmanan et al. 2015) permutation methods randomly permute the predictor data, assigning the greatest importance to predictors associated with the greatest drop in RF forecast skill after permutation. Forward (backward) feature selection (e.g., McGovern et al. 2019) involves retraining the RF after adding (removing) the variable that increases (decreases) RF skill the most (least). Partial-dependence plots (PDPs; Friedman 2001; Molnar 2019; McGovern et al. 2019) show how the prediction varies with a given predictor when all other predictors are held constant at their mean values. Impurity importance (e.g., Breiman 2001; Louppe et al. 2013; McGovern et al. 2019) measures how effectively each predictor sorts the training samples based on the target variable (e.g., the occurrence of an observed storm report), on average, after each split. Impurity importance is frequently quantified by computing the mean change in Gini or entropy score (e.g., McGovern et al. 2019) after each split, with greater scores indicating more effective sorting based on the target variable. These are all global methods because they explain RFs' overall behavior (e.g., Molnar 2019). Local methods, meanwhile,

Corresponding author: Eric D. Loken, eric.d.loken@noaa.gov

DOI: 10.1175/WAF-D-21-0138.1

explain specific predictions. Examples include individual conditional expectation plots (Goldstein et al. 2015), which are PDPs for individual predictions; locally interpretable model-agnostic explanation (LIME; Ribeiro et al. 2016), which uses an easily interpretable model (e.g., linear regression) trained on perturbed predictor data to explain a model's local predictions; Shapely values (Shapley 1953; Molnar 2019; Lundberg et al. 2019), which use game theory to determine each predictor's fair contributions to the final prediction; and TI, a method that combines features of impurity importance and Shapley values.

In this paper, the most important predictors for RF severe weather hazard forecasts are determined using TI and by evaluating RFs trained on different subsets of predictors, since these methods are computationally feasible and easy to understand. TI is also used to identify the learned relationships between RF predictors and observed severe weather. Two ways of obtaining predictors from CAE variables are compared: using individual-member CAE variables at the point of prediction (to potentially learn relationships from individual ensemble members) and using ensemble-mean variables at multiple spatial points (to potentially learn spatial relationships). Through this analysis, this paper seeks to determine the best way to condense simulated ensemble data during preprocessing and understand how RFs leverage CAEs to create skillful severe weather hazard (i.e., severe hail, wind, and tornado) forecasts. A focus on these hazards is adopted herein because they are highly impactful to lives and property (NCEI 2021) yet extremely difficult to predict, owing to their small scale relative to typical CAE grid spacing. This study also intends to build on the work of Loken et al. (2020) by investigating how RFs can attain such strong performance for next-day severe weather hazard prediction.

The remainder of the paper is organized as follows: section 2 describes the methods and datasets, section 3 presents the results, section 4 analyzes a representative case study forecast, section 5 summarizes and discusses the results, and section 6 concludes the paper and offers suggestions for future work.

2. Methods

a. Datasets

The forecast and observational datasets contain 653 days from April 2018 to May 2020 (Table 1). As in Loken et al. (2020), the analysis domain covers the contiguous United States (CONUS), and verification is performed on a grid with approximately 80-km horizontal spacing (Fig. 1a) to match the verification scales used by the SPC (i.e., 40 km from a point). Next-day forecasts (lead times of 12–36 h, valid from 1200 to 1200 UTC) are analyzed.

As in Loken et al. (2020), observed local storm reports (LSRs) are used for training and verifying RF forecasts. Unfiltered LSRs from the SPC website (SPC 2021a) are used for wind, hail, and 2019/20 tornadoes, while 2018 tornado LSRs were obtained from the SPC Storm Events Database (SSED; SPC 2021c) because it provides a more accurate and complete summary of tornado events. The spatial distribution of hail, wind, and tornado LSRs over the full dataset is depicted in Figs. 1b–d.

TABLE 1. HREFv2.1 initialization dates.

Month	2018	2019	2020	Total
January	—	2–23, 25–31	—	29
February	—	1–28	—	28
March	—	1–31	—	31
April	5–30	1–15, 17–30	27, 29–30	58
May	1–16, 18–31	1–31	1–29	90
June	1–6, 9–30	1–30	—	58
July	1–10, 13–31	1–31	—	60
August	1–4, 7–31	1–31	—	60
September	1–15, 17–30	1–26, 28–30	—	58
October	1–31	1–31	—	62
November	1–5, 8–20, 22–30	1–30	—	57
December	1–31	1–31	—	62
Total	260	361	32	653

RF forecasts are trained based on predictors from the High-Resolution Ensemble Forecast System, version 2.1 (HREFv2.1; Jirak et al. 2018; Roberts et al. 2020), an operationalized version of the Storm-Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012, 2016). Like the SSEO, the HREFv2.1 is an assemblage of diverse, individually tuned convection-allowing models (CAMs). The SSEO (Jirak et al. 2016) and HREFv2.1 have demonstrated high degrees of skill for the prediction of severe convection, owing to their relatively large member diversity compared to other ensemble designs (Roberts et al. 2020). Indeed, the diversity, skill, and operational status of HREFv2.1 make it ideal for this study, which seeks to shed light on the optimal use of diverse convection-allowing ensembles for severe weather prediction. HREFv2.1 contains 10 members, which all use approximately 3-km horizontal grid spacing. Collectively, the members use two dynamic cores, four microphysics schemes, and three boundary layer parameterizations. Five of the members are initialized at 0000 UTC, while the other five members are initialized at 1200 UTC the previous day. Full ensemble specifications are given Table 2.

b. RF method overview

RFs are ensembles of decision trees (Breiman 1984), which work by recursively splitting a dataset based on the predictor and value that maximizes a dissimilarity metric (e.g., information gain) during training. Because individual decision trees are prone to overfitting (e.g., Gagne et al. 2014), RFs include multiple unique decision trees grown independently based on a random subset of the training data, with each node's "optimal split" determined from a random subset of predictors. After training, RFs can predict the probability of an unseen testing sample belonging to a certain class (e.g., being associated with an LSR) by running the sample through each tree in the forest. RF probabilities are the mean fraction of training samples associated with the given class at the relevant leaf node across all trees. As in Loken et al. (2019, 2020), RFs are created using random forest classifiers from the Python module Scikit-Learn (Pedregosa et al. 2011).

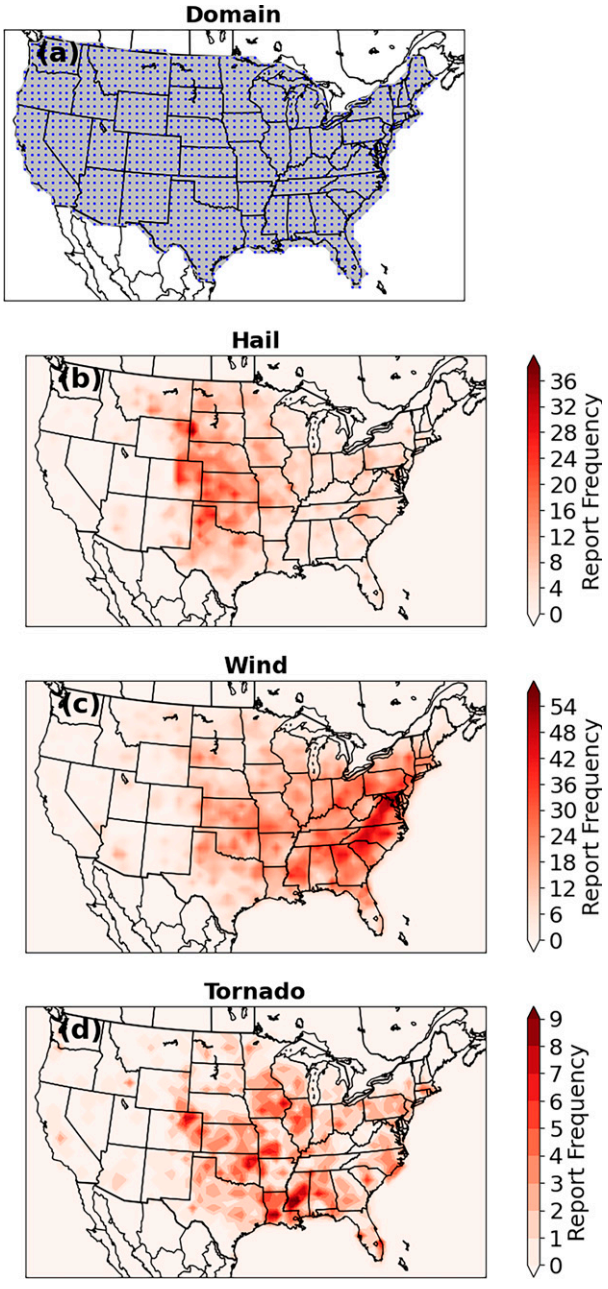

FIG. 1. (a) Verification domain (gray shading) and 80-km grid points (blue dots). (b) Distribution of severe hail reports in the observational dataset. (c),(d) As in (b), but for severe wind and tornado reports, respectively.

c. *RF interpretability and the tree interpreter module*

TI analyzes the path of a testing sample through each tree in a RF and records how each predictor impacts the training sample purity (i.e., the proportion of training samples associated with an LSR) at each node in the testing sample's path. Ultimately, TI sums each predictor's contribution over all nodes in each tree and reports the mean impact of each predictor over all trees in the RF. Like impurity importance, TI measures how well predictors split the training samples at each node; however, TI quantifies the impact of each split using the change in the training sample climatology instead of information gain or change in the Gini index. Moreover, TI is a local method and so only considers the splits made along the path (of each tree) taken by a testing sample. Thus, TI only measures variable importance as it applies to a specific set of testing data, unlike impurity importance, which evaluates how predictors influence all splits in the forest, regardless of whether those splits are utilized during testing.

A nice property of TI is that it decomposes the final RF probability into the sum of a bias term (i.e., the overall climatology of the training set) and the contribution from each predictor, a feature it shares with Shapley values. However, TI and Shapley values are computed differently. As discussed above, TI analyzes how each predictor along a testing path contributes to the final RF probability based on how it splits the training samples. In contrast, Shapley values aim to fairly divide the output probability among predictors based on their (weighted) marginal contributions to the final probability (Molnar 2019). These marginal contributions are calculated by computing how the RF probability differs, on average, when a given predictor is added to a set of other predictors (over all possible sets). Because of their different designs, TI and Shapley values have different characteristics, strengths, and weaknesses. For example, TI tends to assign more credit to predictors deeper in the trees, while Shapley values do not (Lundberg et al. 2019). Additionally, unlike TI, Shapley values are designed to satisfy multiple nice properties, including efficiency (i.e., Shapley values from each predictor sum to the final prediction), symmetry (i.e., Shapley values are the same for two predictors with the same marginal contributions for all possible predictor sets), dummy (i.e., Shapley values of 0 are assigned to predictors that do not change the final prediction when added to any set), and additivity (i.e., Shapley values are additive in situations where the final output is additive; Molnar 2019). However, Shapley values are also expensive to compute and must be approximated in practice. Moreover, when predictors are correlated, Shapley value computations can include unrealistic predictor data, since Shapley values involve computing RF probabilities from sets with some predictors "excluded" (i.e., replaced by predictor data from randomly chosen samples, which might not be physical; Molnar 2019). TI does not have this problem because it only analyzes the single set of predictors along each tree's testing path. Here, for simplicity, TI is used to examine each predictor's mean contribution to each forecast probability, domain- and dataset-wide.

Predictors are analyzed singly as well as in groups of similar variables (e.g., all storm-related variables). TI shows how much, on average, each predictor (set) influences RF probabilities positively, negatively, and overall. Greater overall impact on RF probabilities implies greater "importance" of the given predictor to the RF. TI probability contributions are also stratified based on the observed class to determine whether (and how much) predictors appropriately increase or decrease probabilities. TI contributions are plotted against the value of a given predictor for every testing sample in the

TABLE 2. HREFv2.1 specifications. Dynamic cores are from the Advanced Research version of the Weather Research and Forecasting Model (WRF-ARW; Skamarock et al. 2008) and the Nonhydrostatic Multiscale Model on the B grid (NMMB; Janjić and Gall 2012). Initial and lateral boundary conditions (ICs/LBCs) are from the North American Mesoscale Model (NAM; Janjić 2003), operational Rapid Refresh (RAP; Benjamin et al. 2016), and the National Centers for Environmental Prediction's Global Forecast System (GFS; Environmental Modeling Center 2003). Microphysics parameterizations include the Thompson (Thompson et al. 2008), WRF single-moment 6-class (WSM6; Hong and Lim 2006), Ferrier et al. (2002), and Ferrier–Aligo (Aligo et al. 2018) schemes. Planetary boundary layer (PBL) parameterizations include the Mellor–Yamada–Nakanishi–Niino (MYNN; Nakanishi and Niino 2004), Mellor–Yamada–Janjić (MYJ; Janjić 2002), and Yonsei University (YSU; Hong et al. 2006) schemes. HRW refers to the High-Resolution Window model run. Note that the HREFv2.1 used herein differs slightly from that described in Roberts et al. (2020) in that the time-lagged HRRR member is initialized at 1200 UTC instead of 1800 UTC (i.e., a 12- instead of 6-h time lag).

Member	Dynamic core	ICs/LBCs	Microphysics	PBL	Initialization time
HRRR	WRF-ARW	RAP/RAP	Thompson	MYNN	0000 UTC
HRRR-12	WRF-ARW	RAP/RAP	Thompson	MYNN	1200 UTC
HRW ARW	WRF-ARW	RAP/GFS	WSM6	YSU	0000 UTC
HRW ARW-12	WRF-ARW	RAP/GFS	WSM6	YSU	1200 UTC
HRW NMMB	NMMB	RAP/GFS	Ferrier	MYJ	0000 UTC
HRW NMMB-12	NMMB	RAP/GFS	Ferrier	MYJ	1200 UTC
HRW NSSL	WRF-ARW	NAM/NAM	WSM6	MYJ	0000 UTC
HRW NSSL-12	WRF-ARW	NAM/NAM	WSM6	MYJ	1200 UTC
NAM NEST	NMMB	NAM/NAM	Ferrier–Aligo	MYJ	0000 UTC
NAM NEST-12	NMMB	NAM/NAM	Ferrier–Aligo	MYJ	1200 UTC

dataset to show how different values of a predictor influence RF probabilities.

TI also contains a function that assesses how multiple predictors interact to influence RF probabilities. This function ascribes the change in training data purity at a given node to the combination of predictors *at and above* the given node in the testing path, which results in more accurate contribution values. However, this process of assessing multivariate contributions is very computationally expensive. Therefore, the function is only run on the testing data associated with an LSR. For each severe weather hazard, a scatterplot shows the probability contribution from the top two-variable combinations for each sample in the testing dataset.

d. *Creating RF forecasts*

1) PREDICTOR FIELDS

Here, the RFs consider 32 input fields from the HREFv2.1 as well as latitude and longitude. Each field is categorized as a storm, environment, index, or latitude/longitude variable (Table 3). Nineteen of these represent derived fields (denoted

TABLE 3. Predictor fields. The temporal aggregation strategy for each variable is noted in parentheses. An asterisk denotes a derived quantity.

Simulated storm	Simulated environment		Simulated index	Lat/lon
1-km reflectivity (24-h max)	0–3-km storm relative helicity (24-h max)	MUCAPE (24-h mean)	Supercell composite parameter* (24-h max)	Latitude
Echo top (24-h max)	0–1-km storm relative helicity (24-h max)	MUCIN (24-h mean)	Significant tornado parameter* (24-h max)	Longitude
Upward vertical velocity (24-h max)	2-m temperature (24-h mean)	SB/MUCAPE ratio* (24-h mean)	Significant hail parameter* (24-h max)	—
Downward vertical velocity (24-h min)	2-m dewpoint temperature (24-h mean)	700–500-hPa lapse rate* (24-h mean)	0–1-km energy helicity index* (24-h max)	—
2–5-km updraft helicity (24-h max)	2-m and 925-, 850-, 700-, and 500-hPa dewpoint depression* (24-h mean)	Critical angle proxy* (at time of max STP)	0–3-km energy helicity index* (24-h max)	—
0–3-km updraft helicity (24-h max)	10-m–500-hPa wind shear magnitude* (24-h mean)	Max 10-m wind speed (24-h max)	Product of (MUCAPE) × (10-m–500-hPa wind shear magnitude)* (24-h max)	—
Number of grid points with at least 30 dBZ simulated reflectivity [at time of max 2–5-km updraft helicity (if nonzero) or upward vertical velocity]	10-m–925-hPa wind shear magnitude* (24-h mean)	10-m wind direction (at time of maximum 10-m wind speed)	Lifted index (24-h min)	—

by an asterisk in Table 3). The most complex derived variables are described in the appendix.

2) DATA PREPROCESSING

Preprocessing is required to reduce the dimensionality of the dataset to make ML computationally feasible. The general method of preprocessing is similar to that described in Loken et al. (2020). First, simulated HREFv2.1 data are aggregated in time by computing a 24-h maximum, minimum, or mean, depending on the variable (Table 3). Next, all forecast variables are remapped to the approximately 80-km verification grid using the method described in Loken et al. (2020). Namely, for the variables using temporal maximum (minimum) aggregation, remapping is done by assigning each 80-km grid box the maximum (minimum) value from all the 3-km points falling inside of it. For the variables using temporal mean aggregation, remapping is done using a neighbor budget method (Accadia et al. 2003). Ultimately, RF probabilities are output and analyzed on this 80-km grid, as in Loken et al. (2020). Using an 80-km grid for 12–36-h lead time severe weather prediction and verification is appropriate because 1) predictability is unlikely to exist at smaller scales for those lead times, 2) local severe weather reports are likely underreported at the smallest spatial scales (e.g., if the native 3-km grid were used instead), and 3) an 80-km grid cell covers approximately the same area as the 40-km radius used by the SPC for verification—although the use of an upscaled grid makes verification much more computationally efficient. Moreover, the use of an 80-km grid reflects the verification procedure used by Loken et al. (2020), who showed that similar RF-based severe weather forecasts were skillful compared to SPC and updraft-helicity-based forecasts at 12–36-h lead times.

Two different methods are used to obtain RF predictors in the final step of preprocessing. Because HREFv2.1 is a highly diverse CAE with members designed to be skillful on their own, the first method involves using individual member fields at the point of prediction as predictors. The RFs trained in this way will be subsequently referred to as individual member (IM) RFs. The second method uses predictors from each field's ensemble mean. To keep the number of predictors approximately equal, the RFs trained using the second method consider predictors at the point of prediction *plus* the 8 nearest grid points. Therefore, the RFs trained in this way will be subsequently referred to as (3×3) ensemble mean (EM) RFs.

To help account for the spatial uncertainty in the placement of simulated storms in the IM RFs (which only consider simulated CAE data at a single grid point), all storm fields (except for n30dbz) are spatially smoothed using a two-dimensional isotropic Gaussian kernel density function:

$$\upsilon = \sum_{n=1}^{N} \frac{\upsilon_n}{2\pi\sigma^2} \exp\left[-\frac{1}{2}\left(\frac{d_n}{\sigma}\right)^2\right], \quad (1)$$

where υ is the spatially smoothed value at a given point, N is the number of points in the analysis domain, υ_n is the raw value at point n, d_n is the distance between the nth point and the given point, and σ is the standard deviation of the Gaussian kernel. Here, σ is always taken to be 120 km for simplicity. Unlike in Loken et al. (2020), this value is not optimally tuned for each field and hazard. Rather, 120 km is chosen based on past experience with 2–5-km updraft helicity (UH2–5km); it is thought to be large enough to enhance probability of detection (POD; i.e., correctly forecasting observed LSRs; e.g., Wilks 2011) but small enough to preserve some sharpness and resolution. Importantly, the smoothing is only done for the storm variables in the IM RF. The EM RF uses unsmoothed storm variables because it considers predictors at multiple spatial points. Although spatial smoothing biases the IM RFs' storm field predictors lower (by the nature of the smoothing), the smoothed fields retain a strong relationship with observed severe weather and eliminate the need for the constituent ensemble member to place storms perfectly.

Missing ensemble forecast data are also handled during preprocessing. Because the time-lagged HRRR member only extends to forecast hour 24 (as opposed to 36), it is excluded from the ensemble mean. For the IM RFs, the member is included but uses 12- rather than 24-h temporal aggregation (excluding the HRRR member from the IM RFs does not appreciably change the results presented herein). Additionally, the two NAM members do not forecast radar echo top (RETOP), 0–3-km UH (UH0–3km), critical angle, or significant tornado parameter (STP). Therefore, the IM RFs do not include NAM versions of these variables as predictors, and the EM RFs use an 8-member ensemble mean for these variables.

To help determine how storm, environment, and index variables influence RF skill, IM and EM RFs are trained using all available predictors as well as different subsets.

3) RF TRAINING

All RFs are trained using Scikit-Learn and use the same set of hyperparameters for simplicity: 200 trees, a maximum depth of 15, and 20 minimum samples per leaf node. These hyperparameters are selected based on previous experience with forecasting precipitation and severe weather. Sensitivity tests (not shown) indicate that the above combination of hyperparameters frequently produces highly skilled forecasts. With one exception (when trees are restricted to a depth of 5 or less), altering the hyperparameters within reasonable ranges (i.e., from 50 to 500 trees, from 5 to 50 maximum depth, and from 1 to 50 minimum samples per leaf node) does not appreciably impact RF skill. Moreover, in all cases examined, varying the hyperparameters impacts EM and IM RF skill similarly; thus, small differences in hyperparameters are not expected to appreciably change the results presented herein.

As in Loken et al. (2019, 2020), k-fold cross validation is used to train and verify the RF forecasts, since this approach allows RF forecasts to be trained using a (comparatively) large training dataset. Here, 16 folds are used: the first 13 folds contain 41 days each, and the final 3 folds each contain

40 days. Sixteen folds are chosen as a compromise between computational expense and training dataset length. Sensitivity tests (not shown) suggest that using more folds slightly increases RF skill (due to the usage of a larger training set) but at the cost of greater computational expense. However, varying the number of folds from 4 to 64 impacts the verification metrics of the EM and IM RF forecasts similarly, so the results herein are likely not overly sensitive to the number of folds.

Forecasts are verified on the pooled testing data from each of the 16 folds, which enables verification to be done on the full 653-day dataset. Importantly, TI analysis for a given day is done using the RF from the appropriate fold. Thus, the TI results are aggregated from multiple (but appropriate) RFs.

e. Verification

RF forecasts are evaluated using area under the relative operating characteristics curve (AUC; e.g., Wilks 2011), BSS (e.g., Wilks 2011), performance diagrams (Roebber 2009), and attributes diagrams (Hsu and Murphy 1986).

AUC measures the ability of a forecast system to discriminate between yes events (e.g., the occurrence of severe hail) and no events (e.g., no occurrence of severe hail). Since AUC depends on probability of false detection, it is sensitive to the number of correct nulls. Thus, for severe weather, AUCs above 0.9 are not uncommon (Loken et al. 2020). Here, AUC is computed using the "roc_auc_score" function in Scikit-Learn, which uses the trapezoidal approximation.

Another metric that assesses forecast quality is the Brier score (BS; e.g., Wilks 2011), which measures the magnitude of forecast probability errors. BS is negatively oriented, so 0 (1) is the best (worst) possible score. As with AUC, trivial correct nulls can artificially improve the BS. To account for this effect, the BSS (e.g., Wilks 2011) is used herein. Essentially, the BSS compares the BS of a given forecast to that of a reference forecast. As in Loken et al. (2020), the reference here is a constant forecast of (domain-wide) observed climatological frequency for the given severe weather hazard during the 653-day dataset. Unlike the BS, the BSS is positively oriented. BSSs of 1 (below 0) indicate perfect (negative) skill. This paper plots BSS against AUC to efficiently show both metrics on a single graphic. Points closer to the upper right-hand corner of this plot indicate more skillful forecasts.

Performance diagrams (Roebber 2009) plot POD against success ratio (SR) and additionally display lines of constant bias and critical success index (CSI). These four metrics are all optimized at a value of 1; therefore, more skillful forecasts appear closer to the upper right-hand corner of the diagram. Here, performance diagrams are created by binarizing each set of forecasts at the following probability levels: 0%, 1%, 2%, 5%–15%, …, 85%–95%, 95%–100%.

Finally, attributes diagrams are used to measure reliability—or how well a forecast system's probabilities correspond with observed event relative frequencies. Perfectly reliable forecasts fall along the 1:1 diagonal line on the attributes diagram. Forecasts that contribute positively (negatively) to the BSS fall above (below) the no-skill line, and forecasts that have no resolution are along the horizontal climatology line. Attributes diagrams are created by binning the forecasts using the same forecast probability levels used to create the performance diagrams.

3. Results

a. RF verification

Performance diagrams (Figs. 2a–c) show that the EM RFs have the same or greater CSI compared to the IM RFs at all probability levels tested for all three hazards. Differences are largest for severe wind (Fig. 2b) and at probability values above 15% (Figs. 2a–c).

EM RFs trained with different predictor subsets achieve different levels of skill. For all three hazards, environment-only RFs are clearly inferior (Figs. 2d–f). Index-only RFs clearly outperform environment-only RFs for hazards with a relevant index predictor (i.e., hail and tornadoes; Figs. 2d,f), while storm-only RFs perform nearly as well as all-predictor RFs for all hazards. RFs that use index- *and* storm-related predictors (i.e., the no-environment RFs) have greater CSI than the storm-only RFs for tornadoes at most probability levels (Fig. 2f) and similar CSI to the all-predictor RFs for hail and wind (Figs. 2d,e).

All forecasts have good reliability for all three hazards (Figs. 2g–i). The larger deviations from perfect reliability seen at the higher forecast probabilities are likely due to small sample size. Notably, these deviations happen at comparatively smaller probability levels for the environment-only RFs, owing to those RFs' reduced sharpness. The storm-only RFs tend to produce the sharpest forecasts for all three hazards.

AUC and BSS values from the differently configured RFs reflect some of the key findings above, namely, that EM RFs tend to be superior to corresponding IM RFs; environment-only RFs are generally inferior; and the no-environment, all-predictor, and storm-only RFs tend to be the top-performing configurations for all hazards (Figs. 2j–l). Interestingly, the no-environment RF for severe wind and tornadoes has the highest BSSs and either a better (Fig. 2j) or similar (Fig. 2l) AUC compared to the corresponding all-predictor RF. This suggests that at least modest benefits can be obtained by using both storm *and* index predictors.

b. Influence of predictors on RF probabilities

1) HAIL

For severe hail prediction, storm-related variables exert the most absolute impact on RF probabilities; this is true of both IM and EM RFs (Figs. 3a–d). Overall, the storm variables tend to appropriately decrease RF probabilities when no LSR is present (Figs. 3a,b) and increase probabilities when an LSR is present (Figs. 3c,d). Index variables exert comparatively less impact on RF probabilities but also tend to move

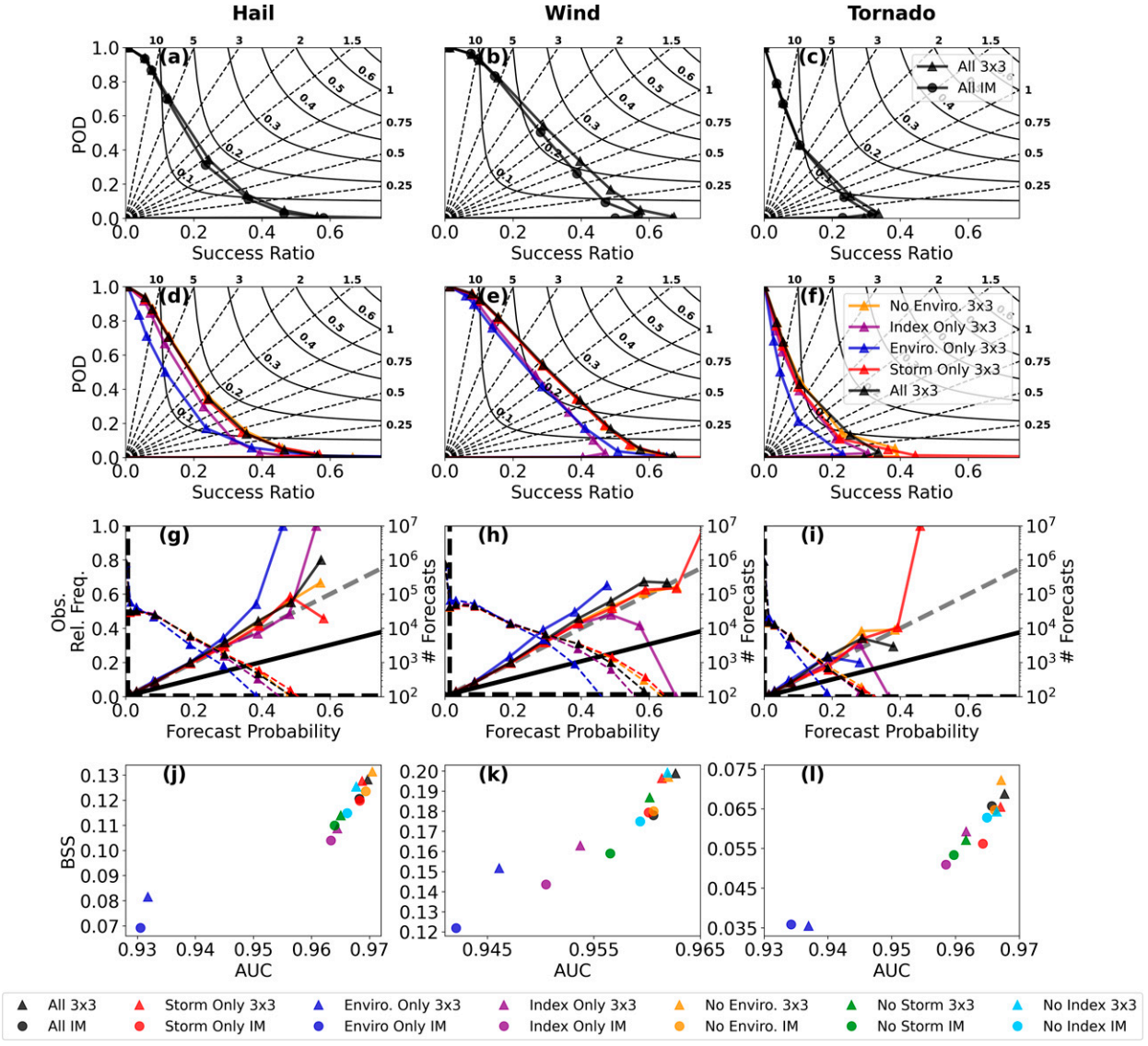

FIG. 2. (a) Performance diagram for all-predictor IM (filled circles) and EM (filled triangles) RFs for severe hail. (b),(c) As in (a), but for severe wind and tornadoes, respectively. Note that the x axis spans from 0 to 0.75. (d)–(f) As in (a)–(c), but for all-predictor (black triangles), storm-only (red triangles), environment-only (dark blue triangles), index-only (purple triangles), and no-environment (yellow triangles) EM RFs. (g) Attributes diagram for the EM RF forecasts listed in (d) for severe hail. The number of forecasts in each forecast probability bin are displayed with a dashed line of the appropriate color. Perfect reliability (dashed gray), no-skill (solid black), and horizontal and vertical climatology lines (dashed black) are also shown. Note that the x axis is truncated at 0.75. (h),(i) As in (g), but for severe wind and tornadoes, respectively. (j) BSS vs AUC plot for severe hail. IM RFs (filled circles) and EM RFs (filled triangles) are displayed. All-predictor (black), storm-only (red), environment-only (dark blue), index-only (purple), no-environment (yellow), no-storm (green), and no-index (light blue) RFs are shown.

probabilities in the appropriate direction. Environment variables impact RF probabilities slightly less (more) than index variables when no (an) LSR is present. However, environmental fields, on average, tend to increase the probabilities when no LSR is present (Figs. 3a,b) and increase probabilities less than the storm and index variables when an LSR is present (Figs. 3c,d). This result is interesting and counterintuitive, since one would expect the most skillful variables to also be the most important. One explanation is that more environmental than index variables exist and thus get selected more frequently as splitting criteria by the RF algorithm. Similar logic can help explain the difference in latitude and longitude contributions between the IM and EM RFs. In the IM RFs, latitude and longitude are each represented once, while in the EM RFs, latitude and longitude are represented nine times (once for each spatial grid point examined), making it more likely that they will be involved in a node's splitting criterion.

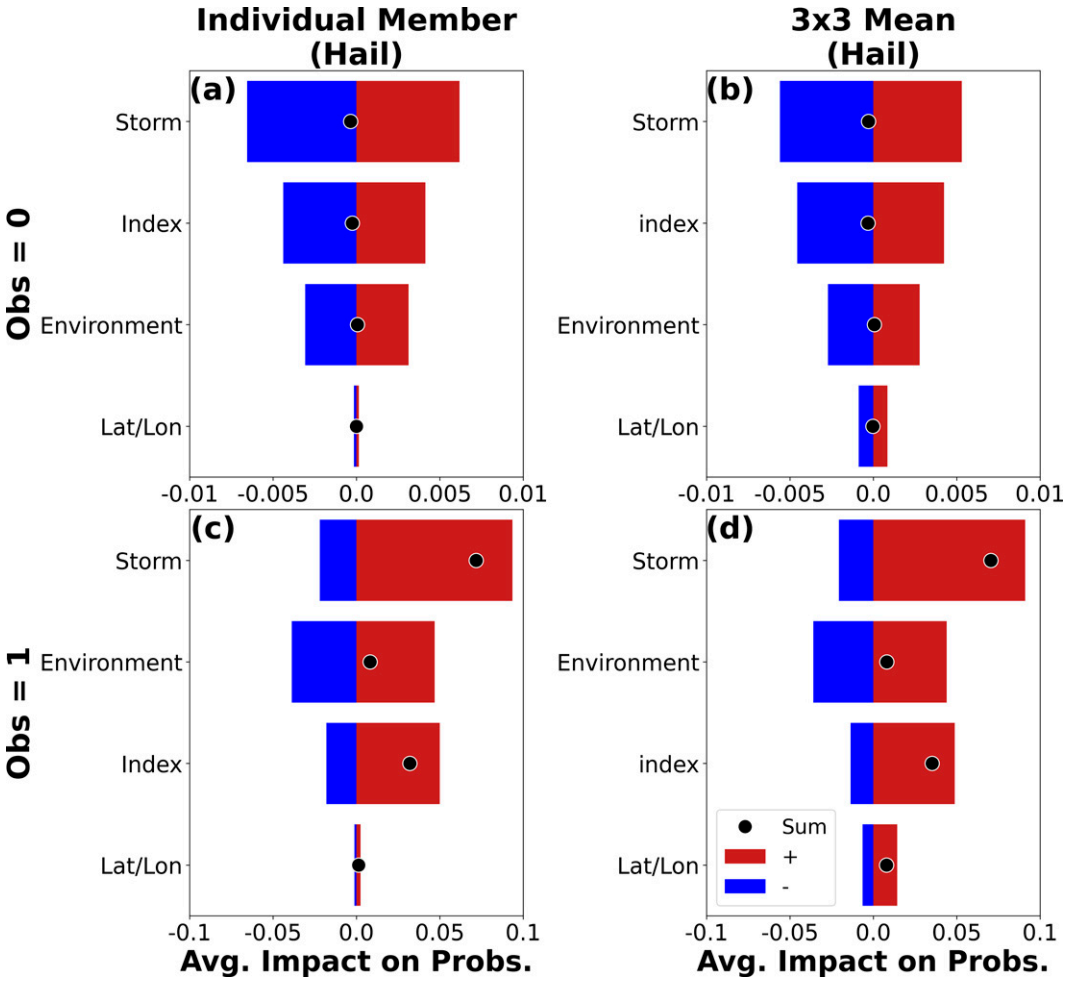

FIG. 3. (a) Mean TI negative (blue), positive (red), and summed (i.e., negative plus positive; black dot) RF probability contributions (per grid point) from storm, index, environment, and latitude/longitude variables in the all-predictor IM severe hail RF. Variable subsets are displayed in order of descending overall importance (i.e., the mean absolute value of contributions). Results are shown for cases not associated with an observed hail storm report. (b) As in (a), but for the all-predictor EM RF. (c),(d) As in (a) and (b), but for cases associated with an observed hail storm report. Note the different *x*-axis scale in (a) and (b) compared to (c) and (d).

For both the IM and EM hail-predicting RFs, the two "most important" predictor fields are UH2–5km and significant hail parameter (SHIP; Figs. 4a–d). Both variables tend to move the RF probabilities appropriately depending on the presence or absence of an LSR. This is a nice result, since UH2–5km has been used to predict severe hail by many previous studies (e.g., Jirak et al. 2014; Gagne et al. 2017; Burke et al. 2020; Loken et al. 2020), and SHIP is designed to indicate environments supportive of significant severe hail. Thus, the RFs emphasize variables that make physical sense.

Most of the environment variables rank low in terms of their relative importance (i.e., how much they influence the RF probabilities). This is somewhat surprising, given the expected relationship between observed hail and most unstable convective available potential energy (MUCAPE) or 700–500-hPa lapse rate. It is speculated that these predictors are relatively unimportant to the RFs because the information they provide is already contained more efficiently in the SHIP.

2) WIND

As with severe hail, storm-related variables exert the most influence on the severe wind probabilities for both types of RFs (Figs. 5a–d). In cases with (without) an LSR, the storm variables exert a greater mean increasing (decreasing) influence on the RF probabilities compared to the environment and index variables (Figs. 5a,b), indicating substantial skill.

In cases without an LSR, the environment and index variables exert a similar influence (Figs. 5a,b); this is relatively unsurprising given that no "wind-specific" index is used as a

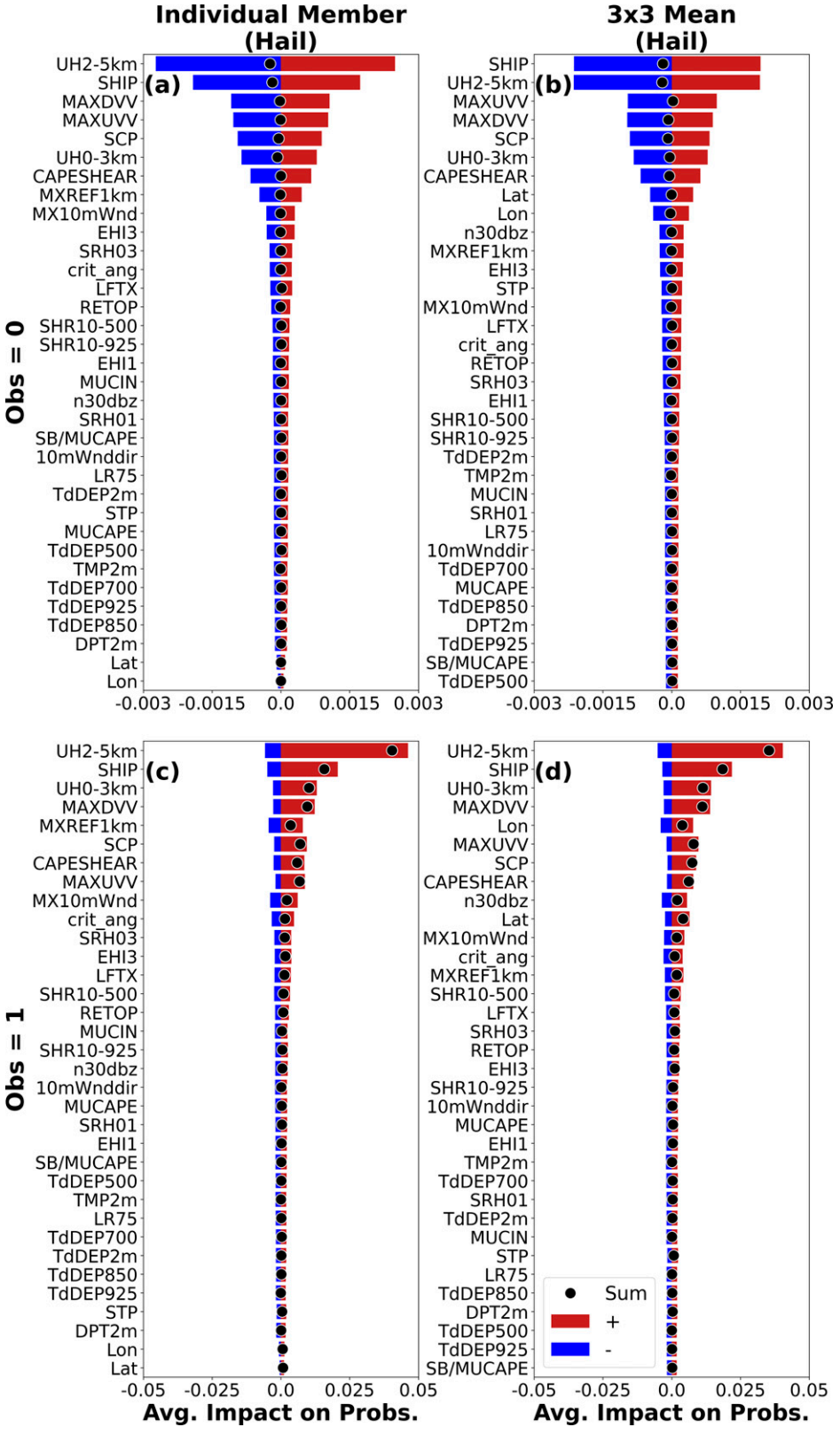

FIG. 4. As in Fig. 3, but for contributions aggregated over the individual predictor fields.

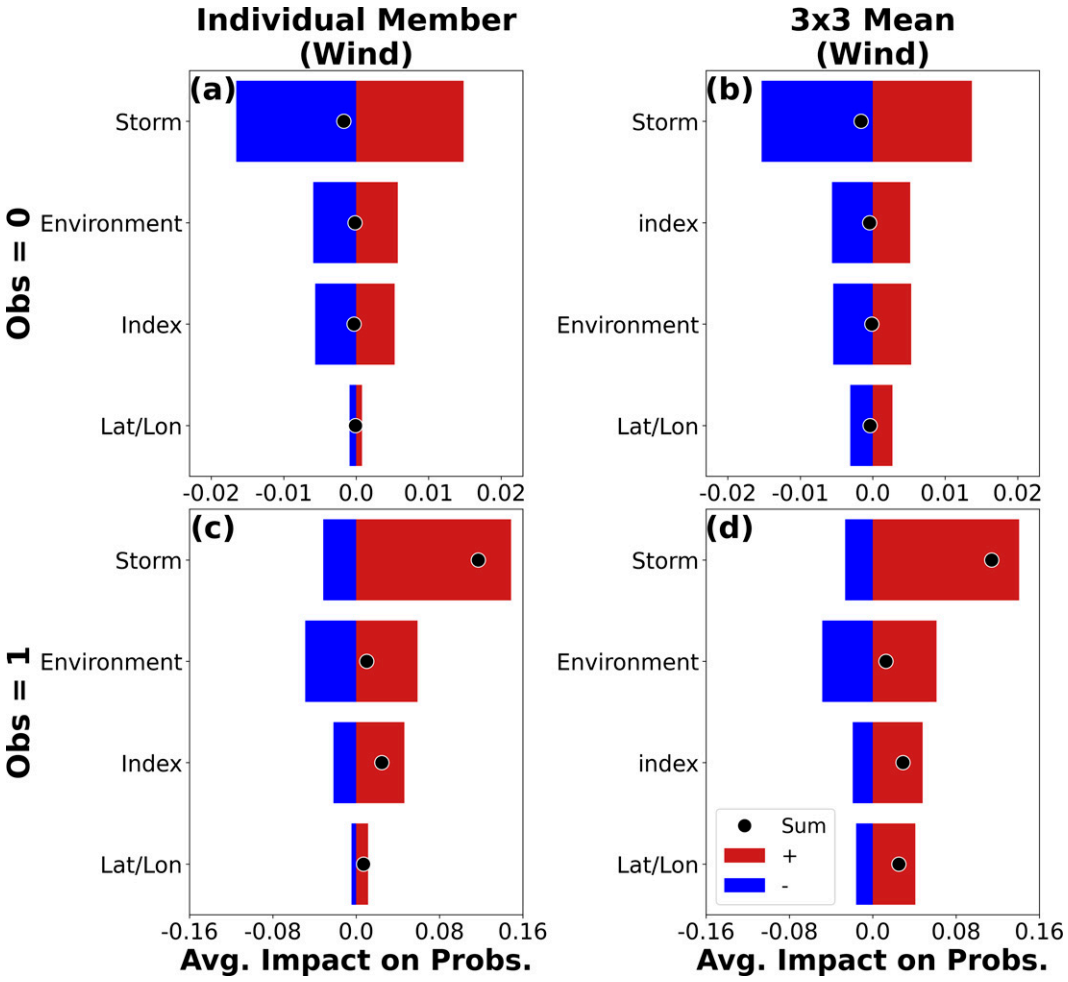

FIG. 5. As in Fig. 3, but for severe wind.

predictor in either RF. However, it is interesting that in cases with an LSR, the environment variables are more "important" but increase the RF probabilities less (on average) than the index variables (Figs. 5c,d). Again, it is possible that this effect is due to the presence of more environment variables (leading to greater TI "importance") and more direct relationships contained in index variables. IM and EM RFs use storm, environment, and index variables similarly, although the EM RFs place much more importance on latitude and longitude predictors compared to the IM RFs.

For both RF configurations, top wind predictor fields include UH2–5km, maximum downward vertical velocity (MAXDVV), maximum upward vertical velocity (MAXUVV), and UH0–3km (Figs. 6a–d). These variables all tend to move the probabilities in the correct direction in instances without (Figs. 6a,b) and with (Figs. 6c,d) an LSR. Latitude and longitude also rank as relatively important predictors for wind, especially in the EM RF. Thus, RFs appear to learn systematic spatial relationships for predicting severe wind. These relationships likely reflect some of the biases present in the severe wind report observation database (Edwards et al. 2018).

3) TORNADO

For tornadoes, storm, environment, and index variables have similar levels of TI importance, in both the IM and EM RFs (Figs. 7a–d). However, while environmental variables *move* the probabilities most in cases with an LSR (Figs. 7c,d), both storm and index variables correctly *increase* the probabilities more in those cases.

Regardless of whether there is an observed LSR, UH0–3km is the most important variable for tornado prediction in both the IM and EM (Figs. 8a–d). This is consistent with Sobash et al. (2019), who found UH0–3km performed better than UH2–5km for predicting tornadoes. Other important predictors include STP, UH2–5km, maximum 1-km simulated reflectivity (MXREF1km), and 0–1-km storm relative helicity (SRH). These fields make physical sense: large values of STP (e.g., Thompson et al. 2002, 2003; Parker 2014) and low-level SRH (e.g., Davies-Jones et al.

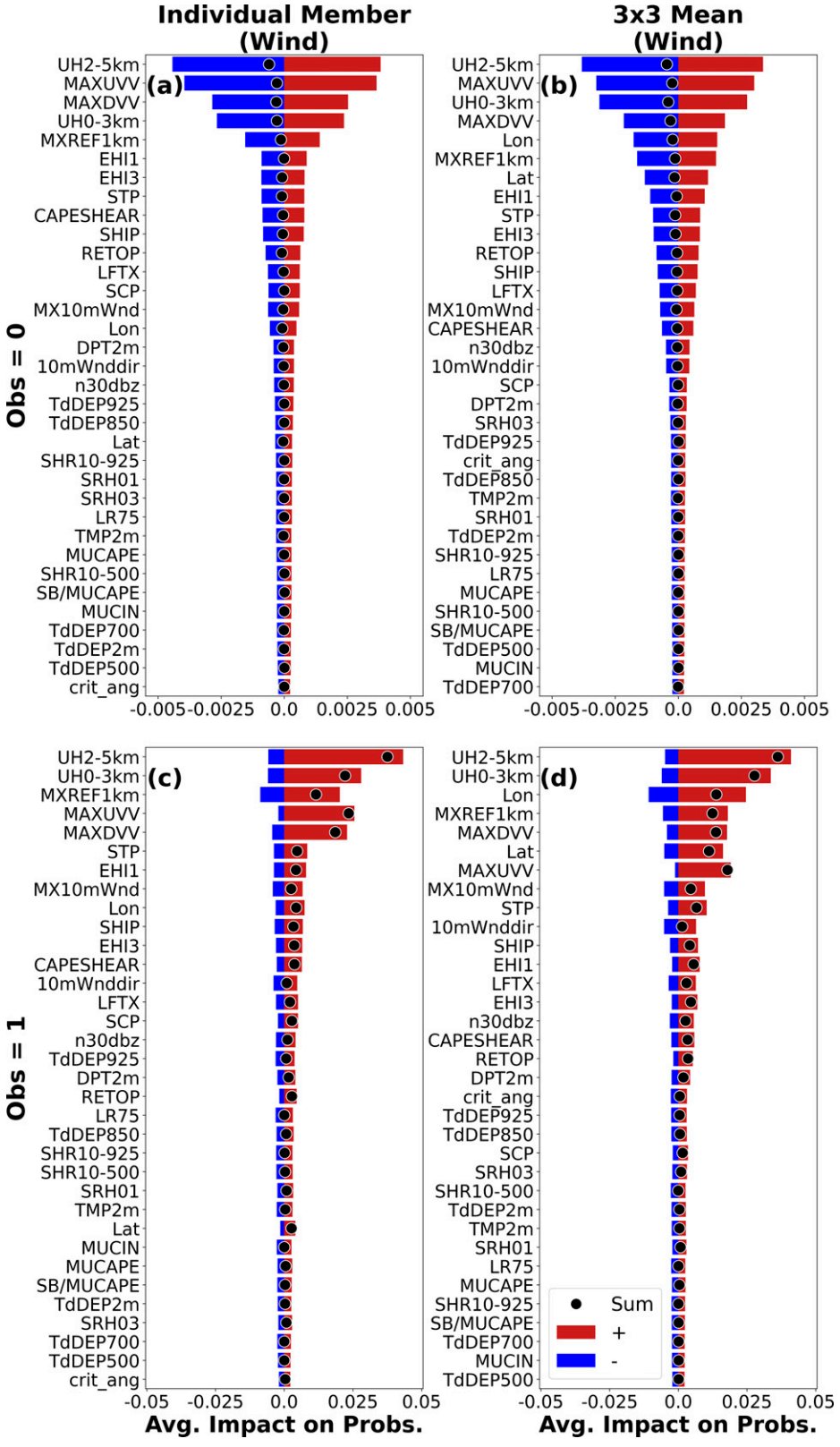

FIG. 6. As in Fig. 4, but for severe wind.

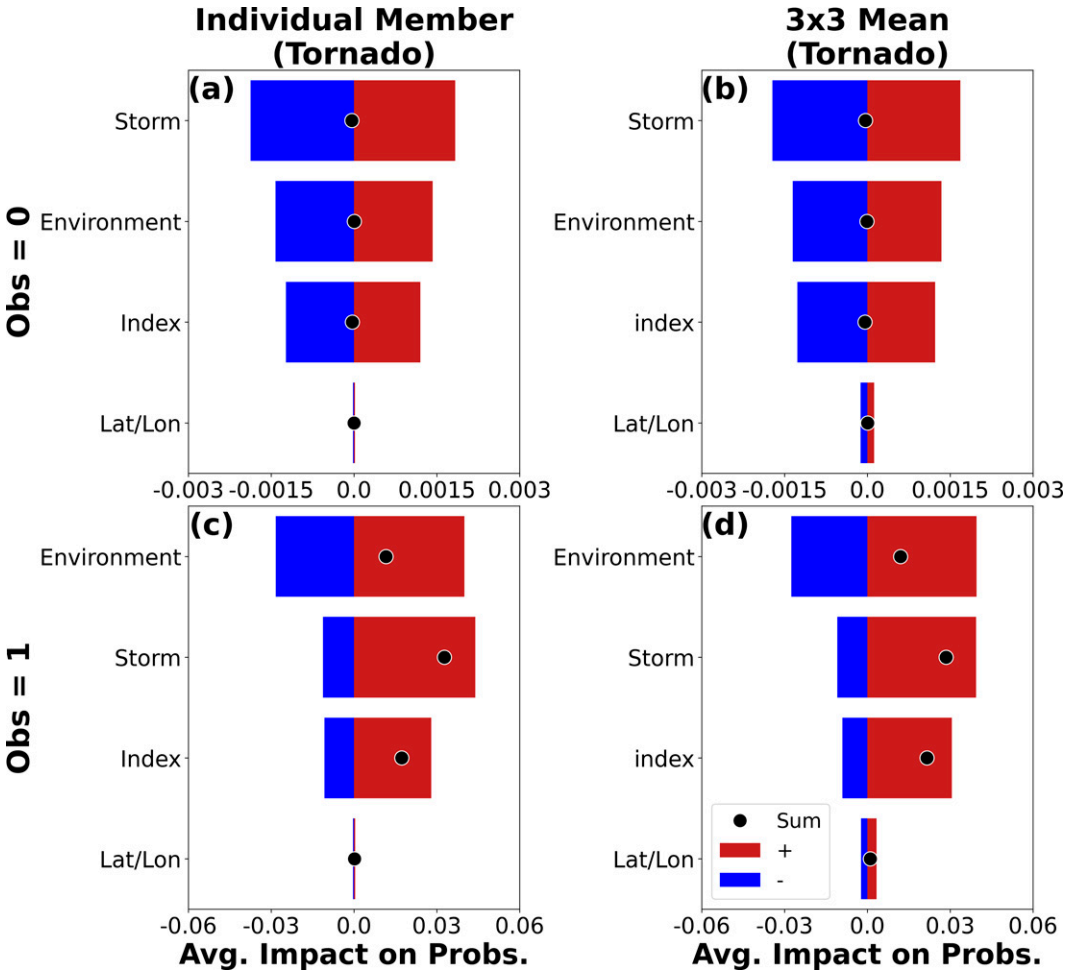

FIG. 7. As in Fig. 3, but for tornadoes.

1990; Johns and Doswell 1992; Rasmussen and Blanchard 1998; Parker 2014; Coffer and Parker 2018) have been associated with environments favorable for tornadoes, large UH2–5km suggests deep rotating updrafts, while high MXREF1km indicates intense storms.

4) MEMBER AND SPATIAL CONTRIBUTIONS

In the IM RFs, the set of non-time-lagged members influences the RF probabilities more than the time-lagged members for all hazards (Figs. 9a–f). Moreover, the nonlagged members decrease (increase) RF probabilities more when no (an) LSR is present, indicating greater skill. This result makes sense given that forecasts with shorter lead times should generally have greater skill and thus be given more "weight" for determining a prediction. Another interesting finding is that, for hail and wind, the NSSL predictors are found to be noticeably more important than predictors from other members (Figs. 9a,b,d,e). It remains unclear if this result reflects some systematically superior characteristic of the NSSL members or a chance occurrence. For tornadoes, the NAM variables are noticeably less important (Figs. 9c,f), likely because the NAM does not include UH0–3km or STP, two of the most important fields for tornado prediction (Fig. 8).

EM RF analysis shows that, for all hazards, the most important predictors are the ones taken from the point of prediction, although the distribution of importance values is not isotropic (Figs. 10a–c). For hail and wind, the storm variables at the point of prediction are noticeably more important than storm variables at surrounding points (Figs. 10d,e). Interestingly, this pattern is much less pronounced for the nonstorm (i.e., environment and index) variables (Figs. 10g,h). A potential interpretation of these results is that the RFs use environment and index variables to assess the environmental conduciveness to severe hail and wind near the point of protection while using storm variables to "pinpoint" where storms are likeliest to occur. The pattern is not as apparent for tornadoes (Figs. 10f,i), however, perhaps due to the greater difficulty of tornado forecasting.

c. *Single-field relationships*

IM RFs heavily rely on UH2–5km to construct forecasts for all 3 hazards (Figs. 4, 6, 8). However, the IM RFs use

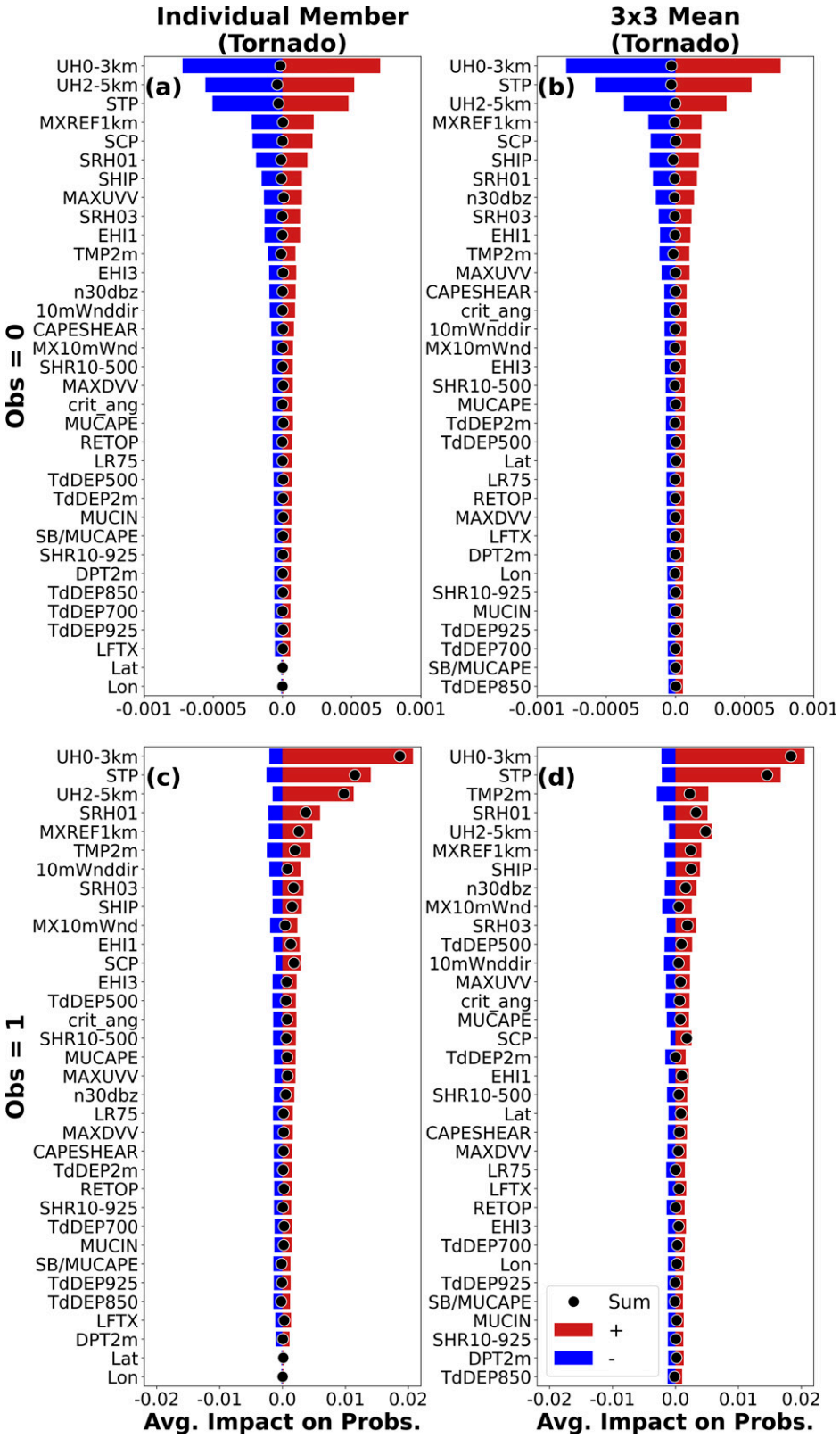

FIG. 8. As in Fig. 4, but for tornadoes.

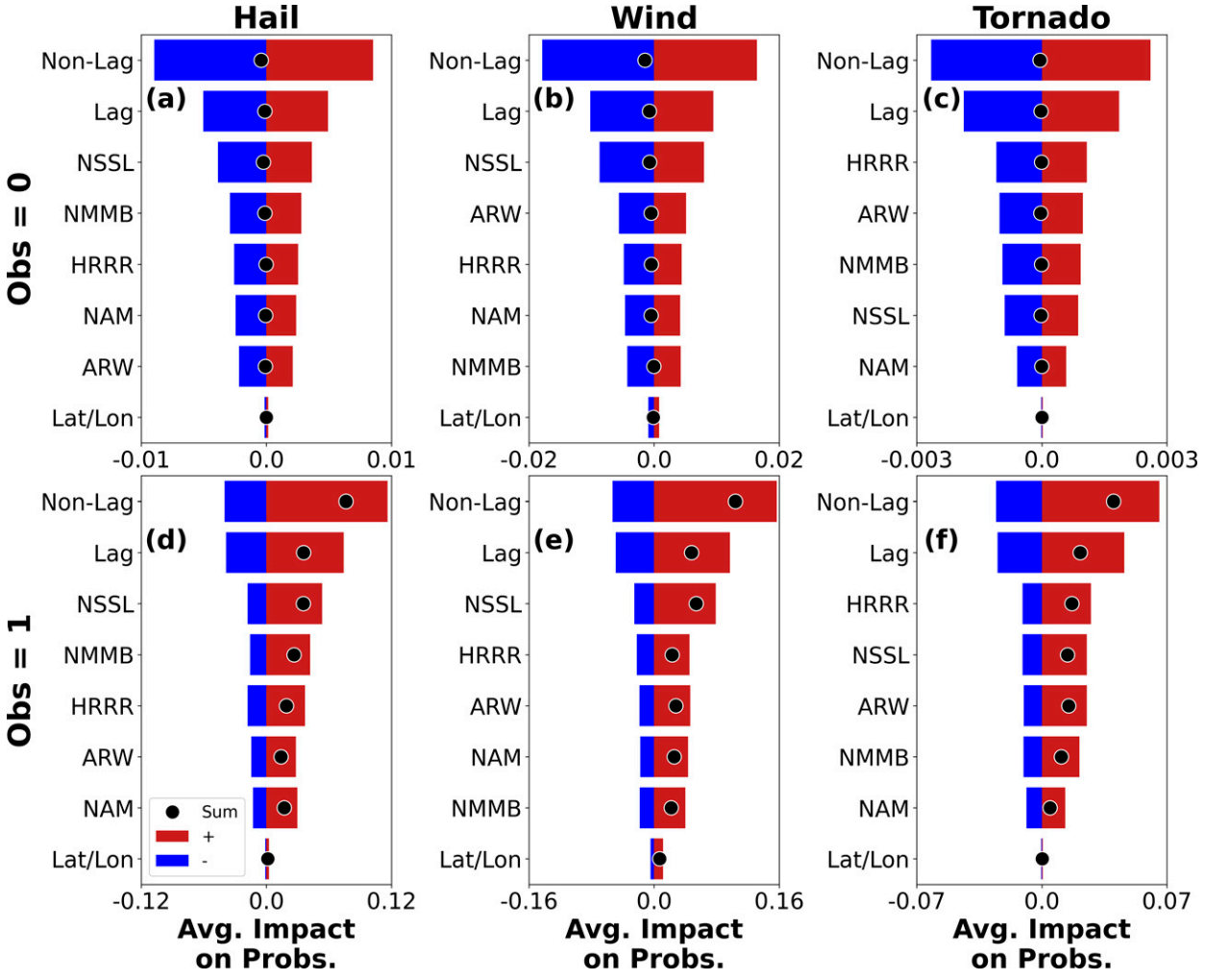

FIG. 9. As in Fig. 3, but for contributions from the non-time-lagged, time-lagged, NSSL, NMMB, HRRR, ARW, NAM, and latitude/longitude predictors from the all-predictor IM RF. Cases not associated with an observed storm report are shown in (a)–(c), while cases with an observed storm report are shown in (d)–(f). Results from (left) severe hail, (center) severe wind, and (right) tornado RFs. Note the different *x*-axis scales in each panel.

UH2–5km from different members in different ways, depending on the hazard (Figs. 11a–r). For example, larger values of (non-time-lagged) HRRR and NSSL UH2–5km tend to result in greater contributions to RF severe hail probability (Figs. 11a,j)—which is not always true of the NMMB and NAM members (Figs. 11g,m). For severe wind, larger UH2–5km tends to increase RF probability contribution only up to a point for most members, although that point varies by member (Figs. 11b,e,h,k,n). For all hazards, the ensemble mean (Figs. 11p–r) has a clearer association with UH2–5km probability contribution compared to any individual member (Figs. 11a–o), showcasing the power of the ensemble mean.

Figure 11 illustrates two other important points. First, while a definite relationship exists between each member's UH2–5km and the RF probability contribution, the sign of the contribution does not necessarily discriminate between events (i.e., LSRs) and nonevents (i.e., no LSRs), likely due to model error. Second, Fig. 11 shows that the same UH2–5km value (for a given member) can contribute differently to the overall RF probabilities depending on the case. This variability is a consequence of other variables interacting with UH2–5km. For example, a UH2–5km value between 25 and 50 $m^2 s^{-2}$ might be favorable or unfavorable for severe hail depending on the environment.

Different fields have different relationships with the probability of observed severe weather, and these relationships vary based on the hazard (Figs. 12a–r). For example, ensemble mean UH2–5km has an "S-shaped" relationship with severe probability for severe hail (Fig. 12a), while the relationship is more "sickle-shaped" for wind (Fig. 12b) and "heavily flattened–S-shaped" for tornadoes (Fig. 12c). Meanwhile, UH0–3km has a more direct relationship with RF probability contribution for tornadoes (Fig. 12f) compared to severe hail (Fig. 12d) or wind (Fig. 12e). As expected, SHIP (Figs. 12g–i) and STP (Figs. 12j–l) are most influential for hail and tornadoes, respectively, while MAXUVV (Figs. 12m–o) has the

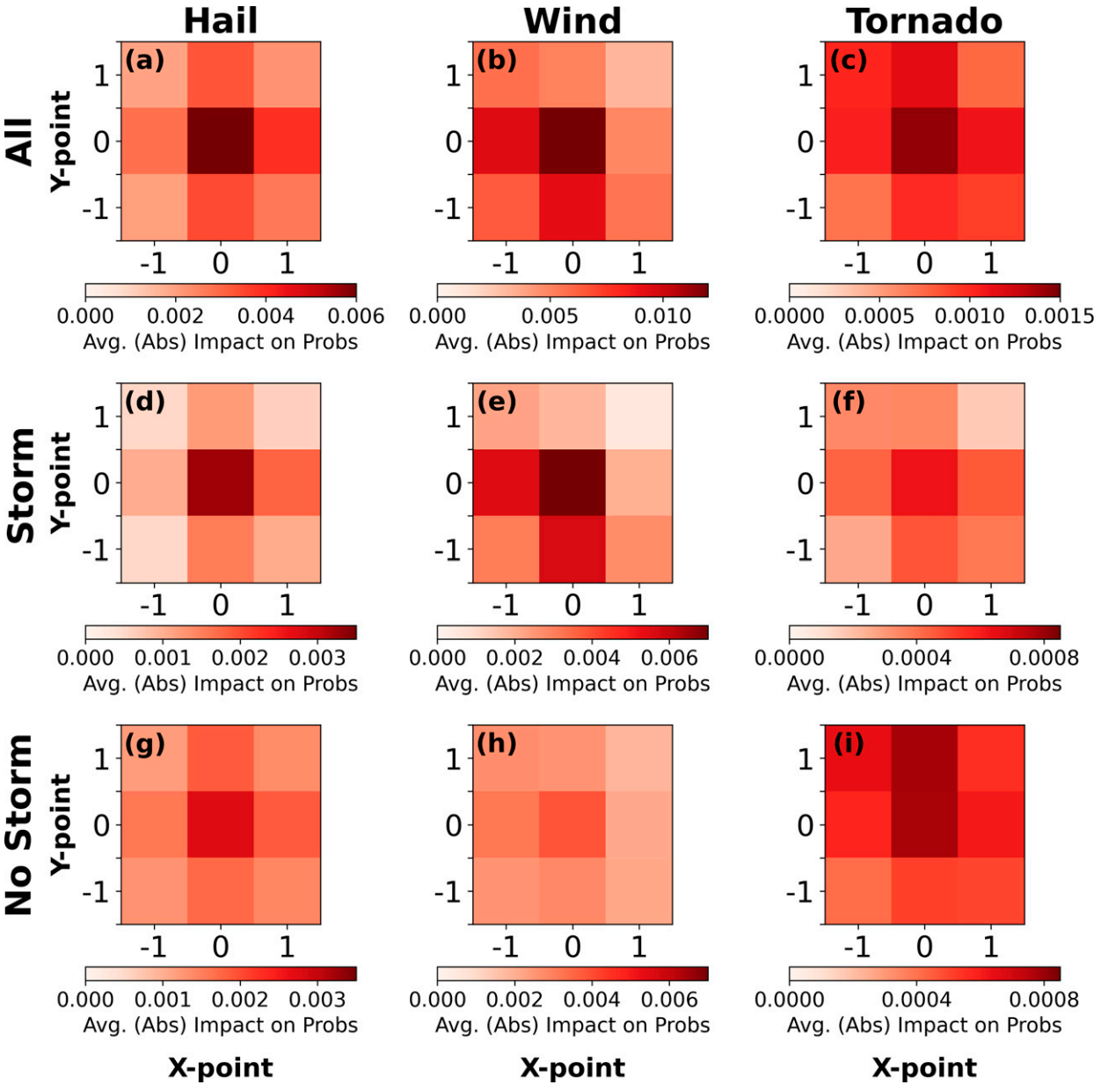

FIG. 10. (a) Mean absolute EM RF probability contributions from all predictors at the point of prediction [(0, 0)] and the 8 closest 80-km grid points for severe hail. (b),(c) As in (a), but for severe wind and tornadoes. (d)–(f) As in (a)–(c), but for only the storm variables. (g)–(i) As in (a)–(c), but for only the nonstorm variables. Note the different color bar scales between columns. Within each column, panels in rows two and three have the same scales.

strongest relationship with hail and wind probabilities. Interestingly, MXREF1km (Figs. 12p–r) has the clearest relationship for severe wind (Fig. 12q), showing negative contributions until approximately 50–55 dBZ and then mostly positive contributions.

d. Multifield relationships

The most important two-variable relationships in the IM RFs involve either two storm predictors or one storm and one index predictor (Figs. 13a–i). For severe hail, the interaction between NSSL UH2–5km and NSSL SHIP is the most important (Fig. 13a). The same value of NSSL UH2–5km (e.g., 25 $m^2 s^{-2}$) can result in negative (weak-to-moderate) hail probability contributions if the SHIP is close to 0 (near 2). Similarly, a SHIP near 0 can result in negative (weak-to-moderate) probabilities if UH2–5km is small [relatively large (e.g., near 100 $m^2 s^{-2}$)]. Thus, simulated storms with strongly (weakly) rotating updrafts in marginal (favorable) simulated environments can result in nonnegligible positive probability contributions. Similar interaction effects are present in most other two-variable combinations described in Fig. 13, although the specific variables involved depends on the hazard. Collectively, the results

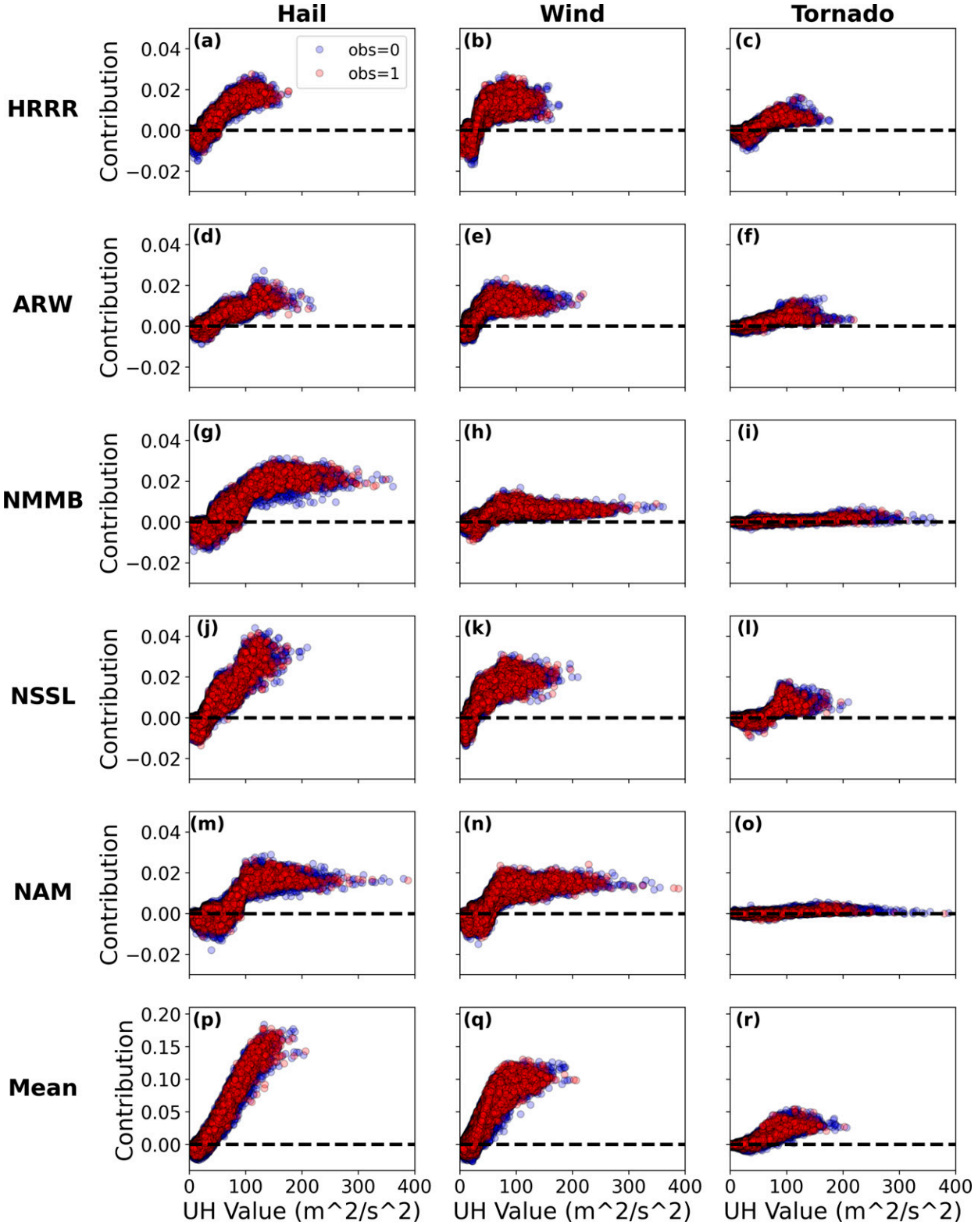

FIG. 11. (a) IM RF probability contributions from the non-time-lagged HRRR member's UH2–5km for severe hail for each sample in the dataset. Samples associated with an (no) observed LSR are colored red (blue). Each point is semitransparent, so darker colors indicate greater sample density. A 0.00 contribution is indicated by a black horizontal line. (b),(c) As in (a), but for severe wind and tornadoes. (d)–(f),(g)–(i),(j)–(l),(m)–(o) As in (a)–(c), but for the non-time-lagged ARW, NMMB, NSSL, and NAM members, respectively. (p)–(r) As in (a)–(c), but for the contributions from all members' UH2–5km graphed against the (10-member) ensemble mean (smoothed) UH2–5km.

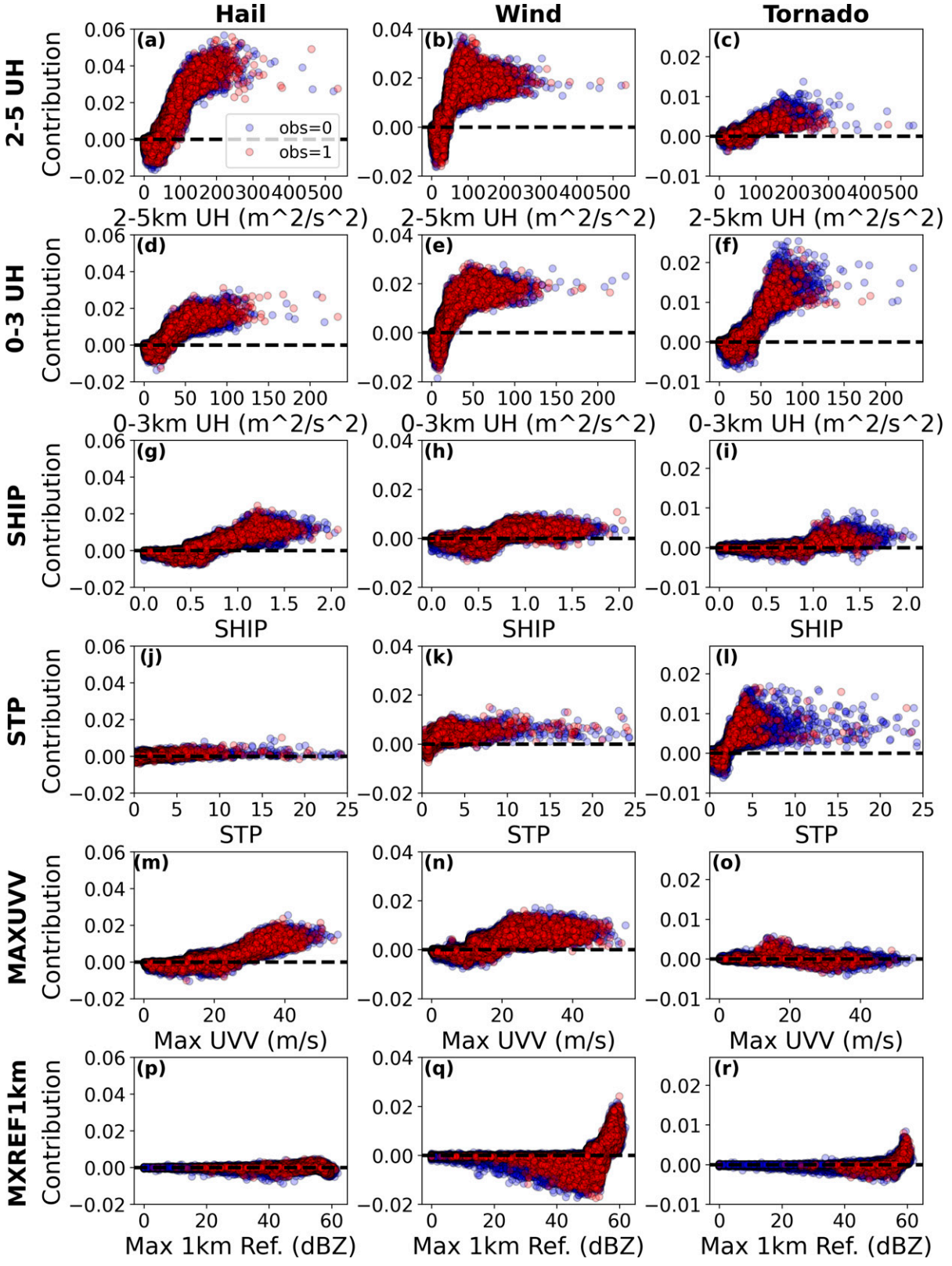

FIG. 12. (a) EM RF contributions from (0, 0) (unsmoothed) mean UH2–5km for each sample in the dataset for severe hail. Samples associated with an (no) observed LSR are colored red (blue). Each point is semitransparent, so darker colors indicate greater sample density. A 0.00 contribution is indicated by a black horizontal line. (b),(c) As in (a), but for severe wind and tornadoes. Note that, unlike Fig. 11p–r, the *x* axis in (a)–(c) refers to the unsmoothed, 9-member ensemble mean UH2–5km. (d)–(f),(g)–(i),(j)–(l),(m)–(o),(p)–(r) As in (a)–(c), but for mean UH0–3km, SHIP, STP, MAXUVV, and spatially smoothed maximum 1-km above-ground simulated reflectivity, respectively.

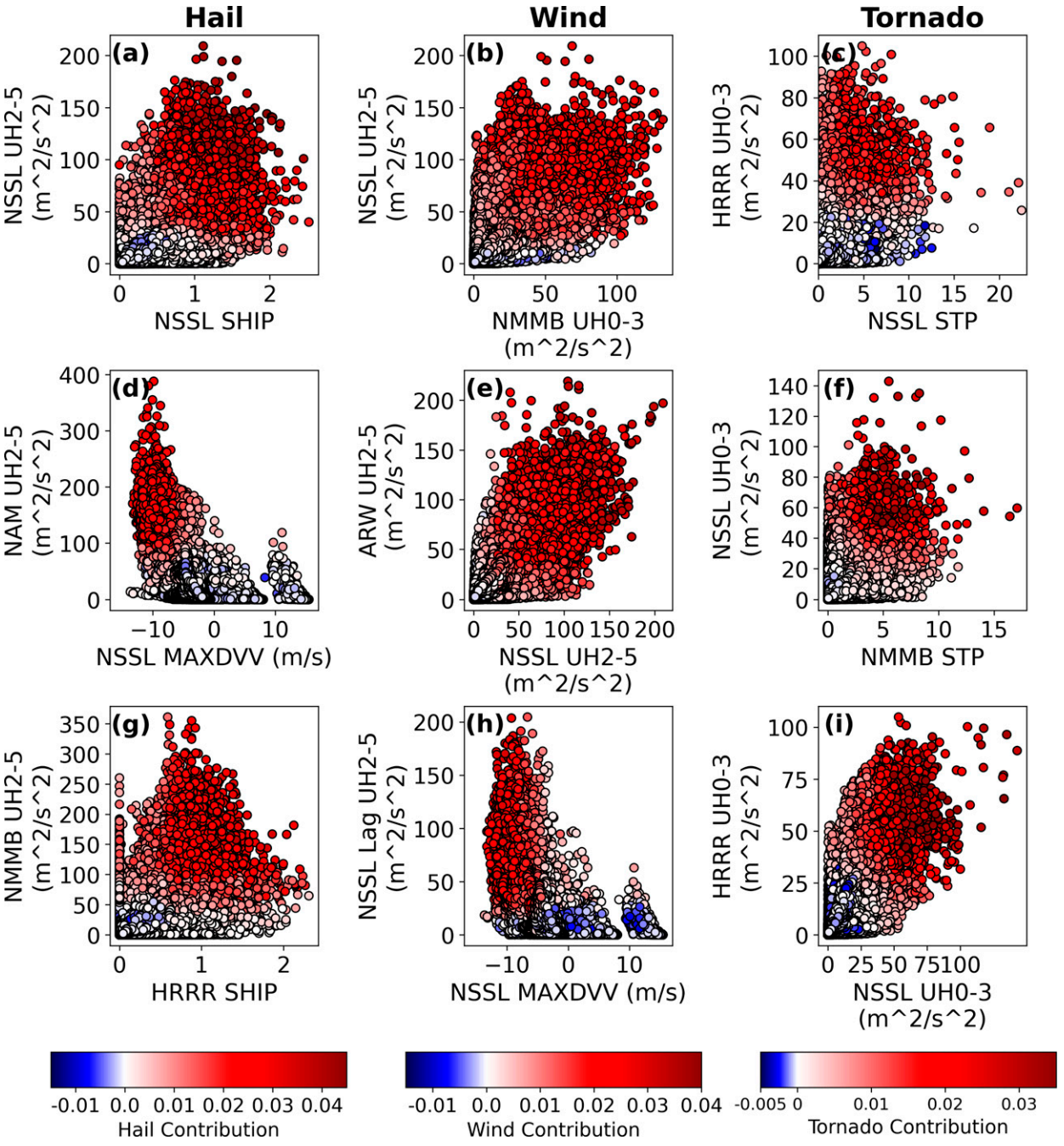

FIG. 13. (a) IM RF probability contribution (shaded dots) resulting from the most important two-variable combination for all samples in the dataset for severe hail prediction. (b),(c) As in (a), but for severe wind and tornado prediction, respectively. (d)–(f) As in (a)–(c), but for the second-most-important two-variable combination for each hazard. (g)–(i) As in (a)–(c), but for the third-most-important two-variable combination for each hazard. Note the different color scales for each hazard.

suggest that using a single-variate UH2–5km (or UH0–3km) threshold (e.g., Sobash et al. 2011, 2016, 2019; Loken et al. 2017, 2020) does not always give the most complete representation of the severe weather threat.

The most important multivariate relationships from the EM RFs (Figs. 14a–i) also include either multiple storm fields or one storm and index field. While the most important EM RF combinations involve variables at or close to the point of prediction, most of the important combinations involve variables at different spatial points (e.g., Figs. 14a–c,e,f,h,i). This is interesting because it suggests an attempt by the RF to account for displacement errors in the simulated storm and/or environment. For example,

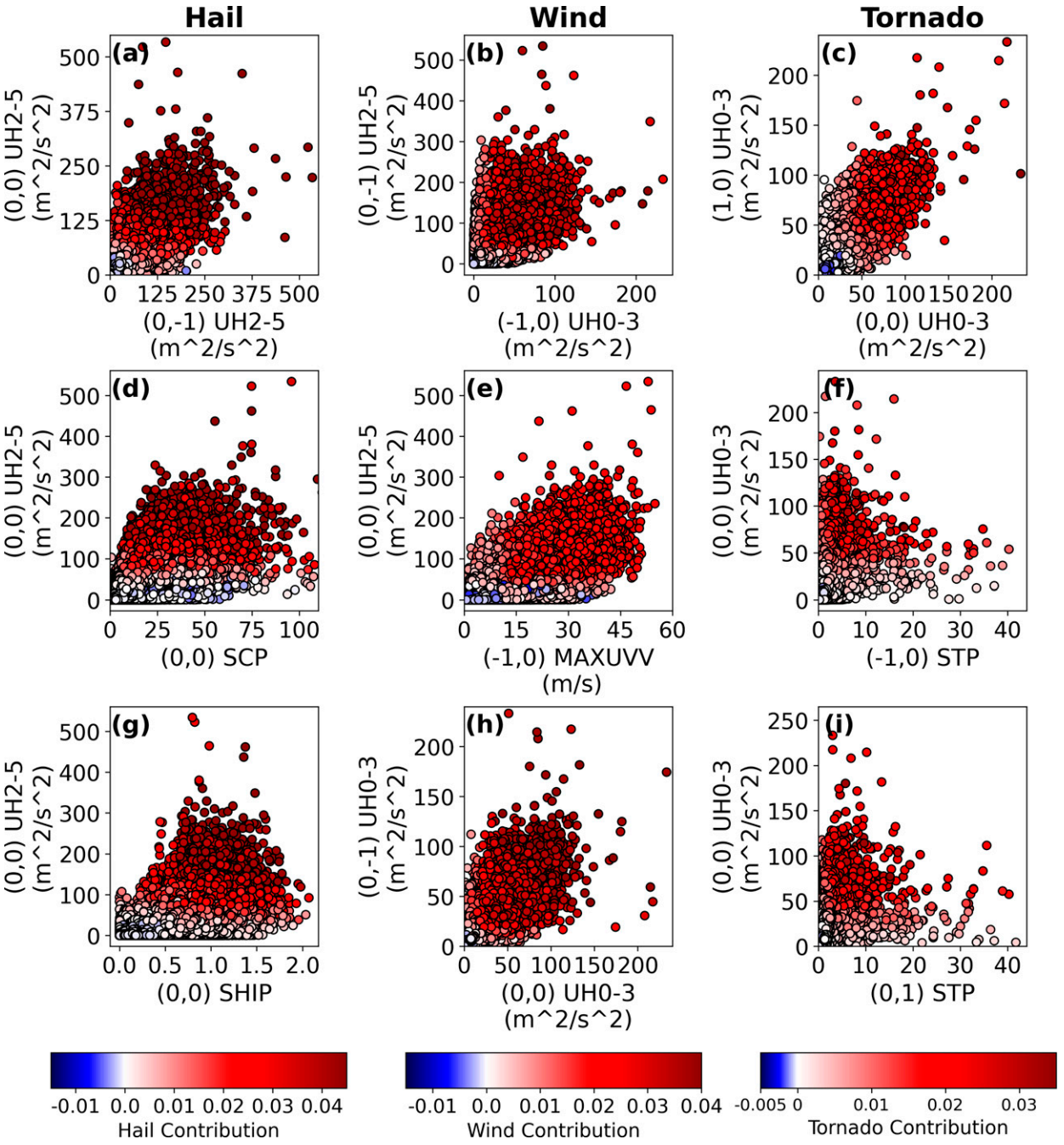

FIG. 14. As in Fig. 13, but for the most important two-variable combinations in the EM RFs.

the RF learns to maximize its probability of severe hail when UH2–5km is large at *and* near the point of prediction (Fig. 14a).

4. Representative case study: 1200 UTC 23 May–1200 UTC 24 May 2020

Four main features helped drive the severe weather on this day: a longwave trough in the western CONUS, a midlevel low and associated surface cyclone in the Upper Midwest, a shortwave trough in the South, and a dryline in the southern High Plains. Figures 15a–l shows some (preprocessed, 9-member) simulated ensemble mean data from HREFv2.1. The temporal mean 2-m temperature (Fig. 15a) and dewpoint temperature (Fig. 15b) fields suggest a (temporal mean) thermal and moisture ridge over the central Plains, downstream of a longwave trough. Daily maximum simulated 10-m wind speeds are highest in western Texas—reaching over 25 m s^{-1} (55.9 mph) there—and southwestern South Dakota (Fig. 15c). Maximum 0–1-km SRH

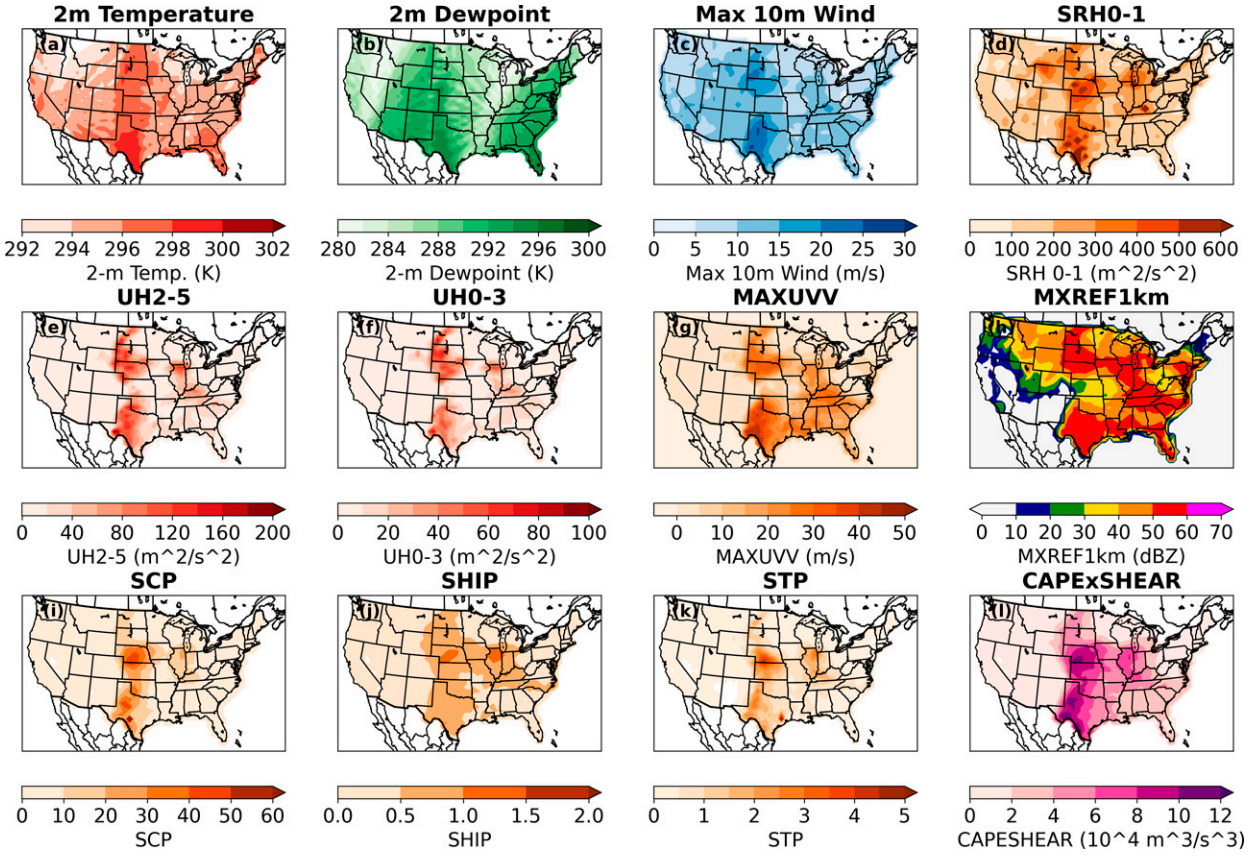

FIG. 15. Preprocessed (9-member) ensemble mean fields for (a) 2-m temperature, (b) 2-m dewpoint temperature, (c) maximum 10-m wind speed, (d) 0–1-km storm relative helicity, (e) 2–5-km updraft helicity, (f) 0–3-km updraft helicity, (g) maximum vertical velocity, (h) spatially smoothed daily maximum 1-km simulated reflectivity, (i) SCP, (j) SHIP, (k) STP, and (l) the product of MUCAPE and 10-m–500-hPa vertical wind shear magnitude, valid 1200 UTC 23 May–1200 UTC 24 May 2020.

is at least 200 $m^2 s^{-2}$ over a large swath of the Great Plains and the Upper Midwest (Fig. 15d). Regions of greater than 80 $m^2 s^{-2}$ UH2–5km are found in the Dakotas, Nebraska, western Oklahoma and western Texas, northern Illinois, and central Kentucky (Fig. 15e). Relatively large values of UH0–3km (Fig. 15f) and MAXUVV (Fig. 15g) are found in these same regions, and maximum simulated reflectivity indicates (simulated) storms over a large portion of the eastern two-thirds of the CONUS (Fig. 15h). Important index variables—including supercell composite parameter (SCP) (Fig. 15i), SHIP (Fig. 15j), STP (Fig. 15k), and the product of MUCAPE and 10-m–500-hPa wind shear (Fig. 15l)—are also elevated throughout much of the Central Plains. STP is maximized on the border of Nebraska and Kansas, but elevated values of STP are also seen in northern Illinois and the Texas Panhandle (Fig. 15k).

IM and EM RF probabilities generally highlight three regions for all three hazards: the northern Great Plains (i.e., from North Dakota to Nebraska), southern Great Plains (i.e., west Texas and western Oklahoma), and parts of the Midwest near northern Illinois (Figs. 16a–f). Additionally, both RFs show a severe wind threat farther south, including 30% or 45% probabilities in central Kentucky and a broad 5% probability for most of the Southeast (Figs. 16c,d). The biggest differences between the RF configurations are the probability magnitudes. For example, the EM (IM) RF has 15% (5%) hail probabilities in northern Illinois and central Kentucky (Figs. 16a,b). Since observed severe hail occurred in northern Illinois, the EM RF has better POD there and is rewarded with a slightly better hail AUC and BS. For severe wind, the EM has higher probabilities in northern Illinois—where a cluster of wind reports was observed—and in central Kentucky and northern North Carolina, where no severe wind LSRs were observed (Fig. 16d). As a result, the EM RF has greater POD in northern Illinois but also more false alarm in regions farther southeast, giving it just slightly worse AUC and BS metrics compared to the IM RF (Figs. 16c,d). The EM tornado RF also has larger probabilities in northern Illinois—giving it better POD there compared to the IM RF—and in southwestern Nebraska—giving it more false alarm there (Figs. 16e,f). Overall, the EM RF has slightly better tornado AUC and BS values.

The IM and EM forecasts use similar fields to construct their forecasts (Figs. 17a–f). The biggest difference is that the EM RFs rely on latitude and longitude more than the IM RFs for severe hail and wind prediction, consistent with Figs. 3–6.

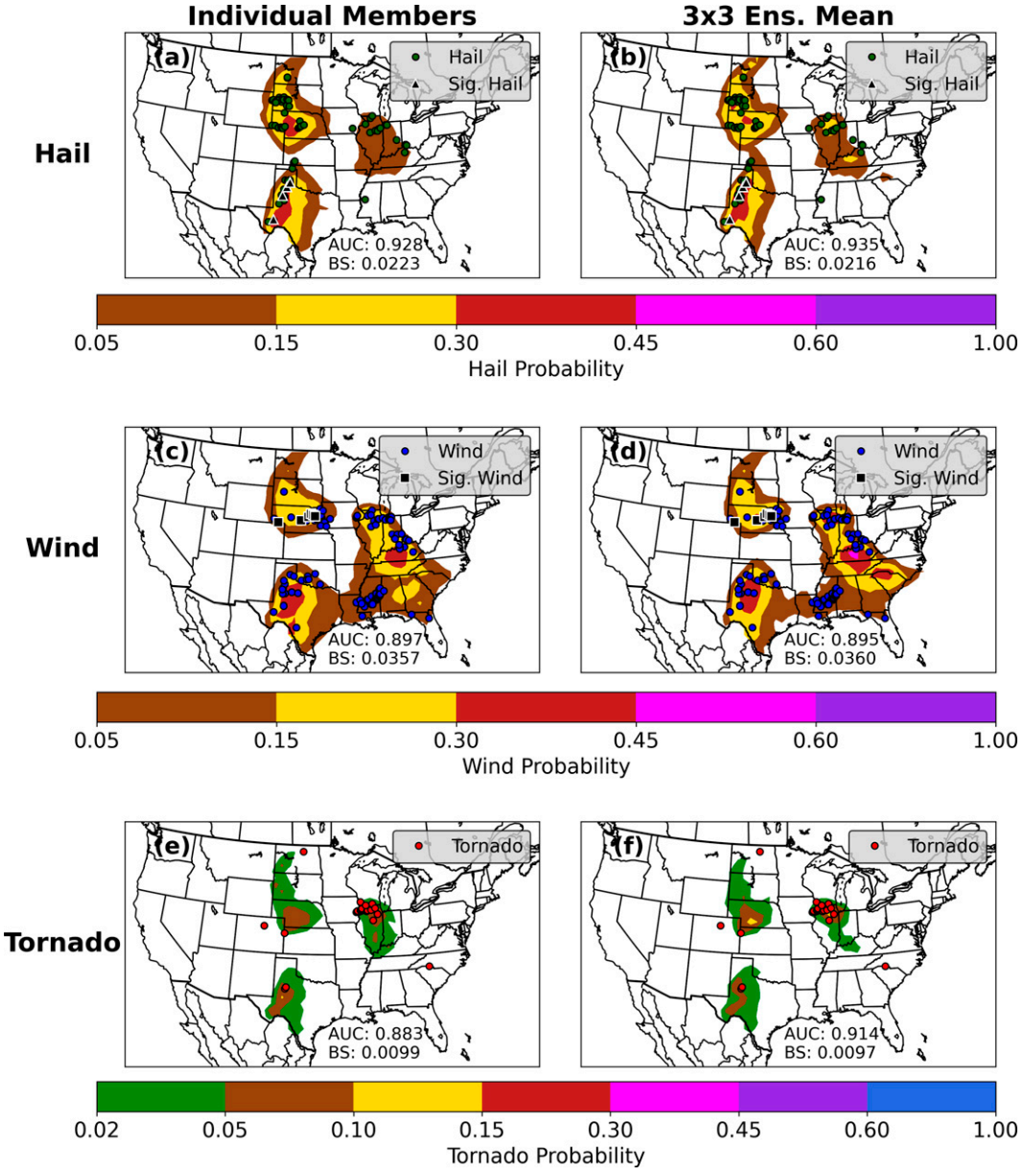

FIG. 16. (a) Severe hail forecast probability from the IM RF (shaded) and observed subsignificant (green dots) and significant (black triangles) hail reports, valid from 1200 UTC 23 May to 1200 UTC 24 May 2020. Individual-day AUC and BS are shown at the bottom of the panel. (b) As in (a), but for the EM RF. (c),(d) As in (a) and (b), but for severe wind forecasts. Observed subsignificant (blue dots) and significant (black squares) are shown. (e),(f) As in (a) and (b), but for tornado forecasts. Observed tornado reports (red dots) are shown. Note that the plotting software applies an automatic linear interpolation between 80-km grid points.

Figures 18 and 19 show the storm, environment, index, and latitude/longitude probability contributions for the IM and EM ensembles, respectively. In both cases, the storm fields tend to exert the greatest influence on the probabilities (Figs. 18a–c and 19a–c). The most obvious difference between the IM and EM RFs is the latitude/longitude contributions for the severe hail and wind forecasts. Unlike the IM RFs (Figs. 18j,k), the EM RFs have large positive contributions for severe hail in most of the Great Plains (Fig. 19j) and large negative (positive) severe wind contributions in the Great Plains (eastern United States) (Fig. 19k).

5. Summary and discussion

In this paper, the Python module TI was used to assess how differently configured RFs use CAE variables to create skillful severe weather forecasts. Two main configurations of RFs were examined: RFs trained on individual-member

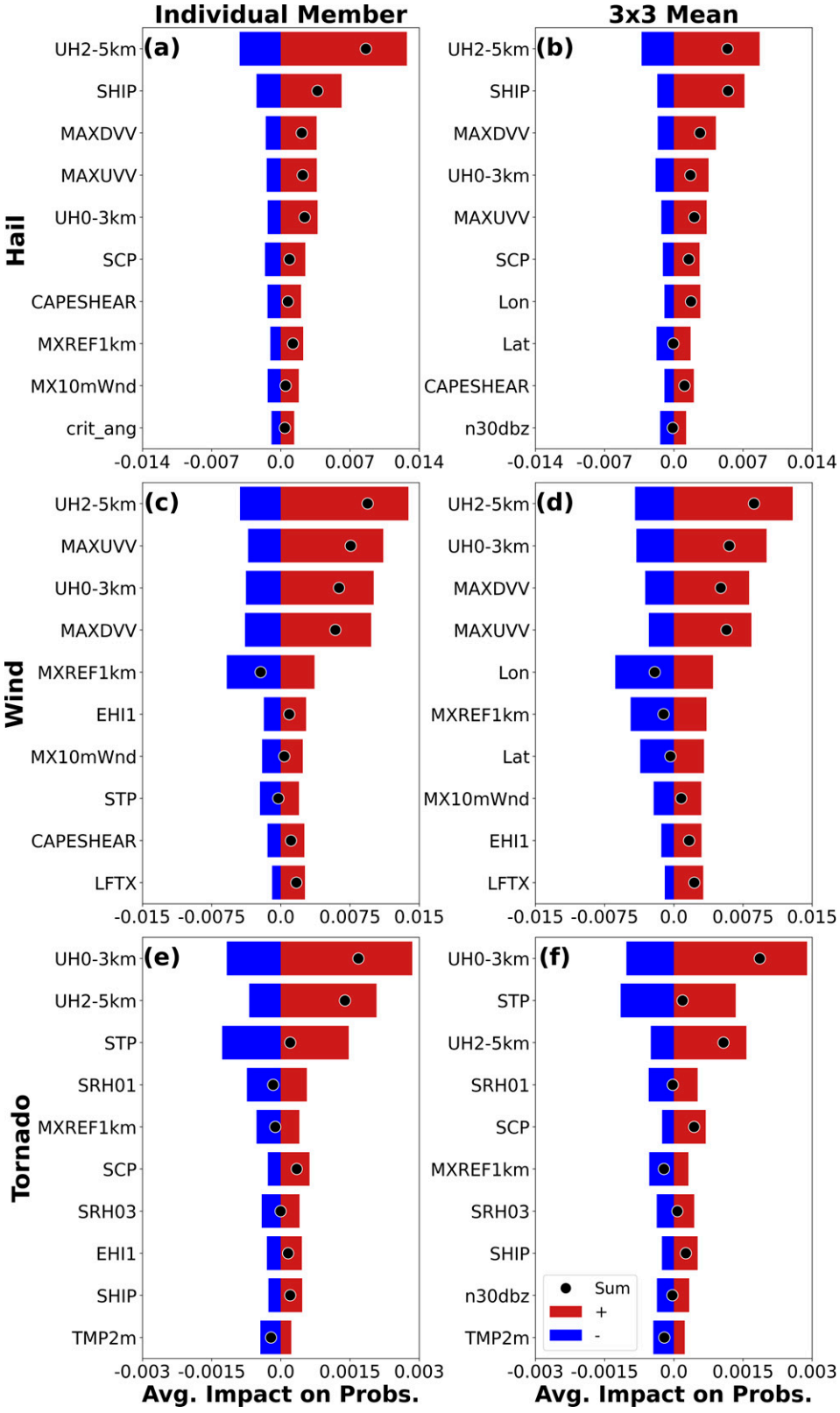

FIG. 17. (a) Mean TI negative (blue), positive (red), and summed (i.e., negative plus positive; black dot) RF probability contributions (per grid point) from the 10 most important fields (aggregated over individual members) for the all-predictor severe hail IM RF, valid from 1200 UTC 23 May to 1200 UTC 24 May 2020. Analysis is done for the entire domain and fields are displayed in descending order of overall importance (i.e., mean absolute value of contributions). (b) As in (a), but for the all-predictor EM RF. (c),(d) As in (a) and (b), but for the severe wind RFs. (e),(f) As in (a) and (b), but for the tornado RFs.

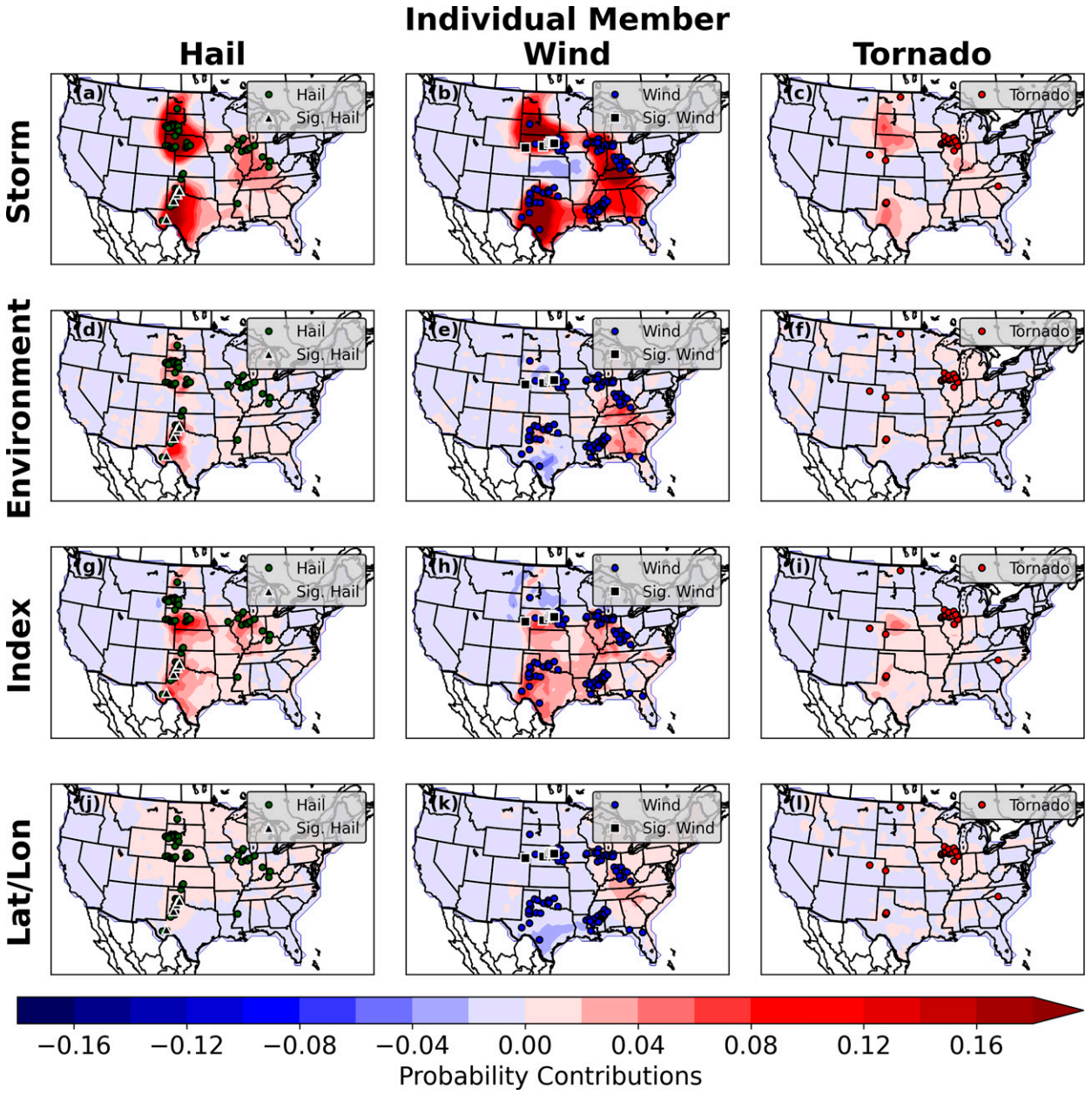

FIG. 18. (a) Aggregated IM RF probability contributions (shaded) from storm-related variables for severe hail prediction, with observed subsignificant (green dots) and significant (black triangles) hail reports overlaid. (b) As in (a), but for severe wind prediction with observed subsignificant (blue dots) and significant (black squares) overlaid. (c) As in (a), but for tornado prediction with observed subsignificant tornado reports (red dots) overlaid. (d)–(f) As in (a)–(c), but for environment variables. (g)–(i) As in (a)–(c), but for index variables. (j)–(l) As in (a)–(c), but for latitude and longitude variables.

predictors using variables at the point of prediction (IM RFs) and RFs trained on ensemble mean predictors using variables at the point of prediction and the 8 closest grid points (EM RFs). For each hazard (severe hail, wind, and tornadoes), IM and EM RFs were trained with the full set of 34 predictor fields as well as various predictor subsets to determine which types of variables contributed most to the RFs' skill.

For all hazards, the EM RFs objectively outperformed the IM RFs when the same fields were used as predictors. Although the skill of ensemble mean fields has long been demonstrated (e.g., Epstein 1969; Leith 1974; Clark et al. 2009; Coniglio et al. 2010), this finding was somewhat unexpected. Rather, it was hypothesized that RFs would be able to identify and exploit unique relationships between individual HREFv2.1 and observed severe weather. However, ensemble

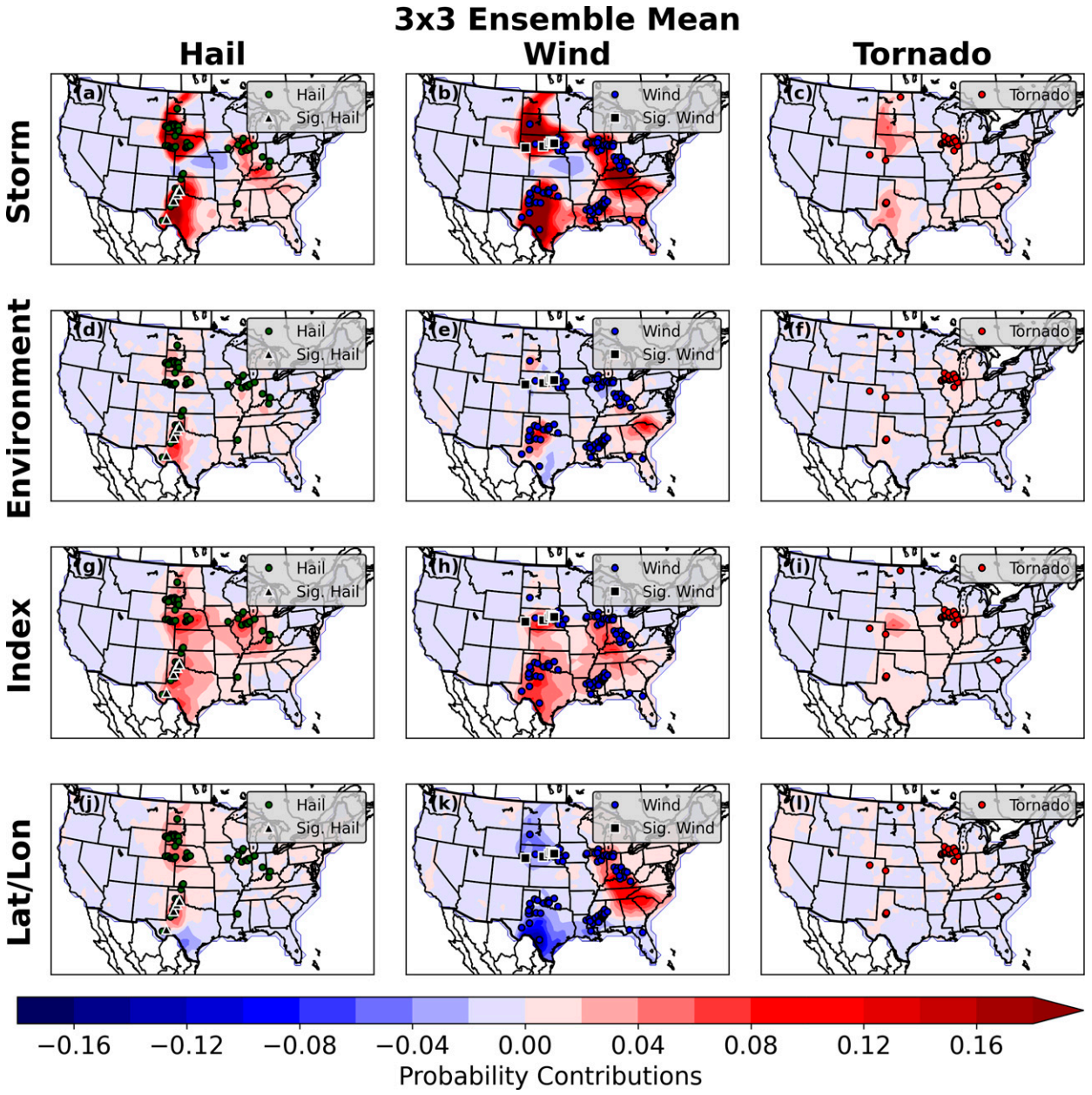

FIG. 19. As in Fig. 18, but for EM RFs.

mean fields generally had clearer relationships with RF probability contribution (e.g., Figs. 11p–r; the pattern also exists for other fields not shown), suggesting that the EM RFs had higher signal-to-noise ratios, which enabled RFs to more easily learn associations between the CAE variables and observed severe weather. Of course, the higher signal-to-noise ratios are likely attributable to the greater skill of ensemble mean fields compared to individual member fields. The EM RFs are also advantageous because they do not require their storm predictors to be spatially smoothed. Thus, the EM RFs require less preprocessing and do not force simulated storms to have an isotropic spatial uncertainty distribution.

To test the impact of the EM RFs' use of predictors from multiple spatial points, a third RF configuration was tested in which ensemble mean predictors were only used at the point of prediction (not shown). The skill of this configuration fell between that of the EM and IM RFs, suggesting that the EM RFs benefitted from *both* using ensemble mean predictors *and* using predictors at multiple spatial points.

Nevertheless, IM RFs were still able to attain a high degree of skill and highlighted similar areas for severe weather on most days compared to the EM RFs (e.g., Fig. 16). Because IM RFs learn relationships from individual member fields, they may provide more insight into optimal ensemble use and

design compared to EM RFs. For example, Fig. 9 suggests that not all members were utilized equally, especially for severe hail and wind prediction, and that different members had different levels of importance for predicting different hazards. It is currently unclear why, exactly, this is the case and how systematic this result is; however, it is a result that merits further attention as it may have implications for model development or ensemble design.

TI importance metrics and verification of the RFs trained on predictor subsets showed that the storm-related variables were the most important. Indeed, RFs trained on only storm predictors were nearly as skillful as RFs trained on the entire set of predictors; this finding held for IM and EM RFs for all three hazards. Interestingly, RFs trained with storm and index variables were slightly more skillful than using all predictors for severe hail and tornado prediction. Meanwhile, RFs using only environment-related predictors always produced the worst verification metrics for all three hazards. Index-only RFs were notably better than environment-only RFs for forecasting severe hail and tornadoes (i.e., when a hazard-specific index variable was available).

Collectively, these results suggest that while nonstorm variables can provide relatively skillful next-day severe weather forecasts (e.g., as in Hill et al. 2020), the storm fields from CAEs provide crucial information that bolsters the forecasting skill at next-day lead times. Thus, it makes sense why the next-day RFs in Loken et al. (2020) performed objectively better relative to SPC human forecasts than the day 1 RFs in Hill et al. (2020).

At the same time, when storm-related fields are not available, results in this study suggest that index variables (e.g., STP, SHIP, the product of MUCAPE and deep-layer shear) can still be used to create skillful severe weather forecasts. This result is consistent with recent climate studies (e.g., Gensini and Brooks 2018; Gensini and de Guenni 2019; Tang et al. 2019) that have associated index variables (e.g., STP, SHIP) from the North American Regional Reanalysis (NARR; Mesinger et al. 2006) with observed severe weather reports to investigate past and/or predicted future U.S. severe weather climatologies. An advantage of index variables is that they require multiple "ingredients" for severe weather to "line up" in space and time, which is a physical requirement for severe weather. This approach may therefore be more useful for predicting severe weather than merely taking a temporal mean of the constituent index fields over the period of interest.

Importantly, both IM and EM RFs emphasized predictors and learned relationships that made physical sense. For example, SHIP was a top predictor for hail, while STP and 0–3-km UH0–3km were top tornado predictors. Additionally, TI analysis found that the UH2–5km from most individual members—as well as the ensemble mean—had an S-shaped relationship with severe weather likelihood, which supports the commonly used method as treating a climatologically large value of UH2–5km as a simulated surrogate severe weather report (e.g., Sobash et al. 2011, 2016, 2019; Loken et al. 2017, 2020; Roberts et al. 2020).

At the same time, results from this paper suggested several reasons why this threshold method may be incomplete. Most importantly, the relationship between UH2–5km and, for example, severe hail is not a perfect step function. With all else equal, larger values of UH2–5km usually suggest larger severe hail probabilities, and there is no threshold below which the probability of severe hail is suddenly 0. Indeed, this study showed that the exact value of UH2–5km, its value at surrounding grid points, and the value of relevant index variables at nearby points are all important for determining severe weather probabilities at a given point. This makes sense intuitively but is hard to encode in an algorithm. Some previous research has attempted to combine UH2–5km and environmental information to improve UH2–5km-based severe weather forecasts, with modest success. For example, Gallo et al. (2016) reduced false alarm from UH2–5km-based tornado forecasts by additionally requiring simulated STP and other environment variables (e.g., lifting condensation level and the ratio of surface-based to most unstable CAPE) to meet certain thresholds. However, the current study suggests that this approach is suboptimal. For example, results herein show that relatively large hail probability contributions can result from small UH2–5km values if SHIP is relatively large (e.g., near 2)—which makes sense due to the possibility of simulated storm initiation or displacement errors. Conversely, severe hail probability contributions can still be positive when SHIP is near 0 if UH2–5km is very large. This type of "thinking" makes sense; essentially the RFs are learning to properly calibrate severe weather probabilities in the face of imperfect, "noisy" predictors.

6. Conclusions and future work

This paper analyzed RF-based severe weather forecast probabilities using TI. Such analysis helped shed light on how differently configured RFs make their forecasts. Having the ability to dissect the "thinking" of a skillful RF can benefit both forecasters and model developers. For example, a forecaster might confidently discount RF guidance when the algorithm emphasizes irrelevant predictors (e.g., in the face of contradictory observations), while unusual learned RF relationships could alert model developers to deficiencies in model parameterizations and/or help researchers design better ensemble prediction systems.

The work presented here provides a foundation for a wide range of future research. One simple but important avenue for future work is to stratify the results by region and season to determine what spatiotemporal relationships are learned and how these relate to the full-domain relationships. It will also be important for future work to investigate *why* predictors are important in certain circumstances, since the current study merely sheds light on *how* RFs produce skillful forecasts. For example, future work should investigate why the NSSL members are more important than the other members for predicting severe hail and wind. Investigating how the importance of different predictors varies at different lead times and spatial scales will also be worthwhile, since this type of analysis should enhance our understanding of severe weather predictability. Indeed, given the results presented here, future work should investigate whether storm fields (and CAEs themselves) might still provide substantial benefits at longer than 36-h lead times. Additionally, future work should determine how much value RF interpretability products provide to RF product users in real-time operational or HWT SFE (e.g., Gallo et al. 2017; Clark et al. 2021) settings.

While this study focused on interpreting RF-based forecasts, other machine learning (ML) methods—such as neural networks (e.g., Sobash et al. 2020) and deep learning (e.g., Lagerquist et al. 2020)—have also been recently used to create skillful severe weather hazard forecasts. Future work may wish to compare interpretability results from these methods to those obtained herein to determine if all ML algorithms use simulated ensemble data in similar ways. There is at least some chance that different ML methods tend to learn (or emphasize) different predictor–predictand relationships, especially given Fleming et al.'s (2021) finding that the mean prediction from multiple ML algorithms outperformed any individual algorithm for water supply forecasts over the American West. Thus, as computing power increases, more work will be needed to enhance, evaluate, and compare interpretability methods for different single- and multialgorithm ML methods.

Acknowledgments. Support for this work was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA–University of Oklahoma Cooperative Agreement NA16OAR4320072, U.S. Department of Commerce. High-performance computing support from Cheyenne (doi:10.5065/D6RX99HX) was provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. We thank Montgomery Flora for providing guidance and discussions on the tree interpreter method. AJC contributed to this work as part of regular duties at the federally funded NOAA/National Severe Storms Laboratory. The statements, findings, conclusions, and recommendations presented herein are those of the authors and do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

Data availability statement. The simulated HREFv2.1 data used in this study are archived internally at the National Severe Storms Laboratory and may be shared upon request. Observed local storm reports are available from the Storm Prediction Center's public archives (https://www.spc.noaa.gov/climo/online/) and Storm Events Database (https://www.spc.noaa.gov/wcm/).

APPENDIX

Select RF Derived Variables

a. SCP

Developed to identify environments supportive of right-moving supercells, SCP (Thompson et al. 2003) is here defined as

$$\mathrm{SCP} = \frac{\mathrm{MUCAPE}}{1000\,\mathrm{J\ kg^{-1}}} \times \frac{\mathrm{SRH03}}{50\,\mathrm{m^2\ s^{-2}}} \times \frac{\mathrm{SHR_{10–500}}}{20\,\mathrm{m\ s^{-1}}} \times \frac{-40\,\mathrm{J\ kg^{-1}}}{\mathrm{MUCIN}}, \tag{A1}$$

where MUCAPE is most unstable convective available potential energy (CAPE; J kg^{-1}); SRH03 is the 0–3-km storm relative helicity (m^2 s^{-2}), $\mathrm{SHR_{10–500}}$ is the magnitude of the vector difference between the 10-m and 500-hPa winds (m s^{-1}), and MUCIN is the most unstable convective inhibition (CIN; J kg^{-1}). Before SCP is calculated, the $\mathrm{SHR_{10–500}}$ term is set to 1 if $\mathrm{SHR_{10–500}}$ is greater than or equal to 20 m s^{-1} or 0 if $\mathrm{SHR_{10–500}}$ is less than 10 m s^{-1}, and the MUCIN term is set to 1 if MUCIN is greater than −40 J kg^{-1}.

b. STP

STP (Thompson et al. 2003) is designed to distinguish between significant and nonsignificant tornadic supercell environments. The STP used here is a fixed-layer version of the updated formulation described in Thompson et al. (2012), namely,

$$\mathrm{STP} = \frac{\mathrm{SBCAPE}}{1500\,\mathrm{J\ kg^{1}}} \times \frac{2000\,\mathrm{m} - \mathrm{LCL}}{1000\,\mathrm{m}} \times \frac{|\mathrm{SRH01}|}{150\,\mathrm{m^2\ s^{-2}}} \times \frac{\mathrm{SHR_{10–500}}}{20\,\mathrm{m\ s^{-1}}} \times \frac{200 + \mathrm{SBCIN}}{150\,\mathrm{J\ kg^{-1}}}, \tag{A2}$$

where SBCAPE is surface-based CAPE (J kg^{-1}), LCL is the lifted condensation level (m) [which is computed here using the approximation 125 × 2-m dewpoint depression (K)], SRH01 is 0–1-km storm-relative helicity (m^2 s^{-2}), $\mathrm{SHR_{10–500}}$ is the magnitude of the vector difference between the 10-m and 500-hPa winds (m s^{-1}), and SBCIN is surface-based CIN (J kg^{-1}). Before the final value of STP is calculated, the following adjustments are made: the LCL term is set to 1 if the LCL is less than 1000 m or 0 if the LCL is greater 2000 m, the deep-layer shear term is set to 1.5 if $\mathrm{SHR_{10–500}}$ is greater than or equal to 30 m s^{-1} or 0 if $\mathrm{SHR_{10–500}}$ is less than 12.5 m s^{-1}, and the SBCIN term is set to 1 if SBCIN is greater than −50 J kg^{-1} or 0 if SBCIN is less than −200 J kg^{-1}.

c. SHIP

Designed to distinguish between significant and nonsignificant hail-producing environments, SHIP (SPC 2021b) is here defined as

$$\mathrm{SHIP} = \frac{\mathrm{MUCAPE} \times \mathrm{MR} \times \mathrm{LR_{700–500}} \times -T_{500}\ (^{\circ}\mathrm{C}) \times \mathrm{SHR_{10–500}}}{42\,000\,000}, \tag{A3}$$

where MUCAPE is the most unstable CAPE, MR is the mixing ratio (g kg^{-1}), $\mathrm{LR_{750–500}}$ is the 700–500-hPa lapse rate (K km^{-1}), T_{500} is the 500-hPa temperature (°C), and $\mathrm{SHR_{10–500}}$ is the magnitude of the vector difference between the 10-m and 500-hPa winds (m s^{-1}). This initial value of SHIP is then modified according to the following rules (executed sequentially):

1) if MUCAPE is less than 1300 J kg^{-1}, $SHIP_{final}$ = SHIP × (MUCAPE/1300 J kg^{-1}), and
2) if $LR_{700–500}$ is less than 5.8 K km^{-1} but greater than 0 K km^{-1}, $SHIP_{final}$ = SHIP × ($LR_{700–500}$/5.8 K km^{-1}), or if $LR_{700–500}$ is greater than 0 K km^{-1}, $SHIP_{final}$ = 0.

Ordinarily, a third condition adjusts the SHIP based on the height of the freezing level (SPC 2021b); however, this is not done here since simulated freezing level height data were not available.

d. EHI

Energy helicity index (EHI) is the product of surface-based CAPE and SRH over a given vertical layer (e.g., 0–1 km or 0–3 km).

e. SB/MUCAPE ratio

The ratio of SBCAPE to MUCAPE is intended to help identify elevated convection. The ratio is set to 1 when MUCAPE is 0.

f. Critical angle proxy

Critical angle, which Esterheld and Giuliano (2008) defines as the angle between the 0–500-m shear vector and 10-m above-ground-level storm-relative inflow, is approximated here as the angle (°) between the 10-m–925-hPa shear vector and the storm-relative 10-m wind.

g. n30dbz

The number of grid points with at least 30-dB*Z* simulated reflectivity (n30dbz) represents the number of native 3-km HREFv2.1 grid points in an approximately 80 km × 80 km box that contain simulated reflectivity of 30 dB*Z* or greater at the time of maximum MAXUVV. This variable is a potential proxy for storm mode.

REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932, https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

Aligo, E. A., B. Ferrier, and J. R. Carley, 2018: Modified NAM microphysics for forecasts of deep convective storms. *Mon. Wea. Rev.*, **146**, 4115–4153, https://doi.org/10.1175/MWR-D-17-0277.1.

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

Breiman, L., 1984: *Classification and Regression Trees*. Wadsworth International Group, 358 pp.

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

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

Clark, A. J., W. A. Gallus, M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, https://doi.org/10.1175/2009WAF2222222.1.

——, and Coauthors, 2021: A real-time, virtual spring forecasting experiment to advance severe weather prediction. *Bull. Amer. Meteor. Soc.*, **102**, E814–E816, https://doi.org/10.1175/BAMS-D-20-0268.1.

Coffer, B. E., and M. D. Parker, 2018: Is there a "tipping point" between simulated nontornadic and tornadic supercells in VORTEX2 environments? *Mon. Wea. Rev.*, **146**, 2667–2693, https://doi.org/10.1175/MWR-D-18-0050.1.

Coniglio, M. C., K. L. Elmore, J. S. Kain, S. J. Weiss, M. Xue, and M. L. Weisman, 2010: Evaluation of WRF Model output for severe weather forecasting from the 2008 NOAA Hazardous Weather Testbed Spring Experiment. *Wea. Forecasting*, **25**, 408–427, https://doi.org/10.1175/2009WAF2222258.1.

Davies-Jones, R., D. W. Burgess, and M. Foster, 1990: Test of helicity as a forecast parameter. Preprints, *16th Conf. on Severe Local Storms*, Kananaskis Park, AB, Canada, Amer. Meteor. Soc., 588–592.

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

Environmental Modeling Center, 2003: The GFS atmospheric model. NCEP Office Note 442, 14 pp., http://www.lib.ncep.noaa.gov/ncepofficenotes/files/on442.pdf.

Epstein, E. S., 1969: The role of initial uncertainties in prediction. *J. Appl. Meteor.*, **8**, 190–198, https://doi.org/10.1175/1520-0450(1969)008<0190:TROIUI>2.0.CO;2.

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

Ferrier, B. S., Y. Jin, Y. Lin, T. Black, E. Rogers, and G. DiMego, 2002: Implementation of a new grid-scale cloud and rainfall scheme in the NCEP Eta Model. Preprints, *19th Conf. on Weather Analysis and Forecasting/15th Conf. on Numerical Weather Prediction*, San Antonio, TX, Amer. Meteor. Soc., 10.1, http://ams.confex.com/ams/SLS_WAF_NWP/techprogram/paper_47241.htm

Fleming, S. W., D. C. Garen, A. G. Goodbody, C. S. McCarthy, and L. C. Landers, 2021: Assessing the new Natural Resources Conservation Service water supply forecast model for the American West: A challenging test of explainable, automated, ensemble artificial intelligence. *J. Hydrol.*, **602**, 126782, https://doi.org/10.1016/j.jhydrol.2021.126782.

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

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

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

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting*, **31**, 273–295, https://doi.org/10.1175/WAF-D-15-0134.1.

——, and Coauthors, 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32**, 1541–1568, https://doi.org/10.1175/WAF-D-16-0178.1.

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

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

Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin, 2015: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *J. Comput. Graph. Stat.*, **24**, 44–65, https://doi.org/10.1080/10618600.2014.907095.

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

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Mon. Wea. Rev.*, **148**, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1.

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

——, Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Wea. Rev.*, **134**, 2318–2341, https://doi.org/10.1175/MWR3199.1.

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

Janjić, Z. I., 2002: Nonsingular implementation of the Mellor–Yamada level 2.5 scheme in the NCEP Meso model. NCEP Office Note 437, 61 pp., http://www.emc.ncep.noaa.gov/officenotes/newernotes/on437.pdf.

——, 2003: A nonhydrostatic model based on a new approach. *Meteor. Atmos. Phys.*, **82**, 271–285, https://doi.org/10.1007/s00703-001-0587-6.

——, and R. Gall, 2012: Scientific documentation of the NCEP Nonhydrostatic Multiscale Model on the B Grid (NMMB). Part 1: Dynamics. NCAR Tech. Note NCAR/TN-4891STR, 75 pp., http://nldr.library.ucar.edu/repository/assets/technotes/TECH-NOTE-000-000-000-857.pdf.

Jirak, I. L., S. J. Weiss, and C. J. Melick, 2012: The SPC Storm-Scale Ensemble of Opportunity: Overview and results from the 2012 Hazardous Weather Testbed Spring Forecasting Experiment. *26th Conf. on Severe Local Storms*, Nashville, TN, Amer. Meteor. Soc., 137, https://ams.confex.com/ams/26SLS/webprogram/Paper211729.html.

——, C. J. Melick, and S. J. Weiss, 2014: Combining probabilistic ensemble information from the environment with simulated storm attributes to generate calibrated probabilities of severe weather hazards. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., P2.5, https://ams.confex.com/ams/27SLS/webprogram/Paper254649.html.

——, ——, and ——, 2016: Comparison of the SPC Storm-Scale Ensemble Of Opportunity to other convection-allowing ensembles for severe weather forecasting. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., 102, https://ams.confex.com/ams/28SLS/webprogram/Session41668.html.

——, A. J. Clark, B. Roberts, B. T. Gallo, and S. J. Weiss, 2018: Exploring the optimal configuration of the High Resolution Ensemble Forecast System. *25th Conf. on Numerical Weather Prediction*, Denver, CO, Amer. Meteor. Soc., 14B.6, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345640.html.

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

Lagerquist, R., A. McGovern, C. R. Homeyer, D. J. Gagne II, and T. Smith, 2020: Deep learning on three-dimensional multiscale data for next-hour tornado prediction. *Mon. Wea. Rev.*, **148**, 2837–2861, https://doi.org/10.1175/MWR-D-19-0372.1.

Lakshmanan, V., C. Karstens, J. Krause, K. Elmore, A. Ryzhkov, and S. Berkseth, 2015: Which polarimetric variables are important for weather/no-weather discrimination? *J. Atmos. Oceanic Technol.*, **32**, 1209–1223, https://doi.org/10.1175/JTECH-D-13-00205.1.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2017: Comparison of next-day probabilistic severe weather forecasts from coarse- and fine-resolution CAMs and a convection-allowing ensemble. *Wea. Forecasting*, **32**, 1403–1421, https://doi.org/10.1175/WAF-D-16-0200.1.

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

——, ——, and C. D. Karstens, 2020: Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. *Wea. Forecasting*, **35**, 1605–1631, https://doi.org/10.1175/WAF-D-19-0258.1.

Louppe, G., L. Wehenkel, A. Sutera, and P. Geurts, 2013: Understanding variable importances in forests of randomized trees. *Conf. on Neural Information Processing Systems*, Lake Tahoe, CA, Neural Information Processing Systems Foundation.

Lundberg, S. M., and Coauthors, 2019: Explainable AI for trees: From local explanations to global understanding. arXiv, 1905.04610, https://arxiv.org/abs/1905.04610.

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

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

Molnar, C., 2019: Interpretable machine learning. A guide for making black box models explainable. GitHub, accessed 19 July 2021, https://christophm.github.io/interpretable-ml-book/.

Nakanishi, M., and H. Niino, 2004: An improved Mellor–Yamada level-3 model with condensation physics: Its design and verification. *Bound.-Layer Meteor.*, **112**, 1–31, https://doi.org/10.1023/B:BOUN.0000020164.04146.98.

NCEI, 2021: U.S. billion-dollar weather and climate disasters. Accessed 7 December 2021, https://www.ncdc.noaa.gov/billions/.

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

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830, https://www.jmlr.org/papers/v12/pedregosa11a.html.

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

Ribeiro, M., S. Singh, and C. Guestrin, 2016: "Why should I trust you?": Explaining the predictions of any classifier. *Int. Conf. on Knowledge Discovery and Data Mining*, San Francisco, CA, Association for Computing Machinery, 1135–1144, https://doi.org/10.1145/2939672.2939778.

Roberts, B., B. T. Gallo, I. L. Jirak, A. J. Clark, D. C. Dowell, X. Wang, and Y. Wang, 2020: What does a convection-allowing ensemble of opportunity buy us in forecasting thunderstorms? *Wea. Forecasting*, **35**, 2293–2316, https://doi.org/10.1175/WAF-D-20-0069.1.

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

Saabas, A., 2016: Random forest interpretation with scikit-learn. Accessed 25 January 2021, https://blog.datadive.net/random-forest-interpretation-with-scikit-learn/.

Schumacher, R. S., A. J. Hill, M. Klein, J. A. Nelson, M. J. Erickson, S. M. Trojniak, and G. R. Herman, 2021: From random forests to flood forecasts: A research to operations success story. *Bull. Amer. Meteor. Soc.*, **102**, E1742–E1755, https://doi.org/10.1175/BAMS-D-20-0186.1.

Shapley, L. S., 1953: A value for n-person games. *Contributions to the Theory of Games II*, H. Kuhn and A. Tucker, Eds., Princeton University Press, 307–317.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., https://doi.org/10.5065/D68S4MVH.

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

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

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

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

SPC, 2021a: Severe weather event summaries. Accessed 26 March 2021, https://www.spc.noaa.gov/climo/online/.

——, 2021b: Significant hail parameter. Accessed 26 March 2021, https://www.spc.noaa.gov/exper/mesoanalysis/help/help_sigh.html.

——, 2021c: Storm Prediction Center WCM page: Severe weather database files (1950–2019). Accessed 26 March 2021, https://www.spc.noaa.gov/wcm/.

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

Thompson, G., P. R. Field, R. M. Rasmussen, and W. D. Hall, 2008: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: Implementation of a new snow parameterization. *Mon. Wea. Rev.*, **136**, 5095–5115, https://doi.org/10.1175/2008MWR2387.1.

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

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

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

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences.* 3rd ed. Elsevier, 676 pp.