UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

THE CREATION AND ANALYSIS OF NEXT-DAY RANDOM FOREST-BASED HIGH-IMPACT WEATHER FORECASTS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

ERIC DAVID LOKEN
Norman, OK
2021

THE CREATION AND ANALYSIS OF NEXT-DAY RANDOM FOREST-BASED HIGH-IMPACT WEATHER FORECASTS

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Adam Clark, Chair

Dr. Steven Cavallo, Co-Chair

Dr. Amy McGovern

Dr. Xuguang Wang

Dr. Michael Richman

Dr. Andrew Fagg

Acknowledgements

In his best-selling book *Outliers*, Malcolm Gladwell argues that "individual" success is seldom due to the efforts of a single individual. Rather, Gladwell asserts that success results from the confluence of favorable situational, cultural, and social factors that provide abundant opportunities for growth. As a graduate student at the University of Oklahoma (OU), I have had numerous opportunities to grow both as a scientist and a person; these opportunities were made possible by many wonderful people that I wish to acknowledge and thank here.

First and foremost, I thank my advisor, Dr. Adam Clark. Adam and I began working together in the fall of 2015, when I was a new Master's student with little prior research experience. Adam not only showed me how to conduct and communicate peer-reviewed research, he did so with everlasting enthusiasm, compassion, and humor. He taught me how to think creatively, write more effectively, and be a better human being. He also continually pushed me to grow as a scientist, encouraging me to apply for the Visitor Program at the Developmental Testbed Center (DTC) in Boulder, CO in the fall of 2018. I was ultimately accepted into the program and benefitted substantially from my time in Boulder.

Indeed, I owe a debt of gratitude to the DTC Visitor Program and wish to thank all the scientists at the DTC, not only for sharing with me stimulating discussions and insightful feedback, but also for their incredible kindness. I especially thank Jamie Wolff for being a wonderful host, as well as Ryan Sobash, David John Gagne, and Xiang Su for lending me their expertise and scientific insights.

Next, I thank the members of my Ph.D. Committee. Drs. Amy McGovern, Andrew Fagg, Michael Richman, Steven Cavallo, and Xuguang Wang have all provided me with the knowledge and tools necessary to complete this dissertation and have shared their feedback on the projects described herein. I give a special thank you to Amy McGovern, whose machine learning (ML) course inspired much of the contents of this dissertation. Amy's passion for ML has been contagious, and I have learned so much from her lectures and meetings.

Next, I thank my parents, Rhonda and Richard Loken, who not only nurtured my scientific curiosity from an early age, but also played a crucial role in supporting me during the numerous challenges that arose during my Ph.D. work. Without their love and support, this dissertation would not have been possible.

Last but not least, I thank my friends and fellow graduate students who have been quick to share their technical expertise and extend their support, companionship, and kindness. In particular, I thank Jon and Laura Labriola, Bo Huang, Montgomery Flora, Elizabeth Smith, Josh Gebauer, Marcus and Rachel Johnson, Addison Alford, Manda Chasteen, Peter McAward, Tyler Bell, Brian Green, Ryan Lagerquist, Amanda Burke, and Ryan Pajela. Their friendship has helped me negotiate a variety of scientific, personal, and logistical challenges during my time as an OU graduate student, and I am very grateful.

Support for this work was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. Additional support for some components of this dissertation was provided by the Developmental Testbed Center (DTC). The DTC Visitor Program is funded by the National Oceanic and Atmospheric Administration, the National Center for Atmospheric Research, and the National Science Foundation. 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. High-performance

computing was also performed at the OU Supercomputing Center for Education and Research (OSCER) at the University of Oklahoma (OU).

Table of Contents

List of Tables

List of Figures

Abstract

Flash floods, tornadoes, damaging winds, and large hail are costly and difficult to predict, even for state-of-the-art, high-resolution numerical weather prediction (NWP) systems. Current operational NWP ensembles have a variety of shortcomings: they are under-dispersive for precipitation, contain biases in precipitation magnitude and convection placement, have suboptimal forecast reliability, and use horizontal grid-spacing too coarse to explicitly depict some high-impact hazards (e.g., severe hail and tornadoes). Thus, post-processing techniques are required to obtain skillful probabilistic hazard forecasts from raw NWP ensemble guidance. Common post-processing methods include the use of proxies (e.g., climatologically large values of 2-5 km updraft helicity; UH2-5km) to represent simulated high-impact weather events and/or the use of spatially smoothed raw ensemble probabilities to improve forecast reliability. However, these methods use limited data and thus tend to be suboptimal. In this dissertation, I develop and analyze a random forest- (RF-) based procedure for obtaining more skillful precipitation and severe weather probabilistic forecasts for next-day lead times (i.e., 12-36 h forecasts valid from 1200 UTC – 1200 UTC). While past studies have used RFs to better predict high-impact weather, my RF procedure is unique because it uses temporally-aggregated, spatially-upscaled, point-based ensemble forecast predictors over the full contiguous United States (CONUS). This method of generating predictors is relatively simple but skillfully accounts for uncertainties in simulated convection timing and placement.

For precipitation and severe weather hazard prediction, I show that my RF procedure improves forecast reliability and resolution relative to top-performing (human and non-human) baseline forecasts. I find that RF post-processing is most beneficial for convection-parameterizing ensembles (which have more initial biases than convection-allowing ensembles) and more-common events (e.g., lighter precipitation thresholds and severe wind and hail compared to tornadoes). For precipitation, I find that RF-based post-processing reduces spatial biases and note that a season of training data is sufficient to produce skillful probabilistic precipitation forecasts for thresholds up to 3-inches. For severe weather, I show that RF-based forecasts have verification metrics similar to or better than corresponding Storm Prediction Center (SPC) day-1 human forecasts for most hazards, seasons, and regions. By discretizing RF forecast probabilities and making SPC probabilities continuous, I show that this result is only partly due to the ability of RFs to generate continuous forecast probabilities.

Through RF sensitivity tests, I find that ensemble mean (EM) predictors are more skillful than individual member (IM) predictors for severe weather forecasting, since EM predictors contain less noise. By conducting additional sensitivity tests and using the Tree Interpreter (TI) Python module, I find that storm predictors are most important for severe weather prediction, followed by index-based predictors, although I note that RFs using both storm and index-based predictors are most skillful. With TI analysis, I show that RFs emphasize different and physically-relevant predictors for each hazard. Further, I demonstrate that RFs learn to implicitly "weigh" multiple appropriate storm and index variables at and near the point of prediction, suggesting that RFs learn to account for model error. Importantly, my work shows that RFs are not constrained by the exceedance (or non-exceedance) of a simple UH2-5km threshold at one point and suggests that RFs can discern between similar ensemble forecast UH2-5km values as well as the same UH2-5km value in different environments.

Chapter 1: General Introduction

1. Introduction

From 2010 to 2020, 81 severe weather- and 18 flood-related events caused over $1 billion in damage each [consumer price index (CPI) adjusted; NCEI 2021]. Collectively, these events cost $251.2 billion and resulted in 1142 deaths (NCEI 2021), making them highly impactful to society. Yet, floods and severe weather (i.e., tornadoes, damaging wind, and large hail) are difficult to predict, even with high-resolution numerical weather prediction (NWP) models, due to uncertainties in initial conditions and model physics (e.g., Roebber et al. 2004; Palmer 2017).

NWP ensembles—first implemented operationally in the early 1990s (Toth and Kalnay 1993; Tracton and Kalnay 1993)—account for initial condition and model uncertainties and provide users with probabilistic forecast guidance (Roebber et al. 2004; Leutbecher and Palmer 2008; Palmer 2017). However, even current convection-allowing ensembles (CAEs; i.e., ensembles whose members do not use convection parameterization) tend to be under-dispersive (e.g., Romine et al. 2014; Schwartz et al. 2014), making their forecast probabilities suboptimal. Moreover, CAEs and convection-allowing models (CAMs) have biases and spatial displacement errors for simulated storms and precipitation (e.g., Johnson and Wang 2012; Herman and Schumacher 2016). Additionally, current operational CAEs lack horizontal grid-spacing fine enough to explicitly simulate tornadoes, severe hail, or microscale severe wind events. Thus, post-processing techniques are required to obtain the most skillful and useful hazard probabilities from CAEs. These techniques can include spatially smoothing raw ensemble probabilities (Sobash et al. 2011; Loken et al. 2017, 2019; Roberts et al. 2019), bias correction through probability matching (Ebert 2001; Clark et al. 2010a,b; Clark 2017; Loken et al. 2019b), and methods for obtaining neighborhood ensemble probabilities (e.g., Schwartz and Sobash 2017; Blake et al. 2018; Roberts et al. 2019).

In the past 5-10 years, machine learning (ML) has become increasingly viable as an alternative method for ensemble post-processing, due in part to growing computing and data storage capacity (e.g., Hamill et al. 2013; Roberts et al. 2019), high-resolution observational data (e.g., Du 2011) and the availability of sophisticated but easy-to-use open-source software (e.g., Scikit-Learn, Pedegrosa et al. 2011; and Keras, Chollet et al. 2015). Conceptually, ML nonlinearly relates predictors (e.g., CAE forecast variables) with predictands (e.g., observed precipitation or local storm reports) using one of a variety of possible algorithms [e.g., neural networks (e.g., Rajendra et al. 2019); support vector machines (Adrianto et al. 2009; Ortiz-Garcia et al. 2014); genetic algorithms (Kishtawal et al. 2003; Wong et al. 2008); random forests (RFs; e.g., Gagne et al. 2014, 2017; Herman and Schumacher 2018c; Burke et al. 2020); etc.]. Of these algorithms, the RF algorithm has several nice properties that make it especially well-suited for CAE post-processing: it can handle biased predictors, its multiple trees make it resistant to over-fitting (e.g., Gagne et al. 2014), it tends to produce reliable output probabilities (Breiman 2001), and it has relatively few hyper-parameters to tune, making it easy to use. Several recent studies have shown the promise of RFs for precipitation (Gagne et al. 2014; Herman and Schumacher 2018c) and severe weather (Gagne et al. 2017; Burke et al. 2020; Hill et al. 2020) prediction.

Despite the promise of the RF technique, many important questions remain regarding its use: How do RF forecasts compare to top-performing non-ML baselines (including human forecasts)? Does RF-based post-processing benefit all ensembles equally? What is the best way to design predictors for next-day precipitation and severe weather prediction? What relationships do RFs learn between CAE variables and observed severe weather? The purpose of this dissertation is to investigate how RFs can be used (in conjunction with NWP ensembles) to improve high-impact weather forecasts by developing, evaluating, and analyzing RFs for next-day precipitation and severe weather prediction.

2. Research background: A brief history of machine learning and its application to weather forecasting

According to Schmidhuber (2015), the roots of ML can be traced to at least the early 19th century, when Legendre (1805) and Gauss (1809, 1821) developed early linear regression techniques based on the method of least squares. More than a century later, McCulloch and Pitts (1943) published their conception for a neural network (NN) algorithm, which was inspired by the human nervous system. However, early NNs predated backpropagation and thus tended to be inefficient and unstable, making their practical use infrequent (Schmidhuber 2015).

In meteorology, early precursors to ML were linear regression- (LR-) based dynamical-statistical techniques developed in the 1950s (e.g., Malone 1955; Klein et al. 1959) alongside emerging electronic NWP technology (e.g., Charney et al. 1950; Bolin 1955; Bergthorsson et al. 1955). One of the earliest studies was Malone (1955), who used LR to predict 24-h sea-level pressure and surface temperature based on previous-day pressure and temperature. Though he mentioned his results were “not spectacular” (p. 812), he noted that his LR forecasts gave positive skill and seemed useful. Similarly, Klein et al. (1959) used multivariate LR to skillfully forecast 5-day mean surface temperatures from NWP-based 5-day mean 700hPa heights two days earlier and 5-day mean surface temperatures four days earlier.

In the 1960s and 1970s, dynamical NWP models became more sophisticated[1], and these more complex models, in turn, facilitated more skillful dynamical-statistical methods. Building on the work of Klein et al. (1959), Glahn and Lowry (1972) related NWP forecast output with observed variables of interest using multivariate LR. With this method, which they termed Model Output Statistics (MOS), they found they could skillfully predict variables such as probability of precipitation, surface wind, surface temperature, and cloudiness. In doing so, they laid the groundwork for future, ML-based dynamical post-processing methods.

Automated methods for growing decision trees—which would eventually become the basis for the random forest (RF; Breiman 2001)—were developed in the 1970s and 1980s (Hssina et al. 2014). The Iterative Dichotomiser 3 (ID3; Quinlan 1979, 1983, 1986) was an early automated decision tree algorithm that used information theory to classify nominal data. While ID3 did not support missing or continuous data, it served as a basis for future decision tree algorithms. For example, Classification and Regression Trees (CART; Breiman 1984) enabled continuous predictors and supported regression-based prediction tasks in addition to classification. Unlike ID3, CART determined the splitting criteria at each node by maximizing the reduction in Gini Index after each split. Eventually, Quinlan (1993) developed C4.5, an updated version of ID3 that supported continuous data, missing and differently-weighted predictors, and tree pruning after creation (Hssina et al. 2014).

[1] For example, Bushby and Timpson (1968)'s 40-km grid-spacing model integrated the inviscid, frictionless, and hydrostatic equations of motion and considered moisture and latent heating—a substantial improvement from Charney et al. (1950)'s hindcasts, which were obtained from integrating the barotropic vorticity equations on a grid with 736-km horizontal spacing.

In weather forecasting, the use of manually-designed decision trees predated the development of the automated ID3, CART, and C4.5 algorithms. These manual decision trees were used as forecast aids at least as early as the mid-1970s, as Dvorak (1975) developed a decision tree for forecasting tropical cyclone intensity from satellite images. Similarly, Colquhoun (1987) conceived of a decision tree-type algorithm to aid with forecasting thunderstorms, severe thunderstorms, and tornadoes based on physical principles. Mills and Colquhoun (1998) extended the decision tree framework in Colquhoun (1987) to use NWP forecast variables for predicting areas of thunderstorms, severe thunderstorms, and tornadic thunderstorms as well as their hazards (e.g., flash floods, downbursts, strong winds, etc.). They found that, while their algorithm could not predict individual thunderstorms, it could help alert forecasters to areas of severe and tornadic thunderstorm potential based on NWP model output. While the "decision trees" in Dvorak (1975), Colquhoun (1987), and Mills and Colquhoun (1998) did not use the ML algorithms developed by Quinlan (1979, 1986, 1993) or Breiman (1984), they showed the promise of tree-based rule systems for NWP-based high-impact weather prediction. Thus, these studies helped pave the way for ML-related tree-based techniques to take hold in meteorology in the late 2000s and beyond.

Modern ML-based NN technology advanced substantially during the 1980s and 1990s, as backpropagation—first described by Linnainmaa 1970—was rediscovered and used for efficiently training NNs (e.g., Werbos 1981; Parker 1985; LeCun 1988). Naturally, this innovation led to a wide range of meteorological applications for NNs, including their use for the prediction of tornadoes (Marzban and Stumpf 1996), precipitation (e.g., Hall et al. 1999), and severe hail size (Marzban and Witt 2001). However, while NNs during this time showed promise, their skill was constrained by limited computing resources and a lack of large labelled datasets for training (e.g., Schultz et al. 2021).

Indeed, the modern form of another ML algorithm, the support vector machine (SVM; Vapnik and Cortes 1995), was developed during the mid-1990s and tended to compare favorably with NNs through the mid-2000s (e.g., Chollet 2018). For example, Liong and Sivapragasam (2002) compared SVMs with NNs for flood forecasting and found that the SVMs had better performance while being simpler and more interpretable. Thus, SVMs were implemented for a variety of meteorological applications, including the prediction of tornadoes (e.g., Trafalis et al. 2003, 2004, 2005; Adrianto et al. 2009), typhoon rainfall (Lin et al. 2009), and floods (e.g., Han et al. 2007).

In the mid-2000s, RFs began attracting the attention of the meteorological community, as Deloncle et al. (2007) used the technique to predict weather regime transitions. Shortly thereafter, Gagne et al. (2009) used k-means clustering with decision trees for storm type classification from simulated radar data.

In the 2010s, greater computing power, the advent of graphical processing units (GPUs), and wider availability of large historical datasets allowed for ML techniques to become more complex and skillful (Schultz et al. 2021). Indeed, these advances have recently enabled NNs to be run with many hidden layers in an approach known as deep learning (DL; e.g., Schultz et al. 2021). In the past 5 years, DL has achieved superhuman performance in a variety of domains including chess, shogi, and go (Silver et al. 2018). In meteorology, DL has been used to obtain skillful results in a variety of tasks ranging from synoptic-scale front detection (Lagerquist et al. 2019) to short-term tornado prediction (Lagerquist et al. 2020) to satellite-based prediction of intense convection (Cintineo et al. 2020).

Classic ML methods, such as the RF, have also benefited from the greater computing and data storage capacity. Indeed, RFs have recently demonstrated substantial skill for a variety of prediction problems, including aviation turbulence (Williams 2014), initiation of mesoscale convective systems (MCSs; Ahijevych et al. 2016), subfreezing road temperatures (Handler et al. 2020), precipitation (Gagne et al. 2014; Herman and Schumacher 2018c), and severe weather (Gagne et al. 2017; Hill et al. 2020; Burke et al. 2020; Flora et al. 2021). Most of these studies used predictors from NWP forecast data, although they used different NWP models (or ensembles) and generated predictors in different ways. For example, Gagne et al. (2014) trained their RFs using CAE forecast variables from a random subset of grid points, while Gagne et al. (2017), Burke et al. (2020), and Flora et al. (2021) generated predictors based on simulated CAE storm objects. Meanwhile, Herman and Schumacher (2018c) and Hill et al. (2020) used convection-parameterizing ensemble predictors at surrounding grid points, with Herman and Schumacher (2018c) utilizing principal component analysis (e.g., Wilks 2011) to limit the feature space and reduce the correlation between predictors. While these studies all achieved a fair degree of skill, the optimal design and specific performance characteristics of RF-based algorithms remain unknown. For example, it is unclear how RF-based post-processing benefits convection-parameterizing vs. convection-allowing ensembles, how much data is required to achieve sufficient forecast skill, how RF-based forecasts compare to top-performing human and automated baselines, and how best to generate RF predictors from ensemble data. Moreover, while Breiman (2001) and others (e.g., McGovern et al. 2019b) have suggested techniques for interpreting RF output, very few previous studies have specifically focused on dissecting the relationships learned by skillful RFs. This dissertation seeks to fill these knowledge gaps for next-day precipitation and severe weather prediction.

3. Research questions and hypotheses

Three research components have been implemented to meet the goal of this dissertation, which is to explore the use of RFs for next-day precipitation and severe weather prediction. In the first component, RFs are designed to predict probability of precipitation at four thresholds from 0.1- to 3-inches. Two RFs are trained for each threshold: one using predictors from the convection-parameterizing Short-Range Ensemble Forecast System (SREF; Du et al. 2015) and one using predictors from the convection-allowing High-Resolution Ensemble Forecast System, Version 2 (HREFv2; Jirak et al. 2018; Roberts et al. 2019). RF forecasts are compared against each other as well as raw and spatially-smoothed ensemble probabilistic precipitation forecasts. The dataset consists of 496 days from April 2017 to November 2018. The primary research questions (Q1.1 – Q1.3) associated with the first component are:

> *Q1.1: What (if any) benefits does RF-based post-processing provide relative to spatially smoothing raw ensemble probabilities (i.e., a top non-ML post-processing method) for different precipitation thresholds from 0.1- to 3-inches?*
>
> *Q1.2: Does RF post-processing benefit CAEs or convection-parameterizing ensembles more?*
>
> *Q1.3: How much training data is required to generate useful RF probabilistic precipitation forecasts at thresholds from 0.1- to 3-inches?*

The hypotheses (H1.1 – H1.3) corresponding to the above research questions are as follows:

> *H1.1: RF-based probabilistic precipitation forecasts will have reduced spatial biases as well as better discrimination ability, sharpness, and resolution compared to spatially-smoothed ensemble probabilities. RF probabilities will provide the greatest benefits relative to spatially smoothed ensemble probabilities at the smallest thresholds, which are climatologically most common.*

> *H1.2: RF post-processing will benefit a convection-parameterizing ensemble more than a CAE due to the greater initial bias of the convection-parameterizing ensemble. Indeed, after RF post-processing, a convection-parameterizing ensemble will have better reliability and nearly comparable resolution compared to an un-post-processed (i.e., raw) CAE forecast. However, post-processed CAE forecasts (from either the RF or spatial smoothing method) will be the most skillful due to the enhancement of CAE reliability and resolution. In accordance with H1.1, RF CAE forecasts will be more skillful than spatially smoothed CAE forecasts.*

> *H1.3: Approximately one year of training data will be required to obtain skillful RF-based precipitation forecasts for the 3-inch forecasts, with less data required as the threshold decreases.*

H1.1 and H1.2 are tested using standard verification metrics for probabilistic forecasts, including: area under the relative operating characteristics curve (AUC; e.g., Wilks 2011), Brier

Score (BS) components (Wilks 1995), Brier Skill Score (BSS; Wilks 1995), performance diagrams (Roebber 2009), and attributes diagrams (Hsu and Murphy 1986). Spatial biases are assessed using a method based on Clark et al. (2010a) and Marsh et al. (2012). H1.3 is tested by re-training RFs with varying subsets of the original training dataset.

In the second research component, RFs are designed to predict severe weather hazards based on data from the Storm-Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012) over 629 days from April 2015 to July 2017. The primary research questions (Q2.1 and Q2.2) associated with the second research component are as follows:

Q2.1: How do next-day RF-based probabilistic severe weather hazard forecasts compare to corresponding Storm Prediction Center (SPC) human forecasts and calibrated, spatially-smoothed 2-5km updraft helicity (UH2-5km) forecasts?

Q2.2: In what seasons and regions do the RF severe weather hazard forecasts perform best?

The corresponding hypotheses (H2.1 and H2.2) are:

H2.1: For all severe and significant severe weather hazards (including any-severe and any-significant-severe categories), RFs will have better discrimination ability, BSS, reliability, and resolution than corresponding calibrated UH2-5km-based forecasts. However, RFs will have worse discrimination ability, BSS, and resolution than corresponding (discrete and continuous) SPC human forecasts. Continuous RFs will *have better reliability and resolution than discrete (i.e., binary) SPC significant severe hazard forecasts, but discrete RF forecasts will not perform better than discrete significant severe SPC forecasts.*

H2.2: The RF forecasts will perform best in the seasons and locations for which severe weather climatological frequency is maximized. For tornadoes and severe hail, this is expected to be the central U.S. during the spring and summer. For severe wind, this is expected to be the eastern U.S. during the summer.

H2.1 and H2.2 are tested using similar verification metrics as described in the first research component (i.e., AUC, BS components, BSS, performance and attributes diagrams). RF forecasts are compared to SSEO hazard-calibrated, spatially smoothed UH2-5km and (discrete and continuous) 0600 UTC SPC human forecasts. H2.2 is addressed by stratifying forecasts by region and season.

In the third research component, differently-configured tornado-, severe wind-, and severe hail-predicting RFs are created from High-Resolution Ensemble Forecast System, Version 2.1 (HREFv2.1; Roberts et al. 2020) data. The dataset includes 653 days from April 2018 to May 2020. The primary research questions (Q3.1 – Q3.3) associated with the third component are as follows:

Q3.1: When using an RF to forecast next-day severe weather hazards, does greater forecast skill result from using individual member predictors at a single grid point or ensemble mean predictors at multiple spatial points?

Q3.2: What CAE predictors do RFs emphasize to make probabilistic next-day severe weather hazard forecasts?

Q3.3: What multi- and single-variate relationships do RFs learn to make skillful severe weather hazard forecasts?

The corresponding hypotheses (H3.1 – H3.3) associated with the third component are:

H3.1: Greater forecast skill will result from providing an RF with individual member predictors at a single grid point.

H3.2: RFs will emphasize storm variables, but index and environment variables will also be important since simulated storms (and their attributes) do not always correspond with observed storms.

H3.3: RFs will learn to emphasize different variables for each hazard (e.g., significant hail parameter [SHIP; SPC2021b] and UH2-5km for severe hail; significant tornado parameter [STP; Thompson et al. 2012] and 0-3 km updraft helicity [UH0-3km] for tornadoes). For all hazards, RFs will learn positive—but nonlinear—relationships between many storm variables (e.g., UH2-5km, simulated reflectivity, maximum upward vertical velocity, etc.) and observed severe weather probability. Indeed, it is hypothesized that many of these variables will have an "S-shaped" relationship with severe weather *probability. However, RFs are also expected to learn (and use) important relationships between multiple variables/predictors and observed severe weather.*

H3.1 is tested using similar verification metrics used in the first two research components (e.g., AUC, BSS, attributes diagrams, and performance diagrams). H3.2 and H3.3 are tested using the Tree Interpreter Python module (Saabas 2016).

4. Dissertation organization

The research questions outlined in the preceding section are addressed in three papers, with each paper assigned to a distinct dissertation chapter. Chapter 2 contains the paper associated with the first research component, *Postprocessing next-day ensemble probabilistic precipitation forecasts using random forests*, which has been accepted by *Weather and Forecasting*. Chapter 3 comprises the second paper, *Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests*, which has also been accepted by *Weather and Forecasting*. Chapter 4 presents the third paper, *Comparing and interpreting differently-designed random forests for next-day severe weather hazard prediction*, which will be submitted to *Weather and Forecasting*. Finally, Chapter 5 provides a general discussion of all three research components with respect to the research questions and hypotheses posed above, summarizes the key lessons from this dissertation, and offers suggestions for future work.

Chapter 2: Post-Processing Next-Day Ensemble Probabilistic Precipitation Forecasts Using Random Forests

A paper published in *Weather and Forecasting*

Eric D. Loken[1,2,3], Adam J. Clark[2,3], Amy McGovern[2], Montgomery Flora[1,2,3], Kent Knopfmeier[1,3]

[1]Cooperative Institute for Mesoscale Meteorological Studies, The University of Oklahoma, Norman, Oklahoma
[2]School of Meteorology, University of Oklahoma, Norman, Oklahoma
[3]NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

Abstract

Most ensembles suffer from under-dispersion and systematic biases. One way to correct for these shortcomings is via machine learning (ML), which is advantageous due to its ability to identify and correct nonlinear biases. This study uses a single random forest (RF) to calibrate next-day (i.e., 12–36-h lead-time) probabilistic precipitation forecasts over the contiguous United States (CONUS) from the 16-km grid-spacing Short-Range Ensemble Forecast System (SREF) and the 3-km grid-spacing High-Resolution Ensemble Forecast Version 2 (HREFv2). Random forest forecast probabilities (RFFPs) from each ensemble are compared against raw ensemble probabilities over 496 days from April 2017 – November 2018 using 16-fold cross validation. RFFPs are also compared against spatially-smoothed ensemble probabilities since the raw SREF and HREFv2 probabilities are overconfident and under-sample the true forecast probability density function. Probabilistic precipitation forecasts are evaluated at four precipitation thresholds ranging from 0.1-in. to 3-in.

In general, RFFPs are found to have better forecast reliability and resolution, fewer spatial biases, and significantly greater Brier Skill Scores and areas under the relative operating characteristic curve compared to corresponding raw and spatially-smoothed ensemble probabilities. The RFFPs perform best at the lower thresholds, which have a greater observed climatological frequency. Additionally, the RF-based post-processing technique benefits the SREF more than the HREFv2, likely because the raw SREF forecasts contain more systematic biases than those from the raw HREFv2. It is concluded that the RFFPs provide a convenient, skillful summary of calibrated ensemble output and are computationally feasible to implement in real-time. Advantages and disadvantages of ML-based post-processing techniques are discussed.

1. Introduction

Over the past 20 years, increases in computing resources have reshaped the state of numerical weather prediction (NWP) in several key ways: by enabling skillful high-resolution ensemble forecasts (e.g., Xue et al. 2007; Jirak et al. 2012, 2016, 2018; Roberts et al. 2019; Clark et al. 2018; Schwartz et al. 2015, 2019); by increasing the capacity to run and store models for research and operations (e.g., Hamill and Whitaker 2006; Kain et al. 2010; Hamill et al. 2013; Clark et al. 2018; Roberts et al. 2019); and by reducing the time required to perform complex analyses, enabling more—and more frequent—high-resolution NWP products (e.g., Kain et al. 2010; Gallo et al. 2017, 2019; Roberts et al. 2019). These changes have led to large improvements in NWP quality and value, particularly for fields related to convection. For example, the higher resolution associated with convection-allowing models (CAMs; i.e., those that explicitly simulate convection and run with horizontal grid-spacing ≤ ~4-km) has improved forecasts of storm initiation, evolution, and mode compared to convection-parameterizing models (e.g., Kain et al. 2006). Meanwhile, convection-allowing ensembles (CAEs) provide further benefits by accounting for uncertainties in initial conditions and/or model physics (e.g., Roebber et al. 2004; Leutbecher and Palmer 2008; Clark et al. 2009) and conveying forecast uncertainty information to the user (e.g., Palmer 2017). Despite ensembles' higher computational cost, their benefits have been well documented at both convection-parameterizing (e.g., Epstein 1969; Leith 1974; Du et al. 1997; Stensrud et al. 1999; Wandishin et al. 2001; Bright and Mullen 2002; Clark et al. 2009) and convection-allowing (e.g., Coniglio et al. 2010; Loken et al. 2017; Schwartz et al. 2017) resolutions.

Nevertheless, CAMs and CAEs still have biases in the placement, timing, and magnitude of precipitation-producing weather systems (e.g., Davis et al. 2006; Kain et al. 2008; Weisman et al. 2008; Herman and Schumacher 2016, 2018c). Additionally, CAEs remain relatively expensive to run and thus typically have small ensemble membership (e.g., Schwartz et al. 2014; Clark et al. 2018). While small ensembles (e.g., consisting of 10-30 members) have been found to deliver nearly as much forecast skill as larger ensembles (e.g., up to 50 members; Clark et al. 2011; Schwartz et al. 2014; Sobash et al. 2016b), they can under-sample the forecast probability density function (PDF; e.g., Schwartz et al. 2010, 2014; Roberts et al. 2019), potentially leading to degraded reliability and under-dispersion, especially in the absence of neighborhood evaluation or post-processing methods (Schwartz et al. 2014). Indeed, most CAMs and CAEs are currently under-dispersive (e.g., Romine et al. 2014). One method to increase CAE spread is to increase the diversity of the ensemble membership, which can be achieved by using members with multiple dynamic cores, analyses, boundary layer schemes, microphysics parameterizations, and initialization periods (e.g., the Storm-Scale Ensemble of Opportunity; Jirak et al. 2012, 2016; and the High-Resolution Ensemble Forecast System, Version 2; Jirak et al. 2018; Roberts et al. 2019). While diverse, informally-designed ensembles can produce skillful forecasts (Jirak et al. 2016; Jirak et al. 2018; Clark et al. 2018; Schwartz et al. 2019), their skill comes with several notable drawbacks. One is that the ensemble members tend to cluster around multiple solutions based on their dynamic core (e.g., Schwartz et al. 2019). This member clustering can cause the ensemble mean forecast to fall outside of the clusters of member solutions (see Fig. 1 in Schwartz et al. 2019) and can adversely affect the quality of the ensemble probabilities, since each member's solution is not equally likely to occur (Schwartz et al. 2019). Another potential consequence of multi-model, multiple-physics CAEs is an artificial inflation of ensemble spread due to the existence of systematic biases between ensemble members (Eckel and Mass 2005; Clark et al. 2010b; Loken et al. 2019b). These shortcomings are typically resolved using one or more post-processing techniques, including isotropic (e.g., Sobash et al. 2011, 2016b; Loken 2017, 2019; Roberts et al. 2019) or anisotropic (e.g., Marsh et al. 2012) spatial smoothing of the raw forecast probability field, recalibration of forecast probabilities (e.g., Hamill et al. 2008), probability matching techniques (e.g., Ebert 2001; Clark et al. 2010a,b, 2017; Loken et al. 2019b), and various neighborhood-based methods to construct ensemble probabilities (e.g., Schwartz et al. 2010; Blake et al. 2018; Roberts et al. 2019; Schwartz and Sobash et al. 2017).

Another avenue for post-processing is machine learning (ML; e.g., McGovern et at. 2017). Conceptually, ML algorithms identify patterns in historical data and use these patterns to correct for systematic ensemble biases. This idea is not new; dynamical-statistical methods have existed since at least the 1950s (e.g., Malone 1955; Klein et al. 1959). One example of a well-performing traditional technique is Model Output Statistics (MOS; Glahn and Lowry 1972), which relates NWP output to observed variables of interest (e.g., observed precipitation). ML-based post-processing methods work similarly; however, while MOS techniques tend to be based on linear regression (e.g., Glahn and Lowry 1972), ML techniques are not necessarily linear. A variety of ML approaches, other than regression, have been applied to weather prediction since the 1980s and include: artificial neural networks (ANNs; e.g., Key et al. 1989; Marzban and Stumpf 1996; Kuligowski and Barros 1998; Hall et al. 1999; Manzato 2007; Rajendra et al. 2019), support vector machines (e.g., Ortiz-Garcia et al. 2014; Adrianto et al. 2009), clustering algorithms (e.g., Baldwin et al. 2005), genetic algorithms (e.g., Szpiro 1997; Kishtawal et al. 2003; Wong et al. 2008), and decision tree-based methods (Breiman 1984, 2001; Herman and Schumacher 2018c).

Although the ML algorithms mentioned above are not "new"—the random forest (RF) technique utilized herein was described nearly 20 years ago by Breiman (2001)—enhanced computing power and storage capacity have facilitated the successful application of ML to NWP in recent years (e.g., McGovern et al. 2017 and works cited therein). Indeed, as computing power and storage continue to increase, the role ML plays in NWP post-processing is likely to grow as well. Especially as forecasters confront an ever-increasing deluge of data (e.g., Carley et al. 2011; McGovern et al. 2017; Karstens et al. 2018), ML or other post-processing techniques may be desired to quickly and effectively summarize information from NWP products. Therefore, this paper seeks to address important basic questions regarding the application of ML techniques in general—and the RF algorithm in particular—to NWP post-processing. Considerations include what, if anything, a ML approach provides relative to simpler forms of post-processing (e.g., 2-dimensional spatial smoothing) and how feasible it would be to implement ML-based predictions operationally. Specifically, the costs and benefits of an RF-based approach are considered relative to 2-dimensional isotropic spatial smoothing for two multi-model, multi-analyses, multi-physics ensembles: the convection-parameterizing Short-Range Ensemble Forecast System (SREF; Du et al. 2015) and the convection-allowing High-Resolution Ensemble Forecast System, Version 2 (HREFv2; Jirak et al. 2018; Roberts et al. 2019). A focus on precipitation is adopted herein due to its importance as a sensible weather field related to convection and the high economic and human impacts of heavy-precipitation events (e.g., NCEI 2019). The next-day (i.e., 1200 UTC – 1200 UTC) time frame is selected due to its relative simplicity and to match operational Day 1 products issued by the Weather Prediction Center (WPC).

The remainder of this paper is organized as follows: section 2 details the methods and datasets used herein, section 3 describes the results and presents two case studies for analysis, section 4 summarizes and discusses important findings, and section 5 concludes the paper and outlines avenues for future work.

2. Methods

a. Datasets

Forecast data from the SREF and HREFv2 are considered over 496 common days, spanning April 2017 – November 2018 (Table 2.1).

Month	**2017**	**2018**
January	-	1-31
February	-	1-28
March	-	1-10, 14-17, 19-20, 22-26
April	28	7-30
May	1-2, 4-5, 7-10, 13-23, 26-31	1-31
June	1, 6-7, 9, 11-13, 15, 17-25	1-7, 10-30
July	3-6, 15-16, 18-19, 22-24, 30-31	1-31
August	1-10, 12-15, 17-30	1-5, 8-31
September	1-10, 13-15, 17-30	1-30
October	1-24, 26-31	1-31
November	1-30	1-4, 6, 9-13
December	1-31	-

Table 2.1 Forecast valid dates for each ensemble.

The analysis domain for both ensembles covers the contiguous United States (CONUS; Fig. 2.1), and the analysis period covers 24-h (1200 UTC – 1200 UTC the next day). Details on each ensemble's configuration are given below.

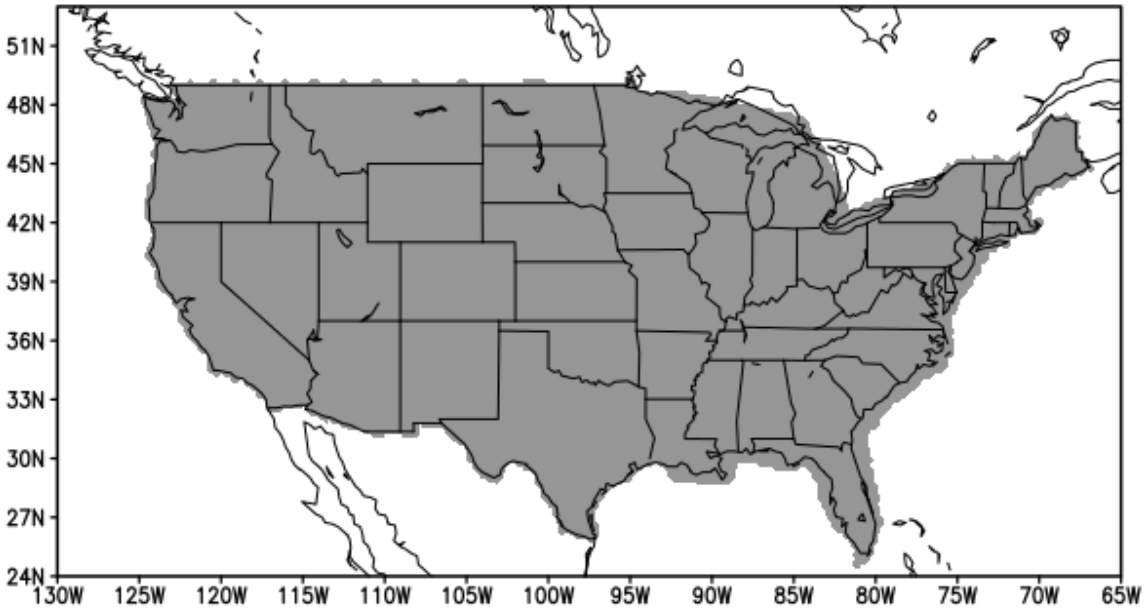


Figure 2.1 Analysis domain for each ensemble.

The SREF is a 26-member convection-parameterizing ensemble in which half of the members use the Advanced Research Weather Research and Forecasting (WRF-ARW; Skamarock et al. 2008) dynamic core and half use the dynamic core from the Nonhydrostatic Multiscale Model on the B grid (NMMB; Janjić and Gall 2012). The SREF uses 16-km horizontal grid-spacing and runs four cycles per day at 0300-, 0900-, 1500-, and 2100-UTC (Du et al. 2015), with forecast fields output every 3 hours. This study uses 15-39-h forecasts from the 2100 UTC initialization. Due to storage and data availability constraints, the SREF analyses herein are output to a grid with 32-km horizontal grid-spacing (NCEP grid 221). SREF configuration details are summarized in Table 2.2.

Member	ICs	LBCs	Conv.	PBL	Microphys.
arw_ctl	RAP	GFS	KF	YSU	WSM6
arw_p1	RAP	GEFS13	Grell	MYNN	Thompson
arw_n1	RAP	GEFS14	BMJ	MYJ	Ferrier
arw_p2	RAP	GEFS15	BMJ	MYJ	Thompson
arw_n2	RAP	GEFS16	KF	YSU	Ferrier
arw_p3	GFS	GEFS17	KF	YSU	Thompson
arw_n3	GFS	GEFS18	Grell	MYNN	WSM6
arw_p4	GFS	GEFS19	KF	YSU	Ferrier
arw_n4	GFS	GEFS20	BMJ	MYJ	WSM6
arw_p5	NDAS	GEFS1	KF	YSU	WSM6
arw_n5	NDAS	GEFS2	Grell	MYNN	Ferrier
arw_p6	NDAS	GEFS3	Grell	MYNN	Thompson
arw_n6	NDAS	GEFS4	BMJ	MYJ	Thompson
nmmb_ctl	NDAS	GFS	BMJ (old shal)	MYJ	Ferrier hi-res
nmmb_p1	NDAS	GEFS1	BMJ (new shal)	MYJ	Ferrier hi-res
nmmb_n1	NDAS	GEFS2	SAS	GFS	WSM6
nmmb_p2	NDAS	GEFS3	BMJ (old shal)	MYJ	WSM6
nmmb_n2	NDAS	GEFS4	SAS	GFS	Ferrier hi-res
nmmb_p3	GFS	GEFS5	BMJ (new shal)	MYJ	WSM6
nmmb_n3	GFS	GEFS6	SAS	GFS	Ferrier hi-res
nmmb_p4	GFS	GEFS7	BMJ (old shal)	MYJ	Ferrier hi-res
nmmb_n4	GFS	GEFS8	SAS	GFS	WSM6
nmmb_p5	RAP	GEFS9	BMJ (new shal)	MYJ	Ferrier hi-res
nmmb_n5	RAP	GEFS10	SAS	GFS	WSM6
nmmb_p6	RAP	GEFS11	BMJ (old shal)	MYJ	WSM6
nmmb_n6	RAP	GEFS12	SAS	GFS	Ferrier hi-res

Table 2.2 SREF member specifications, adapted from Du et al. (2015). Initial conditions (ICs) are taken from the operational Rapid Refresh (RAP; Benjamin et al. 2016), the National Center for Environmental Prediction's (NCEP's) Global Forecast System (GFS), and the North American Mesoscale Model Data Assimilation System (NDAS). IC perturbations are derived using a blend of Global Ensemble Forecast System (GEFS) and SREF analyses. Lateral boundary conditions (LBCs) are from the GFS and GEFS members. Convective parameterizations include the Kain-Fritsch (KF; Kain 2004), Grell (1993), Betts-Miller-Janjic (BMJ; Betts 1986; Janjic 1994), and simplified Arakawa-Schubert (Han and Pan 2011) schemes. Planetary boundary layer (PBL) schemes include the Yonsei University (YSU; Hong et al. 2006), Mellor-Yamada-Nakanishi-Niino (MYNN; Nakanishi and Niino 2004, 2006), Mellor-Yamada-Janjić (MYJ; Janjić 2002) parameterizations as well as that used in the GFS. Microphysics schemes include the WRF single-moment 6-class (WSM6; Hong and Lim 2006), Thompson et al. (2004), and Ferrier et al. (2002) schemes.

The HREFv2 originates from the Storm Prediction Center's Storm-Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012, 2016, 2018), which was developed as a collection of individual CAMs with different dynamic cores, analyses, initialization times, microphysics, and boundary layer parameterizations. Although the HREFv2 and SSEO use "ad-hoc," informal designs, they have consistently outperformed other CAEs (Jirak et al. 2016, 2018; Schwartz et al. 2019). Indeed, the strong performance of the HREFv2 led to its implementation as the National Weather Service's first operational CAE on 1 November 2017 (Jirak et al. 2018; Roberts et al. 2019). Despite the drawbacks arising from its informal design (e.g., unequal likelihood, member clustering, maintenance difficulties; Schwartz et al. 2019), it remains a "high-quality-baseline" (Schwartz et al. 2019) for CAE performance.

The HREFv2 comprises 8 members, with half the membership composed of 12-h time lagged runs (Jirak et al. 2018; Roberts et al. 2019). The non-lagged (time-lagged) members are initialized daily at 0000 UTC (the previous day at 1200 UTC). All members use approximately 3-km horizontal grid-spacing and collectively contain two dynamic cores, two microphysics schemes, and two boundary layer parameterizations. Forecast fields are output hourly from each member. 12-36-h HREFv2 forecasts are used herein. Full details of HREFv2 configuration are given in Table 2.3.

National Center for Atmospheric Research/Earth Observing Laboratory (NCAR/EOL) Stage IV precipitation data (Lin 2011) are used for observations. While the dataset has known deficiencies, especially in regions of complex terrain where radar coverage is sparse and/or inaccurate (e.g., Hitchens et al. 2013; Herman and Schumacher 2016; Herman and Schumacher 2018b), the dataset has high-resolution (~4.8-km grid-spacing) coverage over the full CONUS, making it the preferred observational dataset.

Member	Model Core	IC/LBCs	Microphysics	PBL
HRW NSSL	WRF-ARW	NAM/NAM -6h	WSM6	MYJ
HRW NSSL -12h	WRF-ARW	NAM/NAM -6h	WSM6	MYJ
HRW ARW	WRF-ARW	RAP/GFS -6h	WSM6	YSU
HRW ARW -12h	WRF-ARW	RAP/GFS -6h	WSM6	YSU
HRW NMMB	NMMB	RAP/GFS -6h	Ferrier-Aligo	MYJ
HRW NMMB -12h	NMMB	RAP/GFS -6h	Ferrier-Aligo	MYJ
NAM CONUS Nest	NMMB	NAM/NAM	Ferrier-Aligo	MYJ
NAM CONUS Nest -12h	NMMB	NAM/NAM	Ferrier-Aligo	MYJ

Table 2.3 HREFv2 member specifications. HRW and NAM refer to High Resolution Window and North American Mesoscale Model runs, respectively. The "-12h" in the first column indicates a 12-h time lagged member (i.e., 1200 UTC initialization the previous day instead of 0000 UTC initialization). Initial conditions and lateral boundary conditions (IC/LBCs) are taken from the NAM, Rapid Refresh (RAP), and/or Global Forecast System (GFS), as indicated. A "-6h" indicates that the model from which the IC/LBCs are derived was initialized 6-h before the given HREFv2 member. Microphysics schemes include the WRF single-moment 6-class (WSM6; Hong and Lim 2006) and the Ferrier-Aligo (Aligo et al. 2018) schemes, while boundary layer parameterizations include the Mellor-Yamada-Janjić (MYJ; Janjić 2002) and Yonsei University (YSU; Hong et al. 2006) schemes.

b. Obtaining raw and spatially smoothed ensemble forecasts

Raw SREF and HREFv2 forecast probabilities are computed by first remapping each member's 24-h (1200 UTC – 1200 UTC) quantitative precipitation forecast to NCEP grid 215, which has approximately 20-km horizontal grid-spacing. The remapping is done using a neighbor budget method (Accadia et al. 2003), a nearest-neighbor averaging method that approximately conserves total precipitation. Upscaling to 20-km saves significant computational expense and better matches scales at which predictability should exist at 12-36-h lead times. After upscaling, the fraction of ensemble members exceeding a given precipitation threshold is calculated at each point on the 20-km grid. Four 24-h precipitation thresholds are considered: 0.1-, 0.5-, 1-, and 3-in (i.e., 2.54-, 12.7-, 25.4-, and 76.2-mm).

Given the under-dispersive properties of most CAEs, a 2-dimensional, isotropic Gaussian kernel density function (e.g., Sobash et al. 2011, 2016b; Loken et al. 2017, 2019; Roberts et al. 2019) is often applied to a CAE's raw forecast probability field as a simple but effective means of increasing forecast spread and reducing over-forecasting bias. Since most CAEs are overconfident and under-dispersive, spatial smoothing typically enhances reliability and resolution, but over-smoothing can degrade reliability and sharpness (Sobash et al. 2011, 2016b; Loken et al. 2017, 2019; Roberts et al. 2019). In this study, as in Loken et al. (2019), the following equation is applied to the (remapped) SREF and HREFv2 raw ensemble forecast probabilities to create isotropic spatially smoothed forecast probabilities:

$$\mathrm{f} = \sum_{\mathrm{n}=1}^{\mathrm{N}} \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2}\left(\frac{\mathrm{d_n}}{\sigma}\right)^2\right] \qquad (2.1),$$

where f is the forecast probability at a given point, N is the number of points where at least one ensemble member exceeds the given precipitation threshold, d_n is the distance from the current point to the *n*th point, and σ is the standard deviation of the Gaussian kernel. Importantly, σ controls the degree of spatial smoothing and must be tuned appropriately to produce skillful forecasts. Herein, σ is chosen such that the resulting collection of daily, CONUS-wide forecast probabilities minimizes the Brier Score (BS; e.g., Wilks 1995) over the training dataset. The BS can be expressed as:

$$\mathrm{BS} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{f_i} - \mathrm{o_i})^2 \qquad (2.2),$$

where N is the total number of forecast-observation pairs (i.e., the number of grid points in the domain multiplied by the number of days in the dataset), $\mathrm{f_i}$ is the forecast probability at the *i*th grid point, and $\mathrm{o_i}$ is the binary observation at the *i*th grid point.

c. Random forest-based forecasts

While the umbrella of machine learning includes many popular and powerful algorithms, the random forest (RF; Breiman 2001) algorithm has some important advantages that make it the preferred technique in this study. Namely, RFs do not require standardized inputs, they have relatively few hyper-parameters to tune, they are parallelizable and thus relatively fast to run, and previous studies (e.g., Gagne et al. 2014; Herman and Schumacher 2018a,c) have found that they perform well for precipitation prediction.

The building blocks of RFs are individual decision trees (Breiman 1984). Decision trees recursively split a dataset by selecting, at each node, the variable and threshold that maximizes a dissimilarity metric (e.g., information gain) until a stopping criterion is reached (e.g., the number of dataset samples falls below a specified amount, the tree reaches a certain depth, etc.). Once the splitting criteria are determined for each node using the training data, the tree can be used for prediction on a testing dataset by sorting testing samples through the tree. Testing probabilities are given by the fraction of training samples associated with an observed event of interest at the terminal node, or “leaf node,” into which a testing sample is classified. One drawback of individual decision trees is that they tend to be overly sensitive to small variations in the training dataset (e.g., Gagne et al. 2014). RFs provide a solution to this so-called “brittleness” (Gagne et al. 2014) by growing multiple trees, which are unique due to the introduction of stochasticity into the training process. Specifically, each tree in the RF uses a subset of training samples determined by bootstrap resampling (i.e., resampling with replacement; e.g., Wilks 1995) the full set, and splits at each node are determined by considering a random subset of variables. In the RF framework, testing probabilities of event occurrence are simply the mean testing probabilities from each tree. Although the RF’s multiple trees may make it more difficult for humans to interpret RF output probabilities, the RF method is generally attractive since it is resistant to overfitting and tends to produce outputs with low bias (e.g., Breiman 2001). More details on the RF technique can be found in Herman and Schumacher (2018c), McGovern et al. (2017), and Gagne et al. (2014).

Herein, 18 (20) fields are used as inputs into the RF algorithm to obtain SREF (HREFv2) RFFPs (Table 2.4).

Predictor Variable	**Atmospheric Level**
Temperature	500-, 700-, 850-hPa, and 2-m AGL
Dewpoint Temperature	500-, 700-, 850-hPa, and 2-m AGL
Max. Hourly Simulated Reflectivity*	1 km AGL
CAPE	Surface-based
CIN	Surface-based
PWAT	Entire Column
Max. Hourly Simulated UH*	2-5 km AGL
Max. Hourly U, V Wind	10 m AGL
Max. Hourly Upward Vertical Velocity (UVV), Downward Vertical Velocity (DVV)	100-1000 hPa (400-1000 hPa for NAM members of HREFv2)
Forecast 24-h Precipitation	Surface
Lat., Lon.	N/A

Table 2.4 Predictor variables from each ensemble. Asterisks denote variables used for the HREFv2 RFFPs only. Due to limited computing resources, all predictors except for latitude and longitude represent 24-h temporal mean ensemble mean quantities.

These fields include variables that represent a point's meteorological environment, variables that have an obvious direct relationship with observed precipitation, and latitude and longitude, which are designed to account for spatially-varying precipitation climatology. Simulated 2-5 km updraft helicity (UH) is also included as a predictor given its relationship to sustained rotating updrafts and severe weather occurrence (e.g., Kain et al. 2008; Sobash et al. 2011; Loken et al. 2017), since supercells or mesoscale convective systems that produce elevated values of simulated UH may also produce localized heavy rainfall (e.g., Nielsen and Schumacher 2018). The SREF uses two less fields compared to the HREFv2 since the SREF does not output forecasts of simulated reflectivity or UH.

Predictors are derived from ensemble forecast grid-point values on the 20-km grid. Originally, predictors included forecasts from each ensemble member, since it was hypothesized that the RF algorithm could learn and correct for each member's individual systematic biases. However, simply using the ensemble mean value of each variable produced RFFPs that were at least as skillful as those made using predictors from each member. Moreover, using only ensemble mean forecast values made it computationally feasible for the RF to consider predictors from multiple points in space, potentially allowing the RF to identify and correct nonlinear systematic spatial biases. Therefore, ensemble mean forecast values from points (on the 20-km grid) within an approximately 100-km box surrounding the forecast point (i.e., forecast values from the forecast point and the 24 closest points) are used as predictors. Notably, there is no spatial averaging of the values used beyond the neighbor budget interpolation to the 20-km grid.

Further necessary reductions in dataset dimensionality are achieved through preprocessing the raw ensemble data. First, a temporal mean is taken over the 8 3-hourly (24 1-

hourly) forecast fields each day at each native grid point for the SREF (HREFv2). While useful information is undoubtedly lost using this method, the temporal mean provides an overall summary of the simulated meteorological conditions during the relevant 24-h period, which is hypothesized to be sufficient for skillful RF probabilistic precipitation forecasts on next-day time scales. Each day's temporal mean forecasts are then remapped to the 20-km verification grid. Finally, 10% (i.e., 2,130) of the (remapped) points in the analysis domain are randomly sampled without replacement and added to the dataset for training each day (note that the full domain is still used for testing).

Randomly sampling the domain in this manner, as in Gagne et al. (2014), accomplishes two main objectives: it reduces the computational expense of the algorithm by appreciably shrinking the size of the training dataset, and it decreases the likelihood of including multiple highly-correlated grid points in the training set, reducing the chance of RF over-fitting (i.e., fitting on noise rather than actual, systematic patterns in the data). A sampling rate of 10% is greater than that used by Gagne et al. (2014) but is chosen to balance the tradeoff between computational expense and RFFP skill, which increased only slightly at sampling rates beyond 10% in sensitivity tests from 0.5-70% (not shown). All data preprocessing steps are summarized in Fig. 2.2.

After the data has undergone preprocessing, a random forest classifier from the Python module Scikit-Learn (Pedregosa et al. 2011) is used to train the ensemble RFs and create RFFPs. Based on hyper-parameter sensitivity tests (not shown), the random forest classifier requires: 200 trees, a maximum tree depth of 15 levels, at least 20 samples per leaf node, the minimization of entropy for splits, and the consideration of $\sqrt{\mathrm{n}}$ predictors (where n is the total number of predictors in the dataset) at each node. Separate RFs are trained for each precipitation threshold,

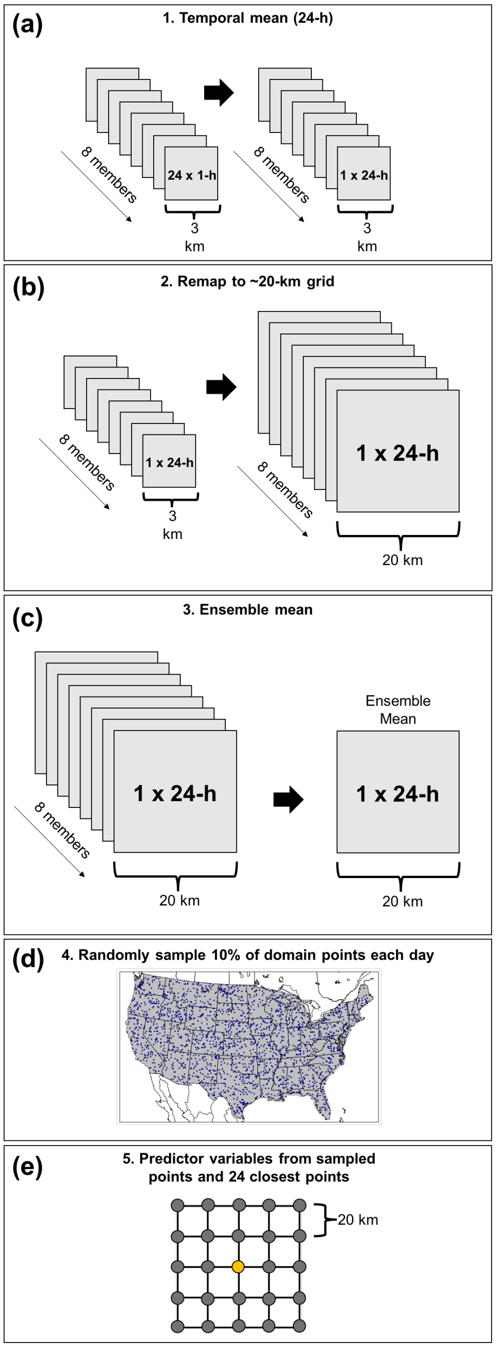


Figure 2.2 Schematic illustrating the data preprocessing steps for the 8-member HREFv2. Note that the SREF follows a similar procedure but has 26 members and starts on a coarser native grid. (a) The temporal mean is taken over 24-h at each native grid point for each ensemble member. (b) The temporally-averaged data is remapped to an approximately 20-km grid. (c) An ensemble mean is taken at each 20-km grid point. (d) 10% of the domain is randomly sampled for training. (e) Training data consists of the predictor variables at each sampled point (yellow) and the 24 closest 20-km points.

but all RFs use the same hyper-parameters. Importantly, since each threshold forecast is created independently, there is no guarantee of consistency between the probabilities of different threshold exceedance. However, the use of different RFs for different thresholds enables a more direct comparison of how the RF technique performs at each threshold individually and allows for different types of precipitation events to be predicted from trees/forests with different, potentially more appropriate structures.

Unlike many previous studies (e.g., Gagne et al. 2014; Herman and Schumacher 2018c), separate RFs are *not* trained for each season and/or geographic region. Using a single RF to represent the entire CONUS year-round likely sacrifices forecast skill, since locations have different time- and space-varying climatologies (e.g., Schumacher and Johnson 2006). However, using a single RF considerably simplifies the prediction and maintenance processes of RF-based post-processing. For example, with multiple regional RFs, RFFPs may be un-physically discontinuous near the border of two regions, requiring additional post-processing. Moreover, multiple RFs require more computing power to train (or retrain) and run when making daily predictions. Additionally, it is hypothesized that the inclusion of latitude and longitude coordinates as well as seasonally-varying environmental variables (e.g., temperature) may help a single RF implicitly account for time- and space-varying precipitation climatologies. This single-RF approach, while perhaps less efficient than a multi-RF approach with explicit dataset filtering, may be advantageous for precipitation prediction since spatially- and seasonally-distant training data (e.g., forecast precipitation) may have at least some relevance for all forecast points. However, the single-RF approach may be less appropriate to use in problem domains where distant training data is less relevant to a given forecast point.

d. Verification

16-fold cross validation with 31 days per fold is used to verify the forecasts. Verification metrics are computed on the full set of 496 forecasts derived from each fold's testing set. To facilitate a fair comparison between the RFFPs and spatially smoothed forecasts, the σ that minimizes the BS over each fold's training set is used to create the spatially smoothed forecasts; hence, σ varies by fold (Fig. 2.3).

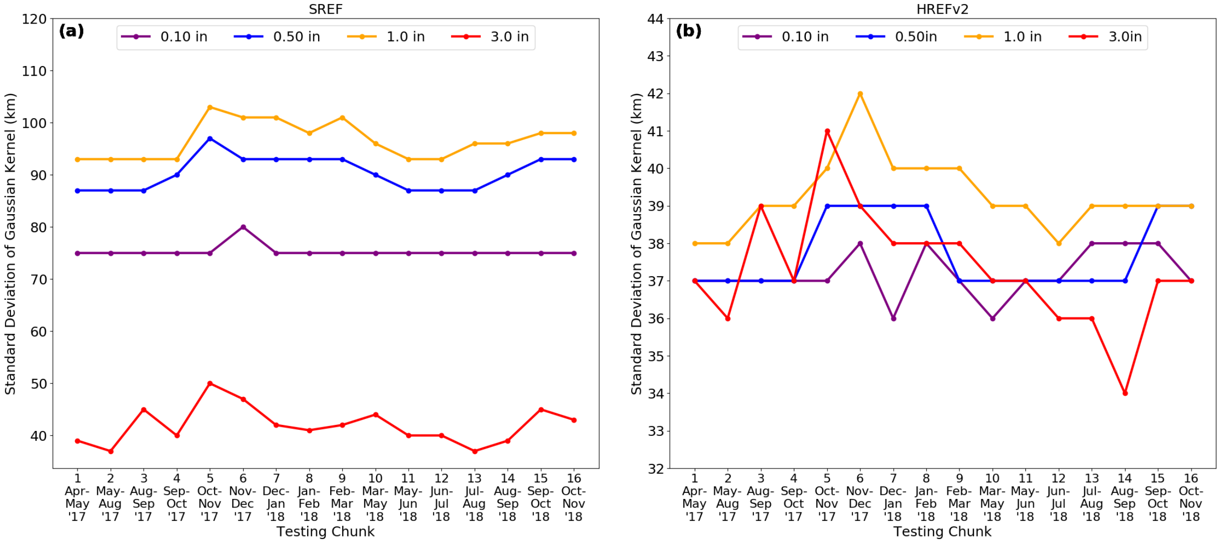


Figure 2.3 Relationship between the standard deviation of the Gaussian kernel (i.e., σ) and testing fold for (a) the SREF and (b) the HREFv2. In each plot, 0.1-, 0.5-, 1-, and 3-inch forecasts are depicted in purple, blue, gold, and red, respectively. The range of dates included in each fold are listed on the x-axis. Note the different y-axis scales.

Verification metrics are computed over the full domain (Fig. 2.1) as well as over five distinct regions (Fig. 2.4), which are based on combinations of the regions defined by Bukovsky (2011). These regions have distinct temperature and precipitation climatologies.

An important strategy for evaluating probabilistic forecasts is the creation of 2 x 2 contingency tables (e.g., Wilks 1995), which are derived from binarizing the forecast at various probability thresholds. Verification metrics such as probability of detection (POD), probability of

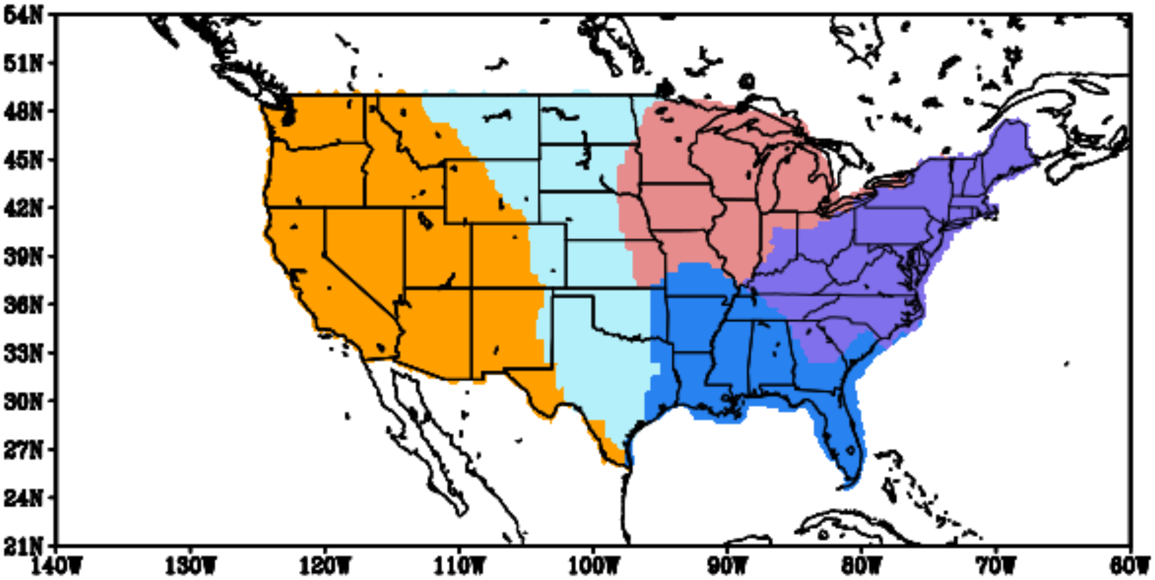


Figure 2.4 The five regional analysis regions, which include the West (gold), Great Plains (light blue), Upper Midwest (salmon), South (royal blue), and East (purple).

false detection (POFD), success ratio (SR), bias, and critical success index (CSI) can then be obtained (e.g., see equations 3-7 in Loken et al. 2017). These metrics form the basis of other forecast evaluation tools used herein, such as the ROC curve (Mason 1982) and performance diagram (Roebber 2009). ROC curves plot POD against POFD at multiple forecast probability thresholds (here, 1, 2, and 5-95% in intervals of 5%). Area under the ROC curve (AUC) provides a measure of forecast discrimination ability, with values of 1 (0.5) indicating a perfect (random) forecast. Since AUC is not sensitive to forecast reliability (Wilks 2001), attributes diagrams (Hsu and Murphy 1986; Wiilks 1995) measure reliability by grouping forecasts into k bins based on forecast probability and plot the mean observed relative frequency of each bin against the bin's probability. Herein, 11 bins are used [0, 5%), [5-15%), …, [85-95%), and [95-100%]. Perfectly reliable forecasts fall along a diagonal line with a slope of 1 passing through the origin. Over- (under-) forecasts fall below (above) the perfect reliability line. Horizontal and vertical lines are plotted at the sample climatological relative frequency, while a “no-skill” line is plotted halfway between the horizontal climatology line and the line of perfect reliability. Points above (below) the no-skill line contribute positively (negatively) to the Brier Skill Score when a reference forecast of climatology is used (Wilks 1995).

Performance diagrams (Roebber 2009) plot POD against SR and include lines of constant bias and CSI. Herein, forecasts are plotted on performance diagrams at each of the 21 probability levels used to create the ROC curves. The most skillful forecasts fall closest to the upper right-hand corner of the plot, where POD, SR, bias, and CSI are all optimized.

The BS (e.g., Wilks 1995) measures the magnitude of the forecast probability errors and can be decomposed into reliability, resolution, and uncertainty components (Murphy 1973; Wilks 1995). The BS is a negatively-oriented score, so a score of 0 (1) indicates perfect (no) skill. One disadvantage of the BS is that it is sensitive to the observed climatological frequency of the event being verified. The Brier Skill Score (BSS) helps account for this effect by comparing the BS to that of a reference forecast, which is often a forecast of climatology. The BSS is defined as:

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{ref}}}{0 - \mathrm{BS}_{\mathrm{ref}}} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}} \quad (2.3),$$

where, herein, $\mathrm{BS}_{\mathrm{ref}}$ is the BS obtained by always forecasting the underlying climatological frequency associated with the entire dataset. The BSS is a positively-oriented score, with possible values from -∞ to 1. A BSS of 0 (1) indicates no (perfect) skill relative to the reference forecast.

A one-sided paired permutation test (e.g., Good 2006) is used herein to test whether the AUC and BSS of a given set of forecasts (e.g., the RFFPs) is significantly greater than a second set of forecasts (e.g., the spatially smoothed probabilities). The general procedure is the same for both AUC and BSS. Individual-day forecasts are randomly permuted between the two forecast systems 10,000 times to create a null distribution of metric differences. The actual difference between the two forecast systems' skill metrics is then compared to the null distribution to obtain a p-value. In the AUC paired permutation test, contingency table elements are randomly permuted rather than the AUC values themselves since individual-day AUC values can be very sensitive to small changes in contingency table elements (Hamill 1999). The final AUC values (and AUC differences) for each iteration are computed based on the permuted contingency table elements. In the same manner, individual-day BSs rather than BSSs are permuted, and BSSs (and BSS differences) for each iteration are computed based on the collective permuted BSs.

Spatial biases are assessed using an approach outlined by Clark et al. (2010a) and Marsh et al. (2012). Conceptually, whenever a yes forecast is issued within the domain, the spatial distribution of yes observations within a 500 x 500 km box is tabulated relative to the yes forecast point and the results are composited over the entire dataset. However, these yes observations are only added to the composite if they fall within the analysis domain. While this method can yield artificially anisotropic contributions to the composite near the domain boundaries, tests (not shown) have indicated that, overall, this method does not appreciably bias the center of the distribution. Thus, in the absence of systematic spatial biases, the center of the distribution should be located at the yes forecast point.

In this study, a yes observation is defined as the Stage IV data exceeding a quantitative precipitation threshold (e.g., 0.1-, 0.5-, 1-, or 3-in) on the verification grid, while a yes forecast is defined as the forecast exceeding a probability threshold that, to the nearest 1%, optimizes frequency bias. Defining yes forecasts in this way allows for a clean comparison between forecasts by removing bias magnitude but still allowing for spatial biases. Table 2.5 shows the forecast probability thresholds and their corresponding frequency biases.

Precipitation Threshold	**Ensemble/Forecast**	**Forecast Probability Threshold (%)**	**Frequency Bias**
0.1 in.	SREF, Raw	62	1.029
	HREF, Raw	38	1.040
	SREF, Smooth	55	1.011
	HREF, Smooth	43	0.996
	SREF, RF	44	0.991
	HREF, RF	43	1.007
0.5 in.	SREF, Raw	47	0.957
	HREF, Raw	38	0.896
	SREF, Smooth	38	0.990
	HREF, Smooth	35	0.989
	SREF, RF	33	0.992
	HREF, RF	35	0.987
1 in.	SREF, Raw	35	1.049
	HREF, Raw	26	1.139
	SREF, Smooth	29	1.007
	HREF, Smooth	29	1.005
	SREF, RF	26	1.011
	HREF, RF	28	0.998
3 in.	SREF, Raw	20	1.045
	HREF, Raw	26	0.812
	SREF, Smooth	18	0.990
	HREF, Smooth	20	1.022
	SREF, RF	17	0.970
	HREF, RF	20	1.030

Table 2.5 Forecast probability thresholds used to (approximately) optimize frequency bias for each forecasting system at each precipitation threshold. Actual values of frequency bias are reported in the fourth column.

One drawback of ML-based post-processing techniques is that they assume the underlying dynamical models do not change with time and must be retrained whenever developers implement changes. An important question, therefore, is: how long of a dataset is required for ML to perform adequately? To address this question, RFs are re-trained and re-evaluated using a dataset comprising the first 62, 124, 248, and 372 days (i.e., the first 1/8, 1/4, 1/2, and 3/4) of the full dataset, respectively. These RFs use the same hyper-parameters as described previously. Although this approach is suboptimal, sensitivity tests suggest that the BSS varies only slightly with different hyper-parameters; moreover, the set of hyper-parameters used previously was deemed close enough to optimal to make using a constant set of hyper-parameters worth the reduced computational expense. As with the full dataset, k-fold cross validation is used to evaluate the forecasts, with 31 forecasts per fold.

This method of assessing the relationship between forecast skill and dataset length is not perfect due to the temporal-varying precipitation climatology. For example, one potential issue is that the smallest datasets, which have fewer folds, get verified only against testing data from the same season as the training data. As more data is added, the size of the training set increases, but the training set starts to include data from other times of the year relative to the test set. Therefore, it is possible that this “new” training data adds only limited value to each testing fold. Additionally, the uncertainty of the forecast itself changes with time due to seasonal variations in climatology, such that, as more dates are added to the dataset, the overall forecast difficulty (and thus, objective skill) changes depending on what dates are added. Despite these deficiencies, the results give useful preliminary insight into the feasibility of adopting ML-based techniques operationally.

3. Results

a. Traditional verification metrics over the full domain

1) ROC METRICS

All forecasts have good discrimination ability, as indicated by ROC diagrams (Fig. 2.5a,d,g,j,m,p,s,v) and AUC (Fig. 2.6a-d). Even the worst-performing forecast system (i.e., the raw SREF ensemble for the 3-in. threshold; Fig. 2.4d) has an AUC of 0.80. Nevertheless, for all thresholds (all but the 3-in. threshold), the SREF (HREFv2) RFFPs have significantly greater AUC than the corresponding raw and smoothed ensemble probabilities ($p < 0.0001$; Fig. 2.7a,c,e,g). The SREF RFFPs also have significantly greater AUC than the raw HREFv2 probabilities ($p < 0.0001$; Fig. 2.7a,c,e,g).

Interestingly, the raw SREF forecast probabilities often have greater AUC compared to the raw HREFv2 forecast probabilities, even though the HREFv2 is a CAE that performs subjectively better than the SREF. This behavior likely reflects the insensitivity of the AUC to bias (thus negating the SREF's poor reliability; e.g., Fig. 2.5b,e,h,k,n,q; Fig. 2.6i-k) as well as the larger membership of the SREF, which enables the raw SREF to issue more unique forecast probabilities and thus have more unique "points" on its ROC curve, possibly increasing AUC.

2) RELIABILITY

The raw SREF and HREFv2 probabilities suffer from substantial over-forecasting bias at all precipitation thresholds, with the raw SREF forecasts generally having the worst reliability (Fig. 2.5b,e,h,k,n,q,t,w; Fig. 2.6i-l). The 0.1-in. raw SREF forecasts (Fig. 2.5b) have particularly poor reliability, as the reliability curve falls below the no skill line for multiple forecast probability bins. Meanwhile, the raw HREFv2 reliability curves contain "gaps" (Fig. 2.5e,k,q,w) since, with only 8 members, the HREFv2 is unable to issue probabilities in all bins. Spatially smoothing the raw ensemble forecasts improves reliability and removes the gaps from the raw

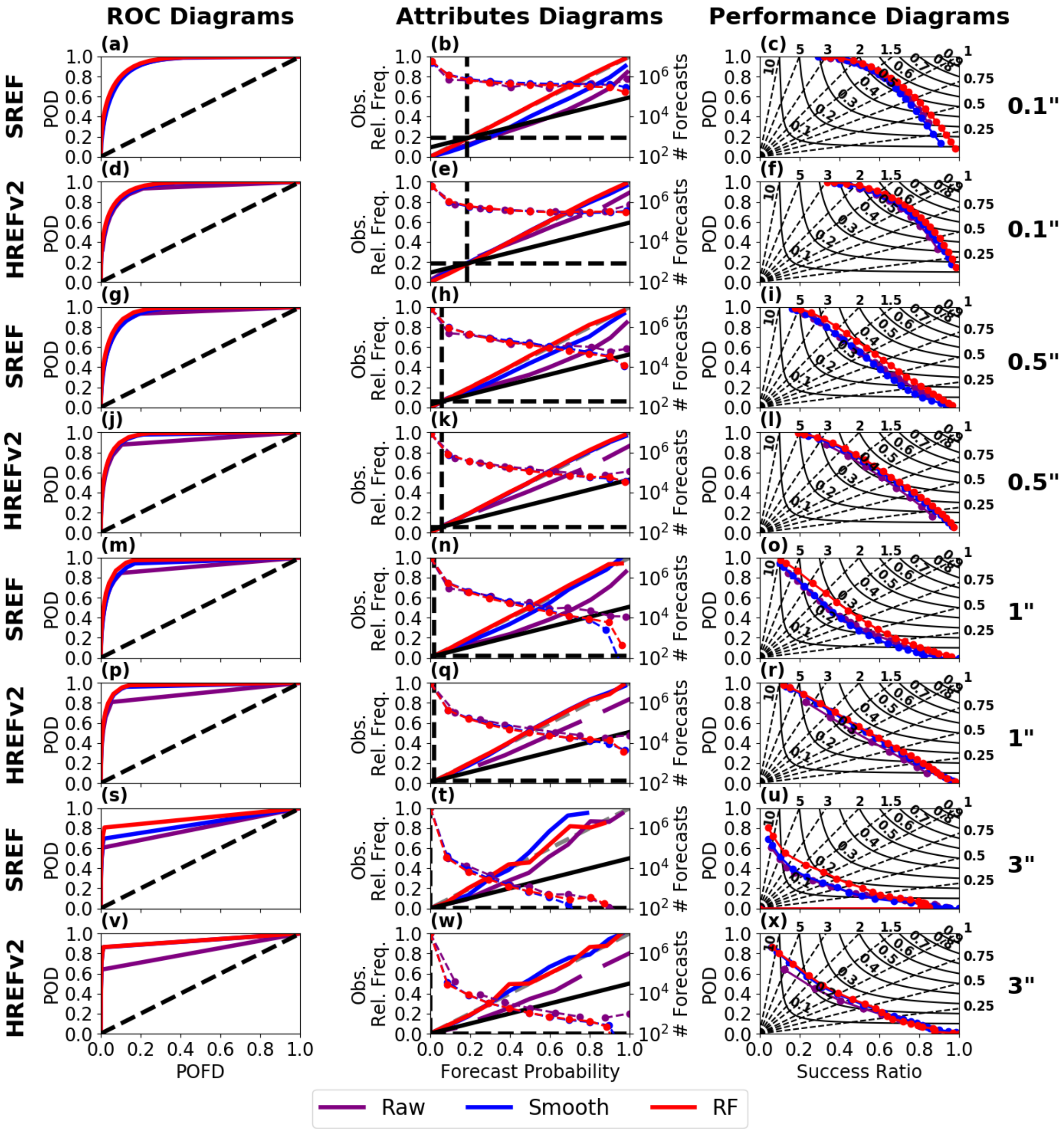


Figure 2.5 (a) ROC curve for the SREF at the 0.1-in. threshold for raw (purple), smooth (blue), and RF (red) forecasts. The black dashed line indicates a random forecast. (b) Attributes diagram for the SREF at the 0.1-in. threshold for the same forecasts as in (a). Black dashed lines indicate the relative frequency of the sample climatology, the solid black line is the "no skill" line, and the dashed gray line represents perfect reliability. The number of forecasts in each probability bin are indicated by the colored dashed lines with filled circles. (c) Performance diagrams for the SREF at the 0.1-in. threshold for the same forecasts as in (a). Lines of constant bias are dashed, while lines of constant CSI are solid. Each of 21 forecast probability levels are indicated by filled circles. (d)-(f) As in (a)-(c) but for the HREFv2. (g)-(l) As in (a)-(f) but for the 0.5 in. threshold. (m)-(r) As in (a)-(f) but for the 1-in. threshold. (s)-(x) As in (a)-(f) but for the 3-in. threshold.

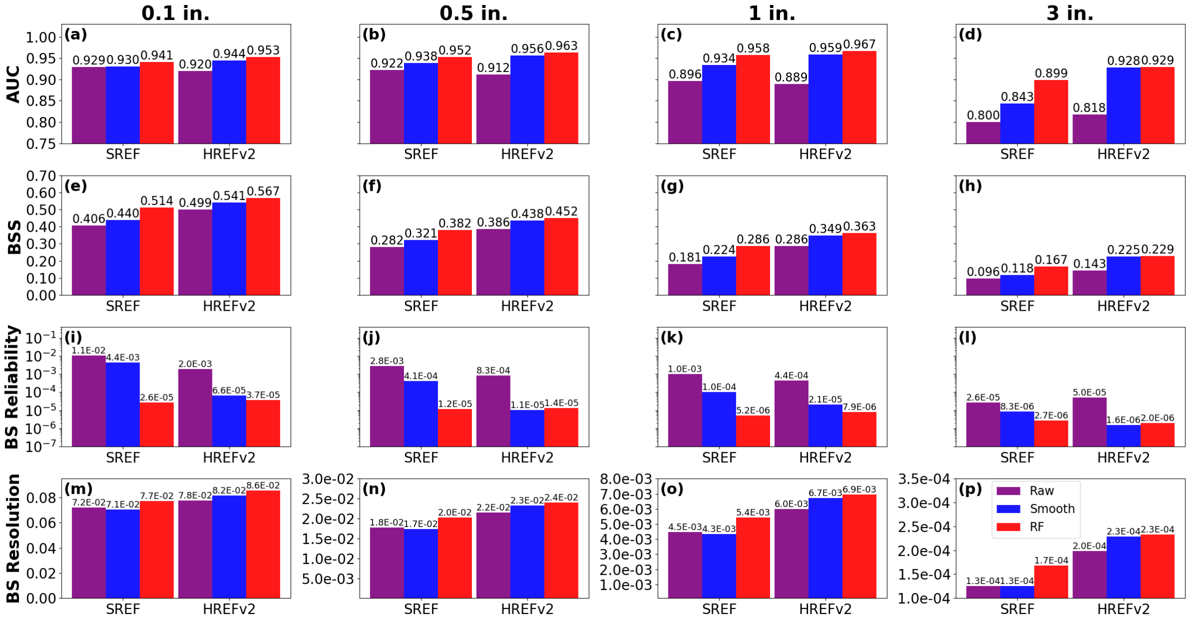
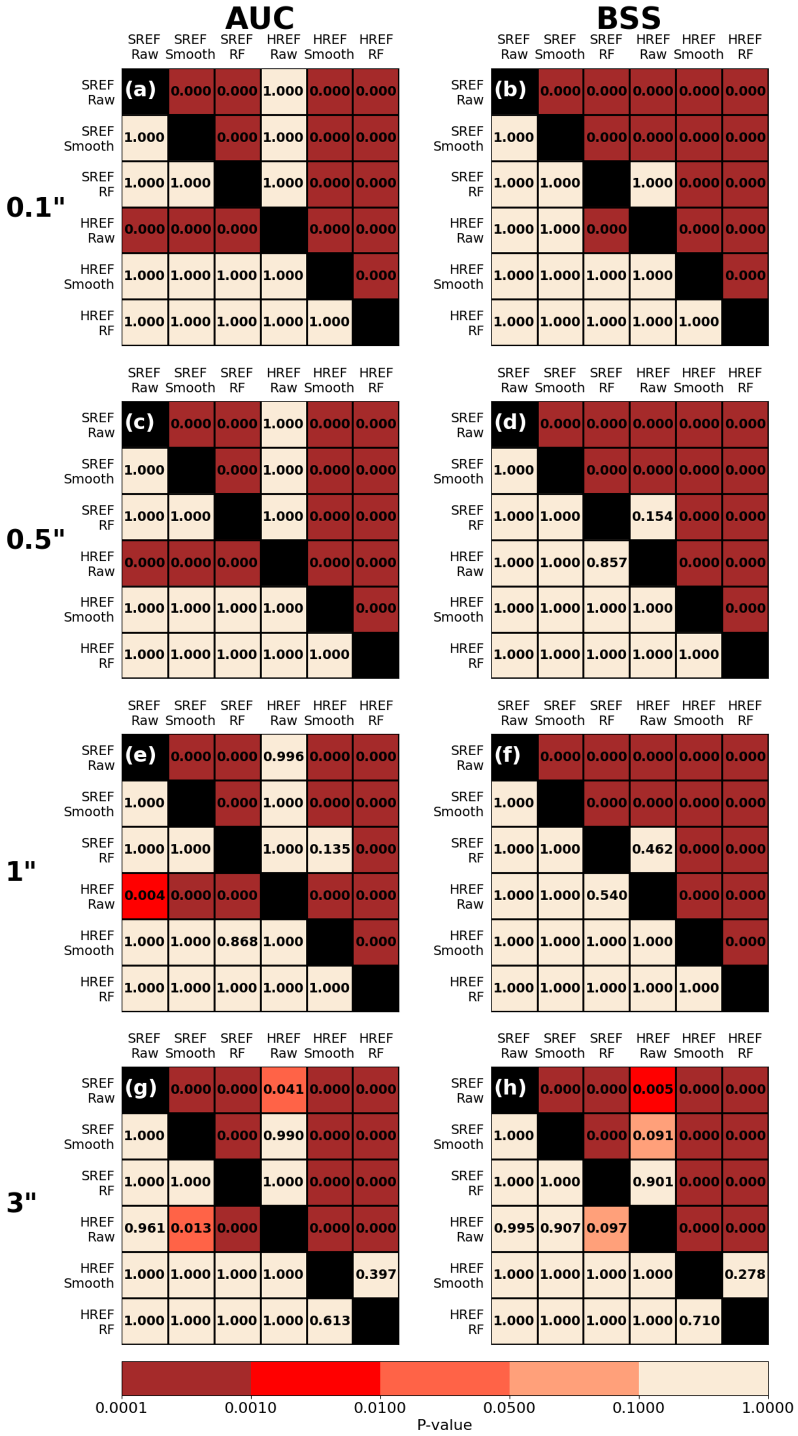


Figure 2.6 (a) AUC for SREF and HREFv2 raw (purple), smooth (blue), and RF forecasts (red) for the 0.1-in. threshold. (b)-(d) As in (a) but for the 0.5-, 1-, and 3-in. thresholds, respectively. (e)-(h) As in (a)-(d) but for BSS. (i)-(l) As in (a)-(d) but for the reliability component of the BS. (m)-(p) As in (a)-(d) but for the resolution component of the BS. Note the different y-axes for (m)-(p), and note that lower values of BS reliability are better.



(a)

	SREF Raw	SREF Smooth	SREF RF	HREF Raw	HREF Smooth	HREF RF
SREF Raw		0.000	0.000	1.000	0.000	0.000
SREF Smooth	1.000		0.000	1.000	0.000	0.000
SREF RF	1.000	1.000		1.000	0.000	0.000
HREF Raw	0.000	0.000	0.000		0.000	0.000
HREF Smooth	1.000	1.000	1.000	1.000		0.000
HREF RF	1.000	1.000	1.000	1.000	1.000	

(b)

	SREF Raw	SREF Smooth	SREF RF	HREF Raw	HREF Smooth	HREF RF
SREF Raw		0.000	0.000	0.000	0.000	0.000
SREF Smooth	1.000		0.000	0.000	0.000	0.000
SREF RF	1.000	1.000		1.000	0.000	0.000
HREF Raw	1.000	1.000	0.000		0.000	0.000
HREF Smooth	1.000	1.000	1.000	1.000		0.000
HREF RF	1.000	1.000	1.000	1.000	1.000	

(c)

	SREF Raw	SREF Smooth	SREF RF	HREF Raw	HREF Smooth	HREF RF
SREF Raw		0.000	0.000	1.000	0.000	0.000
SREF Smooth	1.000		0.000	1.000	0.000	0.000
SREF RF	1.000	1.000		1.000	0.000	0.000
HREF Raw	0.000	0.000	0.000		0.000	0.000
HREF Smooth	1.000	1.000	1.000	1.000		0.000
HREF RF	1.000	1.000	1.000	1.000	1.000	

(d)

	SREF Raw	SREF Smooth	SREF RF	HREF Raw	HREF Smooth	HREF RF
SREF Raw		0.000	0.000	0.000	0.000	0.000
SREF Smooth	1.000		0.000	0.000	0.000	0.000
SREF RF	1.000	1.000		0.154	0.000	0.000
HREF Raw	1.000	1.000	0.857		0.000	0.000
HREF Smooth	1.000	1.000	1.000	1.000		0.000
HREF RF	1.000	1.000	1.000	1.000	1.000	

(e)

	SREF Raw	SREF Smooth	SREF RF	HREF Raw	HREF Smooth	HREF RF
SREF Raw		0.000	0.000	0.996	0.000	0.000
SREF Smooth	1.000		0.000	1.000	0.000	0.000
SREF RF	1.000	1.000		1.000	0.135	0.000
HREF Raw	0.004	0.000	0.000		0.000	0.000
HREF Smooth	1.000	1.000	0.868	1.000		0.000
HREF RF	1.000	1.000	1.000	1.000	1.000	

(f)

	SREF Raw	SREF Smooth	SREF RF	HREF Raw	HREF Smooth	HREF RF
SREF Raw		0.000	0.000	0.000	0.000	0.000
SREF Smooth	1.000		0.000	0.000	0.000	0.000
SREF RF	1.000	1.000		0.462	0.000	0.000
HREF Raw	1.000	1.000	0.540		0.000	0.000
HREF Smooth	1.000	1.000	1.000	1.000		0.000
HREF RF	1.000	1.000	1.000	1.000	1.000	

(g)

	SREF Raw	SREF Smooth	SREF RF	HREF Raw	HREF Smooth	HREF RF
SREF Raw		0.000	0.000	0.041	0.000	0.000
SREF Smooth	1.000		0.000	0.990	0.000	0.000
SREF RF	1.000	1.000		1.000	0.000	0.000
HREF Raw	0.961	0.013	0.000		0.000	0.000
HREF Smooth	1.000	1.000	1.000	1.000		0.397
HREF RF	1.000	1.000	1.000	1.000	0.613	

(h)

	SREF Raw	SREF Smooth	SREF RF	HREF Raw	HREF Smooth	HREF RF
SREF Raw		0.000	0.000	0.005	0.000	0.000
SREF Smooth	1.000		0.000	0.091	0.000	0.000
SREF RF	1.000	1.000		0.901	0.000	0.000
HREF Raw	0.995	0.907	0.097		0.000	0.000
HREF Smooth	1.000	1.000	1.000	1.000		0.278
HREF RF	1.000	1.000	1.000	1.000	0.710	

Figure 2.7 (a) P-values from one-sided paired permutation significance tests for AUC for the 0.1-in. threshold. (b) As in (a) but for BSS. (c)-(d) As in (a)-(b) but for the 0.5-in. threshold. (e)-(f) As in (a)-(b) but for the 1-in. threshold. (g)-(h) As in (a)-(b) but for the 3-in. threshold. Each square reports the p-value associated with testing whether the forecast displayed across the top row has a significantly greater metric than that from the forecast displayed along the left-hand column.

HREFv2 reliability curves (Fig. 2.5b,e,h,k,n,q,t,w). The RF technique tends to produce even better (i.e., near-perfect) forecast reliability for both ensembles at most thresholds (Fig. 2.6i-l).

3) PERFORMANCE DIAGRAMS

Performance diagrams suggest that the skill of the RFFPs matches or exceeds that of the other sets of forecasts at all four precipitation thresholds (Fig. 2.5c,f,i,l,o,r,u,x). The SREF RFFPs clearly outperform corresponding raw and smoothed SREF forecasts (Fig. 2.5c,i,o,u), while the HREFv2 RFFPs have the greatest relative performance at the 0.1-in. threshold (Fig. 2.5f). At the other thresholds (Fig. 2.5l,r,x), the HREFv2 RFFPs and smoothed probabilities demonstrate similar skill, which noticeably exceeds that of the raw HREFv2 probabilities.

One interesting characteristic of the SREF performance diagrams (Fig. 2.5c,i,o,u) is that the second-best performing probabilities (in terms of CSI) tend to be from the raw SREF (e.g., Fig. 2.5c,i,o). This is because the smoothed SREF probabilities require a relatively large amount of spatial smoothing to optimize the BS (Fig. 2.3a), and this degrades resolution (Fig. 2.6m-p). Hence, for the SREF forecasts, a main advantage of the RF technique is that it calibrates the raw ensemble probabilities while improving—rather than sacrificing—resolution.

4) BSS AND BS COMPONENTS

With only one exception (i.e., the smoothed 3-in. HREFv2 probabilities), the RFFPs have significantly greater BSSs ($p < 0.0001$) than the corresponding raw and smoothed ensemble probabilities (Fig. 2.7b,d,f,h). At the 0.1-in. threshold, the SREF RFFPs even have a significantly greater BSS than the raw HREFv2 probabilities ($p < 0.0001$; Fig. 2.7b), which is remarkable given the much coarser horizontal grid-spacing of the SREF. The RF-based approach improves the BSS by simultaneously enhancing forecast reliability and resolution (Fig. 2.6e,i,m).

The RFFPs provide the greatest increase in BSS relative to the corresponding raw and smoothed ensemble forecasts at the smallest precipitation thresholds (Fig. 2.6e-h), likely because the smallest thresholds have the greatest climatological frequency (Fig. 2.8). More occurrences of yes observations in the training dataset make it easier for the RF to identify the systematic relationships between the predictors and observations.

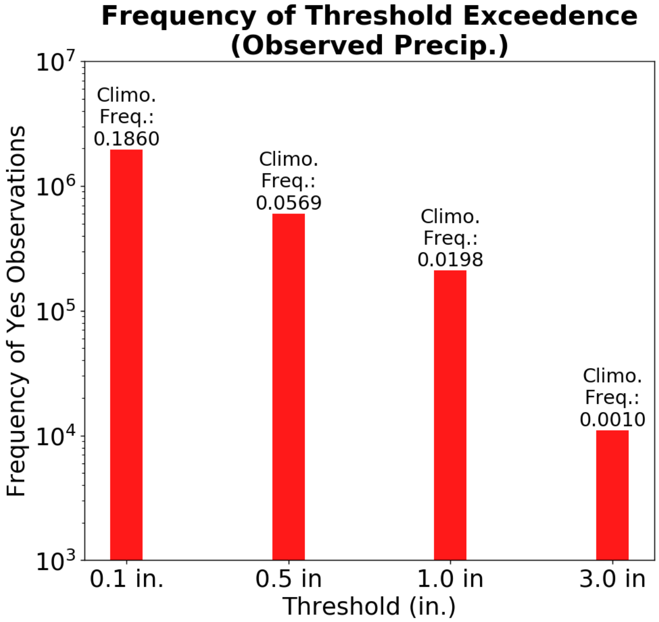


Figure 2.8 Number of "yes" observations (i.e., instances when the observed 24-h precipitation exceeds the given threshold) at the 0.1-, 0.5-, 1-, and 3-in. thresholds. The corresponding relative frequency, abbreviated as "Climo. Freq.," is displayed above each bar. Note the logarithmic y-axis.

RFFPs always have better resolution than the corresponding raw and smoothed ensemble forecast probabilities (Fig. 2.6m-p) and nearly always have better reliability (Fig. 2.6i-l). It is also noteworthy that the RFFPs increase resolution relative to the spatially smoothed ensemble forecasts, both in cases where the 2-dimensional spatial smoothing technique degrades (e.g., the SREF forecasts) and enhances (e.g., the HREFv2 forecasts) reliability.

b. Regional results

Similar results are obtained when forecasts are verified regionally. For the SREF, the RF-based approach improves the BSS in every region at every threshold compared to the raw and smoothed ensemble forecasts (Fig. 2.9a-d). These greater BSSs can be attributed to both better reliability and resolution (Fig. 2.9a-d). Importantly, the RF approach appears to improve the BSS and BS components approximately equally for each region at each threshold (with a few exceptions; e.g., the West region benefits disproportionately at the 1-in. threshold). This finding suggests that a single, CONUS-wide RF can learn enough spatial information such that the benefits to RF-based post-processing are not confined to a single region.

The same general findings also apply to the HREFv2: at each threshold, each region benefits from the RF-based post-processing approximately equally (Fig. 2.10a-d). Of course, these benefits are most pronounced for the lower thresholds, consistent with the full-domain findings presented above. Regardless, the results suggest that, for a given threshold, a single, CONUS-wide RF can provide reliability and resolution benefits to forecasts in all regions, despite each region having different climatological frequencies of threshold exceedance (e.g., Fig. 2.9-2.10).

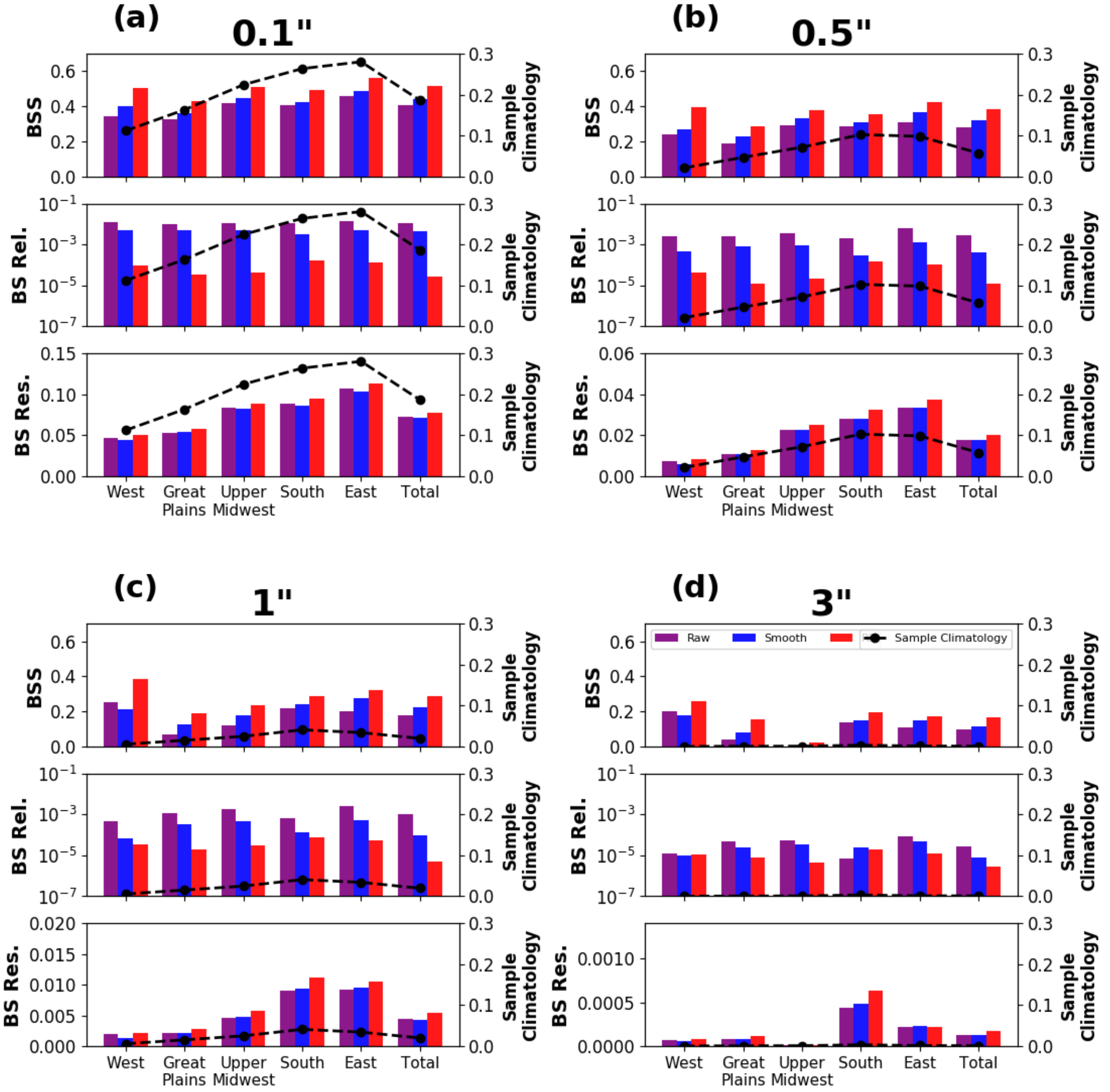


Figure 2.9 Regional BSS, BS reliability, and BS resolution for the raw (purple), spatially smoothed (blue), and RF-based (red) SREF forecasts at the (a) 0.1-, (b) 0.5-, (c) 1-, and (d) 3-inch thresholds. In each case, the black dashed line indicates the climatological relative frequency of threshold exceedance in the given region. Full domain metrics are also given under the "Total" label.

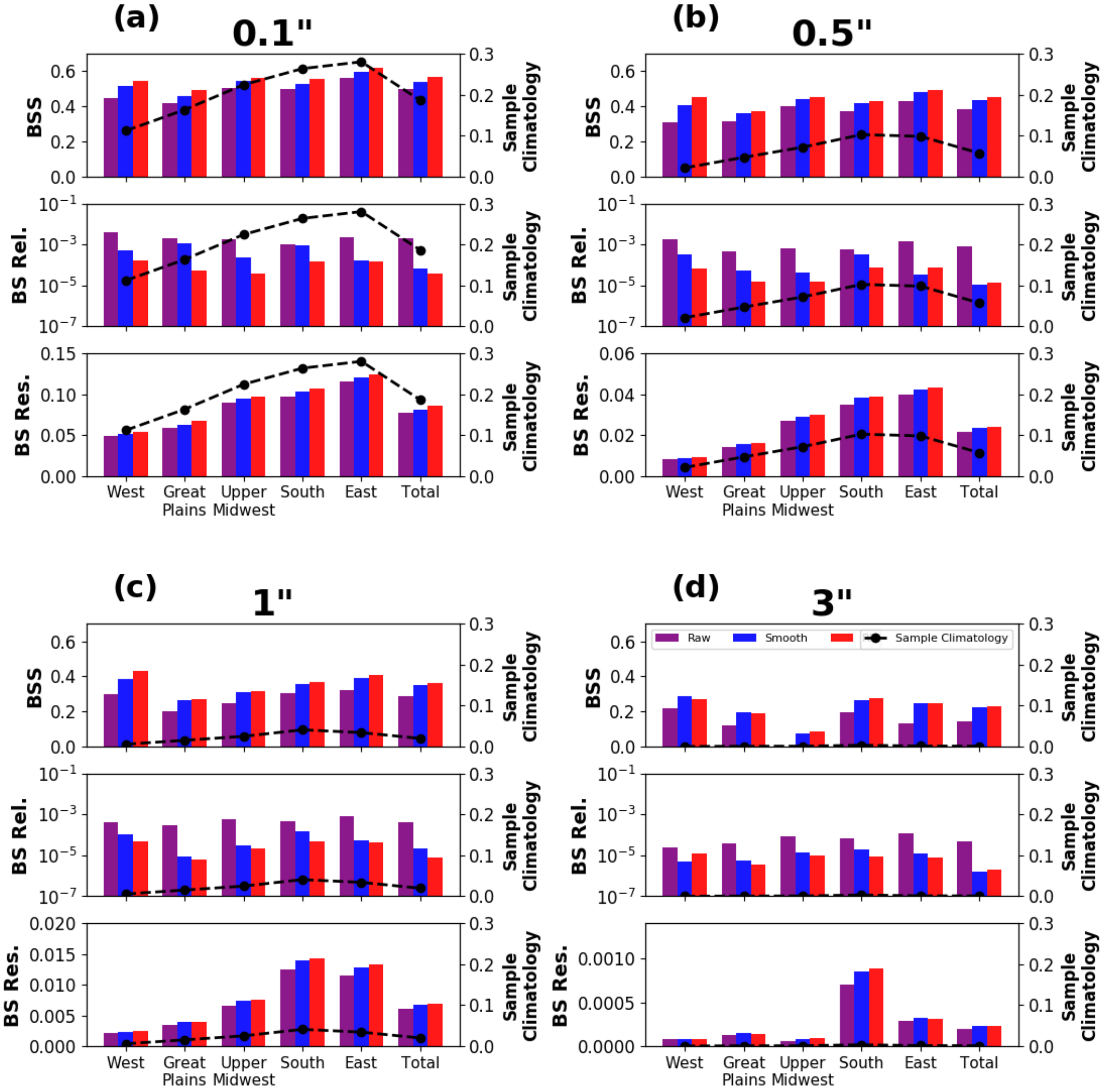


Figure 2.10 As in Fig. 2.9 but for the HREFv2 forecasts. Axes are the same as in Fig. 2.9.

c. Full-domain spatial biases

Full-domain spatial bias magnitudes are small for both ensembles, as the center of the observed conditional distribution seldom falls more than 20-40 km from the yes forecast point (Fig. 2.11a-x).

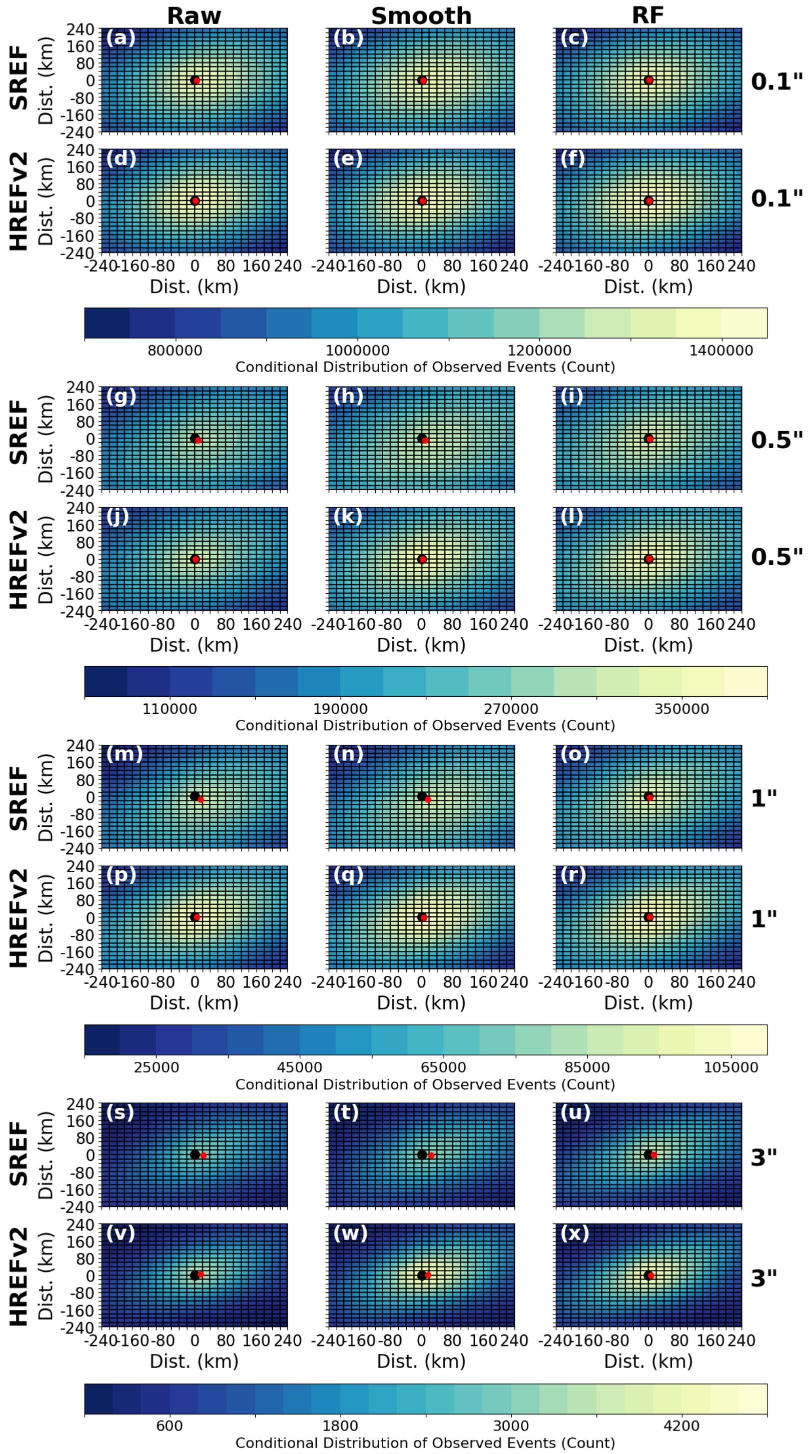


Figure 2.11 (a) Spatial distribution of observed yes events given a yes forecast (see text) at point (0,0) (black dot) for the raw SREF ensemble forecast at the 0.1-in. threshold. The red dot denotes the center of the distribution. (b)-(c) As in (a) but for the SREF-derived smoothed and RF-based forecasts. (d)-(f) As in (a)-(c) but for the HREFv2. (g)-(l) As in (a)-(f) but for the 0.5-in. threshold. (m)-(r) As in (a)-(f) but for the 1-in. threshold. (s)-(x) As in (a)-(f) but for the 3-in. threshold. Note the different color scale used for each threshold.

The spatial biases are greatest for the raw and smoothed SREF forecasts (Fig. 2.11a,b,g,h,m,n,s,t) and for the higher (i.e., 1- and 3-in.; Fig. 2.11m-x) precipitation thresholds. These findings make sense given that the higher thresholds are more likely to be associated with deep convection, which is more difficult to predict—especially for a convection-parameterizing ensemble (e.g., Kain et al. 2006)—due to uncertainties in initiation and evolution. The anisotropy of the conditional distribution of observed yes events seen in Fig. 2.11a-x is consistent with Marsh et al. (2012), who obtained a similar preferred southwest-northeast orientation and explained that it reflects the mean shape and orientation of individual precipitation objects over the full dataset.

One important finding in the present study is that the RF technique helps alleviate spatial biases in the raw and smoothed ensemble probabilities. This result can be seen in two distinct ways. First, the center of the distribution (i.e., the red dot in Fig. 2.11a-x) is closest to the yes forecast point (i.e., the black dot in Fig. 2.11a-x) in the RF plots (i.e., Fig. 2.11c,f,i,l,o,r,u,x). Additionally, difference plots (Fig. 2.12a-p) show that the RF technique tends to add conditional observations in locations that oppose the direction of the spatial bias and/or subtract conditional observations from locations in the same direction of the spatial bias. For example, in the 1-in. raw and smoothed SREF forecasts, the center of the observed distribution falls too far to the southeast of the yes forecast point (Fig. 2.11m,n). In both cases, the RF technique adds conditional observations to the northwest and subtracts conditional observations to the southeast (Fig. 2.12i,j) so that the center of the RF-based conditional distribution of observed yes events is closer to the yes forecast point (Fig. 2.11o).

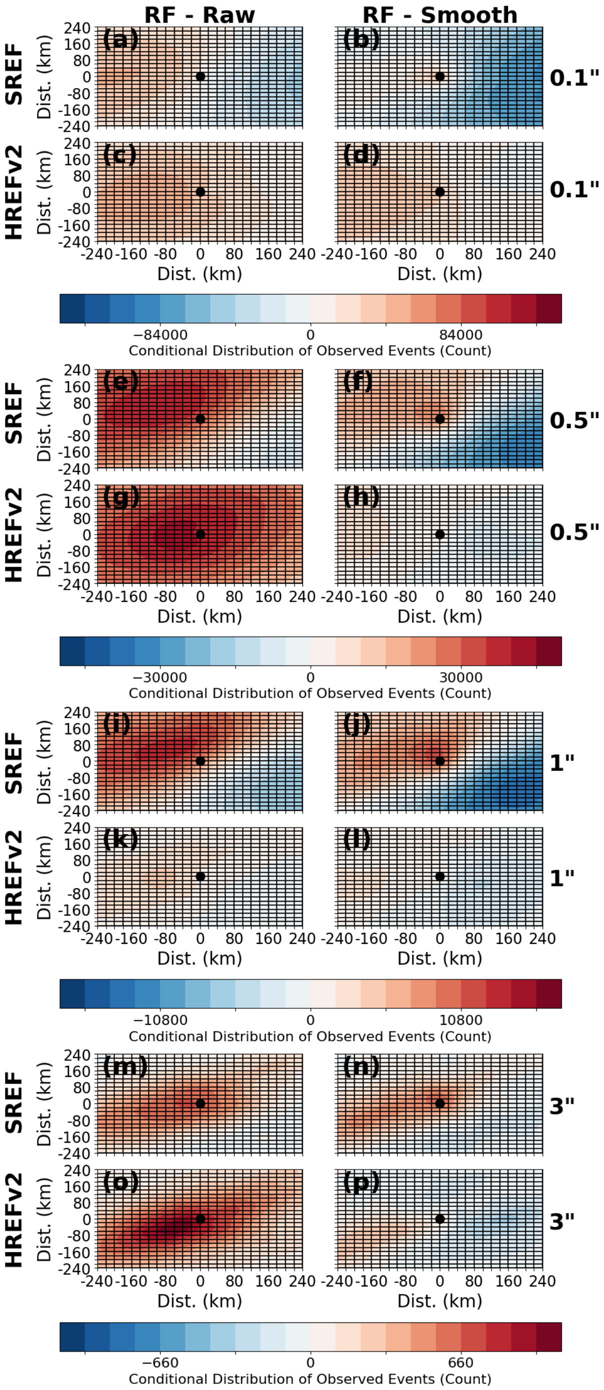


Figure 2.12 (a) Difference between the conditional distribution of yes observed events given a yes SREF-based RF forecast at (0, 0) (black dot) and the conditional distribution of yes observed events given a yes raw SREF forecast at (0, 0) at the 0.1-in. threshold (i.e., Fig. 2.11c minus Fig. 2.11a). (b) As in (a) but subtracting the smoothed SREF distribution from the SREF RF distribution (i.e., Fig. 2.11c minus Fig. 2.11b). (c)-(d) As in (a)-(b) but for the HREFv2. (e)-(h) As in (a)-(d) but for the 0.5-in. threshold. (i)-(l) As in (a)-(d) but for the 1-in. threshold. (m)-(p) As in (a)-(d) but for the 3-in. threshold.

Similar behavior is seen for both ensembles at all thresholds, although the effect is stronger for the SREF since the HREFv2 forecasts have fewer spatial biases. In many cases, the RF approach also adds conditional yes observations to the yes forecast point and surrounding points (e.g., Fig. 2.12g,n,o), which improves the forecast by increasing the conditional probability of a yes observation given a yes forecast.

d. Sensitivity of results to dataset length

The best AUC and BSS values are generally obtained using a dataset of 248 days (Fig. 2.13a,b,d,e). Interestingly, increasing the dataset beyond 248 days results in slightly lower AUCs and BSSs. This finding can potentially be explained by temporal variations in the observed precipitation climatology. For example, since AUC is sensitive to the number of correct

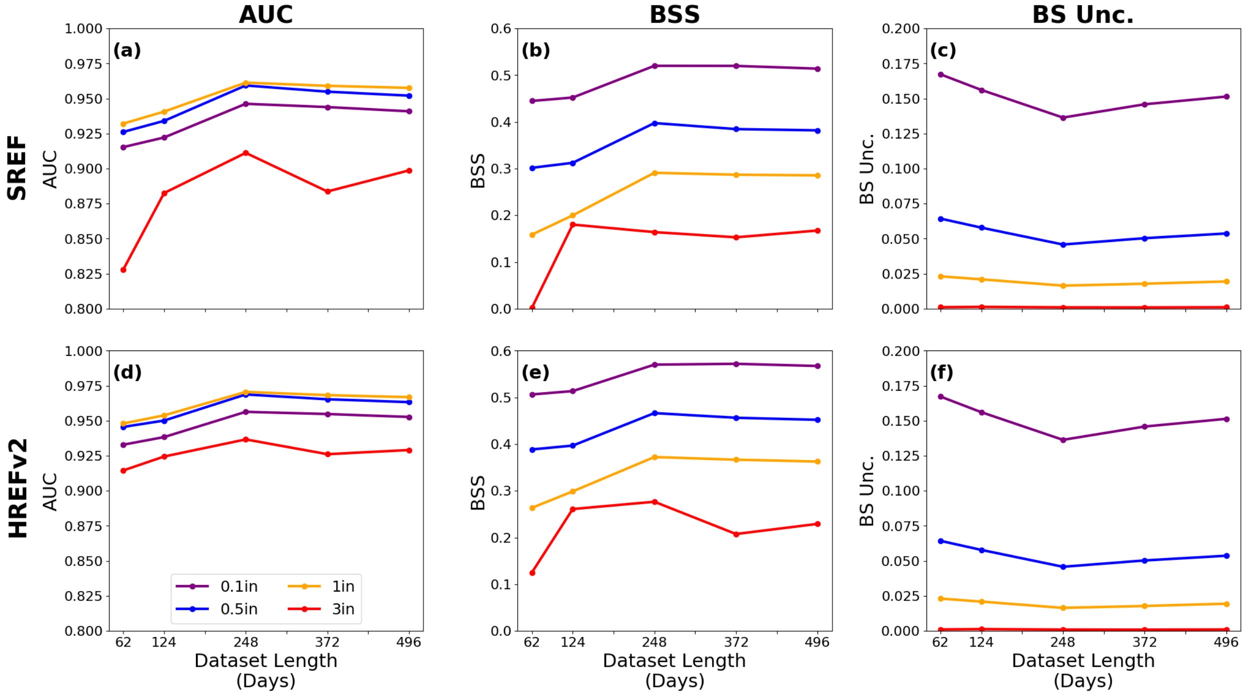


Figure 2.13 (a) AUC as a function of dataset length for the SREF. (b)-(c) As in (a) but for the BSS and uncertainty component of the BS, respectively. (d)-(f) As in (a)-(c) but for the HREFv2.

negatives, AUC may be artificially inflated (deflated) during times of the year with lower (higher) forecast uncertainty. Indeed, this is exactly the pattern that is seen (Fig. 2.13a,c,d,f). The temporal variation in climatology may also help explain the behavior of the 3-in. SREF and HREFv2 BSS curves, which reach a local minimum at 372 days. Although difficult to discern from Fig. 2.13c,f, the 3-in. uncertainty reaches a minimum (maximum) at 372 (124) days. A relatively low (high) forecast uncertainty makes a reference forecast of climatology more (less) skillful and more (less) harshly penalizes small forecast errors. Thus, the BSS decreasing after 124 (248) days for the SREF (HREFv2) may be at least partly explained by the variations in the already-low observed precipitation climatology.

Because these variations in climatology have the potential to "artificially" influence the verification metrics, the results should be interpreted cautiously. Nevertheless, it is likely that the results presented herein aren't due entirely to temporal variations in the dataset climatology, especially since the BSS follows a similar pattern as AUC. For both AUC and BSS, there are obvious gains from increasing the length of the dataset from 62- to 124-days and, in general, additional gains from further increasing the dataset to 248 days. Since each fold's testing set contains 31 days, these findings suggest that a minimum training set length of 93-217 days (i.e., approximately 1-2 seasons) is desirable for adequate performance.

e. Select cases

Two cases are subjectively selected to illustrate the RFFPs' relative performance on individual days.

1) 1200 UTC 2 OCTOBER – 1200 UTC 3 OCTOBER 2017

The heaviest precipitation during this period occurred in a corridor extending from northeastern Minnesota into west-central Kansas ahead of a cold front. Relatively heavy precipitation also occurred in northern Montana downstream of a mid-level shortwave trough, while southern Louisiana and southern Florida experienced weakly-forced tropical showers.

The raw SREF and HREFv2 probabilities performed relatively well at all four thresholds (Fig. 2.14a,d,g,j,m,p,s,v). In general, these probabilities had good sharpness and resolution. However, these raw ensemble forecasts also placed 90-100% probabilities in locations where the observed precipitation did not exceed the threshold (e.g., southern Utah in Fig. 2.14a,d). The spatially smoothed forecasts (Fig. 2.14b,e,h,k,n,q,t,w) helped calibrate the raw forecast probabilities but had reduced sharpness. Meanwhile, the RFFPs (Fig. 2.14c,f,i,l,o,r,u,x) generally had good calibration, sharpness, and resolution. For example, like the 0.5-in. raw SREF probabilities (Fig. 2.14g), the 0.5-in. SREF RFFPs (Fig. 2.14i) exceeded 80% over east-central Minnesota and northern Montana, while the spatially smoothed SREF probabilities (Fig. 2.14h) were less in both areas. Moreover, the 0.5- and 1-in. SREF RFFPs (Fig. 2.14i,o) had less false alarm area over the High Plains compared to the spatially smoothed SREF forecasts (Fig. 2.14h,n). Differences between the HREFv2 smoothed probabilities and corresponding RFFPs were subtler since less spatial smoothing was required to calibrate the raw HREFv2 probabilities. For example, compared to the corresponding smoothed forecasts (Fig. 2.14k,q), the 0.5- and 1-in. HREFv2 RFFPs (Fig. 2.14l,r) had a larger spatial extent of >90% probabilities in the Upper Midwest where observed precipitation exceeded the threshold. The 0.5-in. RFFPs (Fig. 2.14l) also gave slightly lower probabilities in east-central South Dakota but slightly enhanced the probabilities in central Iowa compared to the spatially smoothed probabilities (Fig. 2.14k).

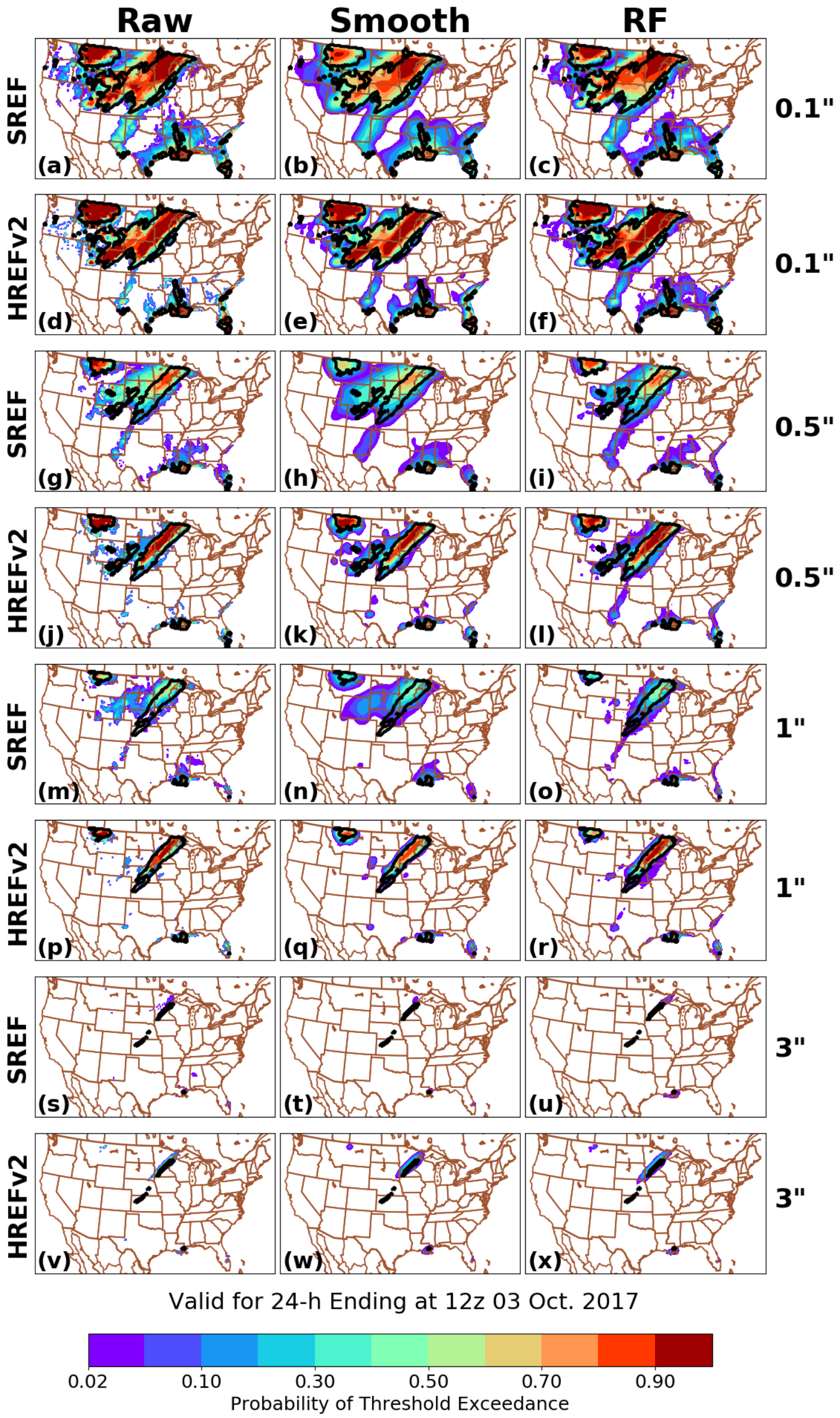


Figure 2.14 (a) 0.1-in. PQPFs from the raw SREF ensemble, valid for the 24-h ending at 1200 UTC on 3 October 2017. The black contours indicate where the observed precipitation exceeded the given threshold. (b)-(c) As in (a) but for the spatially smoothed and RF-based SREF PQPFs. (d)-(f) As in (a)-(c) but for the HREFv2. (g)-(l) As in (a)-(f) but for the 0.5-in. threshold. (m)-(r) As in (a)-(f) but for the 1-in. threshold. (s)-(x) As in (a)-(f) but for the 3-in. threshold.

2) 1200 UTC 22 JUNE – 23 JUNE 2017

Early in this period, elevated storms were ongoing over South Dakota, Minnesota, Wisconsin, and Michigan. Later, surface-based storms formed ahead of a cold front extending from eastern Ontario into central Kansas and eastern Colorado, bringing heavy rainfall to southern Wisconsin, central Michigan, and northern New York. Eastern Colorado and western Kansas also experienced 0.1–0.5-in rainfall associated with post-frontal upslope flow. Meanwhile, Tropical Storm Cindy brought heavy rainfall to the southeastern U.S.

Raw ensemble probabilities from the SREF and HREFv2 (Fig. 2.15a,d,g,j,m,p,s,v) predicted the day's precipitation relatively well, despite several instances of overconfidence (e.g., central Colorado, northeastern Mississippi, and eastern California in Fig. 2.15a; extreme southwestern Kentucky in Fig. 2.15j) and misses (e.g., northwestern Nebraska in Fig. 2.15d; southern Iowa in Fig. 2.15m). Spatially smoothing the raw ensemble probabilities (Fig. 2.15b,e,h,k,n,q,t,w) generally helped improve calibration and POD, but forecasts remained imperfect. For example, 0.1-in. SREF exceedance probabilities (Fig. 2.15b) remained near 1 in southwestern Kentucky and northeastern Mississippi, while the 0.1-in. HREFv2 smoothed probabilities over northwestern Nebraska remained less than 2%. The RFFPs (Fig. 2.15c,f,i.l,o,r,u,x) tended to fix these problems. The 0.1-in. SREF-based RFFPs gave smaller probabilities in northeastern Mississippi (Fig. 2.15c), while the 0.1-in. HREFv2-based RFFPs gave higher (i.e., 2-10%) probabilities in northwestern Nebraska. In general, the RFFPs (Fig. 2.15c,f,I,l,o,r,u,x) had good calibration, sharpness, and resolution. They tended to increase POD and sharpness compared to the spatially smoothed forecasts while only modestly increasing POFD. For example, the HREFv2 1-in. RFFPs (Fig. 2.15r) gave higher probabilities in northern Alabama compared to the raw (Fig. 2.15p) and smoothed (Fig. 2.15q) HREFv2 forecasts while

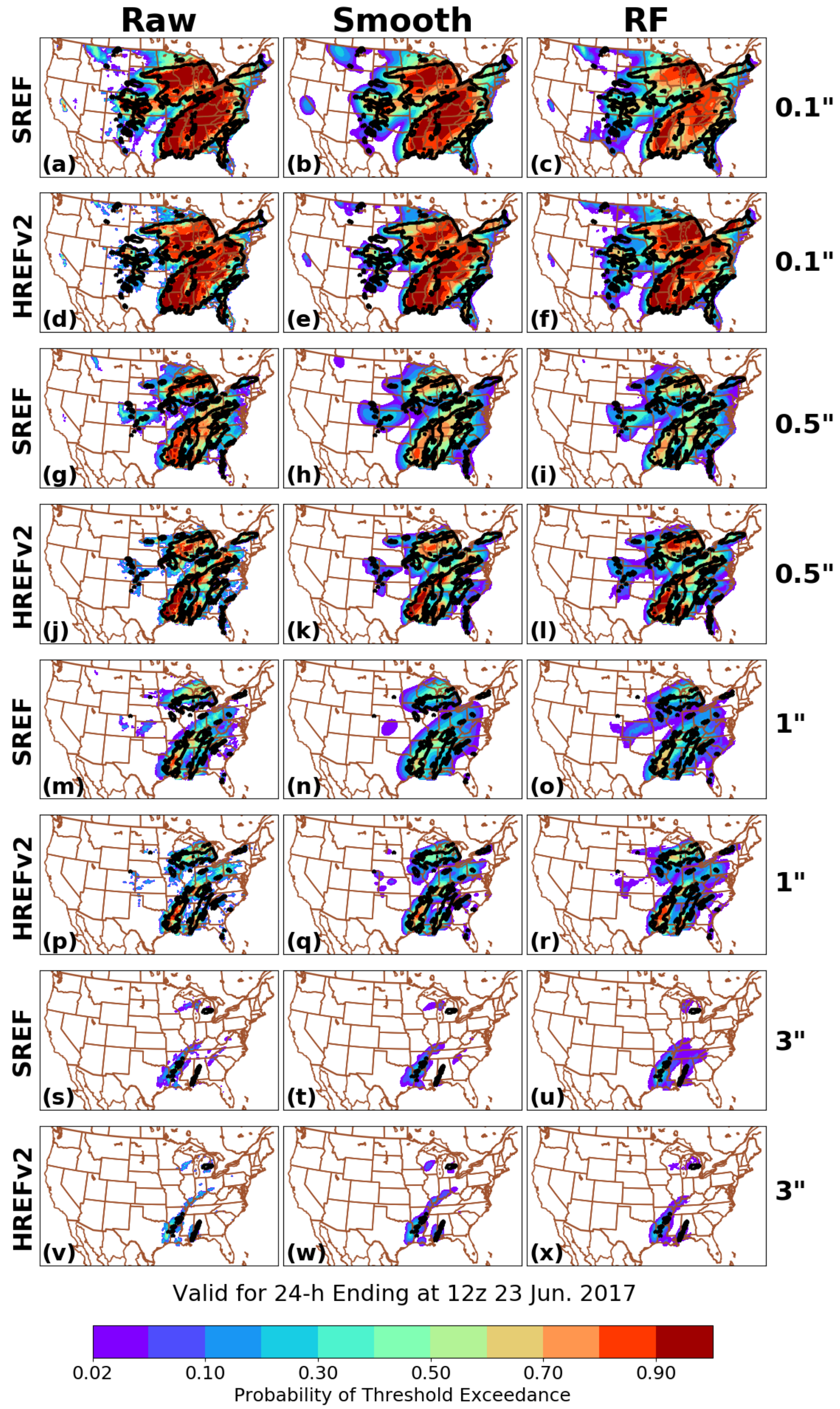


Figure 2.15 As in Fig. 2.14 but for the 24-h period ending at 1200 UTC on 23 June 2017.

the false alarm area increased only slightly. Similarly, the SREF-based 3-in. RFFPs had better POD in central Alabama (Fig. 2.15u) with a false alarm area only slightly greater than the corresponding raw and smoothed ensemble forecasts (Fig. 2.15s-t). While the RFFPs didn't always improve on the raw and smoothed ensemble probabilities (e.g., central Michigan in Fig. 2.15v-x), the general performance of the RFFPs was strong.

4. Summary and Discussion

This paper describes a technique to post-process ensemble probabilistic precipitation forecasts year-round over the contiguous United States (CONUS) using a single random forest (RF). Specifically, the RF-based post-processing is applied to 24-h (1200 UTC – 1200 UTC) probabilistic precipitation forecasts from the Short-Range Ensemble Forecast System (SREF; Du et al. 2015) and the High-Resolution Ensemble Forecast System, Version 2 (HREFv2; Jirak et al. 2018; Roberts et al. 2019) at four precipitation thresholds: 0.1-in. (2.54-mm.), 0.5-in. (12.7-mm.), 1-in. (25.4-mm.), and 3-in. (76.2-mm.). Random forest forecast probabilities (RFFPs) are compared against each ensemble's raw probabilities (i.e., the fraction of members exceeding a threshold) and spatially-smoothed probabilities (i.e., raw ensemble probabilities smoothed in space using an isotropic 2-dimensional Gaussian kernel density function to optimize the Brier Score).

Relative to these baseline forecasts, the RFFPs provide better reliability and resolution, fewer spatial biases, and statistically greater Brier Skill Scores (BSSs) and areas under the relative operating characteristics curve (AUCs). The RFFPs perform best at lower thresholds, which have greater climatological frequencies and thus provide more examples of "yes observations" for the algorithm to use to discern data patterns associated with threshold exceedance. The RF-based post-processing also benefits the SREF more than the HREFv2, a result that makes sense given that the raw SREF contains more systematic biases than the raw HREFv2. The result may also indicate that different ensembles require different sets of predictor variables to achieve the best post-processing benefits. For example, it is possible that, for the HREFv2, the ensemble mean is not as meaningful as an ensemble summary characteristic as it is for the SREF. Similarly, it is possible that the HREFv2 forecast variables contain more small-scale noise than those from the SREF because of the HREFv2's finer horizontal grid-spacing.

The biggest advantage of the RFFPs is that they provide a convenient "summary" product that is calibrated with respect to forecast probability magnitudes and spatial coverage. While near-perfect reliability can also be achieved using 2-dimensional spatial smoothing with the proper value of σ, spatially smoothing ensemble probabilities reduces sharpness (e.g., Sobash et al. 2011, 2016b; Loken et al. 2017, 2019) and potentially sacrifices resolution if too much smoothing is required. Moreover, the "best" value of σ may vary based on geographic location and time of year (e.g., Fig. 2.3), as precipitation uncertainty is reduced where stronger and/or more predictable forcing is present, such as near high terrain (e.g., Blake et al. 2018) or during the cold season (e.g., Schwartz et al. 2019). Thus, while a time- and space-varying σ may be required to properly calibrate forecasts using spatial smoothing, the RF-based approach implicitly accounts for spatial and temporal variations in precipitation uncertainty.

In practice, RFFPs could provide value to forecasters as an ensemble summary product that would eliminate the need for internal forecaster calibration of ensemble biases. Indeed, the RFFPs would fill an important operational need by quickly conveying reliable uncertainty information to the forecaster (Evans et al. 2014). The RFFPs could also be used as an automated "first guess" probabilistic precipitation forecast field, which could increase forecaster efficiency (e.g., Karstens et al. 2018). Importantly, the implementation of RFFPs into operations would be computationally feasible. While training RFs can be expensive, particularly when many predictor variables and training examples are used, using a trained RF to make real-time predictions is cheap. For example, real-time RFFPs are currently being generated from 0000 UTC HREFv2 data. Including the preprocessing step, the RFFPs can be made in 30 minutes or less on a single processor.

Nevertheless, ML-based post-processing has several important drawbacks. Most notably, since ML-based techniques "learn" based on past results, they require quality historical datasets of sufficient length for both the forecast and observations. When modifications are made to the ensemble forecast system, it is often advisable to retrain the RF with forecast data from the new system, since, while the underlying statistical relationships between the forecast and observed variables may generally hold, the optimal splitting thresholds in the RF may change as biases enter or exit the ensemble system. It is an open question (and probably situation-dependent) whether the RF can be retrained simply by adding the new forecast data to the training set (along with the old data) or if the RF should be retrained entirely "from scratch" using only the new data. Fortunately, even if the RF requires retraining from scratch, preliminary results herein suggest that a training set of "only" 93 – 217 days is required to create skillful RFFPs; nevertheless, even 93 days represents a substantial gap between the implementation of the new system and the ability to create skillful RFFPs. Moreover, due to the reduced observed climatological frequency of the higher threshold exceedances, it may be necessary to have more data for the RFFPs to outperform spatially-smoothed ensemble probabilities at the highest thresholds (e.g., 3-in. and greater), which tend to be most impactful in terms of their threat to life and property. Another drawback of the RF-based approach is that the RFFPs are not *always* superior to raw or spatially smoothed ensemble probabilities at every location during every day, and it can be difficult to determine where and why the ML algorithm struggles, particularly in the absence of interpretability information (e.g., partial dependence plots and individual conditional expectation plots, Goldstein et al. 2015; variable importance, McGovern et al. 2017). Therefore, developing and applying useful ML interpretability metrics is an important topic of ongoing research (e.g., Gagne et al. 2019; Herman and Schumacher 2018a). Another important limitation of ML compared to other post-processing techniques is that it can require a substantial degree of hyper-parameter tuning to produce a skillful forecast. Moreover, there are no formal guidelines for constructing the ML model itself, and it can be impossible to know if the model being used is designed optimally. Finally, as with other post-processing techniques, the skill of the RFFPs will ultimately be related to and limited by the skill of the underlying dynamical model (e.g., Gagne et al. 2014). Therefore, while ML-based post-processing techniques can serve as useful tools, they do not eliminate the need for human forecasters and model developers.

5. Conclusion and Future Work

As computing storage and resources continue to increase, opportunities to effectively apply ML to meteorological datasets will undoubtedly become more numerous as well. This paper provides a first attempt at addressing some basic considerations regarding the utilization of machine learning for NWP post-processing. Despite the drawbacks associated with ML-based post-processing, it is found that RFFPs can provide calibrated probabilistic precipitation forecasts whose quality matches or exceeds that of spatially-smoothed ensemble probabilities. Indeed, it is promising that a single RF can attain such forecast quality, especially given the relatively simplistic RF design and short (i.e., < 1.5-year) dataset.

Future work should explore using more complex ML-based techniques for post-processing and/or other RF constructions. For example, in the present study, individual-member forecasts were initially used as predictors, but this implementation consumed too much memory to be feasible. However, if variable importance and/or feature selection (e.g., McGovern et al. 2017; Herman and Schumacher 2018c) were used to strategically reduce the number of predictor variables, predictors from more sources could potentially be incorporated into the algorithm. Including interpretability metrics (e.g., partial dependence plots or individual conditional expectation plots; Goldstein et al. 2015) may also provide value to forecasters using the product in real-time. Given that the precipitation climatology over the CONUS varies in space and time (Schumacher and Johnson 2006), using separate RFs for individual regions and seasons may add further interpretability and skill to the RFFPs. Other ML methods, such as deep learning, may produce better RFFPs and enhance interpretability as well. Because this study examined the impacts of ML-based post-processing on "ad-hoc," multi-model, multi-physics ensembles, future work should investigate how ML-based post-processing affects other, more formally designed ensembles (e.g., the NCAR Ensemble; Schwartz et al. 2015, 2019). Finally, future work may wish to apply the general methods of this study to other prediction problems, such as severe weather, forecasting for longer or shorter time periods, and summarizing ensemble output from multiple NWP sources. It is also recommended that current and future products be evaluated in an operational setting, such as the Flash Flood and Intense Rainfall Experiment (Albright and Perfater 2018) or the NOAA Hazardous Weather Testbed Spring Forecasting Experiment (e.g., Gallo et al. 2017) to more directly assess value to forecasters.

Chapter 3: Generating Probabilistic Next-Day Severe Weather Forecasts from Convection-Allowing Ensembles Using Random Forests

A paper published in *Weather and Forecasting*

Eric D. Loken[1,2]*, Adam J. Clark*[2,3]*, and Christopher D. Karstens*[3,4]
[1]*Cooperative Institute for Mesoscale Meteorological Studies, The University of Oklahoma, Norman, Oklahoma*
[2]*School of Meteorology, University of Oklahoma, Norman, Oklahoma*
[3]*NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma*
[4]*NOAA/NWS/Storm Prediction Center, Norman, Oklahoma*

Abstract

Extracting explicit severe weather forecast guidance from convection-allowing ensembles (CAEs) is challenging since CAEs cannot directly simulate individual severe weather hazards. Currently, CAE-based severe weather probabilities must be inferred from one or more storm-related variables, which may require extensive calibration and/or contain limited information. Machine learning (ML) offers a way to obtain severe weather forecast probabilities from CAEs by relating CAE forecast variables to observed severe weather reports. This paper develops and verifies a random forest- (RF-) based ML method for creating day 1 (1200 UTC – 1200 UTC) severe weather hazard probabilities and categorical outlooks based on 0000 UTC Storm-Scale Ensemble of Opportunity (SSEO) forecast data and observed Storm Prediction Center (SPC) storm reports.

RF forecast probabilities are compared against severe weather forecasts from calibrated SSEO 2-5km updraft helicity (UH) forecasts and SPC convective outlooks issued at 0600 UTC. Continuous RF probabilities routinely have the highest Brier Skill Scores (BSSs), regardless of whether the forecasts are evaluated over the full domain or regional/seasonal subsets. Even when

RF probabilities are truncated at the probability levels issued by the SPC, the RF forecasts often have BSSs better than or comparable to corresponding UH and SPC forecasts. Relative to the UH and SPC forecasts, the RF approach performs best for severe wind and hail prediction during the spring and summer (i.e., March – August). Overall, it is concluded that the RF method presented here provides skillful, reliable CAE-derived severe weather probabilities that may be useful to severe weather forecasters and decision-makers.

1. Introduction

With horizontal grid-spacing less than approximately 4-km, convection-allowing models (CAMs) are important tools for severe weather forecasters, since they adequately resolve the dominant circulations of individual convective storms without convective parameterization (e.g., Weisman et al. 1997; Done et al. 2004). As a result, CAMs more accurately predict storm initiation, evolution, intensity, and mode compared to convection-parameterizing models (e.g., Kain et al. 2006, 2008). Depiction of storm mode is especially useful to severe weather forecasters (e.g., Kain et al. 2006; Clark et al. 2012a) since a storm's morphology is related to its attendant hazards (e.g., Gallus et al. 2008; Duda and Gallus 2010; Schoen and Ashley 2011; Smith et al. 2012). However, CAMs currently lack horizontal grid-spacing fine enough to explicitly simulate individual tornadoes, hailstones, or microscale severe wind events. Therefore, forecasters using CAM guidance must infer simulated severe weather occurrence from modeled storm attributes that are correlated with observed severe weather (e.g., Sobash et al. 2011).

An example of a commonly-used simulated severe storm "surrogate" (Sobash et al. 2011, 2016b, 2019), or proxy, is hourly maximum 2-5 km above ground level updraft helicity (hereafter, UH; e.g., Kain et al. 2008, 2010; Guyer and Jirak 2014; Loken et al. 2017; Sobash et al. 2011, 2016b, 2019). Large values of UH identify not only rotating updrafts associated with supercells, but also the sheared updrafts associated with severe mesoscale convective systems (MCSs; Sobash et al. 2011). As a result, UH has been found to be a skillful predictor of all-hazards severe weather (Kain et al. 2008; Sobash et al. 2011, 2016b, 2019). UH has also been used—generally in conjunction with simulated environmental variables—to forecast tornadoes (Clark et al. 2013; Guyer and Jirak 2014; Gallo et al. 2016; Sobash et al. 2019) and severe wind and hail (Jirak et al. 2014). Other common simulated severe weather proxies include large values of hourly maximum upward vertical velocity (e.g., Roberts et al. 2019), low-level vertical vorticity (e.g., Skinner et al. 2016; Sobash et al. 2019) and UH integrated from 0-1 km above the surface (Sobash et al. 2019).

One major drawback of these proxies is that they require extensive calibration to perform optimally. For example, Sobash and Kain (2017) demonstrated that the best UH threshold to use for all-hazards severe weather prediction varies by location and time of year. Moreover, if binary proxies are smoothed spatially to obtain probabilistic forecasts (e.g., Sobash et al. 2011, 2016b, 2019; Loken et al. 2017), the degree of spatial smoothing must be properly calibrated as well. Too little smoothing results in over-forecasting bias, while too much can yield under-forecasting and degrade sharpness and resolution (e.g., Sobash et al. 2011, 2016b; Loken et al. 2017, 2019a,b). Additionally, these calibrations are CAM- and hazard-dependent. For example, Clark et al. (2012b, 2013) used a larger UH threshold and smaller degree of spatial smoothing to forecast tornado path lengths compared to that used by Sobash et al. (2011) to forecast all-hazards severe weather, while Gagne et al. (2017) used different UH thresholds to predict 25- and 50-mm diameter hail.

Another important drawback of simulated severe weather proxies is that they use limited information to determine the severe weather threat. For example, Clark et al. (2012b) and Gallo et al. (2016) noted that large values of UH may exist in environments that are not conducive to severe weather. However, even when proxies are filtered based on the simulated environment (e.g., Clark et al. 2012b; Jirak et al. 2014; Gallo et al. 2016), the resulting predictions may still be suboptimal since severe weather can still occur in locations with unfavorable simulated environments if the CAM has biases or is not representing the observed environment well. Moreover, the use of environment-based filtering does not mean the resulting prediction has considered all relevant forecast variables.

Another way to extract explicit severe weather guidance from CAMs is to statistically relate multivariate CAM output with the observed occurrence of severe weather. Indeed, this is the general approach of Model Output Statistics (MOS; Glahn and Lowry 1972; Klein and Glahn 1974), which has shown promise for a variety of forecast fields, including: probability of precipitation, maximum and minimum temperatures, cloud coverage, near-surface wind, conditional probability of precipitation, and thunderstorms (e.g., Glahn and Lowry 1972; Klein and Glahn 1974; Carter 1975; Bermowitz 1975; Schmeits et al. 2005; Kang et al. 2011). However, MOS relationships tend to be linear and based on regression while relationships between CAM forecast variables and observed severe weather are likely to be flow-dependent and nonlinear (e.g., Legg and Mylne 2004; Melhauser and Zhang 2012; Torn and Romine 2015; Trier et al. 2015). Thus, machine learning (ML) techniques, which can model nonlinear relationships, may be more appropriate for diagnosing the severe weather threat conveyed by CAM or convection-allowing ensemble (CAE) guidance.

Indeed, recent studies have successfully used ML techniques to create probabilistic precipitation (e.g., Gagne et al. 2014; Herman and Schumacher 2018c; Loken et al. 2019a) and severe weather (e.g., Gagne et al. 2017; Lagerquist et al. 2017; Burke et al. 2020) forecasts based partly or entirely on numerical weather prediction (NWP) predictors. For severe weather prediction, a common approach has been to use predictors associated with storm "objects," which are identified by thresholding a certain simulated storm attribute (e.g., maximum hourly column total graupel mass in Gagne et al. 2017; maximum hourly upward vertical velocity in Burke et al. 2020). Thus, the object identification process "filters out" areas of weaker or non-existent simulated storms. Such an approach is efficient for ML training since it eliminates the need to consider predictors from all grid points but can underperform if there is poor correspondence between simulated and observed storms (Gagne et al. 2017). Conversely, when grid-point-based predictors are used, training takes longer, but higher performance may be achieved when the CAE is imperfect, since the grid-point predictors offer the ML algorithm more (and more relevant) information. Moreover, when grid-point-based predictors and predictands are used, output probabilities are directly given in 2-dimensional (rather than object) space, facilitating user interpretation of ML output.

While grid-point-based methods have been used to obtain skillful probabilistic precipitation forecasts (Herman and Schumacher 2018c; Loken et al. 2019a), they are untested for severe weather prediction. Therefore, this study seeks to develop and evaluate an RF-based method for creating individual-hazard Day 1 (i.e., 1200 UTC – 1200 UTC) severe weather probabilities from grid-point-based CAE forecast output. Due to its skill (Jirak et al. 2016, 2018) and long data archive, the SPC's 7-member Storm Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012, 2016, 2018) is used as the underlying dynamical forecast system. For evaluation against operationally-relevant baselines, the RF-based severe weather forecasts are compared to SSEO UH-based probabilistic forecasts and SPC Day 1 Convective Outlooks (COs) issued at 0600 UTC[2]. While multiple previous studies have applied ML to severe weather prediction, the RF method described herein is unique in that it uses grid-point-based CAE forecast fields as predictors, produces probabilistic forecasts for multiple severe weather hazards over the full contiguous United States (CONUS), and is directly evaluated against top-performing human and NWP baselines.

The remainder of the paper is organized as follows: section 2 describes the methods, section 3 presents the results, section 4 analyzes two representative case studies, section 5 discusses key aspects of the results, and section 6 summarizes and concludes the paper.

2. Methods

a. Datasets

The forecast and observational datasets used herein span 629 days from late April 2015 through early July 2017 (Table 3.1). RF- and UH-based severe weather forecasts are derived from the SSEO (Jirak et al. 2012, 2016), a 7-member CAE with members that use different initial and lateral boundary conditions, initialization times, and microphysics and turbulence parameterizations. Since SPC forecasters began using the SSEO in 2011 (Jirak et al. 2016), its convection-related forecasts have compared favorably with those from other experimental CAEs (Jirak et al. 2016). As a result, the SSEO was ultimately formalized as the High-Resolution Ensemble Forecast System Version 2 (HREFv2), which became the first operational CAE run by

[2] 0600 UTC SPC COs are used because SPC forecasters, like the RFs, have access to 0000 UTC SSEO guidance during that forecast period.

Month	2015	2016	2017	Total
January	-	1-9, 11-13, 15-31	3-27, 29-31	57
February	-	1-29	2-21, 23, 25-26, 28	53
March	-	1-7, 9-11, 13-19, 21-22, 25-31	1-7, 9-11, 13-20, 28-31	48
April	21-26, 28-30	1-10, 12-30	2-4, 6-27, 29-30	65
May	1-10, 12-19, 23-31	1, 3-10, 13-31	1-9, 12, 14-31	83
June	1-9, 11, 13-19, 21-30	1-8, 10-20, 22-23, 25-30	1-6, 8-11, 13-14, 16, 18-22, 24	73
July	1-6, 8-10, 12-16, 20-29	2, 8-14, 16-20, 23-26, 28, 30-31	1-4	48
August	1-3, 5-9, 11-14, 18-31	1-4, 6-10, 12-31	-	55
September	1-9, 11, 14, 19-21, 23-30	1, 3, 5-6, 9-13, 17-22, 24-26, 28	-	41
October	1-6, 9-31	1-2, 4-12, 14-18, 21-24, 27-31	-	54
November	1-3, 5-14, 16-18, 20-23, 25, 27-30	-	-	25
December	1, 4-8, 10, 12-31	-	-	27
Total	216	261	152	629

Table 3.1 SSEO initialization dates.

the National Oceanic and Atmospheric Administration's (NOAA's) Environmental Modeling Center in November 2017 (Jirak et al. 2018; Roberts et al. 2019; Loken et al. 2019b). All SSEO member forecasts are provided on a 4-km contiguous United States (CONUS) domain with 1199 × 799 points. Full SSEO specifications are summarized in Table 3.2.

SSEO forecasts are compared against SPC Day 1 COs, which are issued daily by 0600 UTC and are valid from 1200 UTC to 1200 UTC the following day. These COs include probabilistic forecasts of tornadoes, severe wind [i.e., wind speeds of at least 50 kts (58 mph)],

Member	Dynamic Core	ICs/LBCs	Microphysics	PBL	Initialization Time
NSSL-WRF	WRF-ARW	NAM/NAM	WSM6	MYJ	0000 UTC
HRW ARW	WRF-ARW	RAP/GFS	WSM6	YSU	0000 UTC
HRW ARW (Time-Lagged)	WRF-ARW	RAP/GFS	WSM6	YSU	1200 UTC (12-h time lag)
HRW NMMB	NMMB	RAP/GFS	Ferrier	MYJ	0000 UTC
HRW NMMB (Time-Lagged)	NMMB	RAP/GFS	Ferrier	MYJ	1200 UTC (12-h time lag)
WRF-NMM	WRF-NMM	NAM/NAM	Ferrier	MYJ	0000 UTC
NAM NEST	NMMB	NAM/NAM	Ferrier-Aligo	MYJ	0000 UTC

Table 3.2 SSEO member specifications. Dynamic cores include those from the Advanced Research Weather Research and Forecasting model (WRF-ARW; Skamarock et al. 2008), the Weather Research and Forecasting Nonhydrostatic Mesoscale Model (WRF-NMM; Janjić et al. 2001; Janjić 2003), and the Nonhydrostatic Multiscale Model on the B grid (NMMB; Janjić and Gall 2012). Initial and lateral boundary conditions (ICs/LBCs) are taken 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) as indicated. Microphysics parameterizations include the 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-Janjić (MYJ; Janjić 2002) and Yonsei University (YSU; Hong et al. 2006) schemes. HRW refers to the High Resolution Window model run.

and severe hail (i.e., a maximum hailstone diameter of 1 inch or greater), with probabilities valid for within 25 miles of a point (about a 40-km radius). The COs also denote locations with a 10% or greater probability of observing significant tornadoes [i.e., those with an Enhanced Fujita (EF) rating of 2 or higher], significant severe wind [i.e., wind speeds at least 65 kts (75 mph)], and significant severe hail (i.e., a maximum hailstone diameter of 2 inches or greater) within 25 miles. Individual-hazard probabilities are then used to determine a categorical outlook forecast based on the criteria in Table 3.3.

One limitation of the SPC Day 1 COs is that only certain probability levels (i.e., 2, 5, 10, 15, 30, 45, and 60% for tornadoes; 5, 15, 30, 45, and 60% for severe wind and hail; and 10% for significant severe weather) are contoured. As a result, it is difficult to equitably compare SPC

Individual Hazard Probability	Tornado	Wind	Hail
≥ 2%	Marginal	N/A	N/A
≥ 5%	Slight	Marginal	Marginal
≥ 10%	Enhanced	N/A	N/A
≥ 10% and ≥ 10% Sig.	Enhanced	N/A	N/A
≥ 15%	Enhanced	Slight	Slight
≥ 15% and ≥ 10% Sig.	Moderate	Slight	Slight
≥ 30%	Moderate	Enhanced	Enhanced
≥ 30% and ≥ 10% Sig.	High	Enhanced	Enhanced
≥ 45%	High	Enhanced	Enhanced
≥ 45% and ≥ 10% Sig.	High	Moderate	Moderate
≥ 60%	High	Moderate	Moderate
≥ 60% and ≥ 10% Sig.	High	High	Moderate

Table 3.3 SPC conversion table relating individual hazard probabilities to categorical Day 1 COs. Adapted from https://www.spc.noaa.gov/misc/SPC_probotlk_info.html.

forecasts with the continuous RF- and UH-based forecasts from the SSEO. There are two potential remedies to this problem. The first is to truncate the SSEO-derived forecasts at the same probability levels as used by the SPC. The second is to spatially interpolate the SPC probabilities between contour levels (e.g., Herman et al. 2018). Both methods are used herein. However, in this study, continuous SPC probabilities are created using a method developed at the SPC (Karstens et al. 2019). Herein, raw SPC contours are filled/gridded using a top-hat distribution, such that all grid points enclosed by a contour are assigned that contour value. The gridding procedure is done using the General Meteorological Package (GEMPAK; desJardins et al. 1991) within a 1-degree expanded CONUS domain to negate chronic dampening of probabilities near the edges of the forecast domain. Next, unique probability areas are identified using watershed segmentation (e.g., Lakshmanan et al. 2009), and adjacent probability areas are bilinearly interpolated using a Euclidean distance transformation. Finally, the maximum probability level is assumed to be 25% greater than the maximum non-zero contoured probability level present in the forecast. Continuous SPC probabilities created using this method are henceforth referred to as “full” SPC probabilities, while the raw, discrete SPC probabilities are referred to as “original” SPC probabilities. Importantly, full SPC probabilities do not exist for significant severe weather forecasts, since the SPC only issues a 10% or greater probability contour for significant severe events. Additionally, the SPC does not issue Day 1 outlook probabilities for all-hazards severe or significant severe weather.

Severe weather observations used for verification and RF training are taken from the SPC website (SPC, 2019b) for wind and hail and the SPC Storm Events Database (SSED; SPC, 2019a) for tornadoes. The SSED was required for tornadoes since it displayed information about each tornado’s Enhanced Fujita (EF) rating, necessary for the prediction/verification of significant tornadoes. Unfiltered reports are used to account for all reported instances of severe weather.

b. UH-based forecasts

UH-based probability forecasts for each severe weather hazard are derived from the SSEO. These forecasts are created in the same manner described by Loken et al. (2017). Namely, the fraction of ensemble members exceeding a given UH threshold is noted at each grid point, and that fraction is smoothed using a 2-dimensional isotropic Gaussian kernel density function. Therefore, the UH-based probability, p, at a given grid point can be expressed as:

$$\mathrm{p} = \mathrm{f} * \left(\sum_{\mathrm{n}=1}^{\mathrm{N}} \frac{1}{2\pi\sigma^2} \exp\left[-\frac{1}{2} \left(\frac{\mathrm{d_n}}{\sigma} \right)^2 \right] \right) \qquad (3.1),$$

where f is the fraction of ensemble members exceeding some UH threshold, N is the number of points with at least one member exceeding the threshold, $\mathrm{d_n}$ is the distance between the current grid point and the *n*th point, and σ is the standard deviation of the Gaussian kernel. To determine the combination of UH threshold and σ to use for each hazard, the UH threshold is varied from 10 to 200 m^2s^{-2} in increments of 10 m^2s^{-2} while σ is varied from 30 to 210 km in increments of 30 km. The combination that optimizes the Brier Skill Score (BSS; e.g., Wilks 2011) for a given hazard over the entire dataset is used (right column of Fig. 3.1), with BSS measured relative to a constant forecast of observed hazard climatology during the 629-day dataset. The calibration is done on the 80-km verification grid (see below) rather than the native 4-km grid.

c. Random forest forecasts

1) RF METHOD OVERVIEW

A RF is an ensemble of decision trees (Breiman 2001). Individual decision trees (Breiman 1984) work by recursively splitting a dataset until a stopping criterion is reached (e.g., the tree reaches a specified maximum number of levels, the number of samples at a node falls below a specified threshold, etc.). Splitting criteria are determined by the algorithm during training. Specifically, at each node, the algorithm chooses the threshold and predictor variable that splits the data in a way that maximizes a dissimilarity metric (e.g., information gain, Gini impurity). Class predictions can then be made on unseen data by running a testing example through the tree and analyzing the training samples in the appropriate leaf node (i.e., terminal node). For example, class probabilities are expressed as the fraction of training examples associated with the given class in the leaf node containing the testing example.

Although individual decision trees are human-readable and relatively easy to interpret, they are prone to overfitting, such that small changes to a testing example's predictor variables can produce very different class predictions (e.g., Gagne et al. 2014). The RF algorithm helps remedy this overfitting tendency by growing multiple trees, which are made unique by: 1) growing each tree based on a random subset of training examples, and 2) determining the best split at each node by considering a random subset of predictor variables (Breiman 2001). During testing, RF class probabilities are simply the mean probability from each tree in the RF. In this study, RFs and corresponding RF probabilities are created using random forest classifiers from the Python module Scikit-Learn (Pedregosa et al. 2011). More information on RFs can be found in Loken et al. (2019a) and works cited therein.

2) PREDICTOR VARIABLES

The first step of creating RF-based probabilities is to determine which predictors (or input variables) the RF will consider. Here, predictor variables are based on SSEO forecast fields. However, only a small number of variables relevant to severe weather forecasting are originally stored within the SSEO data archive (i.e., the variables without asterisks in Table 3.4). To enhance RF skill, several predictor variables (i.e., those with asterisks in Table 3.4) are added to these original variables.

Storm Attribute Fields	Environment-related Fields	Other
Max. Hourly Simulated Reflectivity	2-m Temperature	Latitude*
Accumulated 1-h Precipitation	2-m Dewpoint Temperature	Longitude*
Max. Hourly Updraft Speed	2-m Relative Humidity	Smoothed UH probabilities* (created by maximizing AUC)
Max. Hourly UH	MUCAPE	
	CIN	
	0-6 km Shear	
	CAPE × Shear*	

Table 3.4 Predictor variables. Asterisks denote variables that were added during pre-processing.

For example, the product of most unstable convective available potential energy (MUCAPE) and 0-6km wind shear is computed at each native 4-km grid point and stored as a predictor variable.

Latitude, longitude, and smoothed UH probabilities are also added as predictors during preprocessing.

3) DATA PREPROCESSING

While the SSEO contains a limited number of archived forecast fields, there is originally an overwhelming amount of *data* potentially available to the RF, since each SSEO member forecasts each variable at 3-km grid-spacing over the CONUS every hour. To make training the RF computationally feasible, the dimensionality of the SSEO dataset must be reduced through several steps of data preprocessing.

The first preprocessing step is to reduce the temporal dimension of the dataset. This is accomplished by taking a 24-h (1200 UTC – 1200 UTC) temporal maximum (for the storm attribute variables; Table 3.4) or mean (for the environment-related variables) at each 4-km grid point. Next, these temporally-aggregated forecast variables—as well as the observed storm reports—are remapped to an approximately 80-km grid (i.e., NCEP grid 211) to further reduce dataset dimensionality and to match the verification scales used by the SPC. For the storm attribute fields, remapping is done by selecting the maximum forecast value on the 4-km grid within each 80-km grid box. For the environment-related fields, remapping to the 80-km grid is done using a neighbor budget method (Accadia et al. 2003), which approximately conserves the remapped quantity. After remapping, the ensemble mean, maximum, minimum, and standard deviation values are computed for each forecast variable at every 80-km grid point. Additionally, smoothed UH probabilities (to be used as predictors) are derived based on the method in section 2b. However, the UH threshold and standard deviation of the Gaussian kernel combination used is that which maximizes area under the relative operating characteristic curve (AUC; e.g., Wilks

2011; left column of Fig. 3.1) rather than BSS, since AUC is a measure of potential skill after bias calibration (Wilks 2011) and RF outputs typically have low bias (e.g., Breiman 2001).

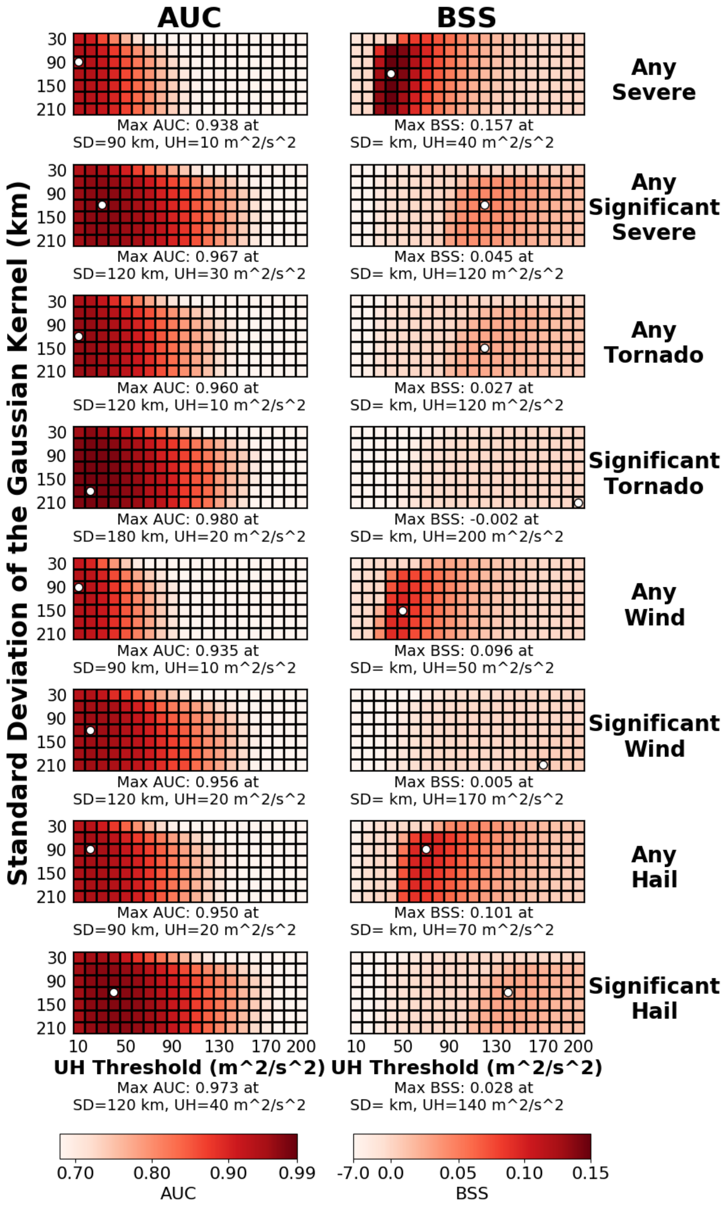


Figure 3.1 Heat maps showing how area under the relative operating characteristics curve (AUC; left column) and Brier Skill Score (BSS; right column) vary with the standard deviation of the Gaussian kernel and UH threshold for UH-based forecasts. Heat maps are for any severe weather hazard (row 1), any significant severe weather hazard (row 2), any tornado (row 3), significant tornadoes (row 4), any severe wind (row 5), significant severe wind (row 6), any severe hail (row 7), and significant severe hail (row 8). In each case, the combination with the highest AUC or BSS is indicated by a white circle and noted below the plot. AUC is used for calibrating smoothed UH RF inputs, while BSS is used for calibrating the smoothed UH forecasts themselves.

After preprocessing, a final set of predictors is obtained for input into the RF. Here, these predictors include the ensemble mean, maximum, minimum, and standard deviation of SSEO forecast fields as well as latitude, longitude, and UH-based probabilities (Table 3.4). For a given grid point prediction, the RF considers these quantities at the 25 closest 80-km grid points.

4) RF PREDICTIONS

The RF gives probabilistic predictions of whether a given 80-km grid box will experience the occurrence of at least one observed severe weather report (all-hazards or individual-hazard) over the 24-h Day 1 CO period (i.e., 1200 UTC – 1200 UTC). Separate RFs are used to predict the occurrence of: all-hazards severe weather, all-hazards significant severe weather, any tornadoes, significant tornadoes, any severe wind, significant severe wind, any severe hail, and significant severe hail. Finally, the predictions from these separate RFs are used to construct an RF-based Day 1 categorical outlook using the same guidelines employed by the SPC (i.e., those in Table 3.3).

5) DISCRETE/TRUNCATED RF PROBABILITIES

To facilitate a fair comparison with the SPC Day 1 outlooks, discrete RF probabilities are created for individual-hazards severe and significant severe weather forecasts using the same probability levels as the SPC (Table 3.3). Discrete RF probabilities (henceforth referred to as truncated RF forecasts) are created by simply converting all continuous RF probabilities between discrete SPC probability levels to the lower probability. For example, continuous severe hail probabilities between 5% (inclusive) and 15% (exclusive) are converted to 5% probabilities, since they would all be contained within a 5% SPC contour. Similarly, for individual-hazard significant severe forecasts, truncated RF probabilities are 10% if the continuous RF probabilities meet or exceed 10% and 0% otherwise.

c. Verification

Probabilistic severe weather forecasts are evaluated over the entire CONUS (Fig. 3.2a) as well over the West, Midwest, and East (Fig. 3.2b), which are defined based on temperature and precipitation climatology and represent an aggregation of regions described in Bukovsky (2011). Forecasts are also analyzed seasonally, with Winter, Spring, Summer, and Fall defined as December–February, March–May, June–August, and September–November, respectively.

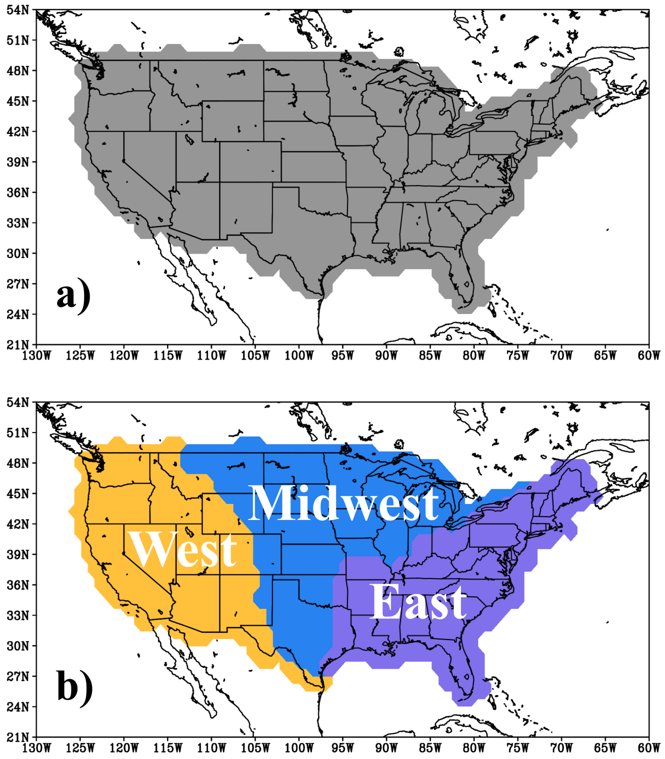


Figure 3.2 (a) Overall analysis domain (gray shading). (b) West (yellow), Midwest (blue), and East (purple) region analysis domains.

Forecasts are verified on the ~80-km NCEP grid 211 to approximately match the verification definitions used by the SPC, which evaluates the occurrence of severe weather within 40-km of a point, and to save computational expense during verification. Continuous RF, truncated RF, original SPC, full/continuous SPC, and (continuous) UH-based probabilities are evaluated and compared against each other whenever possible. Unfortunately, due to the limitations of the SPC forecasts, full SPC probabilities are not created for significant severe weather forecasts, and neither original nor full SPC probabilities exist for all-hazard severe or significant severe forecasts. Additionally, no quantitative verification is performed on the RF- and SPC-based categorical outlooks, since these are not true probabilistic forecasts, but rather summary products that merge probability and intensity information. Forecast evaluation is done using 17-fold cross validation with 37 days per fold. 17 folds are used here to balance the tradeoff between computational expense and training set size and to provide an equal number of days (37) in each fold. As in Loken et al. (2019a), verification statistics are computed over the full set of 629 forecasts derived from each fold's testing set.

Metrics used for verification include: BSS, BS components (e.g., Wilks 1995), attributes diagrams (e.g., Hsu and Murphy 1986), and performance diagrams (Roebber 2009). While AUC is used to set the UH threshold and Gaussian kernel standard deviation for smoothed UH-based predictors, it is not used for forecast evaluation since it is not sensitive to bias and it tends to increase nonlinearly with increasing forecast skill such that two well-performing but differently-skilled forecast systems may have similar AUC values near 1 (Marzban 2004).

The BS (e.g., Wilks 1995), which measures the magnitude of forecast probability errors, can be decomposed into reliability, resolution, and uncertainty components (Murphy 1973; Wilks 1995), and is defined as:

$$\mathrm{BS} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{p_i} - \mathrm{o_i})^2 = \frac{1}{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\left(\mathrm{n_k}(\mathrm{p_k} - \bar{\mathrm{o}}_\mathrm{k})\right)^2 - \frac{1}{\mathrm{N}}\sum_{\mathrm{k}=1}^{\mathrm{K}}\left(\mathrm{n_k}(\bar{\mathrm{o}}_\mathrm{k} - \bar{\mathrm{o}})\right)^2 + \bar{\mathrm{o}}(1 - \bar{\mathrm{o}}) \quad (3.2),$$

where N is the total number of forecast/observation pairs, K is the number of forecast probability bins, $\mathrm{p_i}$ is the forecast probability at point *i*, $\mathrm{o_i}$ is the binary observation (i.e., 0 or 1) at point *i*, $\mathrm{n_k}$ is the number of forecasts in bin *k*, $\bar{\mathrm{o}}_\mathrm{k}$ is the mean observed relative frequency in bin *k*, and $\bar{\mathrm{o}}$ is the overall sample climatological frequency. The three terms on the right of equation (3.2) represent the reliability, resolution, and uncertainty components of the BS, respectively. Meanwhile, the BSS compares the BS to that of a reference forecast, thus enabling a fair comparison for events with different climatological relative frequencies (Wilks 1995). Specifically, the BSS is defined as:

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS_{ref}}}{0 - \mathrm{BS_{ref}}} = 1 - \frac{\mathrm{BS}}{\mathrm{BS_{ref}}} \quad (3.3),$$

where, herein, $\mathrm{BS_{ref}}$ is the BS resulting from always forecasting the observed climatology of the relevant dataset. A BSS of 1 (0) indicates perfect (no) skill relative to the reference forecast. Ninety-five percent confidence intervals (95CIs) for each forecast's BSS values are determined using resampling with replacement (i.e., bootstrapping; e.g., Wilks 2011). Specifically, 629 random samples (with replacement) are drawn from a given forecast's 629 individual-day BS values. The aggregate BS and BSS over the random sample are then computed and stored. After 10,000 iterations of this process, the 95% BSS confidence interval is noted by observing the 2.5- and 97.5-percentile values of the stored BSS distribution.

While the reliability component of the BS provides a single-number summary of how well forecast probabilities correspond with observations, attributes diagrams allow users to assess reliability separately for each of *k* probability bins. Herein, bins are defined by the following probability level ranges: [0-1%), [1-2%), [2-5%), [5-15%), [15-25%), …, [85-95%), and [95-100%]. Perfectly-reliable forecasts fall along a line of slope 1 passing through the origin; over- (under-) forecasts fall below (above) this line. Attributes diagrams also contain horizontal and vertical lines plotted at the sample climatological relative frequency as well as a no-skill line located halfway between the horizontal climatology line and the perfect reliability line. Points above (below) the no-skill line contribute positively (negatively) to the BSS when a reference forecast of climatology is used (Wilks 1995).

Performance diagrams (Roebber 2009) binarize probabilistic forecasts at specific probability levels (herein, 0, 1, 2, 5-95% in increments of 10%, and 100%) and display probability of detection (POD), success ratio (SR), bias, and critical success index (CSI) on a single plot (e.g., see Roebber 2009 equations 1-4). Points falling closer to the upper right-hand corner of the diagram exhibit greater skill, since POD, SR, CSI, and bias are all optimized at a value of 1. Moreover, POD, SR, CSI, and bias are all independent of the number of correct negatives, making the performance diagram a good tool for evaluating forecasts with many trivial correct negatives.

3. Results

a. Full-domain, full-period results

The continuous RF forecasts have the greatest overall BSS values for each of the hazards examined (Fig. 3.3a). Compared to the UH-based forecasts, the continuous RF forecasts give substantially better predictions for all hazards except significant tornadoes (Fig. 3.3a). This is an important result given that UH is a skillful predictor of severe weather (e.g., Kain et al. 2008; Sobash et al. 2011, 2016b, 2019) and is widely used in testbed settings (e.g., Kain et al. 2008;

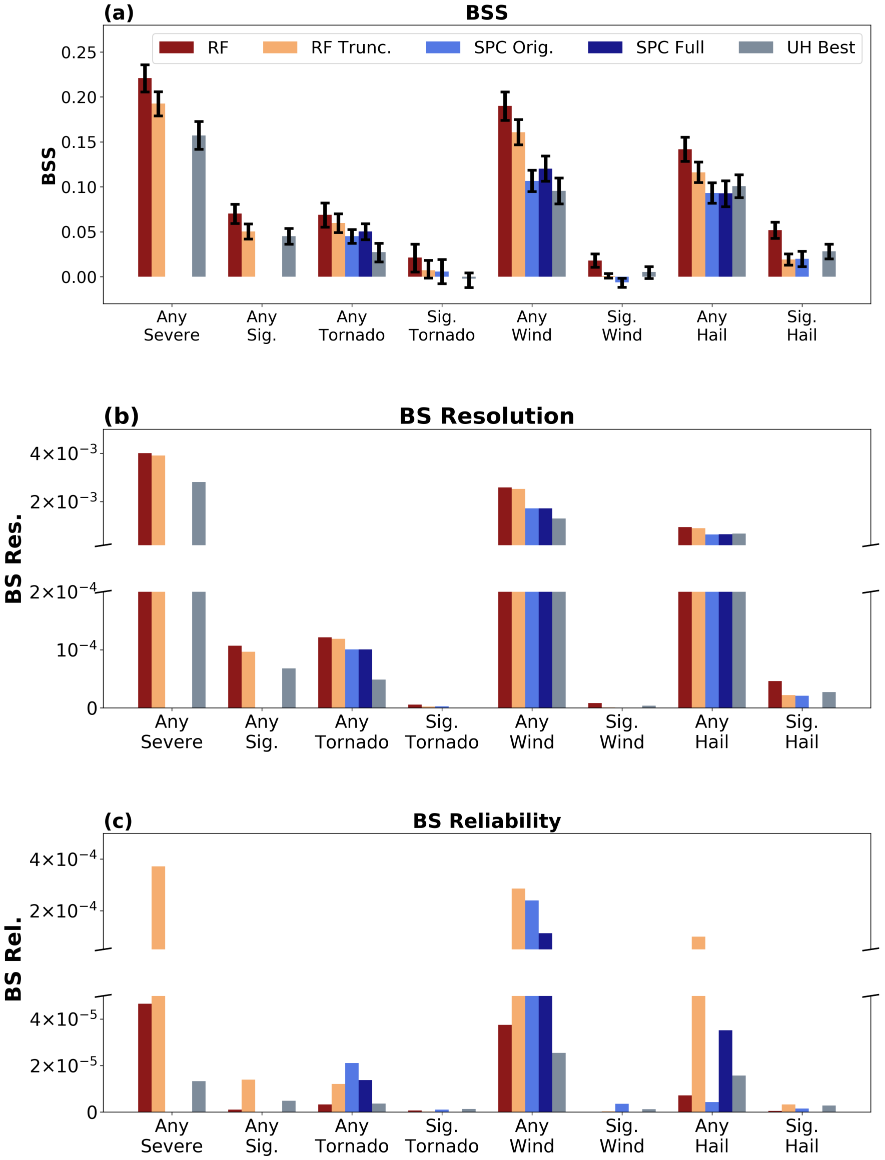


Figure 3.3 (a) CONUS-wide BSS for the full/continuous RF-based probabilities (dark red), truncated RF-based probabilities (yellow), original SPC probabilities (light blue), full/continuous SPC probabilities (dark blue), and UH-based probabilities determined using the optimal standard deviation and UH threshold combination for each hazard (gray). (b)-(c) As in (a) but for the resolution and reliability components of the BS, respectively. Black bars denote 95% confidence intervals in (a). In (b) and (c), axes are scaled differently on either side of the breaks, allowing for easier interpretation of all data. Note that the SPC does not issue forecasts for all-hazards severe or significant severe weather and that full/continuous SPC probabilities are not available for the individual significant severe hazards.

Clark et al. 2012a; Guyer and Jirak 2014; Gallo et al. 2017; Roberts et al. 2019). The continuous RF forecasts always have better resolution (Fig. 3.3b) and frequently—though not always—have better reliability (Fig. 3.3c) than the UH forecasts. Of course, it is likely that the UH-based forecasts would obtain a higher BSS if a time- and space-varying UH threshold were used instead of a constant one (Sobash and Kain 2017). However, calibrating the UH threshold in space and time requires substantially more computational resources compared to a constant threshold calibration. While training a RF is also computationally intensive, the RF considers multiple variables, and its multivariate "calibration" occurs implicitly as the algorithm is run.

The continuous RF forecasts also perform substantially better than the full SPC forecasts for hail and wind but not tornado prediction (Fig. 3.3a), an unsurprising result given this study's lack of tornado-specific predictors [e.g., significant tornado parameter (STP; Thompson et al. 2003), low-level storm relative helicity (e.g., Coffer et al. 2019), etc.]. Thus, it is possible that adding predictors with a stronger correlation to observed tornado and/or low-level mesocyclone occurrence could improve the RF tornado and significant tornado forecasts. However, even without tornado-specific predictors, the continuous RF forecasts have better resolution (Fig. 3.3b) and better (i.e., smaller) reliability values (Fig. 3.3c) than the continuous SPC forecasts for all hazards.

When the continuous RF forecasts are truncated at the probabilities used by the SPC, BSS values are, unsurprisingly, reduced (Fig. 3.3a). Much of this reduction comes from degraded reliability (Fig. 3.3c) rather than decreased resolution (Fig. 3.3b). However, the truncated RF probabilities still have substantially greater BSSs than the original SPC probabilities for severe wind (Fig. 3.3a). Truncated RFs also have higher BSSs relative to the original SPC forecasts for severe hail, with the 95CIs of the two forecasts just barely overlapping. For the significant severe hazards, the truncated RFs do not substantially outperform the original SPC forecasts. However, the continuous RF forecasts do have notably greater BSSs than the original SPC forecasts for significant severe wind and significant severe hail. This outperformance is due to the improved resolution (Fig. 3.3b) and reliability (Fig. 3.3c) that is possible with access to continuous rather than binary (i.e., ≥ 10%) forecast probabilities.

While the RF-based forecasts have the best resolution for all hazards (Fig. 3.3b), they don't necessarily have the best reliability (Fig. 3.3c); however, reliability among all forecasts for all hazards is generally very good (Figs. 3.4-3.5). Large deviations from perfect reliability are typically associated with small sample size in the relevant forecast probability bin(s) [e.g., the UH significant severe weather forecasts (Fig. 3.5a,c,e,g) at higher forecast probabilities and the RF and UH tornado probabilities greater than 30% (Fig. 3.4c)]. Interestingly, both the original and SPC probabilities under-forecast tornadoes (Fig. 3.4c) and severe wind (Fig. 3.4e). For the original SPC forecasts, this under-forecasting is at least partially due to their use of discrete probabilities (i.e., probabilities between two discrete levels are mapped to the lower level). However, the under-forecasting may also reflect a general philosophy of the SPC to emphasize higher-end tornado and wind events, given that SPC categorical outlooks are directly dependent on forecast hazard probability (Table 3.3). For example, it is possible that forecasters may wish to convey a message other than "moderate" or "high risk" to emergency managers or other users when they anticipate higher probabilities of marginally-severe wind (e.g., ~50 kts.) or low-end (e.g., EF0) tornado reports. Similarly, the SPC may wish to have high POD—even at the expense of false alarm—for significant tornadoes and significant severe wind events, which could explain the SPC over-forecasting for these hazards (Fig. 3.5c,e). The SPC does not have the same over-forecasting bias for severe (Fig. 3.4g) and significant severe (Fig. 3.5g) hail, perhaps since these

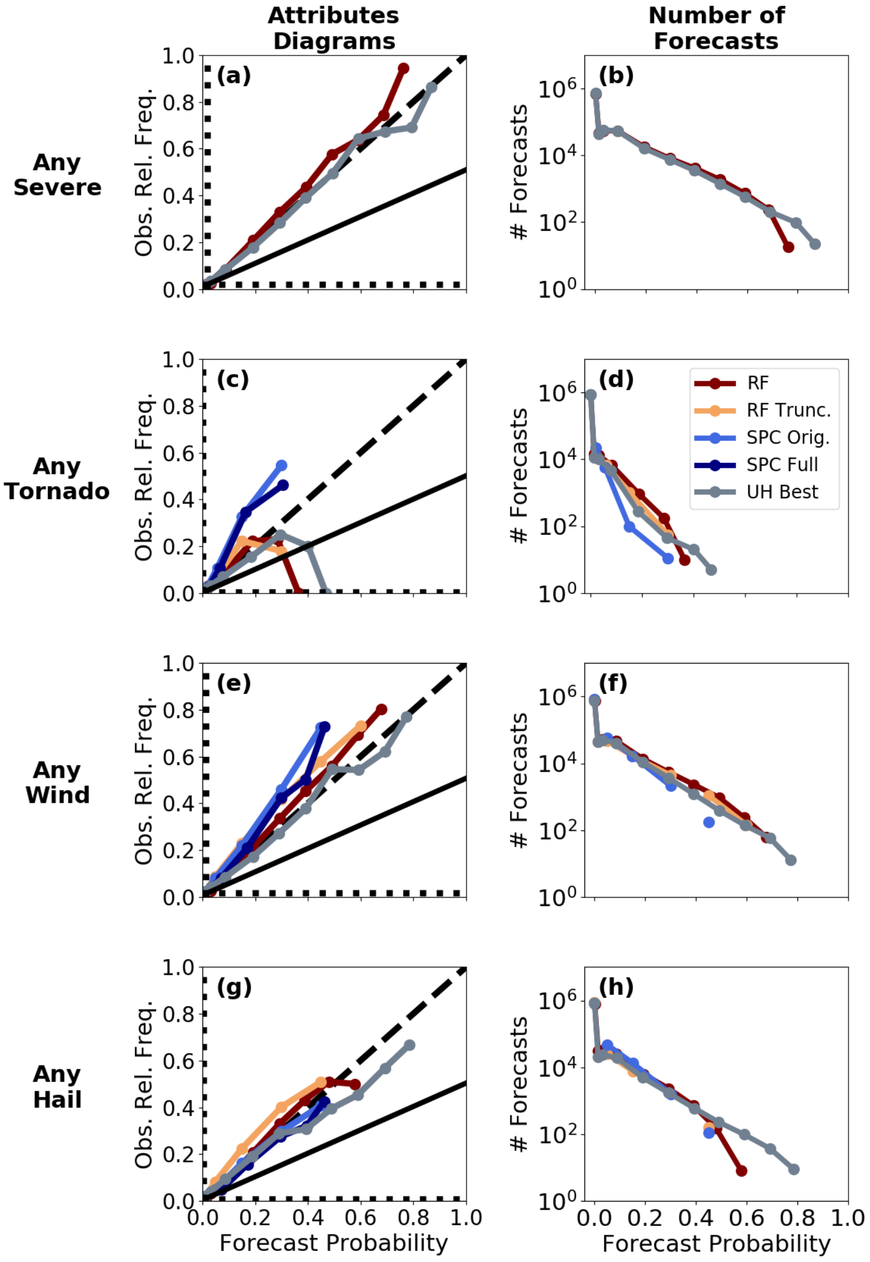


Figure 3.4 (a) Attributes diagrams for the full RF (dark red) and calibrated UH (gray) any-severe weather forecasts. The black long-dashed line indicates perfect reliability, the solid black line indicates the "no skill" line, and the black short-dashed lines represent climatological relative frequency. (b) Number of forecasts in each probability bin for the forecasts in (a). (c)-(d) As in (a)-(b) but for any tornado forecasts. Truncated RF (yellow), original SPC (light blue), and full SPC (dark blue) forecasts are shown in addition to the continuous RF (dark red) and calibrated UH (gray) forecasts. (e)-(f) As in (c)-(d) but for any severe wind forecasts. (g)-(h) As in (c)-(d) but for any severe hail forecasts.

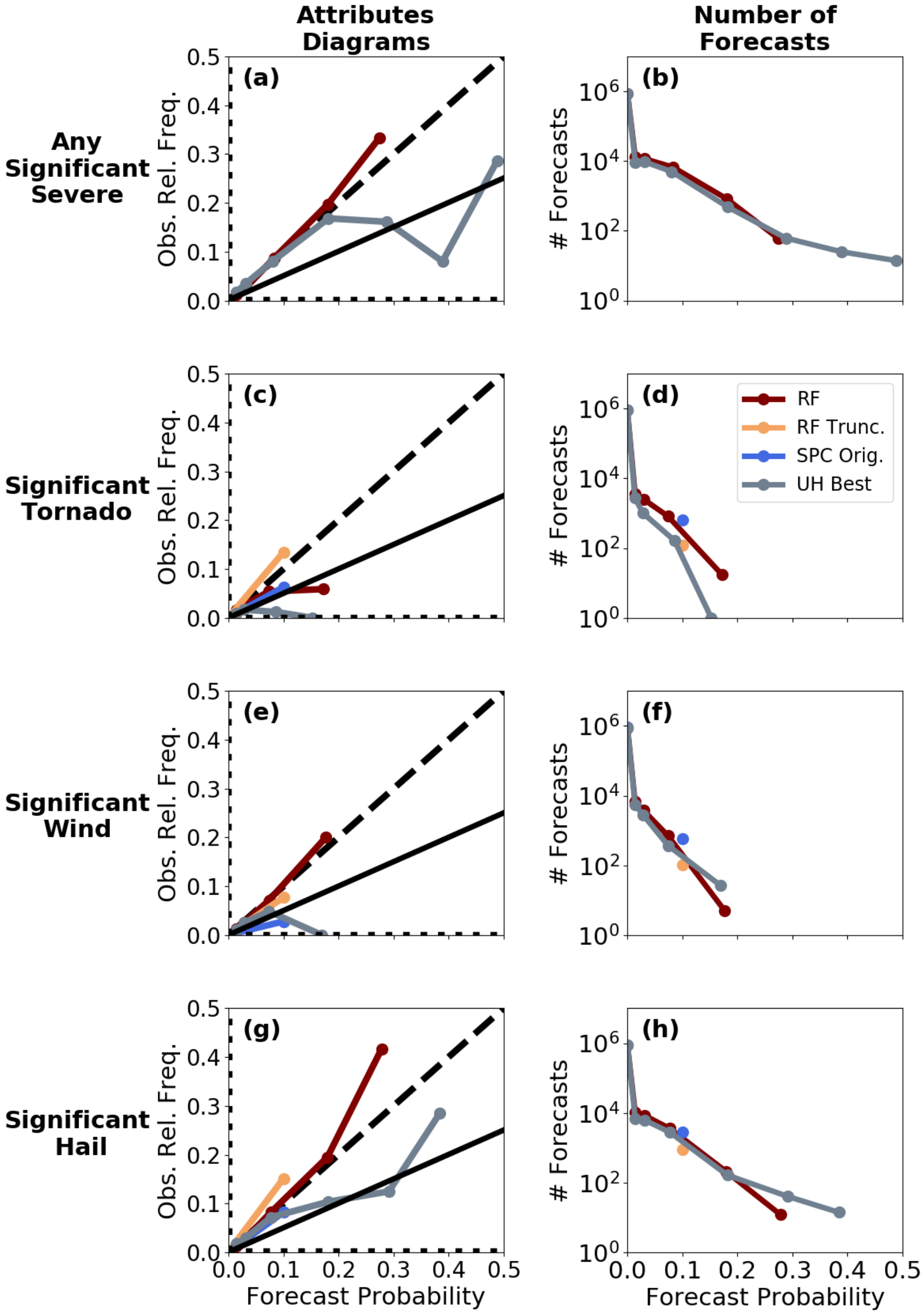


Figure 3.5 As in Fig. 3.4 but for (a)-(b) any significant severe, (c)-(d) significant tornado, (e)-(f) significant severe wind, (g)-(h) and significant severe hail forecasts. Note that, unlike in Fig. 3.4, the x- and y-axes stop at 0.5 and full SPC forecasts are not plotted.

events have less potential for truly devastating impacts. Importantly, the UH and RF forecasts give equal weight to all observed storm reports and do not consider the potential societal impacts of observed severe weather.

As statistical methods, the UH and RF forecasts tend to struggle most for the rarest events, which have the least amount of data. For example, the UH forecasts have good reliability for most hazards but some over-forecasting at higher probabilities for tornadoes (Fig. 3.4c) and significant severe weather hazards (Fig. 3.5a,c,e,g). Meanwhile, the continuous RF forecasts tend to have either near-perfect reliability (e.g., Fig. 3.4g; Fig. 3.5a,e,g) or slight under-forecasting (e.g., Fig. 3.4a,e) at most probability levels for most hazards. Unsurprisingly, the truncated RF forecasts tend to under-forecast relative to the continuous RF forecasts, since—like the original SPC forecasts—all continuous forecast probabilities less than a given discrete level are assigned to the next lowest level. Nevertheless, both the continuous and truncated RF forecasts have excellent reliability for the prediction of all hazards at probabilities with a sufficiently large sample size.

Performance diagrams (Fig. 3.6a-h) generally corroborate the BSS-based results (Fig. 3.3a-c), showing a clear outperformance of the RF-based method for most hazards. For example, the continuous RF forecasts substantially outperform the UH forecasts for all-hazard severe (Fig. 3.6a) and significant severe (Fig. 3.6b) weather at all probability levels. The continuous and truncated RF forecasts also clearly outperform both the SPC and UH forecasts for severe wind (Fig. 3.6e), significant severe wind (Fig. 3.6f), severe hail (Fig. 3.6g), and significant severe hail (Fig. 3.6h). Interestingly, for tornadoes, the RF-based forecasts perform as well as (for the lower forecast probabilities) or slightly worse than (for the higher forecast probabilities) those from the

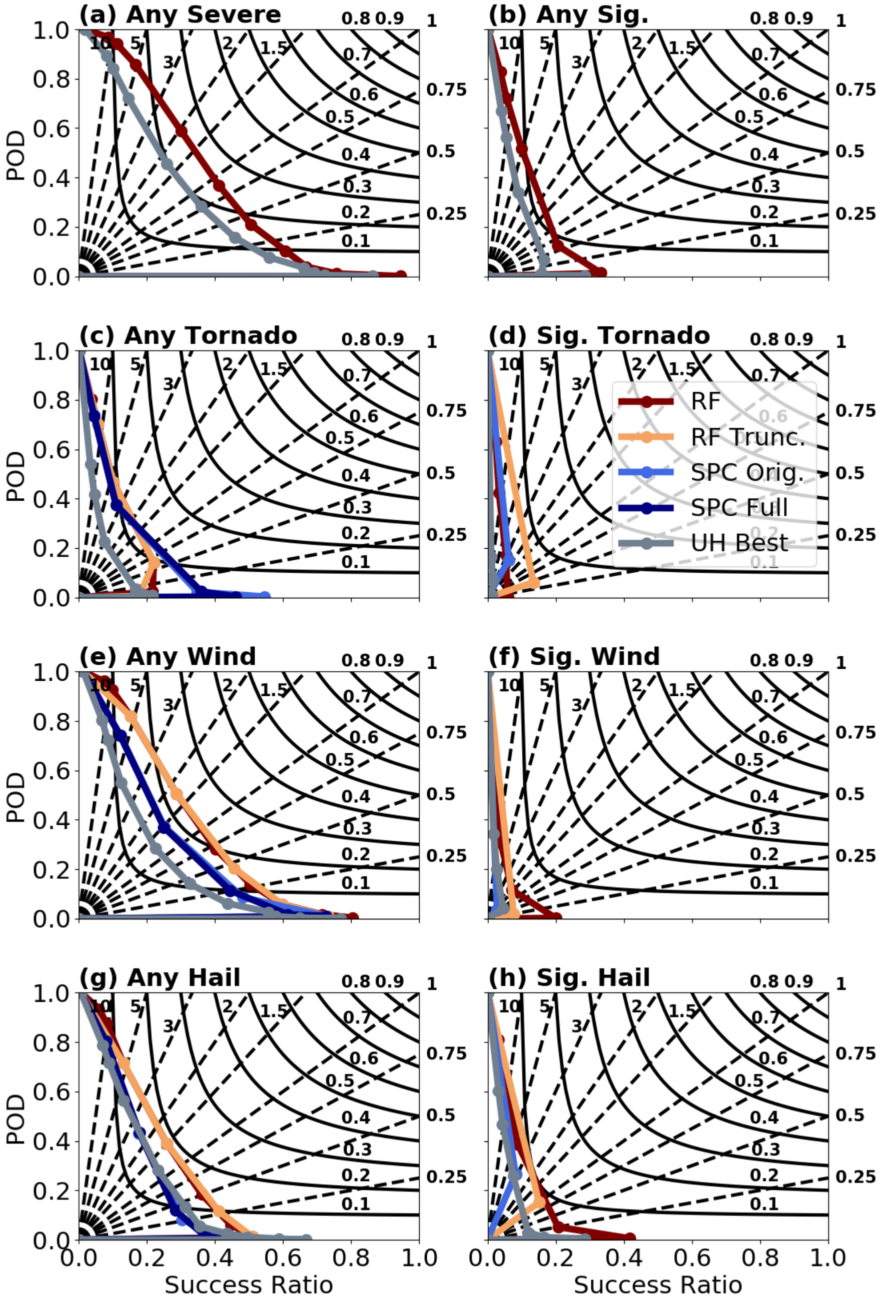


Figure 3.6 Performance diagrams for (a) any severe weather, (b) any significant severe weather, (c) any tornado, (d) significant tornado, (e) any severe wind, (f) significant severe wind, (g) any severe hail, and (h) significant severe hail forecasts. Note that only the full RF (dark red) and calibrated UH (gray) forecasts are shown in (a) and (b). All other panels additionally show original SPC (light blue) and truncated RF (yellow) forecasts. Full SPC (dark blue) forecasts are only shown in (c), (e), and (g).

SPC, with the UH-based forecasts noticeably inferior (Fig. 3.6c). Again, the RF-based forecasts' worse performance for tornado prediction potentially reflects the lack of tornado-specific predictors in this study. For significant severe hazards (Fig. 3.6d,f,h), skill is relatively low for all forecasts, but the RF forecasts have CSI values at least as high as those from SPC and UH forecasts.

In general, the continuous and truncated RF forecasts have similar CSI scores. There is one interesting exception, however: for the significant tornado forecasts, the truncated RF method is associated with a noticeably higher CSI (Fig. 3.6d). The likely cause is the poor reliability of the continuous RF forecasts at greater than 10% probabilities due to small sample size (Fig. 3.5c-d). Because the continuous RF probabilities dramatically over-forecast significant tornadoes above 10% probability, the truncation procedure dramatically improves reliability (Fig. 3.5c) and CSI (Fig. 3.6d) at the 10% level.

b. Seasonal and regional results

Consistent with Sobash and Kain (2017), it is found herein that the "best" UH threshold to use (i.e., the one that maximizes BSS) for all-hazards severe (Fig. 3.7a-b) and significant severe (Fig. 3.7c-d) weather prediction depends on season and region. The best-performing UH threshold is particularly sensitive to region: values of 60, 40, and 30 m^2s^{-2} (140, 110, and 130 m^2s^{-2}) are best for the West, Midwest, and East regions, respectively, for all-hazards severe (significant severe) weather (Fig 3.7b,d). While the best UH threshold does not change much seasonally for the all-hazard severe weather forecasts (Fig. 3.7a), seasonal variations are more apparent for all-hazard significant severe weather forecasts (Fig. 3.7c). Importantly, the continuous RF always outperforms the best UH forecast over a given region or season (Fig. 3.7a-d).

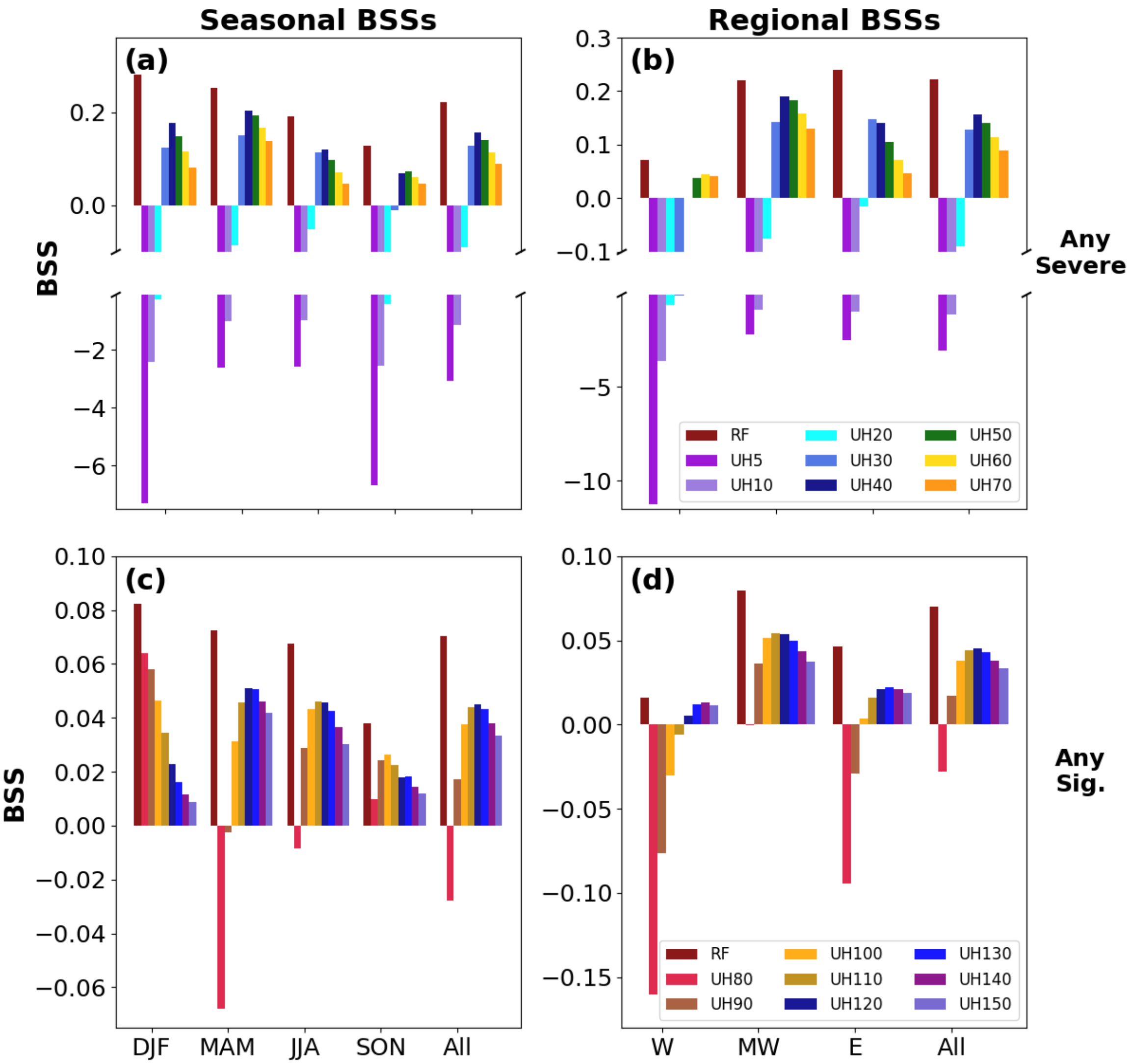


Figure 3.7 (a) BSS of full RF (dark red) and UH-based forecasts for any severe weather. UH forecasts use a Gaussian kernel standard deviation of 120 km and a UH threshold of 5 (dark purple), 10 (light purple), 20 (light blue), 30 (royal blue), 40 (dark blue), 50 (dark green), 60 (yellow), and 70 m^2s^{-2} (orange), respectively. BSSs are computed over the winter (DJF), spring (MAM), summer (JJA), and fall (SON) seasons as well as over the entire year (All). (b) As in (a), but BSSs are computed over the West (W), Midwest (MW), and East (E) regions as well as over the full CONUS (All). (c) As in (a), but forecasts are for any significant severe weather and the UH-based forecasts use thresholds of 80 (light red), 90 (brown), 100 (yellow), 110 (tan), 120 (dark blue), 130 (blue), 140 (purple), and 150 m^2s^{-2} (light purple). (d) As in (c) but BSSs are computed over the regions in (b).

When all forecasts are verified seasonally, a similar pattern emerges: with just one exception (i.e., fall tornado forecasts; Fig. 3.8a), the continuous RF forecasts have the highest BSSs for all hazards during all seasons (Fig. 3.8a-f). Both the continuous and truncated RF forecasts have substantially greater BSSs for summer severe wind prediction compared to either the UH or SPC forecasts (Fig. 3.8c). The continuous RF forecasts also dramatically outperform the best-performing SPC forecast for the prediction of spring and summer severe hail (Fig. 3.8e) and spring significant severe hail (Fig. 3.8f). Additionally, the continuous RF forecasts substantially outperform the UH-based forecasts—but not the continuous SPC forecasts—for spring severe wind (Fig. 3.8c) and winter tornadoes (Fig. 3.8a). However, it should be noted that using a spatiotemporally varying UH threshold would likely improve the BSSs of the UH forecasts (e.g., Sobash and Kain 2017), especially for the winter tornado forecasts. While the continuous RF forecasts generally exhibit noticeably larger BSSs than the other forecasts for significant severe hazards (e.g., Fig. 3.8b,d,f), the seasonal 95CIs are typically quite large for these hazards. Truncated RF forecast BSSs are generally higher—but not dramatically higher—than those from the original SPC forecasts for sub-significant severe weather prediction (i.e., Fig. 3.8a,c,e), although the truncated RF forecasts do have substantially better summer severe wind forecasts. For the significant severe hazards, the truncated RF probabilities have BSSs similar to the original SPC probabilities during each season.

When BSS is tabulated regionally, it is apparent that the RF method struggles at predicting tornadoes (Fig. 3.9a), significant tornadoes (Fig. 3.9b), and significant severe wind (Fig. 3.8d) in the West region. However, for all other hazards and regions (Fig. 3.9a-f), the continuous RF forecasts have the greatest BSSs. Regionally, the RF approach gives the greatest relative benefit for East severe wind prediction (Fig. 3.9c); both the continuous and truncated RF

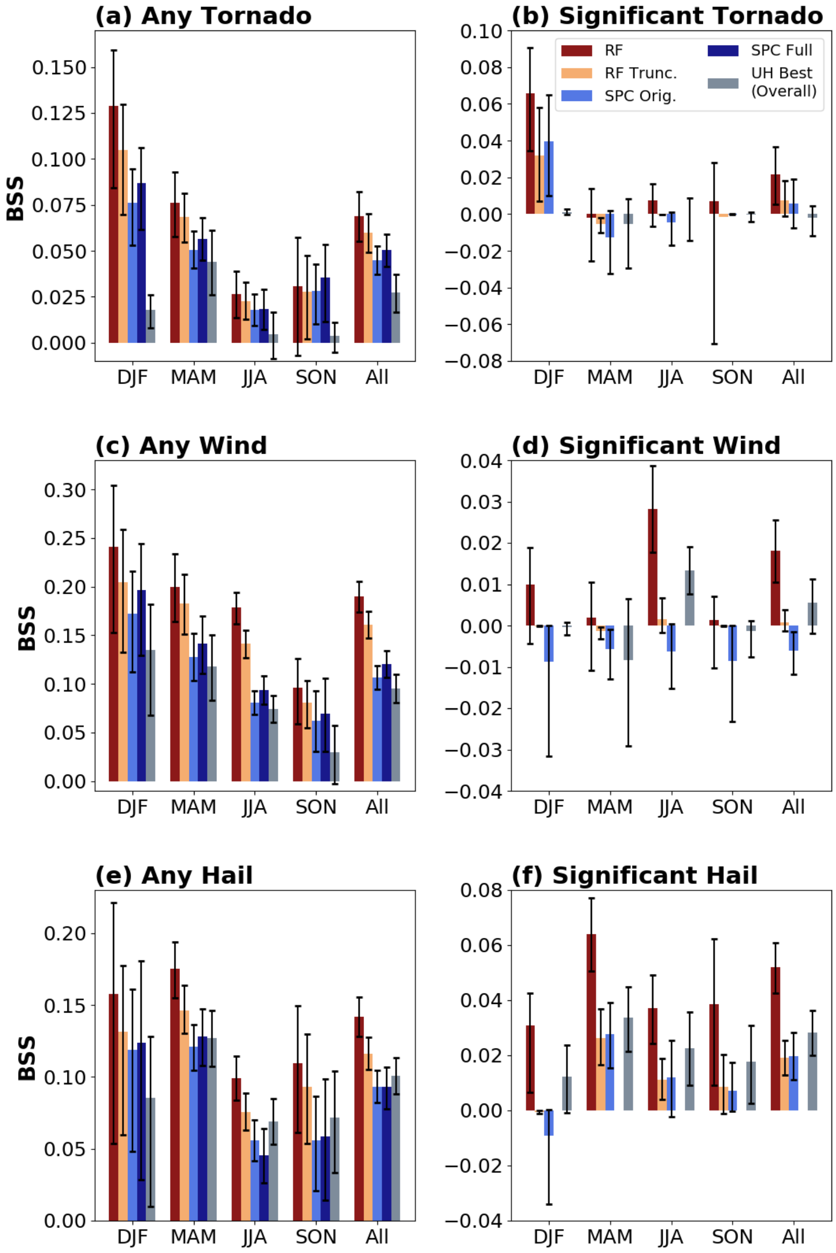


Figure 3.8 (a) BSS for any tornado probabilistic forecasts from the full RF (dark red), truncated RF (yellow), original SPC (light blue), full SPC (dark blue), and spatially-smoothed UH (gray). BSSs are computed over the winter (DFJ), spring (MAM), summer (JJA), and fall (SON) seasons as well as over the entire year (All). Note that the UH-based forecasts use the combination of standard deviation and UH threshold that produces the best BSS over the CONUS over the entire year. Black bars indicate 95% confidence intervals. (b) As in (a) but for significant tornadoes. (c) As in (a) but for any severe wind. (d) As in (a) but for significant severe wind. (e) As in (a) but for any severe hail. (f) As in (a) but for significant severe hail. Note that full SPC probabilities are not shown in the significant severe panels [i.e., (b), (d), and (f)].

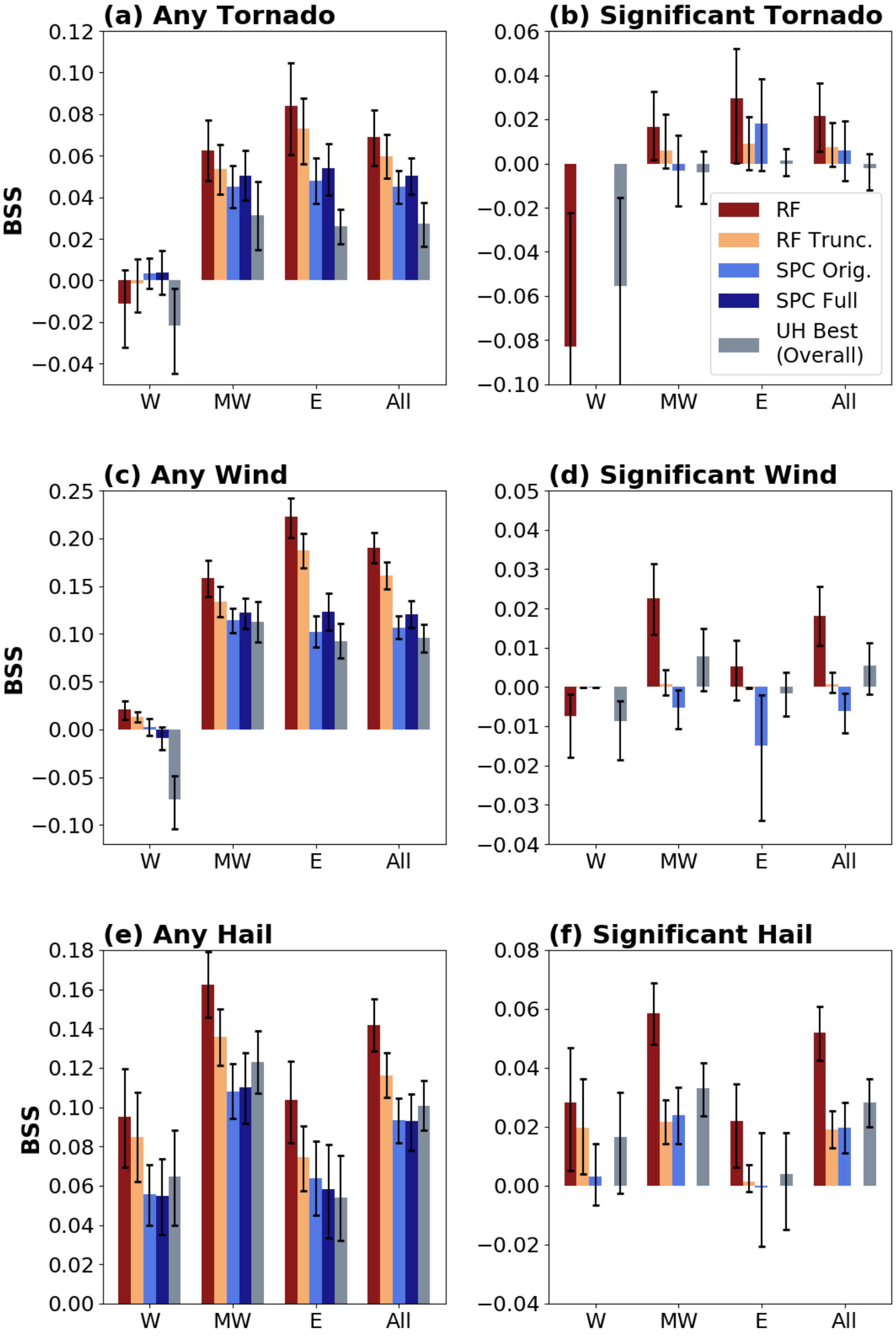


Figure 3.9 As in Fig. 3.8, but BSSs are computed over the West (W), Midwest (MW), and East (E) regions, as well as over the full CONUS (All).

forecasts have substantially greater BSSs than the UH- or SPC-based forecasts. The continuous RF also noticeably outperforms the UH and SPC forecasts for the prediction of West and East severe wind (Fig. 3.9c) and Midwest severe hail (Fig. 3.9e) and significant severe hail (Fig. 3.9f). As with the seasonal verification results, truncated RF significant severe probabilities (Fig. 3.9b,d,f) tend to have similar BSSs to original SPC probabilities for each region.

4. Case Studies

a. 26-27 May 2015

At 1200 UTC on 26 May, a mid-level trough and associated mesoscale convective complex (MCC) was centered over central Iowa. A line of thunderstorms extended along a surface front from the MCC southeast into eastern Mississippi. As the period progressed, the mid-level trough moved northeastward over the Great Lakes region and helped deepen an associated surface cyclone, ultimately leading to several tornado and severe wind reports in Illinois and Wisconsin before 1900 UTC. The cyclone’s cold front also advanced eastward and helped force the development of severe-wind-producing thunderstorms over eastern Alabama, western Georgia, and the Ohio Valley. Farther west, storms initiated along a dryline extending from west-central Oklahoma southward into central Texas by 2300 UTC. These storms produced numerous reports of severe wind and hail, with multiple significant severe hail reports and one significant severe wind report.

The RF and SPC outlooks (Fig. 3.10a,b) had some notable differences on this day, including the RF outlook’s use of the enhanced risk over two locations as well as the RF outlook’s greater areal coverage of slight risk areas. In the Upper Midwest, the RF shifted the 2% and greater tornado probabilities westward compared to the SPC (Fig. 3.11a,b). As a result, the

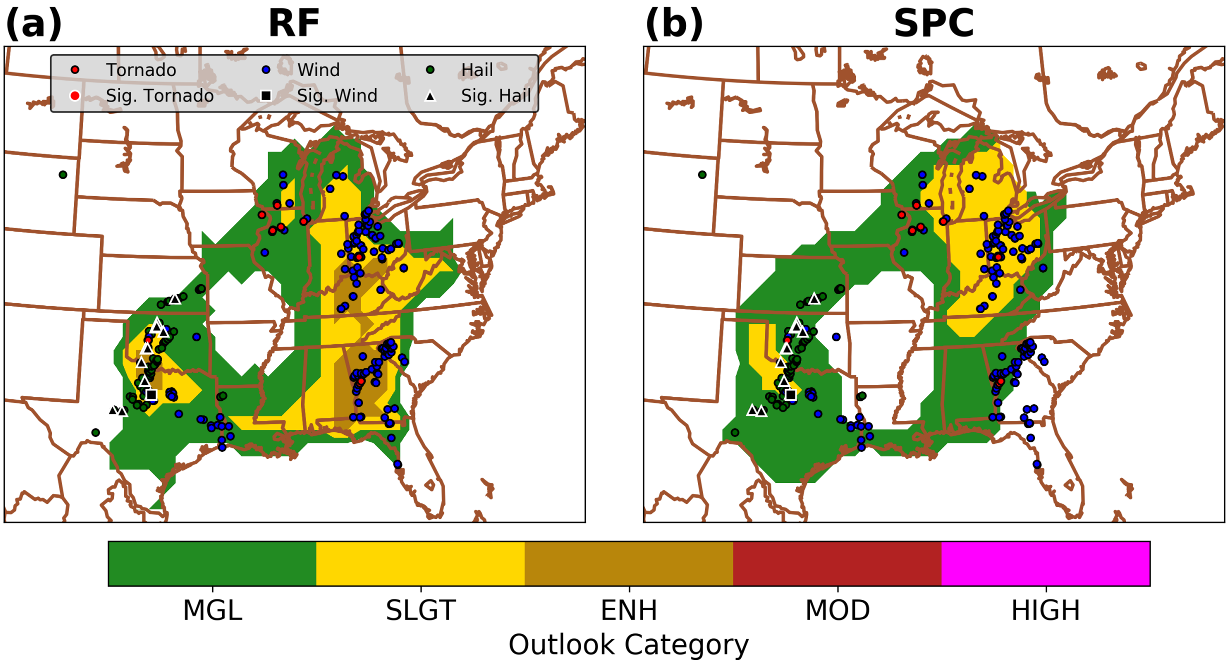


Figure 3.10 Day 1 categorical convective outlook from the (a) RF approach and (b) SPC 0600 UTC forecast, valid for the 24-h period ending at 1200 UTC on 27 May 2015. Small red, blue, and green circles outlined in black represent observed tornado, severe wind, and severe hail reports, respectively. Observed significant tornado, significant severe wind, and significant severe hail reports are represented by white-outlined large red circles, black squares, and black triangles, respectively.

RF had better POD for tornadoes in eastern Iowa, southern Wisconsin, and northern Illinois. Along the Oklahoma-Texas border, the RF issued 10% tornado probabilities with 6% significant tornado probabilities. Ultimately, no significant tornadoes were observed in this region, although multiple tornado reports occurred near the RF's 10% tornado area. Unlike the SPC forecast, the RF forecast issued 30% severe wind probabilities in a region extending from the Ohio Valley to the western Florida Panhandle (Fig. 3.11c,d). Numerous severe wind reports were observed near these locations, giving the RF a better POD. The RF also moved the 15% severe wind area slightly southeastward into southern Oklahoma and northern Texas, which better captured some severe wind reports—including a significant severe wind report—in that region. Notably, the one significant severe wind report fell near the RF's 2% contour for significant severe wind. Indeed,

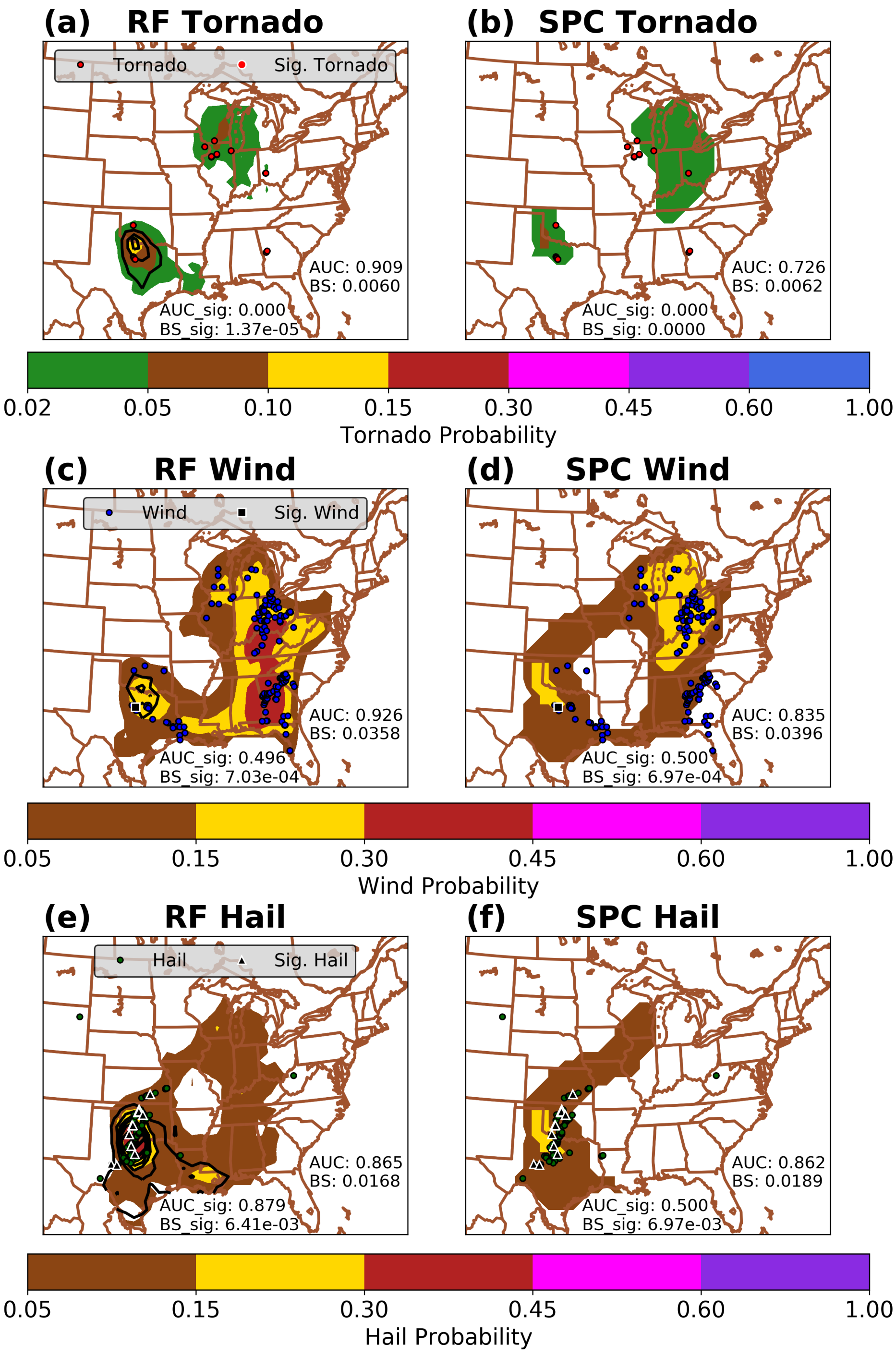


Figure 3.11 (a) RF-based tornado probabilities (shaded) and significant tornado probabilities (contoured every 2% with ≥ 10% probabilities hatched), valid for the 24-h period ending at 1200 UTC on 27 May 2015. (b) As in (a) but for SPC forecasts issued at 0600 UTC. (c)-(d) As in (a)-(b) but for severe wind forecasts. (e)-(f) As in (a)-(b) but for severe hail forecasts. For each hazard, corresponding observed severe weather reports are plotted as described in Fig. 3.9. Note that SPC forecasts do not have significant severe contours less than 10%. Individual-day AUC and BS values are given for each forecast, with overall hazard (significant hazard) metrics given in the lower right corner (at the bottom) of each panel.

one advantage of the RF forecast is its ability to communicate nonzero (but still less than 10%) probabilities for significant severe weather. For severe hail, the RF forecasts gave a much larger 5% area than the SPC (Fig. 3.11e,f) but focused on a similar area for its 15% probabilities. However, unlike the SPC, the RF forecasts produced a large area of 30% severe hail probabilities and indicated a greater than 10% chance of significant severe hail in western Oklahoma and northern Texas. Ultimately, numerous severe and significant severe hail reports occurred in this region. Two significant severe hail reports in central Texas also fell outside of the RF's 10% "hatched area" for significant severe hail but within the RF's 2% significant severe hail contour. However, the RF forecast did have greater false alarm than the SPC in eastern Louisiana (where the RF issued 15% probabilities) and over a large area extending from central Wisconsin to the Gulf Coast (where the RF generally issued 5% probabilities). Nevertheless, the RF generally made improvements over the SPC forecast. A human forecaster with access to the RF probabilities on this day might have had more confidence in a Texas-Oklahoma significant severe hail event and a widespread severe wind event in the Ohio Valley and Southeast.

UH-based probabilities might have only communicated part of this story. For example, compared to RF all-hazard probabilities (Fig. 3.12a), UH-based probabilities (Fig. 3.12b) were much lower over the Ohio Valley and Southeast United States. However, UH all-hazards severe and significant severe probabilities (Fig. 3.11b,d) were generally similar to those from the RF (Fig. 3.12a,c) over the Southern Plains.

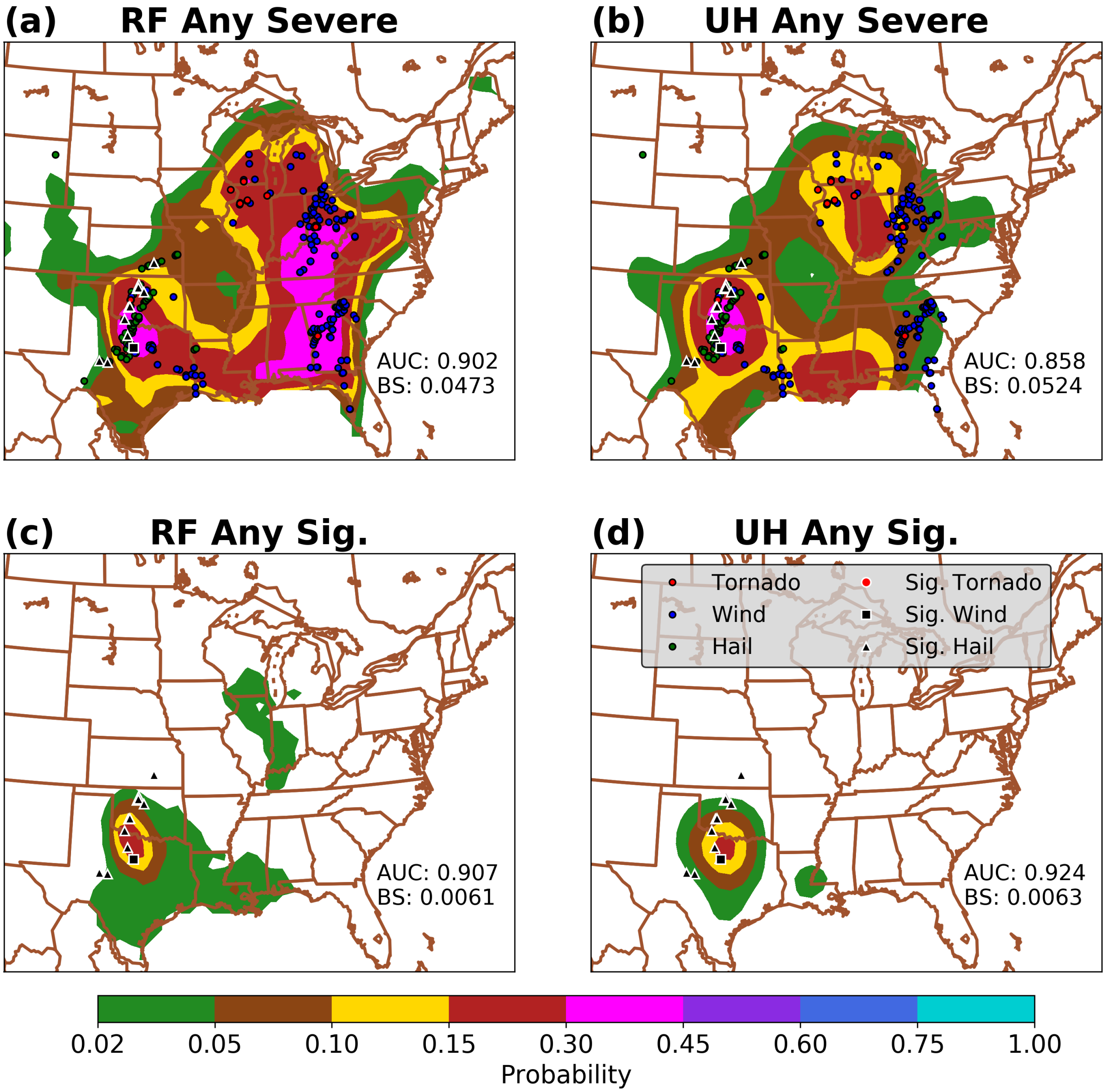


Figure 3.12 (a) RF- and (b) UH-based probabilities of all-hazards severe weather, valid for the 24-h period ending at 1200 UTC on 27 May 2015. Observed severe weather reports are plotted as described in Fig. 3.9. (c)-(d) As in (a)-(b) but all-hazards significant severe weather probabilities are plotted, and only significant severe observed reports are overlaid. Individual-day AUC and BS values are reported in the lower right corner of each plot.

b. 18-19 May 2017

The SPC identified 18 May 2017 as a high-risk day in the Southern Plains (e.g., Fig. 3.13b), with their 0600 UTC outlook highlighting the potential for widespread long-track tornadoes in parts of Oklahoma and Kansas. At the surface, a cyclone was developing in the western Oklahoma Panhandle by 1200 UTC. Strong southerly winds throughout central Texas and Oklahoma advected rich low-level moisture into the Southern Plains, where strong deep-layer vertical wind shear was in place. Storms began forming in the warm sector along the dryline in western Oklahoma and northern Texas by 1830 UTC and quickly became severe. Severe storms also formed along the warm front in central Kansas by 2130 UTC. Meanwhile, in the Northeast, severe hail- and wind-producing storms initiated ahead of a cold front in an unstable, sheared environment.

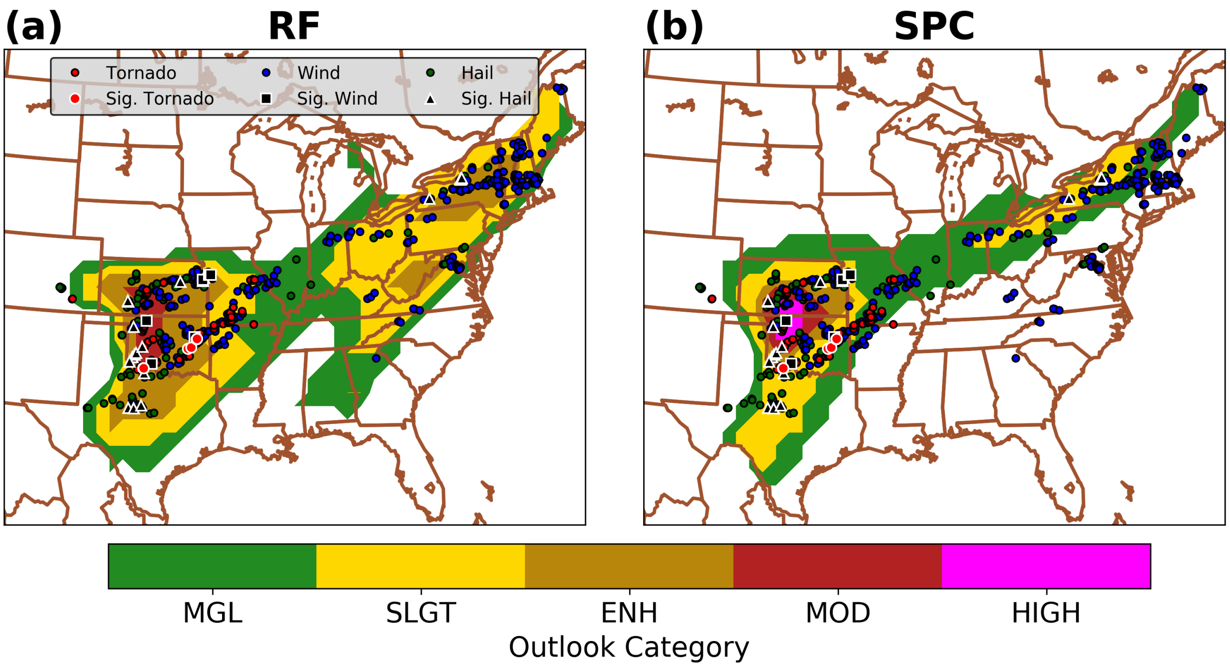


Figure 3.13 As in Fig. 3.10, but valid for the 24-h period ending at 1200 UTC on 19 May 2017.

While the RF and SPC forecasts identified similar threat areas in their outlooks (Fig. 3.13a,b), they issued different maximum outlook categories, with the RF (SPC) issuing a moderate (high) risk in the Southern Plains and an enhanced (slight) risk in the Northeast. Interestingly, although the RF produced smaller tornado probability magnitudes in the Southern Plains (Fig. 3.14a), it gave larger areas of higher-end (i.e., >10%) tornado probabilities there. Indeed, most of the observed tornadoes occurred within these areas of higher-end RF probabilities. The RF tornado forecast also expanded its 2% tornado probabilities farther east compared to the SPC, enabling it to better capture the QLCS tornado reports in Missouri (Fig. 3.14a,b). While the RF and SPC agreed on the area with the largest significant tornado probability (i.e., southern Kansas and northern Oklahoma; Fig. 3.14a,b), most of the observed significant tornadoes fell outside of this region but within/near the RF's 2% significant tornado probability contour. The RF and SPC forecasts had very similar tornado forecasts in the Northeast, with the RF forecasts having slightly less false alarm area.

RF and SPC severe wind probability magnitudes were quite different on this day, with the RF having higher probabilities in both the eastern U.S. and Southern Plains (Fig. 3.14c,d). These higher probabilities led to better POD for the RF in New York, northern Pennsylvania, and southern Oklahoma but greater false alarm in most of West Virginia and northern Texas. The RF also expanded the 15% probability area farther eastward compared to the SPC, giving it greater POD in Arkansas and Missouri.

RF and SPC hail forecasts were similar, although the RF extended the 30% probability area and 10% significant severe hail area farther south into central Texas, where severe and significant severe hail occurred (Fig. 3.14e,f). Additionally, the RF indicated 2% significant severe hail probabilities in New York and Kansas where significant severe hail was observed but

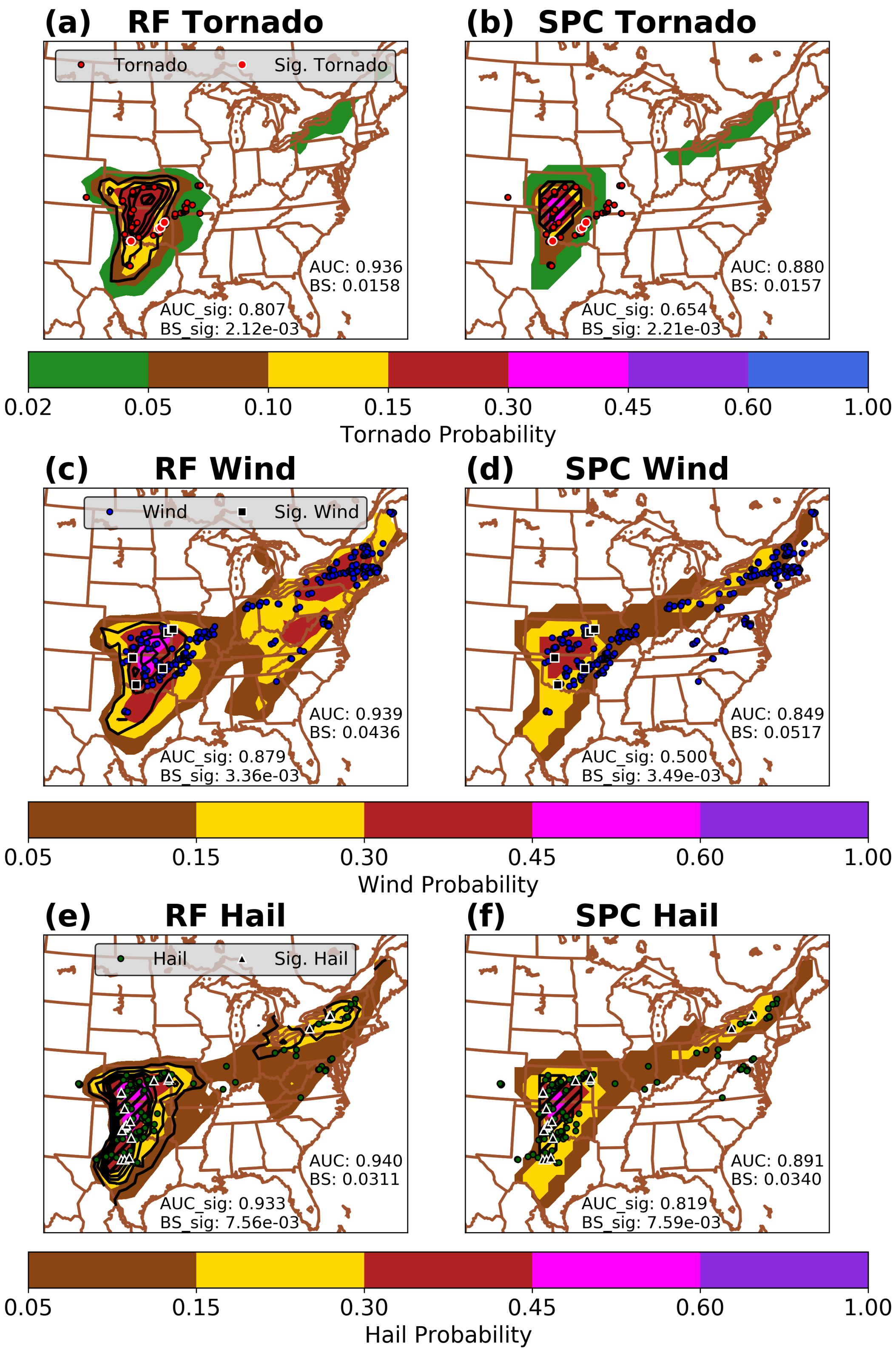


Figure 3.14 As in Fig. 3.11, but valid for the 24-h period ending at 1200 UTC on 19 May 2017.

fell outside of the RF or SPC 10% significant severe hail probabilities. Finally, the RF demonstrated better severe hail POD in Maryland (Fig. 3.14e,f). Overall, the RF-based outlook (Fig. 3.13a) and individual-hazard probabilities (Fig. 3.14a,c,e) compared favorably against the corresponding SPC forecasts on this day.

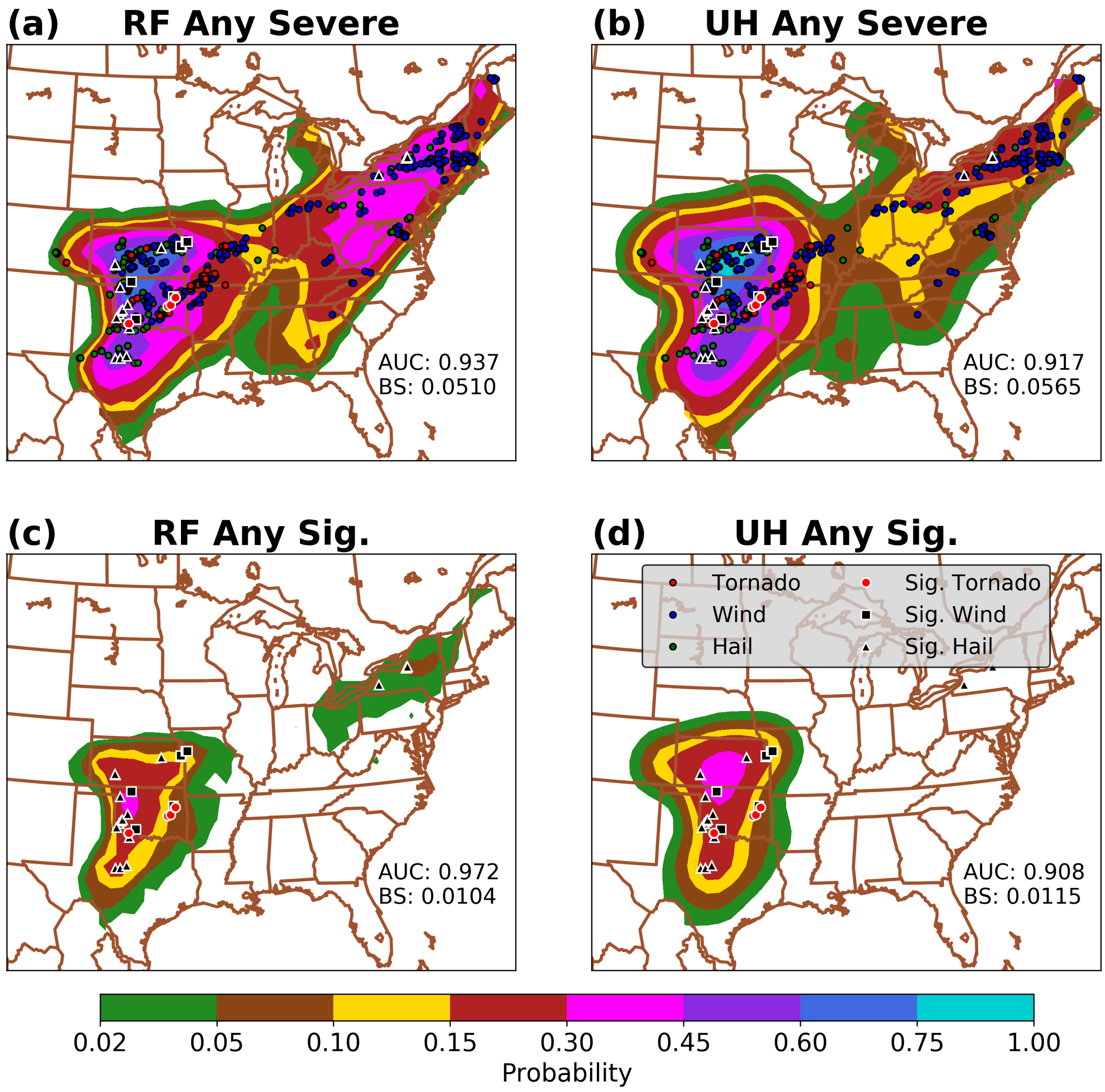


Figure 3.15 As in Fig. 3.12, but valid for the 24-h period ending at 1200 UTC on 19 May 2017.

RF all-hazards severe and significant severe weather probabilities (Fig. 3.15a,c) also compared favorably with UH-based probabilities (Fig. 3.15b,d), especially in the Northeast, where the RF had better POD for severe and significant severe weather. In the Southern Plains, RF and UH forecasts were generally similar. However, it is noteworthy that the RF significant severe forecasts (Fig. 3.15c) shift the maximum probabilities southwest into western Oklahoma, close to a cluster of significant severe reports, while the UH probability maximum is in central Kansas, away from any such cluster (Fig. 3.15d).

5. Discussion

Compared to the SPC forecasts, the RF probabilities frequently highlighted similar areas for severe weather but gave different probability magnitudes. However, the RF forecasts herein occasionally assigned higher probabilities (e.g., greater than 15% or even 30%) to areas outside of the SPC's marginal risk. When this happened, many times the areas with the higher RF probabilities did experience observed severe weather. This occurred most often for severe wind events in the East region. In these instances, it is possible that the differences between the SPC and RF forecasts could be partially explained by biases and non-meteorological artifacts in the severe wind report observations, given the high ratio of estimated to measured severe wind reports in the eastern and southeastern U.S. (Edwards et al. 2018). While the RF algorithm views all observed storm reports equally (i.e., as unbiased, perfect observations) and does not consider storm coverage, density, intensity, or potential societal impacts when constructing its probabilities, SPC forecasters may be mindful of how their forecast probabilities equate to outlook categories (Table 3.3) and may emphasize higher-impact events that pose a greater threat to life and property.

The biggest advantage of the RF method described herein is its ability to create skillful CAE-derived severe weather guidance products analogous to those issued by the SPC. However, it must be emphasized that the goal in creating these RF-based products is not to replace human forecasters but to augment them. Indeed, this augmentation could potentially take a variety of (non-mutually exclusive) forms. First, RF-based forecasts could provide a skillful, reliable first guess (e.g., Karstens et al. 2018) product, which forecasters could modify based on other data sources (e.g., satellite and radar trends, surface analyses, etc.) and their expertise. Such a product could increase forecaster efficiency and facilitate proper forecast calibration (Karstens et al. 2018). Used as a first guess or “last check” product, the RF guidance may also identify a threat area that a forecaster might have overlooked for a given hazard (e.g., significant severe hail in the southern Plains; Fig. 3.11e-f). The RF forecasts may also help simply by providing useful uncertainty information in challenging forecasting situations. Such uncertainty information may be especially valuable for more precisely quantifying the threat of significant severe weather, which is rare but extremely threatening to life and property. Finally, it is conceivable that the RF-based forecasts—when properly interrogated using ML interpretability metrics (e.g., McGovern et al. 2019b)—may give forecasters and researchers insight into ensemble biases or complex relationships between CAE forecast output and observed severe weather. Human forecasters learning from ML would not be unprecedented, as artificial intelligence techniques have recently provided new knowledge to human experts in other complex domains, such as the game of Go (Silver et al. 2016, 2017) and multi-player no-limit Texas Hold’em poker (Brown and Sandholm 2019). A future study is planned to determine how and why RF-based severe weather probabilities differ from human and UH-based forecasts.

6. Summary and Conclusion

This paper used a random forest (RF) to produce CONUS-wide 1200 UTC – 1200 UTC Day 1 Convective Outlooks (COs) and individual-hazard severe weather probabilities from Storm Scale Ensemble of Opportunity (SSEO) forecast output. Temporally-aggregated grid-point-based forecast variables were used as predictors. The grid-point-based approach is advantageous because it allows users to interpret RF output directly in 2-dimensional space and does not require the assumption of perfect correspondence between simulated and observed storms.

Continuous and discrete (i.e., truncated) RF forecasts created herein were compared against calibrated, spatially smoothed 2-5 km updraft helicity (UH) forecasts as well as original and continuous (i.e., full) SPC Day 1 COs issued at 0600 UTC. The continuous RF forecasts almost always produced the highest BSSs, both when the forecasts were verified over the entire dataset and when verification was performed regionally or seasonally. The truncated RF forecasts frequently had the second-highest BSSs and were often better—but never substantially worse—than the corresponding original SPC forecasts. In general, the RF method performed best relative to the SPC and UH forecasts for severe wind and hail prediction in the Midwest and East regions during the spring and summer. All forecasts—including the RF-based ones—generally had very good reliability, while the RF forecasts tended to have the best resolution.

Given the promising results of the RF technique described herein, it is important to evaluate its skill and value to forecasters in an operational environment. To this end, efforts are under way to apply the technique described herein to the operational HREFv2 with the goal of evaluating real-time RF forecasts in future Hazardous Weather Testbed Spring Forecasting Experiments (e.g., Clark et al. 2012a; Gallo et al. 2017). While such formal evaluation is necessary to draw more robust conclusions, it is speculated that real-time RF-based guidance will aid human Day 1 severe weather forecasts by providing forecasters with calibrated CAE-based severe hazard probabilities and outlooks.

Chapter 4: Comparing and Interpreting Differently-Designed Random Forests for Next-Day Severe Weather Hazard Prediction

A paper to be submitted to *Weather and Forecasting*

Eric D. Loken[1,2] and Adam J. Clark[2,3]
[1]Cooperative Institute for Mesoscale Meteorological Studies, The University of Oklahoma, Norman, Oklahoma
[2]School of Meteorology, University of Oklahoma, Norman, Oklahoma
[3]NOAA/OAR National Severe Storms Laboratory, Norman, Oklahoma

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 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, multi-variate relationships that go well beyond thresholding 2-5km updraft helicity.

1. Introduction

Interest in using machine learning (ML) to assist with high-impact weather prediction has markedly increased during the past 3-5 years. Indeed, as of this writing, 71 of the 107 *Weather and Forecasting* articles that include the phrase “machine learning” in the title, abstract, main body text, or as a keyword have been published in 2018 or later (AMS 2021). This strong recent interest is likely driven by a combination of factors, including enhanced computing power; greater data storage capacity; and the availability of free, easy-to-use ML software (e.g., Scikit-Learn, Pedegrosa et al. 2011; Keras, Chollet et al. 2015).

Promising results have also helped fuel the recent enthusiasm for ML research. In the past few years, studies have demonstrated that ML can be used to skillfully predict severe wind (Lagerquist et al. 2017), severe hail (Gagne et al. 2017; Burke et al. 2020), next-hour tornadoes (Lagerquist et al. 2020), convective duration (McGovern et al. 2019a), convective mode (Jergensen et al. 2020), precipitation (Herman and Schumacher 2018c; Loken et al. 2019a), intense convection (Cintineo et al. 2020), and next-day (e.g., Loken et al. 2020; Sobash et al. 2020) and beyond (i.e., day 1-3; Hill et al. 2020) severe weather based on numerical weather prediction (NWP) data and/or radar- and satellite-based predictors. Many of these methods not only attain high objective performance metrics, but also demonstrate an ability to add skill and/or value to human forecasts, particularly in the domain of severe weather prediction. For example, Hill et al. (2020), who used a random forest (RF) with NOAA's Second-Generation Global Ensemble Forecast System Reforecast (GEFS/R; Hamill et al. 2013) predictors to forecast day 1-3 severe weather, found the combination of RF and Storm Prediction Center (SPC) probabilistic forecasts outperformed individual RF or SPC probabilities. Hill et al. (2020) also noted that their day 2-3 forecasts had higher Brier Skill Scores (BSSs) than corresponding SPC forecasts. Meanwhile, Loken et al. (2020) found that next-day RF-based severe and significant severe hazard forecasts—which used convection-allowing predictors from the Storm-Scale Ensemble of Opportunity (SSEO; Jirak et al. 2012)—had higher Brier Skill Scores (BSSs) than corresponding day 1 SPC human forecasts. This pattern held for most hazards in most seasons and regions, even when RF forecast probabilities were discretized to match those used by the SPC. During the 2020 Hazardous Weather Testbed Spring Forecasting Experiment (HWT SFE; Clark et al., 2021), participants evaluated a similar RF method—only applied to the High-Resolution Ensemble Forecast System, Version 2.1 (HREFv2.1; Roberts et al. 2020)—for real-time next-day tornado, severe hail, and severe wind prediction. The participants generally found the method skillful and useful, particularly for severe hail and wind.

The skill achieved by the RF forecasts in Hill et al. (2020), Loken et al. (2020), and Clark et al. (2021) raises several natural and important questions: How do RFs use ensemble data to create skillful forecasts? What relationships does the RF learn between ensemble forecast variables and observed severe weather? Are the current preprocessing techniques optimal? This study seeks to address these questions by comparing multiple differently-configured RFs and interpreting their output using the Python-based tree interpreter module (TI; Saabas 2016), which—to the authors' knowledge—has not yet been formally applied in the domain of meteorology.

Interpreting ML output is often challenging due to the complexity of many ML algorithms, but it can be extremely important. Uncovering the relationships learned by a ML algorithm can confirm the algorithm is working as intended, build trust with product users, and potentially provide new insights into underlying weather prediction tools (e.g., ensembles, satellites, etc.). Herman and Schumacher (2018a) conducted one of the first meteorology-related studies that focused on interpreting ML output. Using Gini importance and logistic regression coefficients, they showed that their RF-based precipitation forecasts learned physical spatiotemporal relationships between GEFS/R variables and observed precipitation. To help researchers and forecasters better understand ML predictions, McGovern et al. (2019b) summarized a variety of ML and deep learning (DL) interpretability techniques, including: impurity importance (Breiman 2001), single- (Breiman 2001) and multi-pass (Lakshmanan et al. 2015) permutation importance, forward and backward feature selection, partial-dependence plots (Friedman 2001), individual conditional expectation plots (Goldstein 2015), saliency maps (Simonyen et al. 2014), gradient-weighted class activation maps (Selvaraju et al. 2017), backward optimization (Olah et al. 2017), and novelty detection (Wagstaff and Lee 2018). A growing number of studies are using these techniques to visualize the relationships learned by skillful ML models. For example, Jergensen et al. (2020) used multi-pass permutation importance, sequential forward selection, and partial dependence plots to determine which variables were most important for ML-based storm classification. They found, based on the permutation importance and partial dependence plots, that storm shape predictors (e.g., storm age, area, eccentricity, and compactness) were among the most important for most of the ML methods examined. Meanwhile, Cintineo et al. (2020)—who developed a convolutional neural network (CNN) to distinguish between intense and ordinary convection in observed satellite images—used saliency maps, layer-wise relevance propagation, and permutation importance to analyze the features learned by their CNN. They found the CNN learned both well-established (e.g., overshooting tops, cold-U thermal patterns, cold rings, etc.) and previously-unrecognized (i.e., strong brightness temperature gradients) satellite features associated with intense convection.

In this paper, TI is used to identify the most important predictors for RF severe weather hazard forecasting as well as the relationships between those 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 ensemble data during preprocessing and understand how RFs leverage CAE data to create skillful severe weather forecasts.

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 4.1). Notably, these dates are not evenly spread throughout the year, with May over-represented and January-March under-represented. As in Loken et al. (2020), the analysis

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

Table 4.1 HREFv2.1 initialization dates.

domain covers the contiguous United States (CONUS), and verification is performed on a grid with approximately 80-km horizontal spacing (Fig. 4.1a) to match the verification scales used by the SPC. Next-day forecasts (lead times of 12-36 hours, valid from 1200 UTC to 1200 UTC) are analyzed herein.

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 2019-2020 tornadoes, while 2018 tornado LSRs were obtained from the SPC Storm

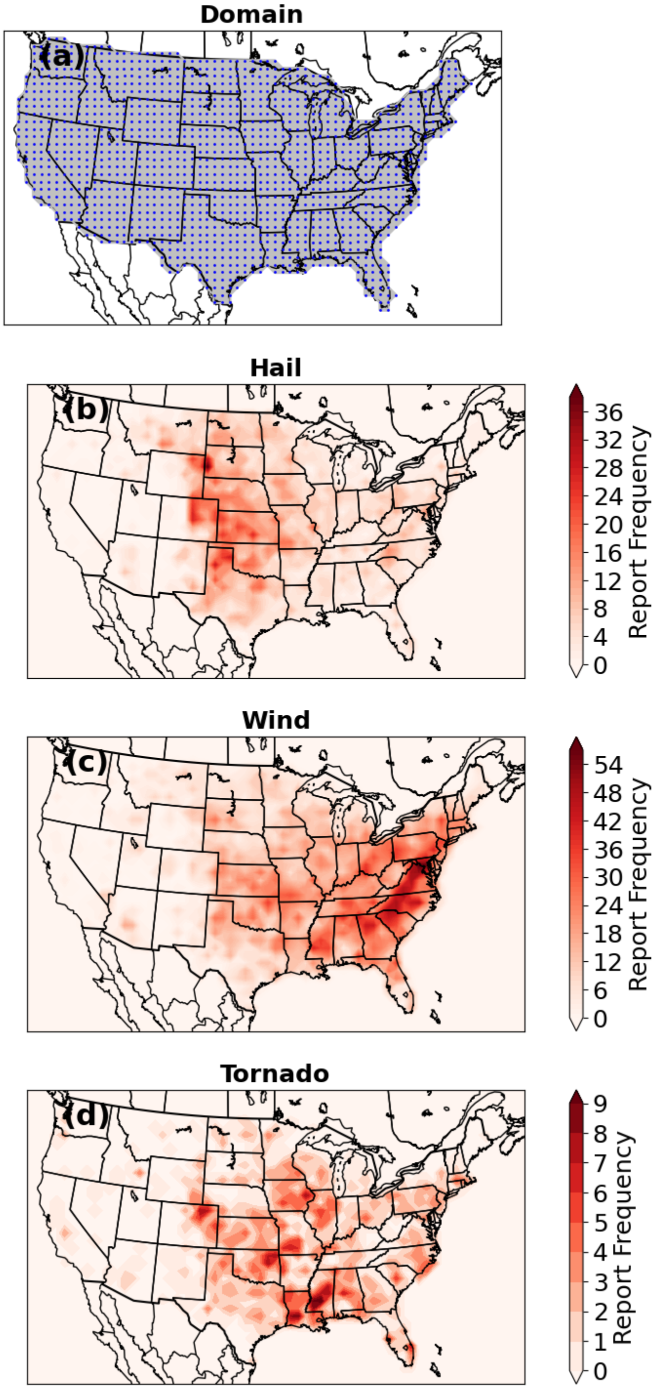


Figure 4.1 (a) Verification domain (gray shading) and 80-km grid points (blue dots). (b) Distribution of severe hail reports in the observational dataset. (c)-(d) As in (b) but for severe wind and tornado reports, respectively.

Events Database (SSED; SPC 2021e). The SSED was used for tornadoes when possible because it provides a more accurate and complete summary of tornado events than that shown on the SPC website. The spatial distribution of hail, wind, and tornado LSRs over the full dataset is depicted in Fig. 4.1b-d.

RF forecasts are trained based on predictors from the HREFv2.1, an operationalized version of the SSEO. Like the SSEO, HREFv2.1 is an assemblage of diverse, individually-tuned convection-allowing models (CAMs). The SSEO (Jirak et al. 2016) and HREFv2.1 (Jirak et al. 2018; Roberts et al. 2020) 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 4.2.

b. RF method overview

RFs are ensembles of decision trees (Breiman 2001), 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

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

Table 4.2 HREFv2.1 specifications. Dynamic cores are from the Advanced Research 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 instead of 1800 UTC (i.e., a 12- instead of 6-h time lag).

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 and computing the fraction of training samples associated with the given class at the relevant leaf node for each tree in the forest. Overall RF probabilities are then simply the mean probabilities from each tree. As in Loken et al. (2019, 2020), RFs are created using random forest classifiers from the Python module Scikit-Learn (Pedegrosa et al. 2011). More details on the RF algorithm can be found in Loken et al. (2019) and works cited therein.

c. RF interpretability and the tree interpreter module

RFs are appealing for real-time weather prediction since they tend to produce reliable probabilistic predictions (Breiman 2001), can handle raw (as opposed to standardized) input data, are computationally efficient, and require little tuning compared to other ML methods. However, it can be difficult to deduce *how* a RF arrives at its prediction given that RFs routinely use hundreds of trees. This study uses a Python package called tree interpreter (TI; Saabas 2016) to examine how various sets of predictors impact RF probabilities in different situations. TI analyzes the path of a testing sample through each tree in a RF and records how each predictor impacts the training data 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 on the training data purity over all trees in the RF.

TI operates similarly to impurity importance (e.g., Breiman 2001; Louppe et al. 2013; McGovern et al. 2019b) in that it measures how effectively splitting criteria sort the training data at each node. However, instead of using information gain or Gini index to classify the impact of each split, TI measures how the underlying training data climatology (i.e., the fraction of training data associated with a "yes" observed event) changes with each split and ascribes the change to the predictor responsible for the split. Additionally, unlike other methods of impurity importance, TI is a local method and so only considers the splits made along the path (of each tree) taken by a testing sample. In contrast, Gini importance does not require a testing sample but instead considers the mean impact of all predictors over all potential splits in each tree in the RF. Because TI tracks the training data climatology associated with only the relevant testing path of each tree, TI enables the final RF output probability to be decomposed into the sum of a bias term (i.e., the overall climatology of the training set) and the contribution from each predictor.

Here, 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 (i.e., either positively or negatively). 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 (i.e., whether the given prediction is associated with an LSR) to determine whether (and how much) predictors tend to appropriately increase or decrease probabilities. Additionally, to investigate the relationships learned by the RFs, TI contributions are plotted against the value of a given predictor for every testing sample in the dataset. This type of analysis shows how a given predictor influences RF probabilities at different values and is in the same spirit as individual conditional expectation (ICE) plots (Goldstein et al. 2015).

TI is also used to assess how multiple predictors interact to influence RF probabilities. This is done by setting the *joint_contribution* keyword to True in the call to the TI predict method. Conceptually, when *joint_contribution* is True, TI ascribes the change in data purity at a given node to the combination of predictors *at and above* the given node in the testing path. Thus, setting *joint_contribution* to True gives more accurate predictor contribution values, since—strictly speaking—the *combination* of predictors at and above a given node is responsible for the resultant training data purity at any given point in the tree. However, because the trees used herein are complex and the *joint_contribution* option must evaluate the contributions from all combinations of predictors found at and above each node in the testing path for all trees in the RF (for each prediction), the process of assessing multivariate contributions is very computationally expensive. Therefore, instead of running the *joint_contribution* TI on all testing data, it is only run on the testing data associated with an LSR. Thus, the variable interactions highlighted in this analysis should be interpreted accordingly. For each severe weather hazard, the three most influential two-variable interactions are analyzed. A scatterplot shows the probability contribution from the relevant two-variable combination for each sample in the testing dataset.

d. Creating RF forecasts

1) PREDICTOR FIELDS

HREFv2.1 has a large archive of forecast variables that may be useful for severe weather prediction. In all, 32 input fields are used as well as latitude and longitude (to enable the RF to learn spatial patterns). The 34 total fields are categorized as storm, environment, index, or latitude/longitude variables, as summarized in Table 4.3.

Twelve of the 34 fields represent derived quantities (denoted by an asterisk in Table 4.3). Many of these derived quantities are straightforward; for example, maximum 10-m wind magnitude and direction are derived from the 10m u- and v- wind components, and wind shear magnitude is computed by taking the vector difference of simulated wind at the two levels (e.g., 10m and either 925hPa or 500hPa). Other fields, including most index variables, are more complicated and deserve greater elaboration, which is provided below.

Simulated Storm	**Simulated Environment**		**Simulated Index**	**Lat/Lon**
1 km Reflectivity (24h max.)	0-3 km Storm Relative Helicity (24h max.)	MUCAPE (24h mean)	Supercell Composite Parameter* (24h max.)	Latitude
Echo Top (24h max.)	0-1 km Storm Relative Helicity (24h max.)	MUCIN (24h mean)	Significant Tornado Parameter* (24h max.)	Longitude
Upward Vertical Velocity (24h max.)	2m Temperature (24h mean)	SB/MUCAPE ratio* (24h mean)	Significant Hail Parameter* (24h max.)	-
Downward Vertical Velocity (24h min.)	2m Dewpoint Temperature (24h mean)	700 – 500 hPa Lapse Rate* (24h mean)	0-1 km Energy Helicity Index* (24h max.)	-
2-5 km Updraft Helicity (24h max.)	2m, 925 hPa, 850 hPa, 700 hPa, 500 hPa Dewpoint Depression* (24h mean)	Critical Angle Proxy* (At time of max. STP)	0-3 km Energy Helicity Index* (24h max.)	-
0-3 km Updraft Helicity (24h max.)	10m – 500 hPa wind shear magnitude* (24h mean)	Max 10m Wind Speed (24h max.)	Product of (MUCAPE) x (10m – 500 hPa wind shear magnitude) * (24h max.)	-
Number of Grid Points With At Least 30 dBZ Simulated Reflectivity (At time of max. 2-5 km Updraft Helicity [if non-zero] or Upward Vertical Velocity)	10m – 925 hPa wind shear magnitude* (24h mean)	10m Wind Direction (At time of maximum 10 m wind speed)	Lifted Index (24h min.)	-

*Table 4.3 Predictor fields. The temporal aggregation strategy for each variable is noted in parentheses. * denotes a derived quantity.*

Thompson et al. (2003) developed the supercell composite parameter (SCP) to identify environments supportive of right-moving supercells. Here, SCP is defined as:

$$\mathrm{SCP} = \frac{\mathrm{MUCAPE}}{1000\ \mathrm{J/kg}} \times \frac{\mathrm{SRH03}}{50\mathrm{m^2/s^2}} \times \frac{\mathrm{SHR}_{10-500}}{20\ \mathrm{m/s}} \times \frac{-40\ \mathrm{J/kg}}{\mathrm{MUCIN}} \quad (4.1),$$

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

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

$$\mathrm{STP} = \frac{\mathrm{SBCAPE}}{1500\ \mathrm{J/kg}} \times \frac{2000\mathrm{m} - \mathrm{LCL}}{1000\mathrm{m}} \times \frac{|\mathrm{SRH01}|}{150\mathrm{m^2/s^2}} \times \frac{\mathrm{SHR}_{10-500}}{20\ \mathrm{m/s}} \times \frac{200 + \mathrm{SBCIN}}{150\ \mathrm{J/kg}} \quad (4.2),$$

where SBCAPE is surface-based CAPE in J/kg, LCL is the lifted condensation level in m (which is computed here using the approximation 125× 2-m dewpoint depression [in Kelvin]), SRH01 is 0-1km storm-relative helicity in m^2/s^2, $\mathrm{SHR}_{10\text{-}500}$ is the magnitude of the vector difference between the 10m and 500 hPa winds (in m/s), and SBCIN is surface-based CIN in J/kg. 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 $\mathrm{SHR}_{10\text{-}500}$ is greater than or equal to 30 m/s or 0 if $\mathrm{SHR}_{10\text{-}500}$ is less than 12.5 m/s, and the SBCIN term is set to 1 if SBCIN is greater than -50 J/kg or 0 if SBCIN is less than -200 J/kg.

The Significant Hail Parameter (SHIP; SPC 2021b) is similar to STP in that it was developed to distinguish between significant and non-significant hail-producing environments. Here, SHIP is defined as:

$$\mathrm{SHIP} = \frac{\mathrm{MUCAPE} \times \mathrm{MR} \times \mathrm{LR}_{700-500} \times -\mathrm{T}_{500}\ (^{\circ}\mathrm{C}) \times \mathrm{SHR}_{10-500}}{42{,}000{,}000} \quad (4.3),$$

where MUCAPE is the most-unstable CAPE, MR is the mixing ratio in g/kg, $\mathrm{LR}_{750\text{-}500}$ is the 700-500 hPa lapse rate in K/km, T_{500} is the 500 hPa temperature (in degrees Celsius), and $\mathrm{SHR}_{10\text{-}500}$ is the magnitude of the vector difference between the 10m and 500 hPa winds (in m/s). This initial value of SHIP is then modified according to the following rules (executed sequentially): 1) if MUCAPE is less than 1300 J/kg, $\mathrm{SHIP}_{\mathrm{final}}$ = SHIP $\times \frac{\mathrm{MUCAPE}}{1300\ \mathrm{J/kg}}$, and 2) if $\mathrm{LR}_{700-500}$ is less than 5.8 K/km but greater than 0 K/km, $\mathrm{SHIP}_{\mathrm{final}}$ = SHIP $\times \frac{\mathrm{LR}_{700-500}}{5.8\ \mathrm{K/km}}$, or if $\mathrm{LR}_{700-500}$ is greater than 0 K/km, $\mathrm{SHIP}_{\mathrm{final}}$ = 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 freezing level height data was not available.

Other derived variables include CAPESHEAR, the product of MUCAPE (in J/kg) and SHR_{10-500} (in m/s), and 0-1 and 0-3 km energy helicity index (EHI1 and EHI3), defined as the product of SBCAPE and SRH over the relevant vertical layer. Critical angle, which Esterheld and Giuliano (2008) defines as the angle between the 0-500m shear vector and 10 m above-ground-level storm-relative inflow, is approximated here as the angle (in degrees) between the 10m – 925 hPa shear vector and the storm-relative 10 m wind. Finally, "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 dBZ or greater at the time of maximum MAXUVV. This variable was added as a potential proxy for storm mode.

2) DATA PREPROCESSING

Data 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, HREFv2.1 data is aggregated in time by computing, for each field, either a 24-h maximum, minimum, or mean, depending on the variable. Storm and index variables use a temporal maximum or minimum as appropriate (e.g., in the case of MAXDVV), while most—but not all—environment fields use a temporal mean. The temporal aggregation strategy for each variable is included in Table 4.3. 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).

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. Predictors are taken from the point of prediction and the 8 nearest grid points. Therefore, the RFs trained in this way will be subsequently referred to as (3x3) ensemble mean (EM) RFs.

To help account for the spatial uncertainty in the placement of simulated storms in the IM RFs (which only consider data at a single grid point), all storm fields except for n30dbz are spatially smoothed using a 2-dimensional isotropic Gaussian kernel density function:

$$\mathrm{v} = \sum_{\mathrm{n}=1}^{\mathrm{N}} \frac{\mathrm{v_n}}{2\pi\sigma^2} \exp\left[-\frac{1}{2}\left(\frac{\mathrm{d_n}}{\sigma}\right)^2\right] \qquad (4.5),$$

where *v* is the spatially-smoothed value at a given point, N is the number of points in the analysis domain, v_n is the raw value at point *n*, d_n is the distance between the *n*th 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 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.

Missing ensemble data is 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), 0-3km UH (UH0-3km), critical angle, or 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 created using all available predictors as well as subsets of predictors. The reduced-predictor RFs use, respectively, only storm, only environment, only index, no storm (i.e., index and environment), no environment (i.e., storm and index), and no index (i.e., storm and environment) predictors.

3) RF TRAINING

All RFs are trained using Scikit-Learn and use the same set of hyper-parameters: 200 trees, a maximum depth of 15, and 20 minimum samples per leaf node. These hyper-parameters are selected based on previous experience with forecasting precipitation and severe weather. A constant set of hyper-parameters is used here for simplicity, since previous sensitivity tests (not shown) have suggested that the RF forecasts are relatively insensitive to variations in these hyper-parameters.

As in Loken et al. (2019, 2020), k-fold cross-validation is used to train and verify the RF forecasts. Here, 16 folds are used: the first 13 folds contain 41 days each, and the final 3 folds each contain 40 days. 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 data 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), Brier Skill Score (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). Buizza et al. (1999) suggests 0.7 is the lower AUC threshold for a useful probabilistic forecast, while 0.8 is the lower threshold for a "good" forecast. However, since AUC depends on probability of false detection, it is sensitive to the number of correct nulls. Thus, for rare events with many trivial correct nulls (e.g., severe weather), AUC may routinely be higher than the "good" benchmark threshold listed in Buizza et al. (1999). Indeed, Loken et al. (2020) showed that AUCs well above 0.9 are not uncommon, even for UH2-5km-based severe weather hazard forecasts, with higher AUCs for the (rarer) significant severe hazards. 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. Like AUC, BS is sensitive to event climatological frequency; trivial correct nulls can artificially reduce (i.e., improve) the BS. To account for this effect, the Brier Skill Score (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. Since both metrics are positively-oriented and optimized at 1, points in the upper right-hand corner of this plot indicate more skillful forecasts.

Performance diagrams (Roebber 2009) plot probability of detection (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. Here, attributes diagrams are created by binning the forecasts using the same forecast probability levels used to create the performance diagrams. More details on AUC, BSS, performance diagrams, and attributes diagrams can be found in Loken et al. (2020) and works cited therein.

3. Results

a. RF Verification

Performance diagrams (Fig. 4.2a-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. The IM and EM RFs tend to have similar CSIs for the smallest probability levels (i.e., up to 5%); differences are more noticeable at the probability values above 15% for all three hazards. Interestingly, the CSI

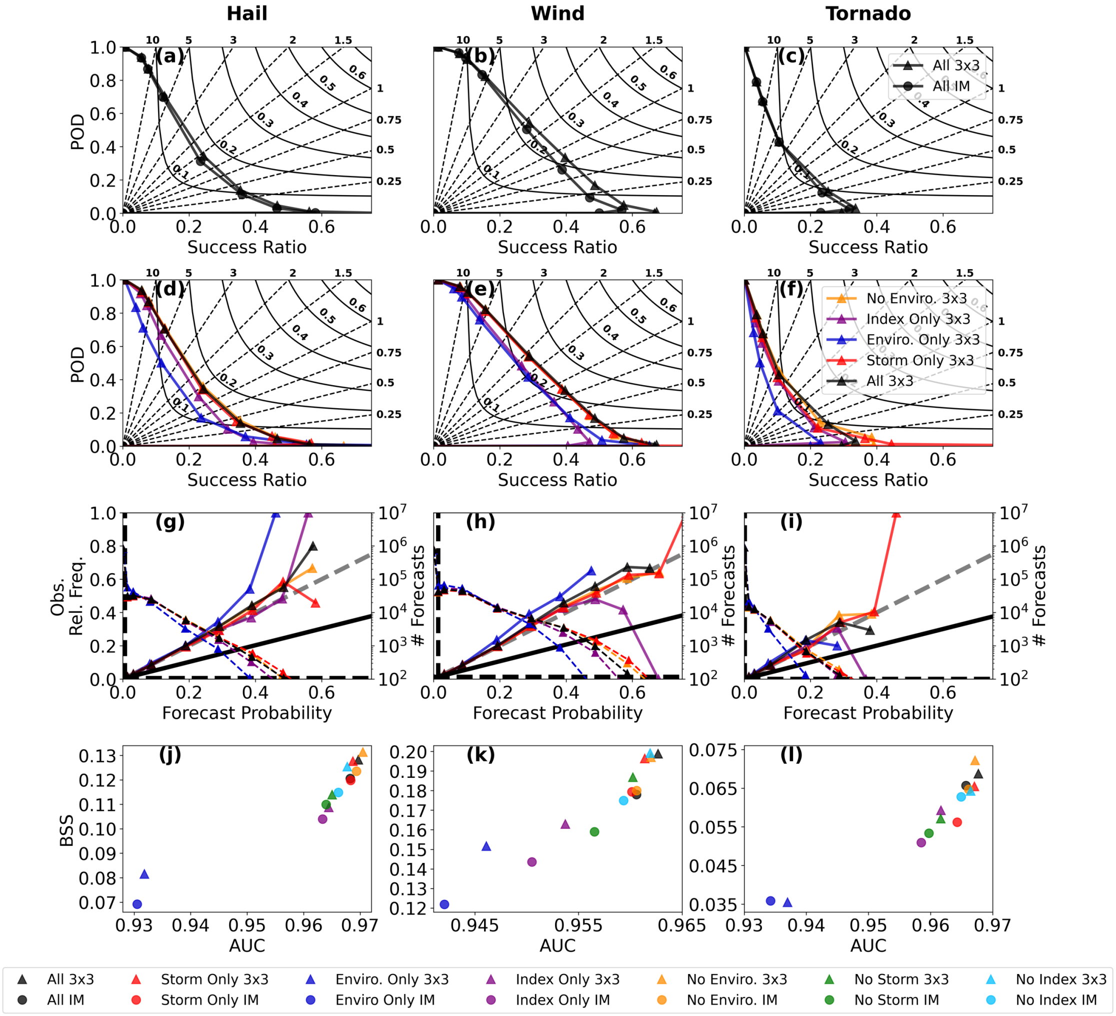


Figure 4.2 (a) Performance diagram for all-predictor IM (filled circles) and EM (filled triangles) RFs for severe hail. (b)-(c) As in (a) but for severe wind and tornadoes, respectively. Note that the x-axis spans from 0 to 0.75. (d)-(f) As in (a)-(c) but for all-predictor (black triangles), storm-only (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.

difference is most pronounced for severe wind (Fig. 4.2b) compared to severe hail (Fig. 4.1a) and tornadoes (Fig. 4.2c). Because the EM RFs are found to be more skillful (and to preserve figure readability), we only show performance and attributes diagrams for the reduced-predictor EM RFs.

EM RFs trained with different predictor subsets achieve different levels of skill. For all three hazards, the RFs trained with only environmental predictors perform quite poorly relative to the other RFs (Fig. 4.2d-f). Relative to the environment-only RFs, the RFs using only index-related predictors achieve notably higher CSIs at most probability levels for severe hail (Fig. 4.2d) and tornado (Fig. 4.2f) prediction. For severe wind prediction, the environment-only and index-only RFs have nearly overlapping performance diagram curves (Fig. 4.2e), which is unsurprising since no wind-specific index is included in the index predictors. Interestingly, for severe wind, the index-only RF performs slightly better at the smaller forecast probabilities, while the environment-only RF performs slightly better at the higher forecast probabilities. For all three hazards, the RFs using only storm-related predictors perform nearly as well as the corresponding all-predictor RFs, suggesting that substantial skill is derived from the storm-related variables. 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. 4.2f). For severe hail and wind, the no-environment, storm-only, and all-predictor RFs have similar performance diagram curves.

All forecasts have good to very good reliability for all three hazards (Fig. 4.2g-i). The large 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 reiterate the findings described above. For all hazards, the EM RFs tend to have greater AUC and BSS values compared to their IM RF counterparts (Fig. 4.2j-l). Additionally, IM and EM RFs trained using only environment predictors have substantially lower AUC and BSS values relative to the other RFs (Fig. 4.2j-l). Meanwhile, the no-environment, all-predictor, and storm-only RFs tend to be amongst the top-performing configurations for all hazards. Interestingly, for severe wind and tornadoes, the EM RF using storm and index predictors (i.e., the no-environment RF) gives the highest BSS and either a better (Fig. 4.2j) or similar (Fig. 4.2l) AUC compared to the all-predictor RF. This suggests that at least modest benefits can be obtained by using both storm *and* index predictors. Pairing storm and environment predictors is typically less skillful, at least for severe hail (Fig. 4.2j) and tornado (Fig. 4.2l) prediction. Moreover, using only index predictors tends to produce greater AUC and BSS values than using only environmental predictors (Fig. 4.2j-l). This result is important because it suggests that merely supplying a RF with the most basic, constituent predictors is inferior to providing a RF with more complex predictors that have already been appropriately "pre-related" with other predictors (during preprocessing). Intuitively, this makes sense; a single "multivariate" predictor that has a direct association with observed severe weather should be easier to learn (from the same amount of training data) than multiple predictors that have weaker individual associations with observed LSRs.

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 (Fig. 4.3a-d).

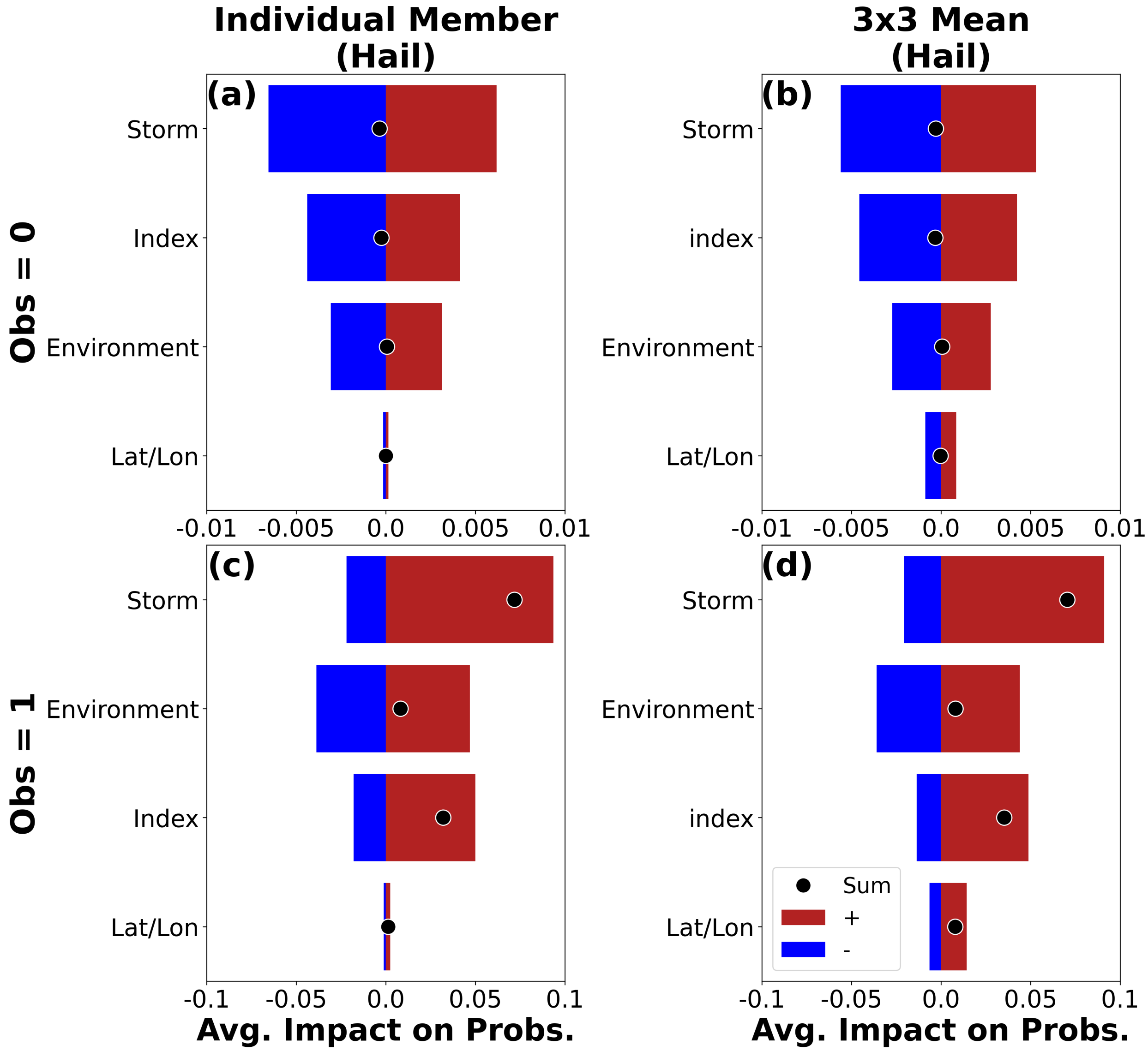


Figure 4.3 (a) Mean TI negative (blue), positive (red), and summed (i.e., negative plus positive; black dot) RF probability contributions (per grid point) from storm, index, environment, and latitude/longitude variables in the all-predictor IM severe hail RF. Variable subsets are displayed in order of 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)-(b) but for cases associated with an observed hail storm report. Note the different x-axis scale in (a)-(b) compared to (c)-(d).

Overall, the storm variables tend to appropriately decrease RF probabilities when no LSR is present (Fig. 4.3a-b) and increase probabilities when an LSR is present (Fig. 4.3c-d). Index variables exert comparatively less impact on RF probabilities but also tend to move probabilities in an appropriate direction based on the presence or absence of an LSR. 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 (Fig. 4.3a-b) and increase probabilities much substantially less than the storm and index variables when an LSR is present (Fig. 4.3c-d). These results align with the verification results from the reduced-predictor RFs that show environment-only RFs are much less skill than RFs that use storm and/or index predictors for severe hail prediction. It is interesting that the environment variables move the RF probabilities more than the index variables when an LSR is present (Fig. 4.3c-d) but are responsible for increasing the probabilities less than the index variables in these cases. Why a notably "inferior" variable would be more "important" is unclear. One possible explanation is that more environmental variables than index variables exist and so get used more often due to the RF algorithm being forced to choose amongst a subset of predictors at each node. For both the IM and EM RFs, the latitude and longitude variables exert the least impact on RF probabilities; however, these variables move the forecast probabilities more in the EM RF (Fig. 4.3b,d) compared to the IM RF (Fig. 4.3a,c). Again, a likely explanation is the number of latitude and longitude variables in each configuration. In the IM RFs, latitude and longitude are each only represented once, while in the EM RFs, latitude and longitude are each represented nine times (once for each spatial grid point examined), making it more likely that those variables are used to split the training dataset at a given RF node. In both the IM and EM RFs, the latitude and longitude, on average, move the probabilities slightly higher in cases with an LSR (Fig. 4.3c-d) but don't tend to move the probabilities much higher or lower in cases without an LSR (Fig. 4.3a-b).

Performing this same analysis for individual fields shows that 2-5 km updraft helicity (UH2-5km) and SHIP (not necessarily in that order) are the top two "most important" predictors for both the IM and EM RFs in all instances (Fig. 4.4a-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 are emphasizing variables that make physical sense.

MAXDVV and MAXUVV are also amongst the top-performing predictors for both RFs. One interesting result is that MAXUVV tends to, on average, exert a positive impact on RF probabilities even in the absence of an LSR (Fig. 4.4a-b). However, it also tends to increase RF probabilities strongly (compared to similarly-important variables) when an LSR is reported (Fig. 4.4c-d). This finding suggests that MAXUVV may be more useful for enhancing POD and less useful for reducing false alarms. Similarly, while UH2-5km was effective at reducing probabilities for instances without an LSR, 0-3km UH is less useful for that purpose (Fig. 4.4a-b) but is useful for appropriately increasing RF probabilities (Fig. 4.4c-d).

Variable rankings are generally similar between the EM and IM methods; however, there are some appreciable differences. The most notable, consistent with Fig. 4.3, is that both latitude and longitude appear much more important in the EM RF. Another noticeable difference is that smoothed 1-km above-ground simulated reflectivity (MXREF1km) is more important in the IM than EM RF, although it is not clear why.

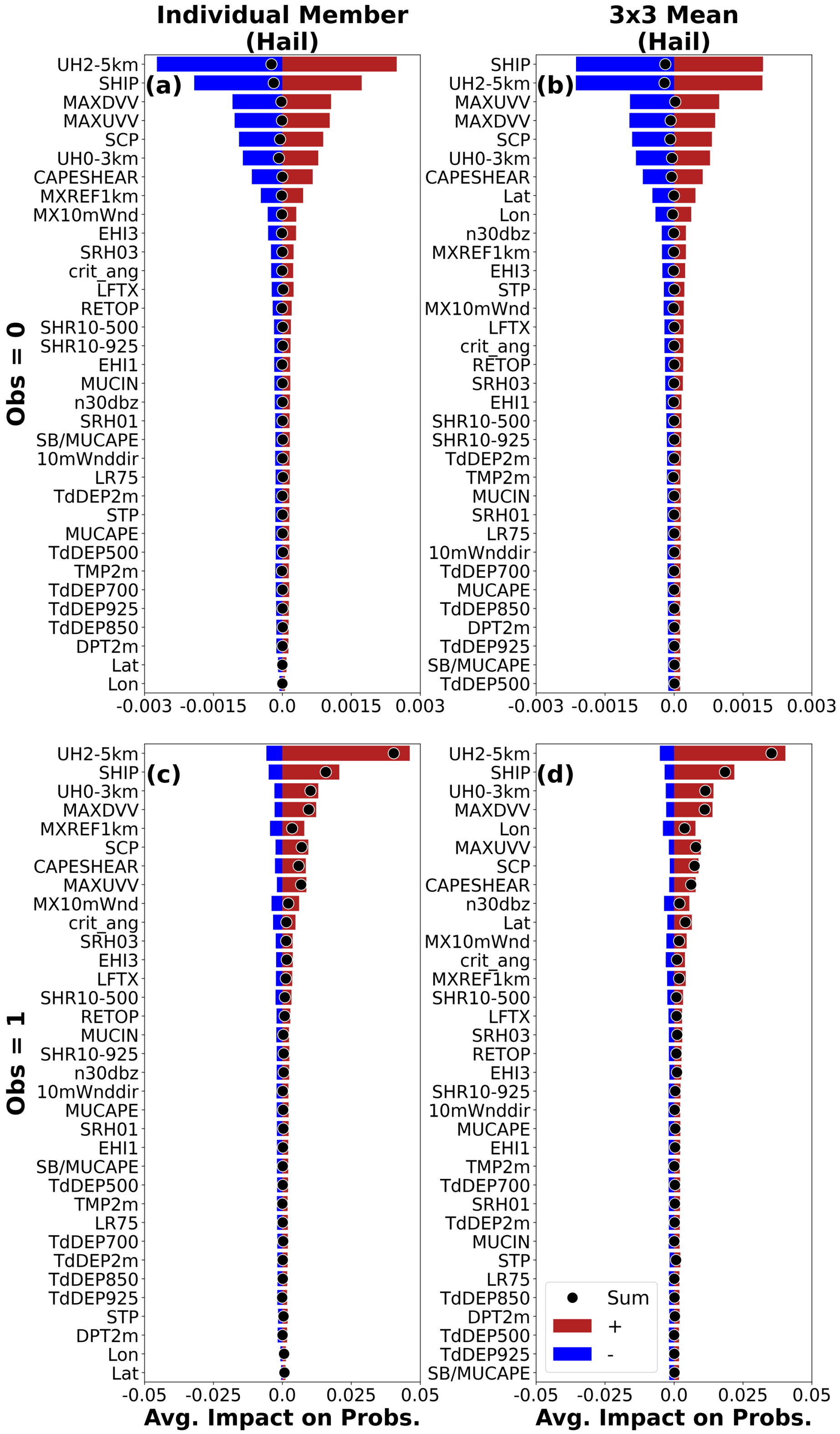


Figure 4.4 As in Fig. 4.3 but for contributions aggregated over the individual predictor fields.

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 MUCAPE or 700-500 hPa lapse rate (LR75). 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 (Fig. 4.5a-d). Indeed, the storm variables are substantially more important than either the environment or index variables for both types of RFs. 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 (Fig. 4.5a-b), indicating substantial skill.

In cases without an LSR, the environment and index variables exert a similar influence (Fig. 4.5a-b); this is relatively unsurprising given that no "wind-specific" index is used as a 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 (Fig. 4.5c-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 [e.g., SCP, EHI (both 0-1 km and 0-3 km), and CAPESHEAR (i.e., the product of MUCAPE and 10m-500hPa wind shear magnitude)].

IM and EM RFs tend to use storm, environment, and index variables similarly. However, as seen in Fig. 4.3 for severe hail, the EM RFs place much more importance on latitude and longitude predictors compared to the IM RFs. Indeed, the average probability contribution from

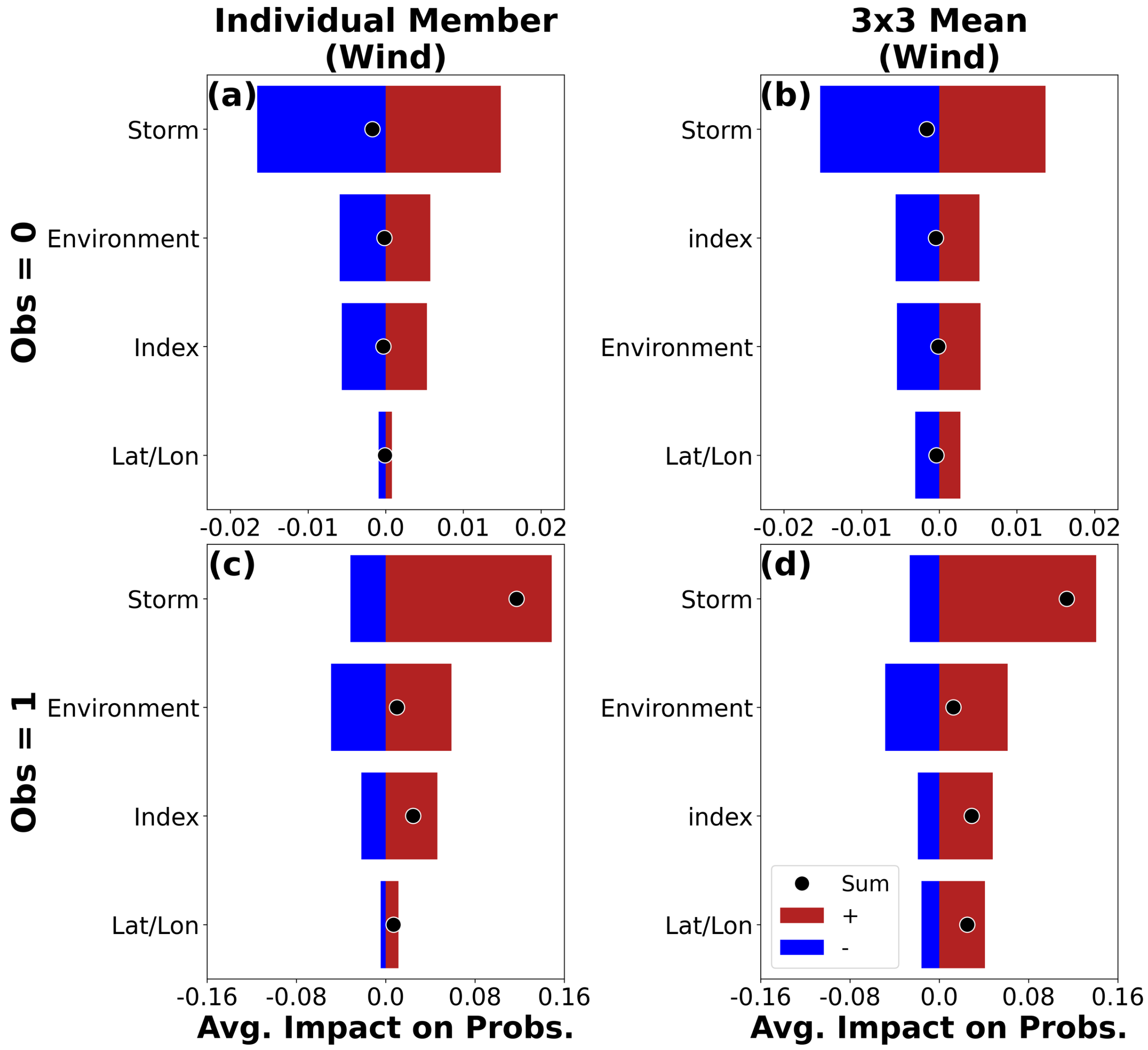


Figure 4.5 As in Fig. 4.3 but for severe wind.

latitude and longitude in the EM RF is substantial, nearly equaling that from the index variables in cases with an observed storm report (Fig. 4.5d).

In terms of specific fields, UH2-5km, MAXDVV, and UH0-3km are amongst the top wind predictors for both RF configurations (Fig. 4.6a-d). These variables all tend to move the probabilities in the correct direction in instances without (Fig. 4.6a-b) and with (Fig. 4.6c-d) an LSR. MAXUVV is another top predictor for severe wind, and—as seen with severe hail—it

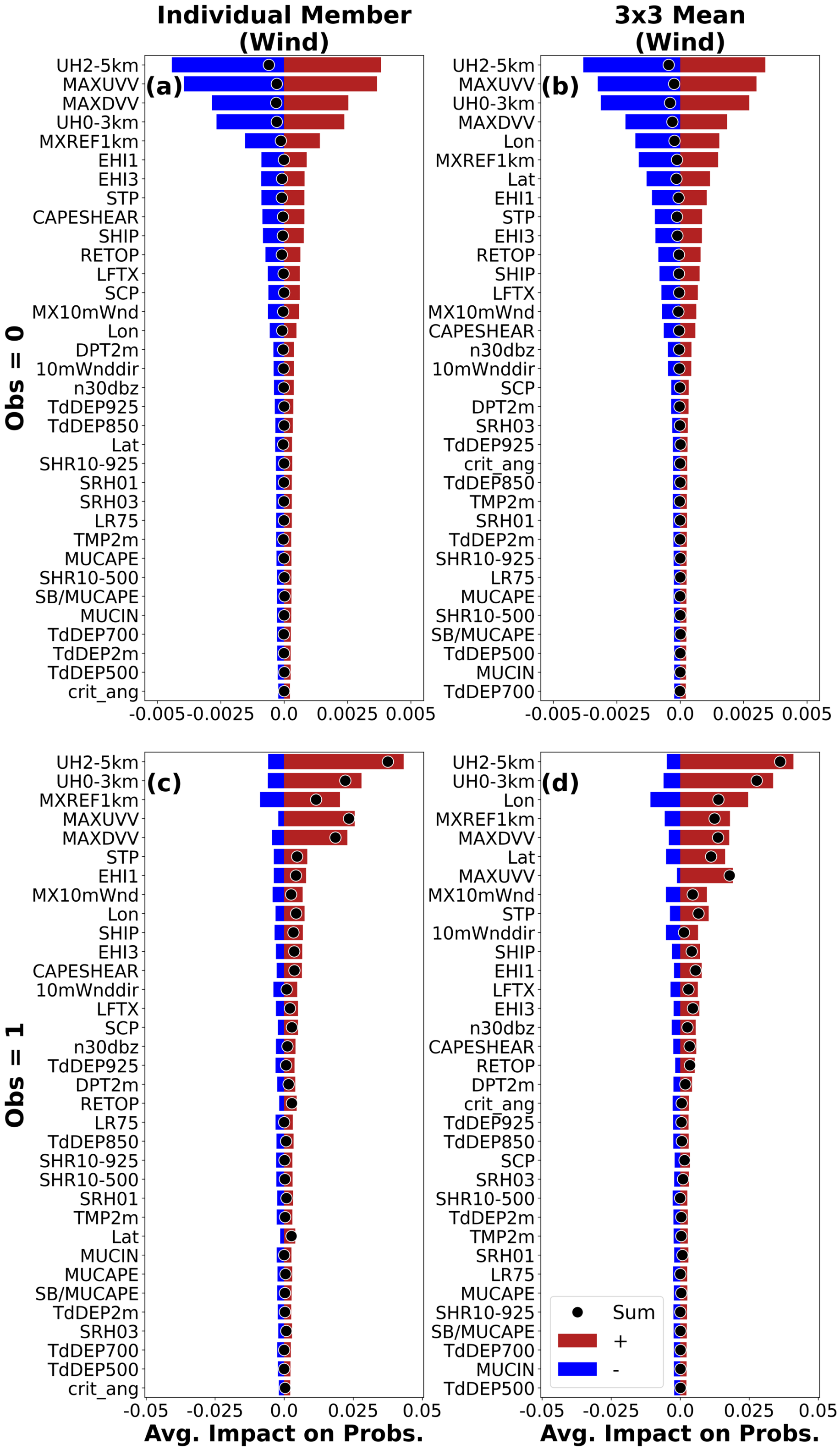


Figure 4.6 As in Fig. 4.4 but for severe wind.

seems well-suited for enhancing POD. While it only modestly decreases probabilities in instances without an LSR (Fig. 4.6a-b), it almost always increases probabilities when an LSR is observed (Fig. 4.6c-d).

The biggest difference between IM and EM RF configurations is the relative influence of the latitude and longitude variables. Indeed, longitude is the 5th (3rd) most important EM RF variable in cases without (with) an LSR (Fig. 4.6b,d)! In contrast, longitude is the 15th (9th) most important predictor to the IM RF without (with) and LSR (Fig. 4.6a,c). However, even for the IM RF, longitude and latitude are still notably more important for severe wind (Fig. 4.6a,c) than severe hail (Fig. 4.4a,c). The relatively high importance of these location variables for severe wind suggests that the RFs are learning systematic relationships between location and observed severe wind. It is likely that these relationships are at least partially due to 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 overall levels of importance, in both the IM and EM RFs (Fig. 4.7a-d). Interestingly, storm variables move the probabilities most in cases with no observed LSR (Fig. 4.7a-b), but environmental variables move the probabilities most in cases with an LSR (Fig. 4.7c-d). However, both storm and index variables tend to correctly increase the probabilities more than the environment variables when an LSR is present (Fig. 4.7c-d), indicating that these variables may be more useful than the environment variables.

The relative importance of storm, environment, and index variables is similar for IM and EM RFs. As with severe hail and wind, latitude and longitude predictors exert more influence on

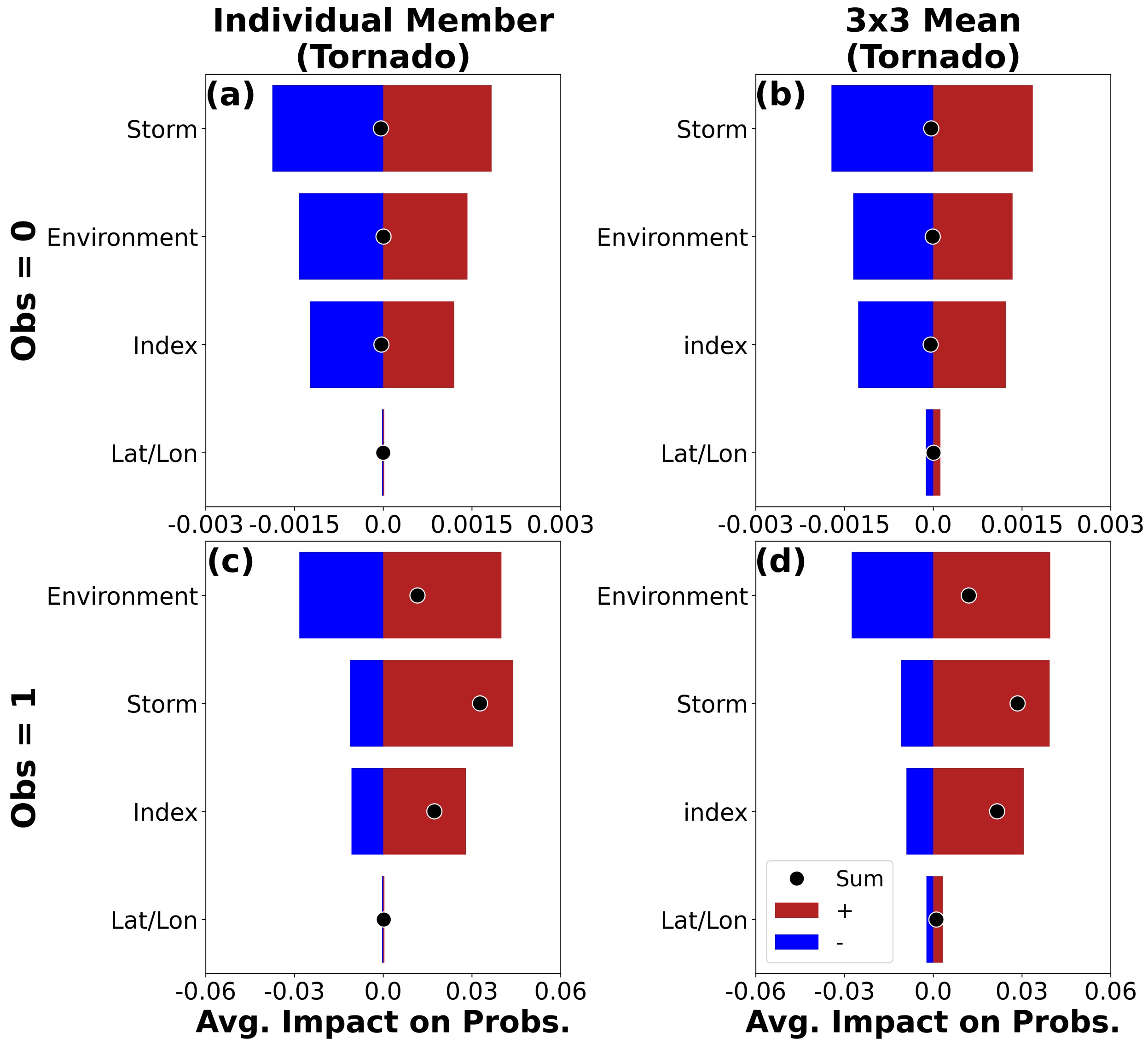


Figure 4.7 As in Fig. 4.3 but for tornadoes.

the EM RF probabilities, but even for the EM RF, latitude and longitude predictors have little influence on the RF probabilities compared to the storm, environment, and index variables.

UH0-3km is the most important variable for tornado prediction in both the IM and EM, regardless of whether there is an observed LSR (Fig. 4.8a-d). This is consistent with Sobash et al. (2019), who found UH0-3km performed better than UH2-5km (but worse than 0-1km updraft

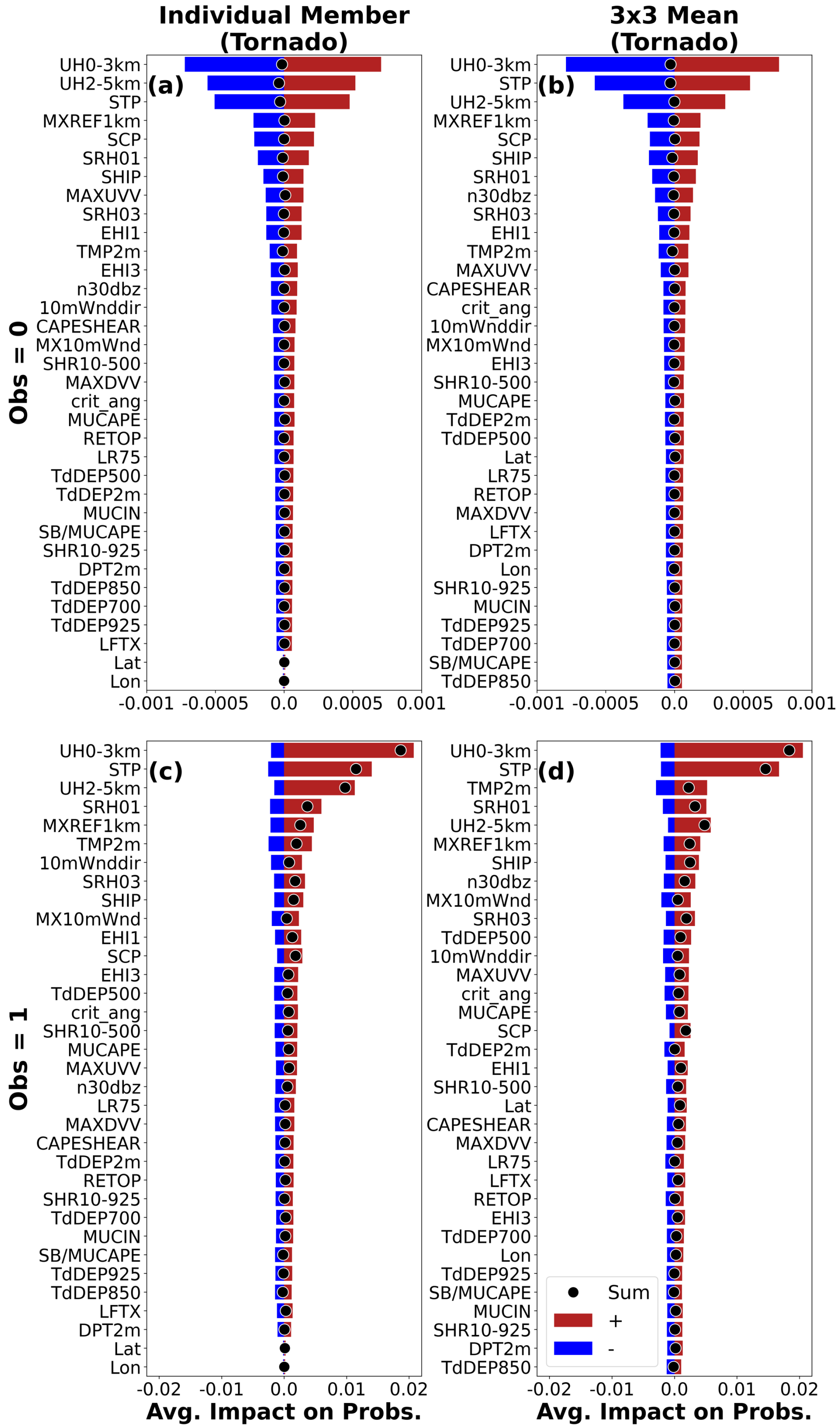


Figure 4.8 As in Fig. 4.4 but for tornadoes.

helicity) for predicting tornadoes. Other important predictors include STP, UH2-5km, MXREF1km, and 0-1km 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. 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, and high MXREF1km indicates intense storms.

One interesting finding is the relatively high importance of 2-m temperature (TMP2m), especially for the EM RF when an observed LSR is present (Fig. 4.8d). While TMP2m is the 3rd most important predictor in this case, it increases the probabilities less than other "less important" variables (e.g., 0-1km SRH and UH2-5km; Fig. 4.8d). However, TMP2m also tends to correctly decrease RF probabilities more than similarly-important variables when no LSR is present (Fig. 4.8b). These results suggest that the RF uses TMP2m more to help reduce false alarms than to increase POD.

Given that high boundary layer relative humidity—and therefore small low-level dewpoint depressions—have been associated with tornadoes (Markowski et al. 2002), it is surprising that 2-m and 925 hPa dewpoint depression (TdDEP2m and TdDEP925, respectively) are relatively unimportant. One potential explanation is that the RF merely considers the 24-h temporal mean dewpoint depression, which may have a weaker association with tornadoes than TdDEP2m or TdDEP925 in the hours directly preceding (potential) tornadogenesis. It should be noted that STP (indirectly) contains information about the TdDEP2m, since the LCLs used in the STP computation are approximated from TdDEP2m. STP is likely a more important tornado predictor than TdDEP2m or TdDEP925 because it efficiently combines information from multiple fields at a common time.

4) MEMBER AND SPATIAL CONTRIBUTIONS

The IM method allows TI to determine the relative importance of different sets of members. Unsurprisingly, for all hazards, the set of non-time-lagged members influences the RF probabilities more than the time-lagged members (Fig. 4.9a-f). The non-lagged members tend to decrease RF probabilities more when no LSR is present (Fig. 4.9a-c) and increase RF probabilities more when an LSR is present (Fig. 4.9d-f).

Interestingly, for wind and hail, the different members are not viewed as equally important. For both hazards, the NSSL variables are noticeably more important than variables from other members (Fig. 4.9a,b,d,e). It is also interesting that the NMMB member variables are second most-important (behind the NSSL variables) for severe hail (Fig. 4.9a,d) but least important (of the member variables) for severe wind prediction (Fig. 4.9b,e). Meanwhile, for tornado prediction, the HRRR, ARW, NMMB, and NSSL variables, respectively, have similar levels of importance, while the NAM variables are notably less important (Fig. 4.9c,f). The smaller importance of the NAM variables for tornadoes is likely attributable to the NAM's lack of UH0-3km and STP, two of the most important fields for tornado prediction (Fig. 4.8).

Because it uses predictors from different spatial points, the EM method allows for the analysis of spatial predictors. Over all variables, it is clear the most important predictors are the ones taken from the point of prediction for all hazards (Fig. 4.10a-c). However, the distribution of importance values over the 3x3 grid is not isotropic. For severe hail, the RF places more importance on predictors to the east of the point of prediction (Fig. 4.10a); for severe wind, more importance is placed on predictors to the west and south of the point of prediction; and for

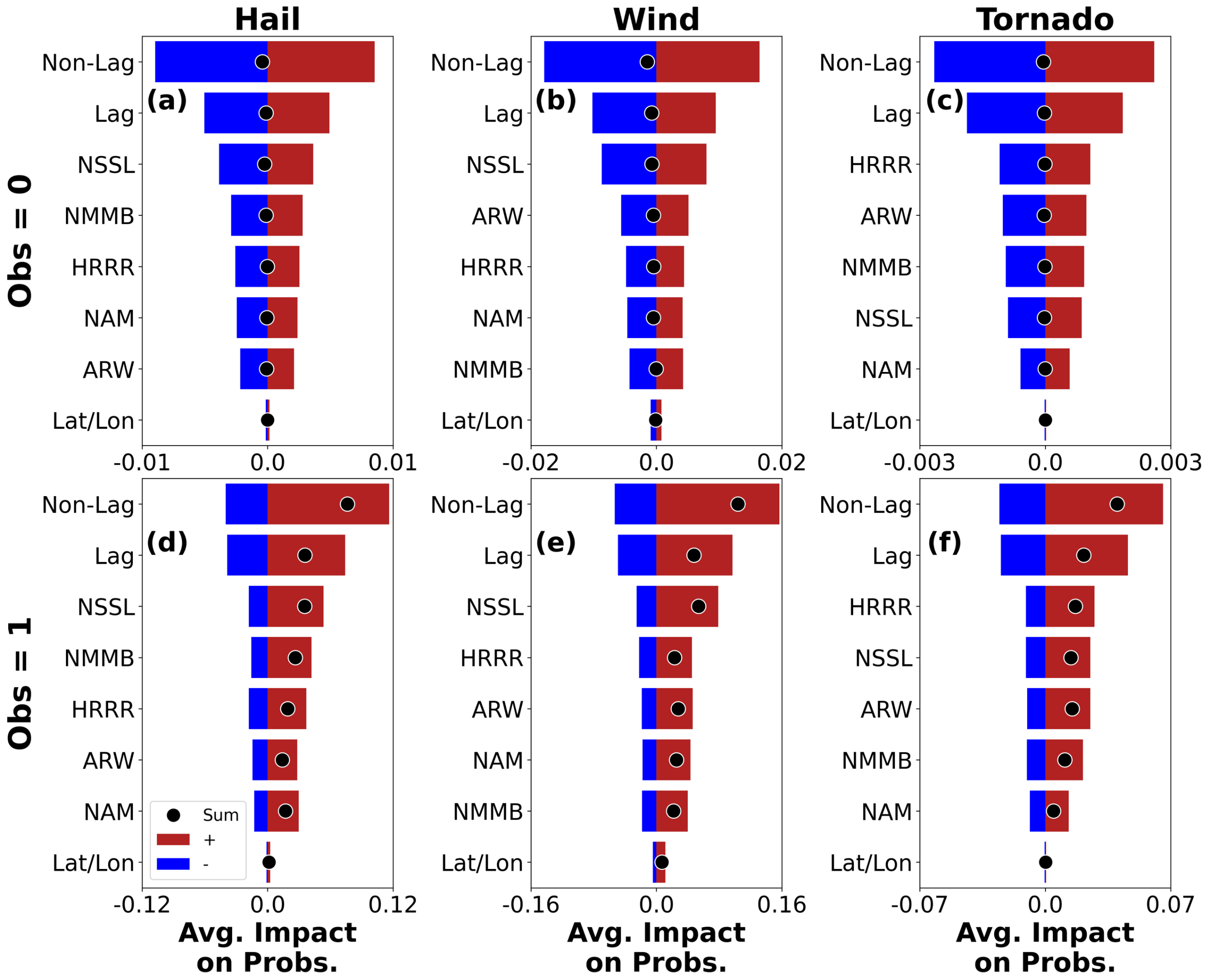


Figure 4.9 As in Fig. 4.3, but for contributions from the non-time-lagged, time-lagged, NSSL, NMMB, HRRR, ARW, NAM, and latitude/longitude predictors from the all-predictor IM RF. Cases not associated with an observed storm report are shown in (a)-(c), while cases with an observed storm report are shown in (d)-(f). Columns 1, 2, and 3 show results from the severe hail, severe wind, and tornado RFs, respectively. Note the different x-axis scales in each panel.

tornadoes more importance is placed on variables to the northwest and southeast compared to the southwest and northeast.

For severe hail and wind prediction, storm variables at the point of prediction are substantially more important than storm variables at surrounding points (Fig. 4.10d-e). This same pattern is not found to the same extent for the non-storm (i.e., environment and index) variables;

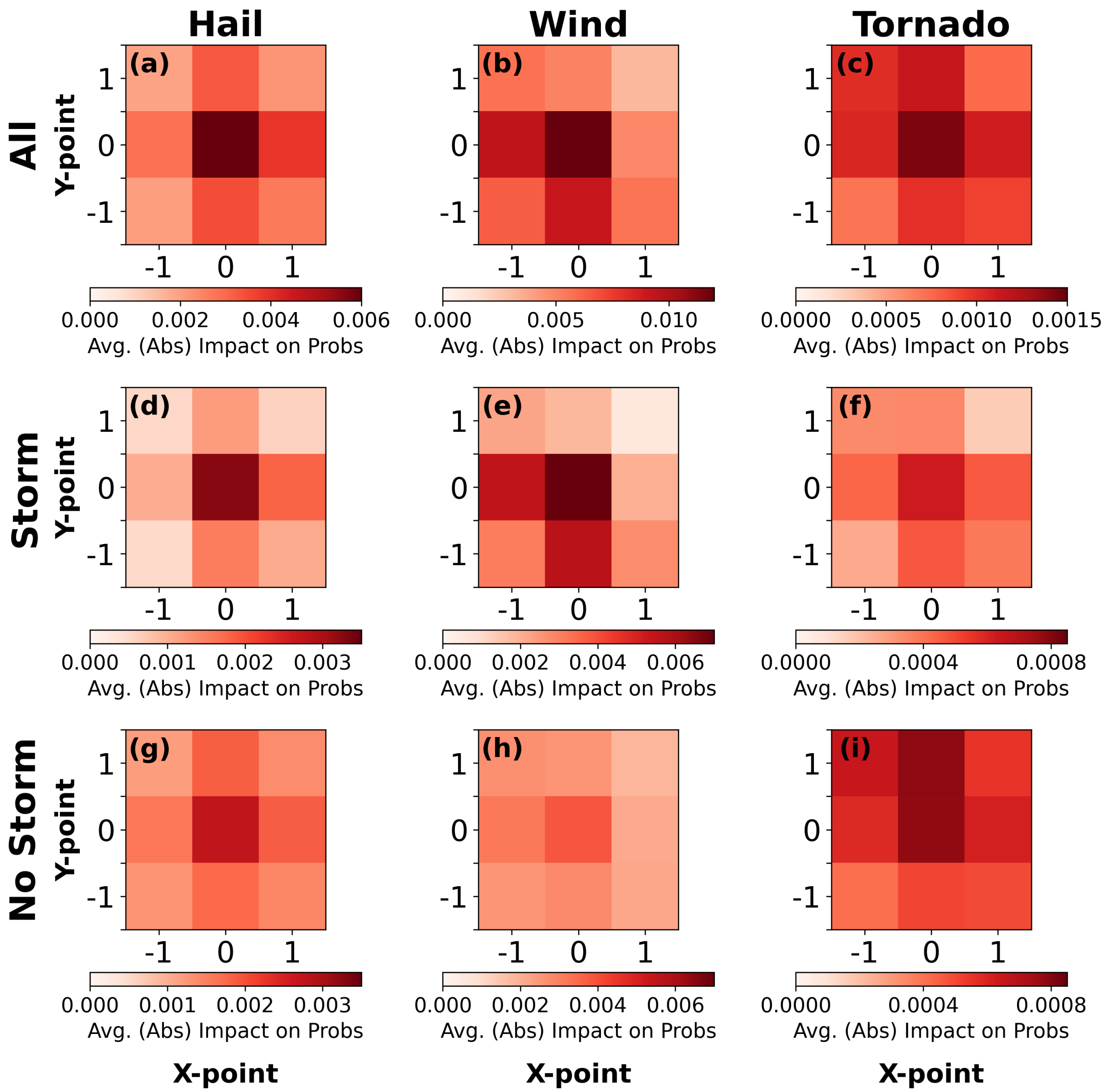


Figure 4.10 (a) Mean absolute EM RF probability contributions from all predictors at the point of prediction [(0,0)] and the 8 closest 80-km grid points for severe hail. (b)-(c) As in (a) but for severe wind and tornadoes. (d)-(f) As in (a)-(c) but for only the storm variables. (g)-(i) As in (a)-(c) but for only the non-storm variables. Note the different color bar scales between columns. Within each column, panels in rows two and three have the same scales.

predictors from the surrounding points are comparatively much more important (Fig. 4.10g-h).

One potential interpretation of these results is that the RFs use environment and index variables to assess the conduciveness of the environment to severe weather around the point of protection, while they use the storm variables to “pinpoint” where storms are most likely to occur. Interestingly, this relationship does not apply as much for tornadoes (Fig. 4.10f,i).

c. Single-field relationships

The hail-, wind-, and tornado-predicting IM RFs learn different relationships between UH2-5km and the likelihood of observed severe weather depending on the ensemble member and forecast hazard (Fig. 4.11a-r). For example, for the non-lagged HRRR and NSSL members, larger values of UH2-5km tend to result in greater contributions to RF severe hail probability (Fig. 4.11a,j), while larger values of UH2-5km from the non-lagged NMMB and NAM members did not always result in greater RF hail probability contributions (Fig. 4.11g,m). For severe wind, increasing UH2-5km tends to increase RF probability contribution only up to a point for most of the non-lagged members (Fig. 4.11b,e,h,k,n). Beyond approximately 50 m^2/s^2, for example, higher values of UH2-5km in the non-lagged ARW member do not further increase the RF severe wind probability contribution (Fig. 4.11e). For tornado prediction, the UH2-5km and RF probability contribution is more muted, consistent with the earlier finding that UH0-3km is a more important tornado predictor (Fig. 4.8). UH2-5km above 100 m^2/s^2 from the non-lagged HRRR and NSSL members appears to have the greatest increasing impact on RF tornado probabilities (Fig. 4.11c,l).

Interestingly, for all hazards, the relationship between the ensemble mean UH2-5km and the total ensemble UH2-5km contribution to the RF probabilities (Fig. 4.11p-r) tends to be clearer than the corresponding relationship for an individual ensemble member (e.g., Fig. 4.11a-o). This result supports the earlier finding that EM RFs are more skillful than IM RFs and

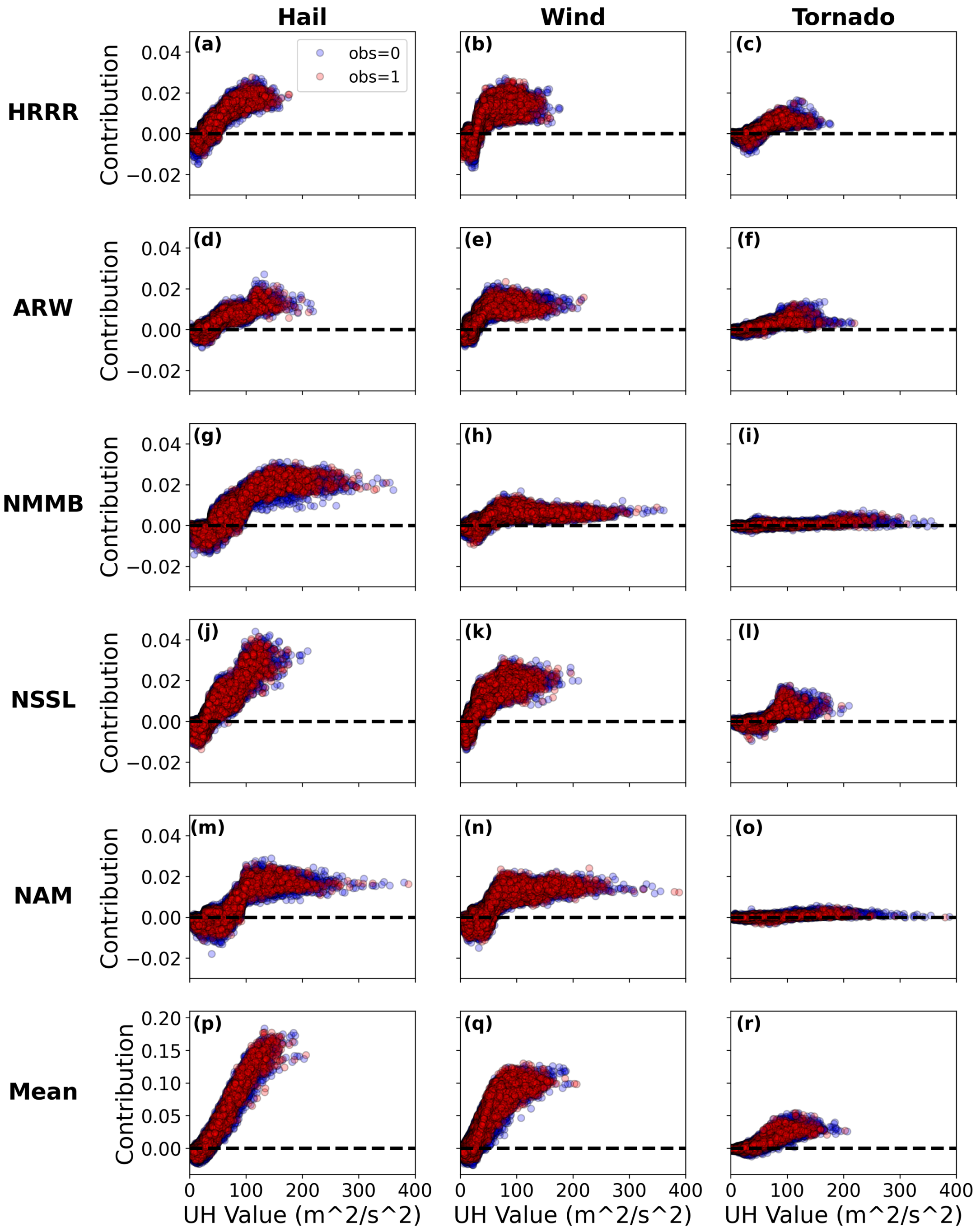


Figure 4.11 (a) IM RF probability contributions from the non-time-lagged HRRR member's UH2-5km for severe hail for each sample in the dataset. Samples associated with an (no) observed LSR are colored red (blue). Each point is semi-transparent, so darker colors indicate greater sample density. A 0.00 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.

suggests the reason is because ensemble mean fields more efficiently and effectively summarize the ensemble information.

Fig. 4.11 also illustrates two other important points. First, while a definite relationship exists between each member's UH2-5km and the contribution to the RF probabilities, the sign of the contribution does not necessarily discriminate well between events (i.e., severe LSRs) and non-events (i.e., no LSRs). Intuitively, this makes sense; sometimes CAEs have errors in storm location, so a large value of UH2-5km (and a large UH2-5km contribution to RF probabilities) is not associated with an observed LSR. Conversely, sometimes, a severe storm is present in the observations but not in any members. ML cannot "fix" these types of errors but can learn what a certain UH2-5km value from a given member (or ensemble mean) means for the probability of observed severe weather (Fig. 4.11). The second important point in Fig. 4.11 is that the same UH2-5km value (for a given member or the ensemble mean) can contribute differently to the overall RF probabilities depending on the case. Indeed, when UH2-5km is small (e.g., near 25-50 m^2/s^2 for many members for many hazards), UH2-5km can contribute either positively or negatively to the RF probabilities. This variability is a consequence of other variables interacting with UH2-5km. For example, a small-to-moderate value of UH2-5km may be favorable or unfavorable for severe hail depending on the environment.

Expectedly, different fields have different relationships with the probability of observed severe weather, and these relationships vary based on the hazard. Fig. 4.12a-r shows some of these relationships learned by the EM RFs. Ensemble mean UH2-5km has an "S-shaped" relationship with severe probability for severe hail (Fig. 4.12a), while the relationship is more "sickle-shaped" for wind (Fig. 4.12b) and "heavily-flattened-S-shaped" for tornadoes (Fig. 4.12c). Meanwhile, UH03-km has the clearest direct relationship with tornadoes (Fig. 4.12f)

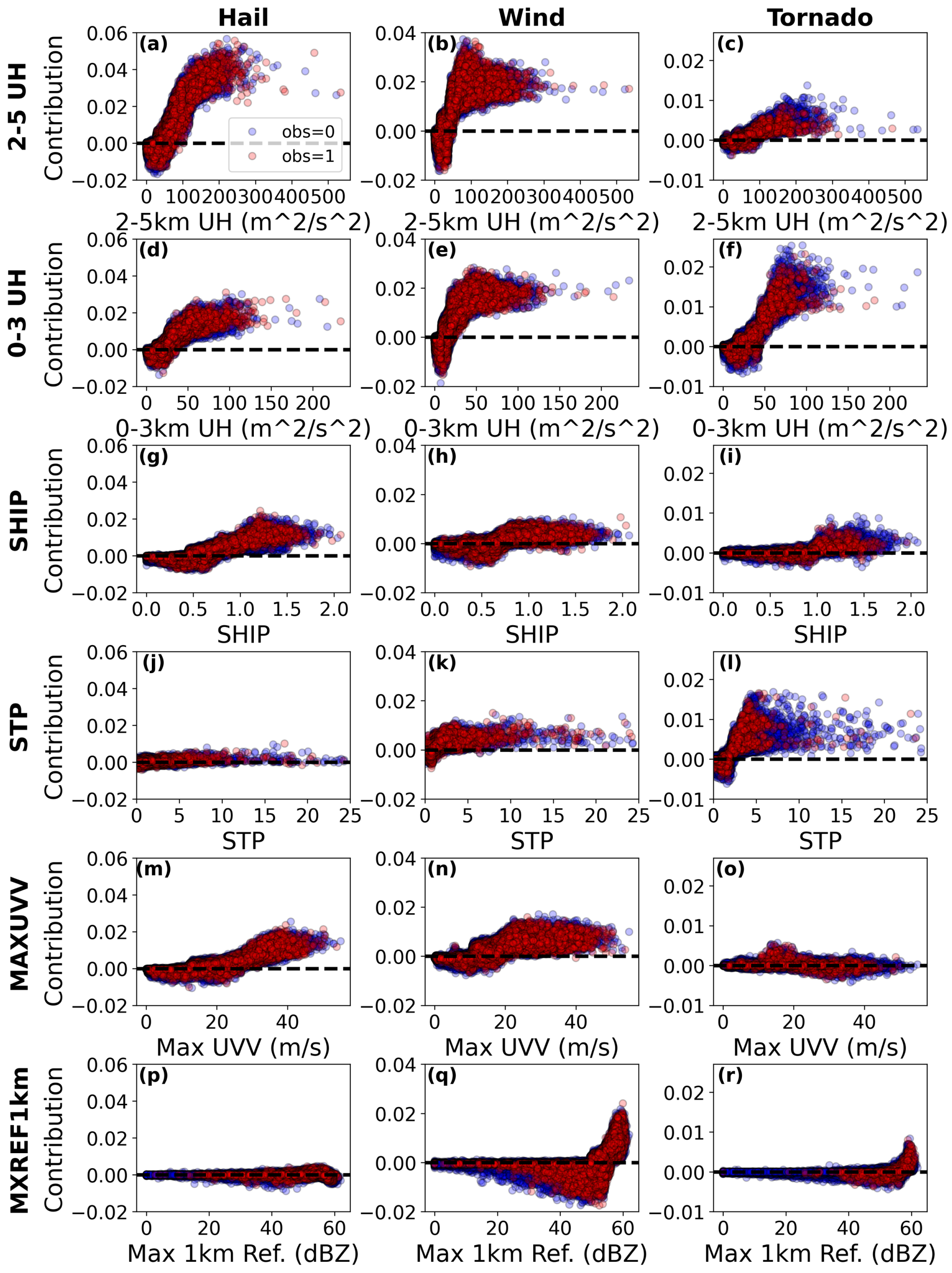


Figure 4.12 (a) EM RF contributions from (0,0) (unsmoothed) mean UH2-5km for each sample in the dataset for severe hail. Samples associated with an (no) observed LSR are colored red (blue). Each point is semi-transparent, so darker colors indicate greater sample density. A 0.00 contribution is indicated by a black horizontal line. (b)-(c) As in (a) but for severe wind and tornadoes. Note that, unlike Fig. 4.11p-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.

compared to severe hail (Fig. 4.12d) or wind (Fig. 4.12e), consistent with the earlier finding that UH0-3km is the most important tornado predictor. All three hazards have a weak relationship between SHIP and RF probability contribution (Fig. 4.12g-i), though the relationship is strongest for severe hail (Fig. 4.12g), as expected. Similarly, STP is related most to tornado probability contributions (Fig. 4.12l) and exerts little impact on hail probability contributions (Fig. 4.12j). MAXXUVV is related most to severe hail probability contribution (Fig. 4.12m), which makes sense as strong updrafts are required to support large hailstones. Interestingly, MXREF1km has the clearest relationship for severe wind (Fig. 4.12q), showing negative contributions until approximately 50-55dBZ and then mostly positive contributions. The relationship in Fig. 4.12q makes sense, since the strongest storms should be associated with all hazards of severe weather; however, it is surprising that MXREF1km does not have a similar effect on severe hail (Fig. 4.12p) or tornado (Fig. 4.12r) probabilities, especially since large reflectivity values are physically associated with a greater likelihood of severe hail.

d. Multi-field relationships

Running TI with *joint_contribution* set to True shows that the greatest contributions to the RF probabilities are due from single predictor variables (not shown). With that said, multivariable interactions are not negligible, as illustrated in Figs. 4.11-4.12. Fig. 4.13a-i shows the RF probability contributions from the three most important two-variable relationships in the IM RFs for each hazard. Interestingly, all these relationships involve either two storm predictors or one storm and one index predictor.

For severe hail, the interaction between NSSL UH2-5km and NSSL SHIP is the most important (Fig. 4.13a). Intuitively, it makes sense that the RF probability contribution should be

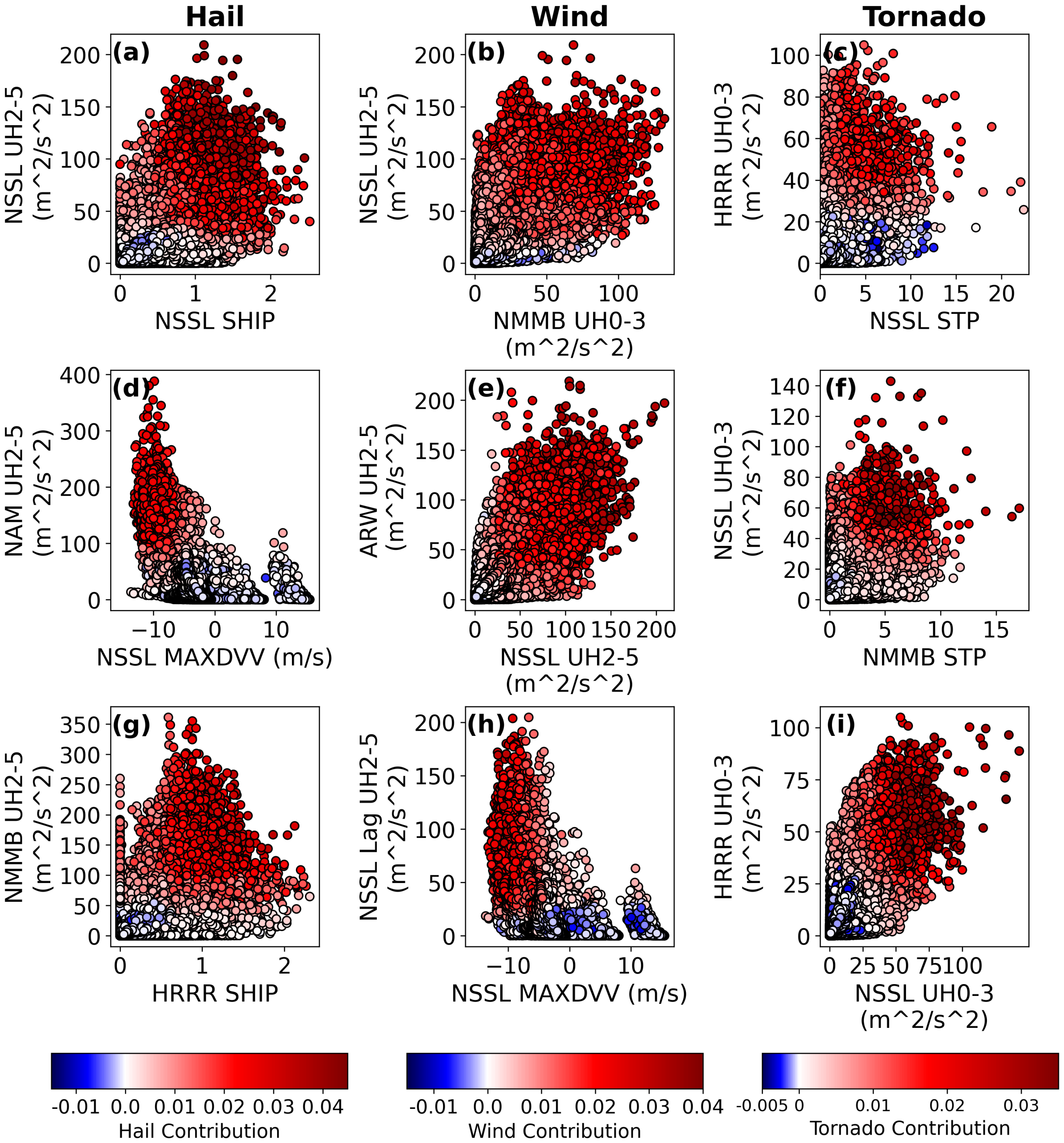


Figure 4.13 (a) IM RF probability contribution (shaded dots) resulting from the most important two-variable combination for all samples in the dataset for severe hail prediction. (b)-(c) As in (a) but for severe wind and tornado prediction, respectively. (d)-(f) As in (a)-(c) but for the second-most-important 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.

maximized when SHIP and UH2-5km are both largest. However, according to Fig. 4.13a, the same value of NSSL UH2-5km (e.g., 25 m^2/s^2) can result in negative hail probability contributions if the SHIP is close to 0 or weak-to-moderate hail probability contributions if the SHIP is near 2. Similarly, a SHIP near 0 can result in negative probabilities if UH2-5km is also small but can result in weak-to-moderate probabilities if UH2-5km is relatively large (e.g., near 100 m^2/s^2). Thus, simulated storms with strongly (weakly) rotating updrafts in marginal (favorable) simulated environments can still result in non-negligible probabilities of observed severe weather. A similar effect is seen in the interaction between NSSL UH0-3km and NMMB STP for tornado prediction (Fig. 4.13f). To a lesser extent, the effect is also present between NMMB UH2-5km and HRRR SHIP for severe hail (Fig. 4.13g) and between HRRR UH0-3km and STP for tornadoes (Fig. 4.13c).

Some important two-variable interactions involve two storm-related variables; for example, NAM UH2-5km and NSSL MAXDVV are an important combination for severe hail prediction (Fig. 4.13d). Very large values of NAM UH2-5km (i.e., 200 m^2/s^2 and above) are nearly always associated with associated with relatively high RF probability contributions. However, when NAM UH2-5km is less (e.g., near 80 m^2/s^2), NSSL MAXDVV plays a large role in modulating the RF probability contributions (Fig. 4.13d). These results suggest that the common practice of using a UH2-5km threshold to forecast severe weather (e.g., Sobash et al. 2011, 2016b, 2019; Loken et al. 2017, 2020) does not always give the most complete representation of the severe weather threat.

Indeed, even when a single field is considered from multiple members (e.g., UH2-5km in Fig. 4.13e; UH0-3km in Fig. 4.13i), a constant threshold would likely still provide incomplete information. First, (imaginary) isolines of constant probability contribution in Fig. 4.13e,i do not have a slope of exactly -1, which is unsurprising given that different members have different climatologies of UH2-5km and UH0-3km (e.g., Roberts et al. 2020). Additionally, in general, there is still a non-zero gradient in probability contribution at relatively large values of UH2-5km and UH0-3km for both members. For example, Fig. 4.13e shows that the probability contribution is larger when both members have UH2-5km near 200 m^2/s^2 compared to when both members have UH2-5km near 100 m^2/s^2, even though both values are relatively large. A similar effect is seen in Fig. 4.13i with UH0-3km.

The most important multivariate relationships from the EM RFs reflect a similar general pattern: the combinations involve either multiple storm fields or one storm field and one index field (Fig. 4.14a-i). Unsurprisingly, the most important combinations also involve variables at or close to the point of prediction. However, most of the combinations involve variables at different spatial points (e.g., Fig. 4.14a,b,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. Intuitively, it makes sense what the EM RFs are learning. For example, the probability of severe hail will be maximized when UH2-5km is large at *and* near the point of prediction (Fig. 4.14a). Especially when the UH2-5km at the point of prediction is low or marginal (e.g., 25 m^2/s^2), the UH2-5km at a neighboring grid point can make a big difference in determining the severe hail probability contribution (Fig. 4.14a).

In general, Figs. 4.13-4.14 suggest that RFs consider the specific values of multiple storm and index variables to construct their probabilities. While UH2-5km is an important predictor for most hazards (Figs. 4.4, 4.6, 4.8), when UH2-5km is marginal (e.g., near or below 50 m^2/s^2 in Fig. 4.13a,g or Fig. 4.14g), other fields (e.g., SHIP) can play an especially important role in quantifying the severe weather threat.

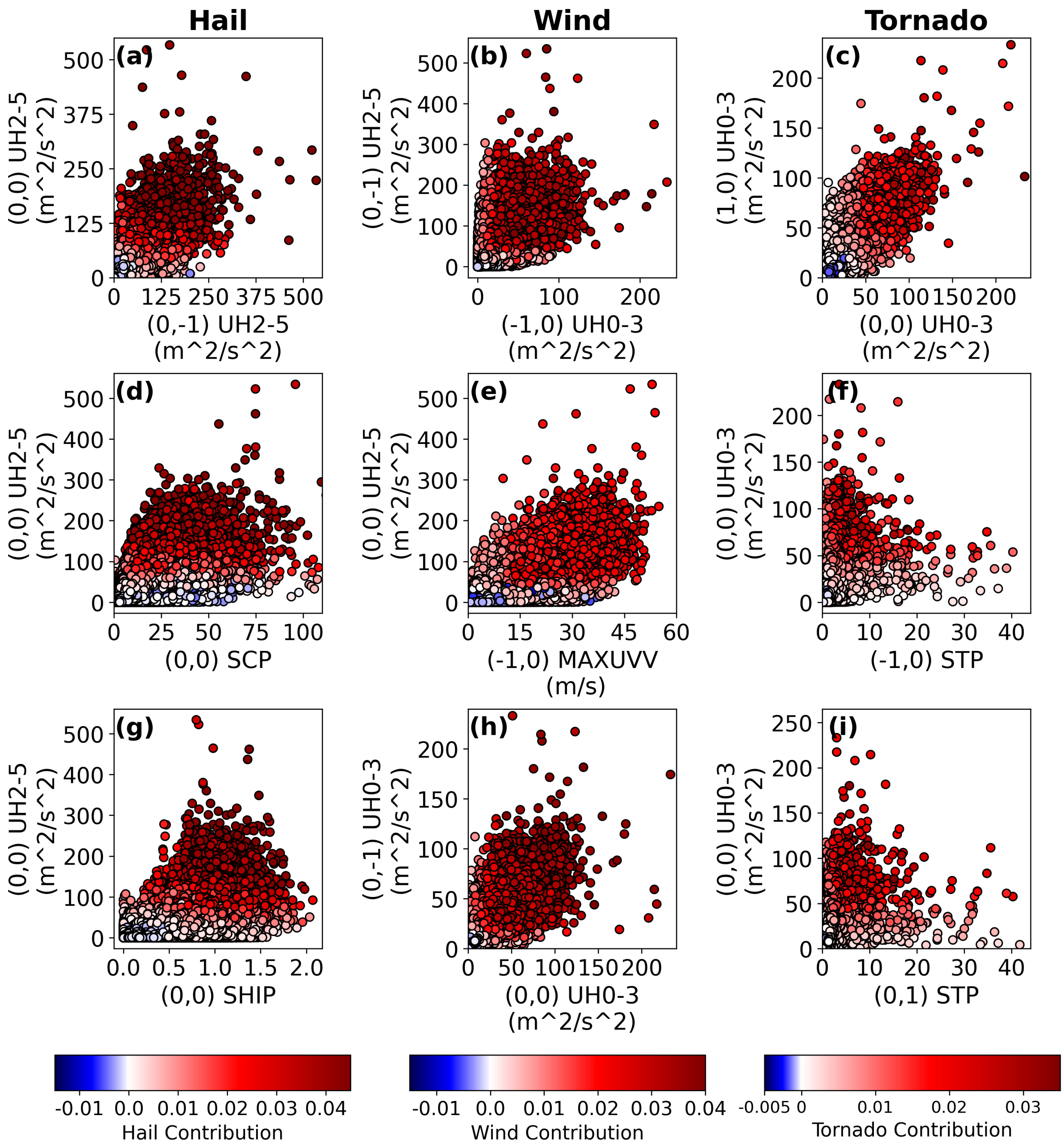


Figure 4.14 As in Fig. 4.13 but for the most important two-variable combinations in the EM RFs.

4. Case Study: 1200 UTC 23 May – 1200 UTC 24 May

To illustrate how TI can be used to dissect an RF forecast on an individual day, a case study is presented for analysis. 23 May 2020 is selected because it is a representative example that involves all severe weather hazards over multiple areas.

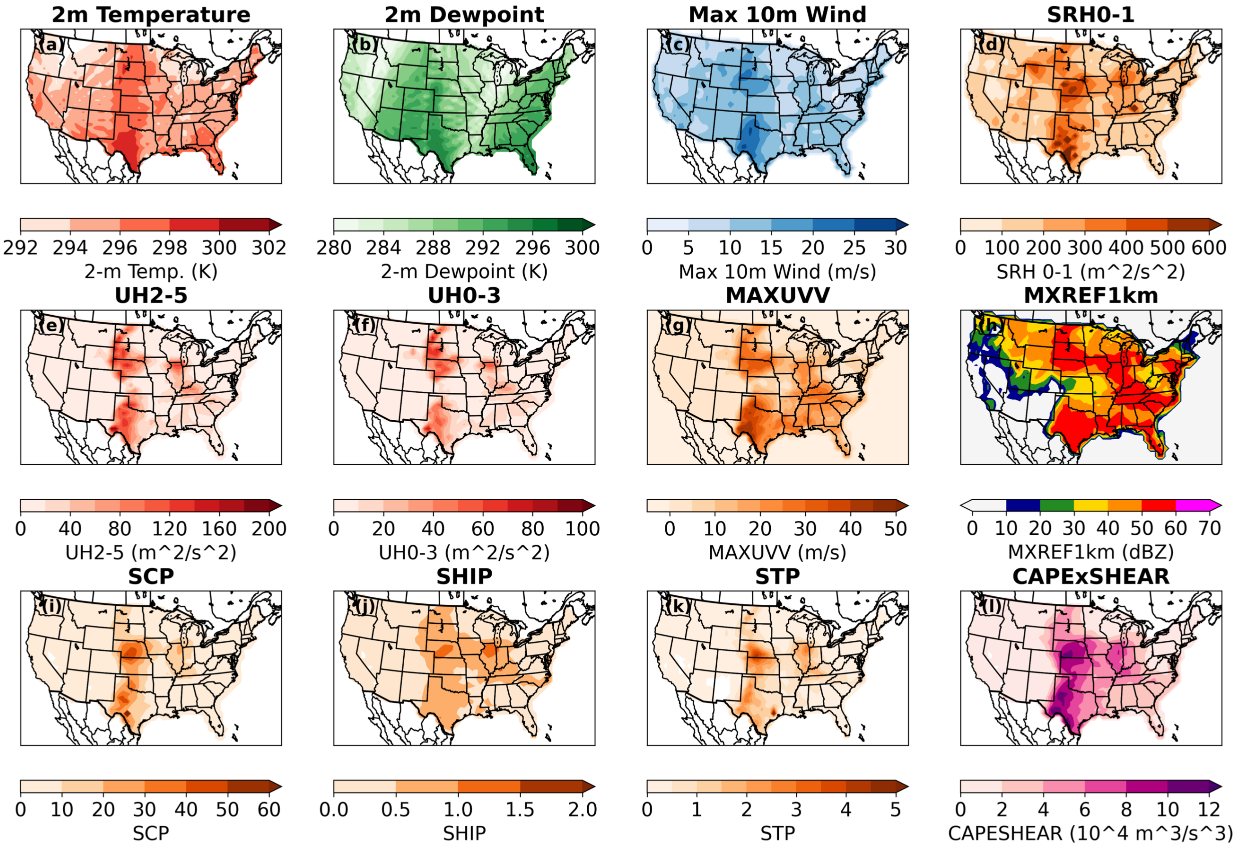


Figure 4.15 Preprocessed (9-member) ensemble mean fields for (a) 2-m temperature, (b) 2-m dewpoint temperature, (c) maximum 10-m wind speed, (d) 0-1km storm relative helicity, (e) 2-5km updraft helicity, (f) 0-3km updraft helicity, (g) maximum vertical velocity, (h) spatially-smoothed daily maximum 1-km simulated reflectivity, (i) SCP, (j) SHIP, (k) STP, and (l) the product of MUCAPE and 10m-500hPa vertical wind shear magnitude, valid 1200 UTC 23 May to 1200 UTC 24 May 2020.

Four main features helped drive the severe weather on this day: a longwave trough in the western CONUS, a mid-level low and associated surface cyclone in the Upper Midwest, a shortwave trough in the South, and a dryline in the Southern High Plains. Fig. 4.15a-l shows some (preprocessed, 9-member) ensemble mean fields from HREFv2.1. The temporal mean 2-m temperature (Fig. 4.15a) and dewpoint temperature (Fig. 4.15b) 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 (55.9 miles per hour) there—and in southwestern South Dakota (Fig. 4.15c). Maximum 0-1km SRH is at least 200 m^2/s^2 over a large swath of the Great Plains and the Upper Midwest (Fig. 4.15d). 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. 4.15e). Relatively large values of UH0-3km (Fig. 4.15e) and MAXUVV (Fig. 4.15f) are found in these same regions, and spatially smoothed simulated reflectivity indicates (simulated) storms over a large portion of the eastern two-thirds of the CONUS (Fig. 4.15h). Important index variables—including SCP (Fig. 4.15i), SHIP (Fig. 4.15j), STP (Fig. 4.15k), and the product of MUCAPE and 10m-500hPa wind shear (Fig. 4.15l)—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. 4.15k).

IM and EM RF probabilities generally highlight three regions for all three hazards: the Upper Great Plains (i.e., North Dakota to Nebraska), the Lower Great Plains (i.e., west Texas and western Oklahoma), and parts of the Midwest near northern Illinois (Fig. 4.16a-f). Additionally, both RFs highlight a severe wind threat farther south, including 30% or 45% probabilities in central Kentucky and a broad 5% probability for most of the Southeast (Fig. 4.16c-d).

The EM and IM forecasts for each hazard are quite similar, highlighting the same general areas. The biggest differences tend to be the probability magnitudes from each RF, with the EM

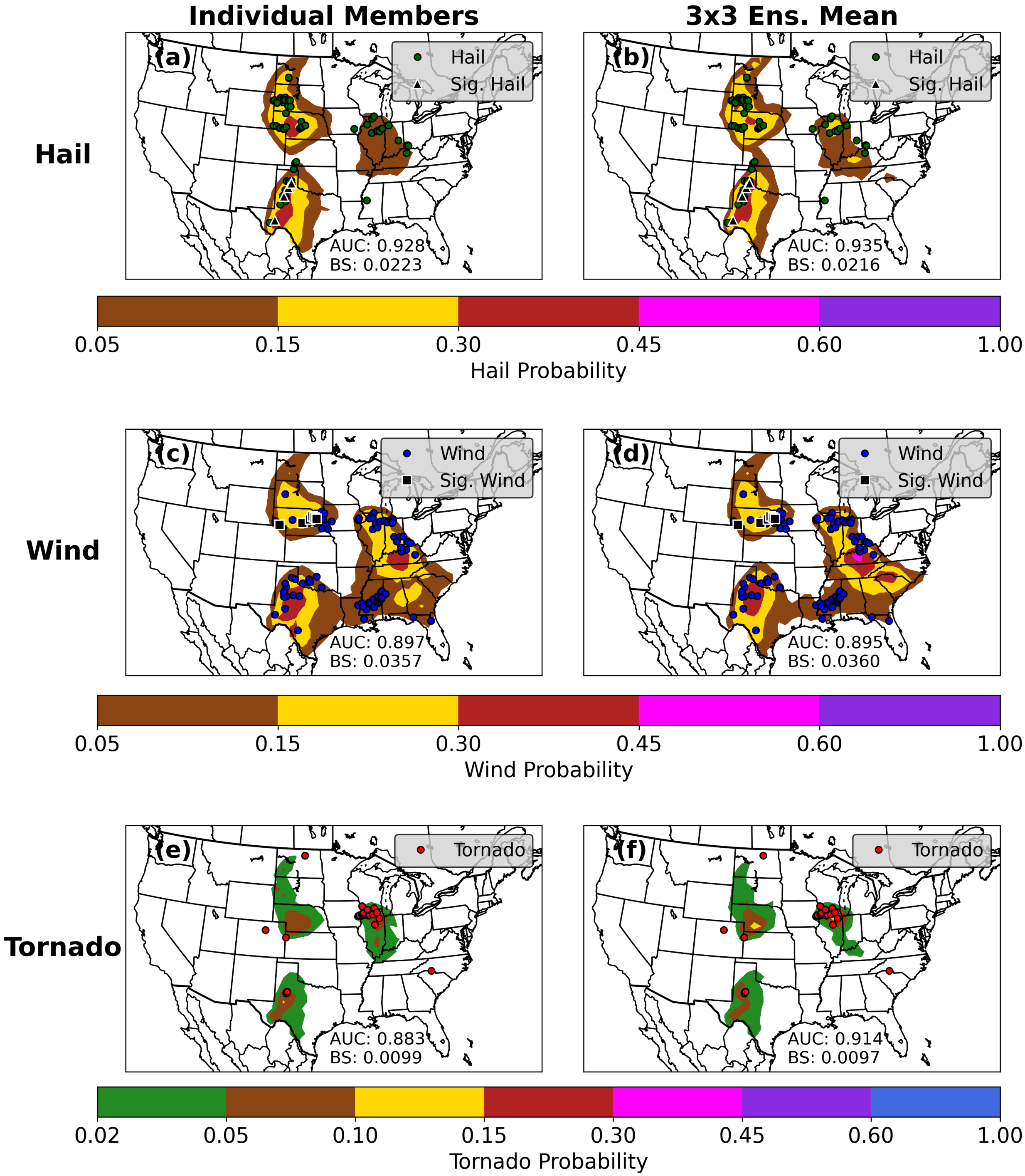


Figure 4.16 (a) Severe hail forecast probability from the IM RF (shaded) and observed sub-significant (green dots) and significant (black triangles) hail reports, valid from 1200 UTC 23 May 2020 to 1200 UTC 24 May 2020. Individual-day AUC and BS are shown at the bottom of the panel. (b) As in (a) but for the EM RF. (c)-(d) As in (a)-(b) but for severe wind forecasts. Observed sub-significant (blue dots) and significant (black squares) are shown. (e)-(f) As in (a)-(b) but for tornado forecasts. Observed tornado reports (red dots) are shown.

usually providing larger probabilities. For example, the EM RF has 15% hail probabilities in northern Illinois and central Kentucky (Fig. 4.16b) compared to 5% probabilities from the IM RF (Fig. 4.16a). 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. 4.16d). 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 (Fig. 4.16c-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 (Fig. 4.16e-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 (Fig. 4.17a-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. 4.3-4.6. Otherwise, similar fields tend to be emphasized, and they tend to impact RF probabilities similarly.

Figs. 4.18-4.19 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 (Fig. 4.18a-c, 4.19a-c), although the storm contribution fields from the EM RFs tend to be less “smoothed,” which makes sense given that—unlike the IM RFs—the EM RFs consider unsmoothed storm predictors from different spatial points. The most obvious difference between the IM and EM RFs is the latitude/longitude contributions for the severe hail and wind forecasts. While the IM RFs have relatively low contributions from latitude/longitude (Fig. 4.18j-k), the EM RFs have large positive

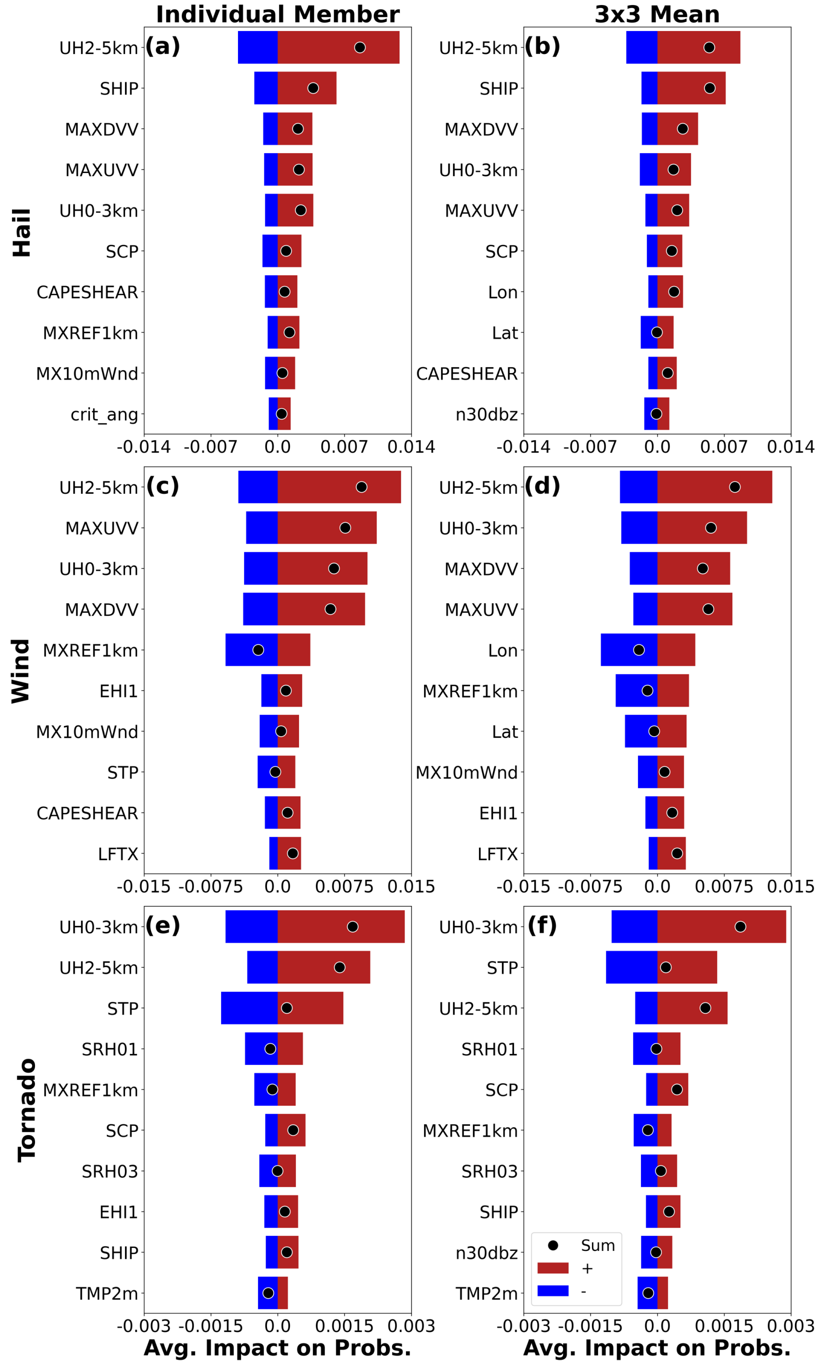


Figure 4.17 (a) Mean TI negative (blue), positive (red), and summed (i.e., negative plus positive; black dot) RF probability contributions (per grid point) from the 10 most important fields (aggregated over individual members) for the all-predictor severe hail IM RF, valid from 1200 UTC 23 May to 1200 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)-(b) but for the severe wind RFs. (e)-(f) As in (a)-(b) but for the tornado RFs.

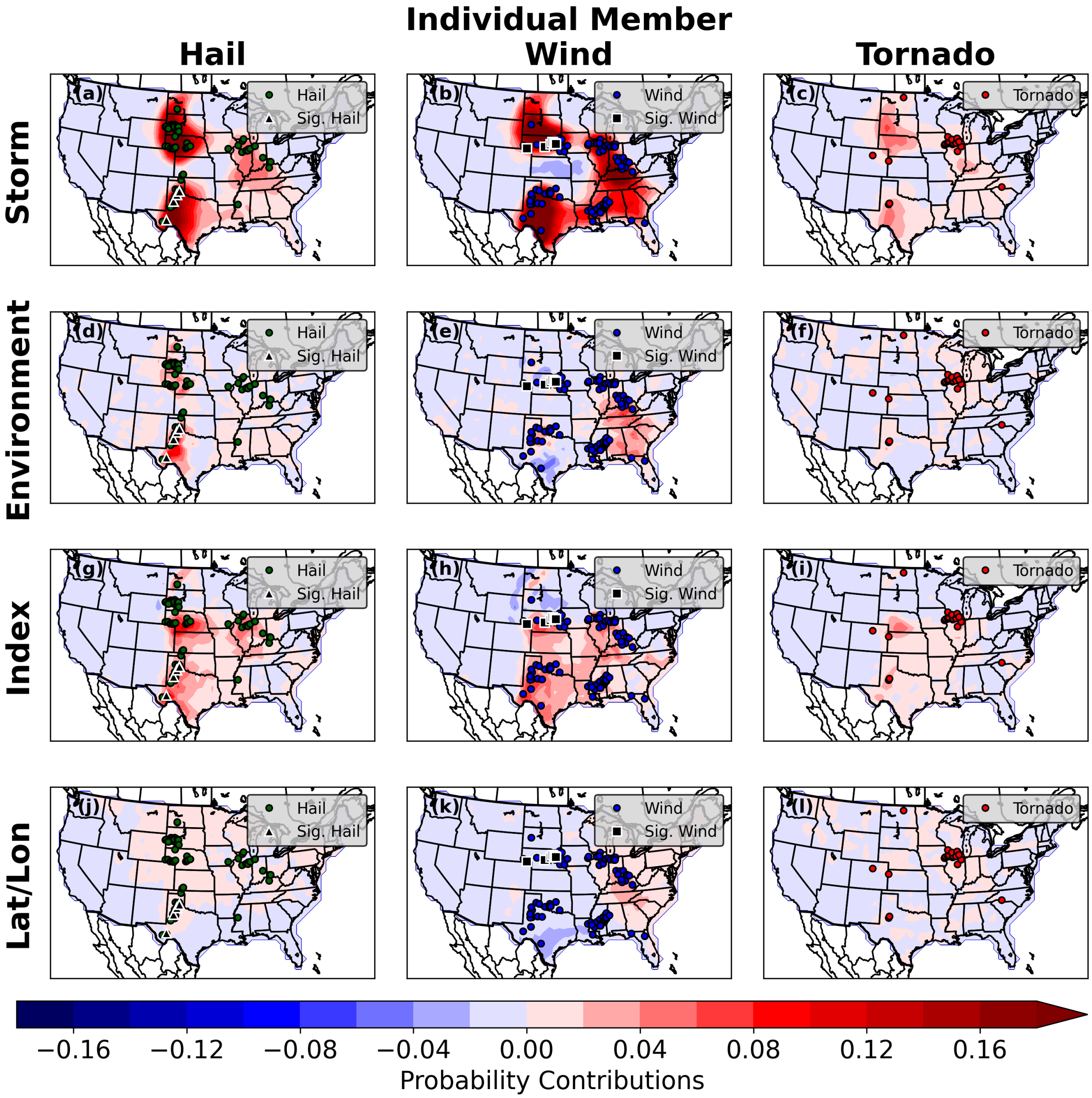


Figure 4.18 (a) Aggregated IM RF probability contributions (shaded) from storm-related variables for severe hail prediction, with observed sub-significant (green dots) and significant (black triangles) hail reports overlaid. (b) As in (a) but for severe wind prediction with observed sub-significant (blue dots) and significant (black squares) overlaid. (c) As in (a) but for tornado prediction with observed sub-significant 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.

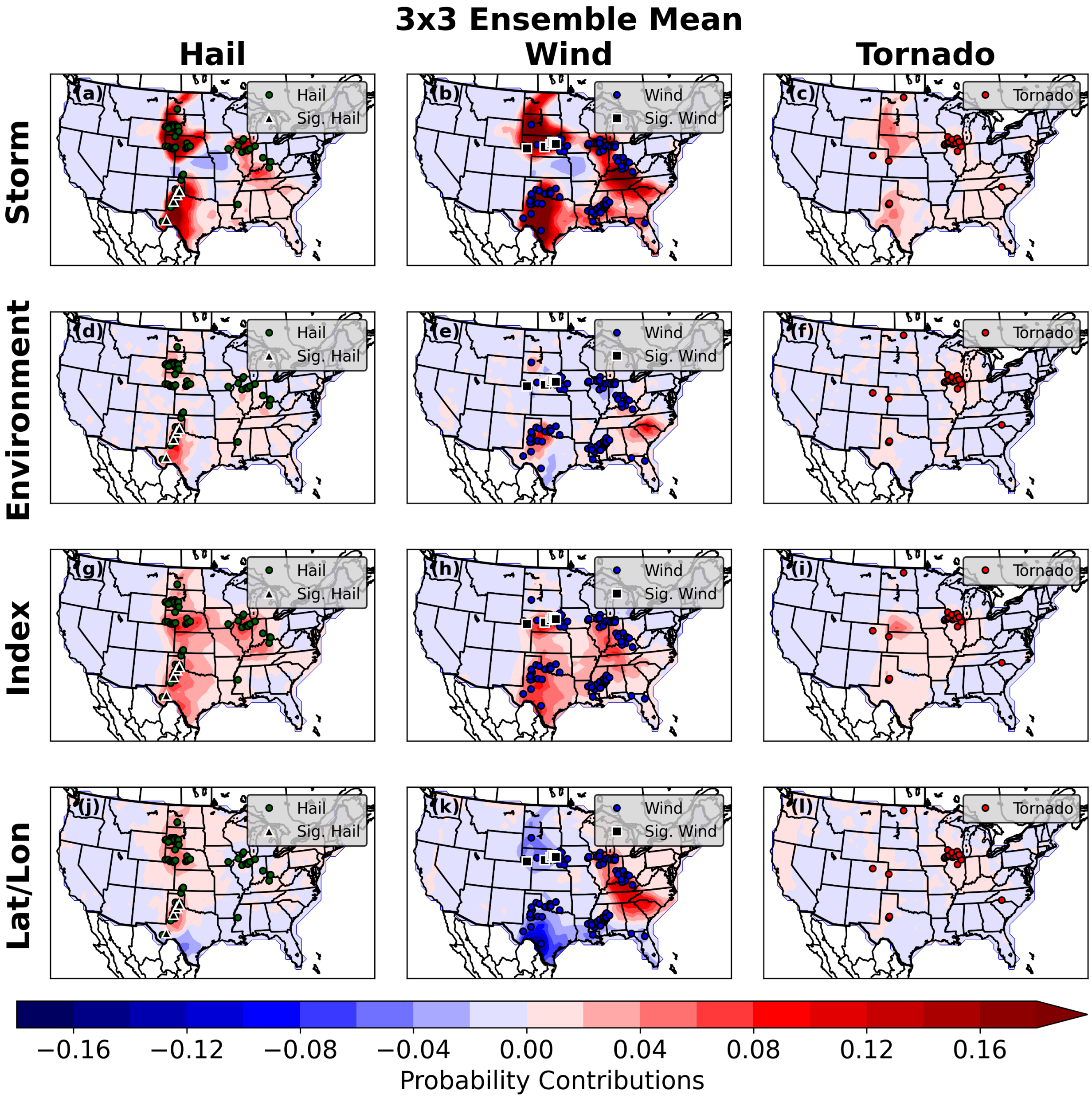


Figure 4.19 As in Fig. 4.18 but for EM RFs.

contributions for severe hail in most of the Great Plains (Fig. 4.19j) and large negative (positive) severe wind contributions in the Great Plains (eastern U.S.) (Fig. 4.19k). These are the spatial patterns seen in the observational dataset (Fig. 4.2b-d).

The largest differences between the EM and IM probabilistic forecasts appear to be driven primarily by the latitude/longitude, environment, and index variables. For example, the EM RF's larger severe wind probabilities in eastern Kentucky is due to greater latitude/longitude and index contributions (Fig. 4.18h,k; Fig. 4.19h,k). Meanwhile, the EM RF's greater severe wind probability in northern South Carolina is due to greater environment and latitude/longitude contributions (Fig. 4.18e,k; Fig. 4.19e,k). The IM and EM RFs have relative similar tornado probability contributions (Fig. 4.18cfil; Fig. 4.19cfil).

5. Summary and Discussion

In this paper, the Python module tree interpreter (TI) was used to assess how differently-configured random forests (RFs) use convection-allowing ensemble (CAE) variables to create skillful severe weather forecasts. Two main configurations of RFs were examined: RFs trained on individual-member 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 32 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 mean fields generally had clearer relationships with RF probability contribution (e.g., Fig. 4.11p-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.

With that said, 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. 4.16). 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. 4.9 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 non-storm variables can provide relatively skillful next-day severe weather forecasts (e.g., as in Hill et al. 2020), the storm fields from convection-allowing ensembles (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 Storm Prediction Center (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, etc.) 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, etc.) 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. Indeed, as ML technology progresses, finding better and more efficient ways to summarize ensemble data during preprocessing will be crucial to obtaining the most skillful RFs.

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 0-3km updraft helicity (UH0-3km) were top tornado predictors. Additionally, TI analysis found that the 2-5km updraft helicity (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, 2016b, 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, etc.), 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 and why the different members’ UH2-5km forecasts have different relationships with severe hail and wind probabilities. As RF and ML tools are applied to more prediction tasks, investigating how the importance of different predictors varies at different lead times and spatial scales will also be important, since this type of analysis should enhance our understanding of severe weather predictability. Indeed, the results presented here (i.e., the strong importance of the storm fields) raise the question of whether the storm fields (and CAEs themselves) might still provide substantial value at longer than 36-h lead times. Certainly, such a question merits further consideration as computing resources continue to increase. Finally, 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.

Chapter 5: General Conclusion

1. General discussion of research hypotheses

Since the state of the atmosphere can never be perfectly described or modeled, high-impact weather events—including floods, tornadoes, severe wind, and severe hail—are inherently uncertain (e.g., Palmer 2017). Accurately quantifying that uncertainty is important for facilitating optimal decision making and more effective weather warnings (e.g., Palmer 2017; Rothfusz et al. 2018; NOAA 2020). Traditionally, ensembles have been used to quantify uncertainty (e.g., Palmer 2017). However, ensembles frequently suffer from under-dispersion and suboptimal reliability (e.g., Romine et al. 2014; Schwartz et al. 2014; Loken et al. 2019b), biases in the magnitude and placement of precipitation and convective systems (Herman and Schumacher 2016), and coarse grid-spacing relative to the hazards predicted (e.g., for severe weather). Thus, ensembles are useful but imperfect prediction tools that require additional methods to quantify high-impact weather uncertainty most effectively.

Recently, machine learning (ML) techniques have emerged as a promising means to quantify ensemble uncertainty. While ML technology itself is not new, better computing resources have recently made the application of ML to meteorological problems much more effective (e.g., Schultz et al. 2021). RFs, in particular, have shown great promise in improving ensemble forecasts for both precipitation (e.g., Gagne et al. 2014; Herman and Schumacher 2018c) and severe weather (e.g., Gagne et al. 2017; Hill et al. 2020). However, much remains unknown regarding the use of ML to improve numerical weather prediction (NWP) models and ensembles. For example, previous research has done little to investigate how RFs benefit different types of ensembles (e.g., convection-parameterizing vs. convection-allowing), how RF forecasts compare to corresponding human (and top-performing non-human) forecasts, what preprocessing methods work best for generating ensemble-based predictors, and how RF forecasts use ensemble data to create skillful forecasts. This dissertation designed three research components to fill these knowledge gaps by: 1) developing probabilistic precipitation and severe weather hazard RFs, 2) comparing those RFs to top-performing human and other non-ML baselines, 3) investigating different strategies for generating predictors from CAE data, and 4) interpreting the relationships learned by RFs.

The first research component developed and evaluated precipitation-predicting RFs based on SREF and HREFv2 forecast variables. The hypotheses associated with the first research component were as follows:

> *H1.1: RF-based probabilistic precipitation forecasts will have reduced spatial biases as well as better discrimination ability, sharpness, and resolution compared to spatially-smoothed ensemble probabilities. RF probabilities will provide the greatest benefits relative to spatially smoothed ensemble probabilities at the smallest thresholds, which are climatologically most common.*

> *H1.2: RF post-processing will benefit a convection-parameterizing ensemble more than a CAE due to the greater initial bias of the convection-parameterizing ensemble. Indeed, after RF post-processing, a convection-parameterizing ensemble will have better reliability and nearly comparable resolution compared to an un-post-processed (i.e., raw) CAE forecast. However, post-processed CAE forecasts (from either the RF or spatial smoothing method) will be the most skillful due to the enhancement of CAE*

reliability and resolution. In accordance with H1.1, RF CAE forecasts will be more skillful than spatially smoothed CAE forecasts.

H1.3: Approximately one year of training data will be required to obtain skillful RF-based precipitation forecasts for the 3-inch forecasts, with less data required as the threshold decreases.

The work done in Chapter 2 was among the first to systematically compare RF-based post-processing to a skillful non-ML post-processing technique (i.e., spatially smoothing raw ensemble probabilities), as most previous work (e.g., Gagne et al. 2014; Herman and Schumacher 2018c) only compared ML techniques to raw ensemble probabilities. Spatially smoothing raw ensemble probabilities is an effective and commonly-used method for improving reliability and discrimination ability, provided the proper degree of smoothing is used (e.g., Loken et al. 2019b; Roberts et al. 2019). However, one negative attribute of spatial smoothing—particularly isotropic smoothing—is that it maintains the general shape of the raw ensemble probabilities and thus has a limited ability to correct for spatial biases. RFs, on the other hand, compute new forecast probabilities at each point based on each point's unique set of predictors. Therefore, I hypothesized, correctly, that RFs would be able to reduce ensemble spatial biases better than post-processing by spatial smoothing (H1.1). Specifically, I found that the RF forecasts moved the center of conditional distribution of observed "yes" events (i.e., observed precipitation exceeding a threshold) closer to the "yes" forecast point, using a method outlined by Clark et al. (2010a) and Marsh et al. (2012). I also found, as hypothesized in H1.1, that RFs had greater AUC, BSS, critical success index (CSI), and better BS resolution and reliability values compared to spatially-smoothed forecasts. This result made sense because, while spatially smoothed forecasts can be calibrated to have near-perfect reliability, smoothing necessarily sacrifices sharpness and assumes a specific (e.g., isotropic) distribution of forecast uncertainty. An advantage of RFs is that they tend to maintain excellent reliability (e.g., Breiman 2001) but can also retain higher forecast sharpness and resolution compared to the spatial smoothing post-processing method, as I found in Chapter 2.

As hypothesized in H1.1, I found that RF forecasts had the greatest increase in BSS relative to the spatially smoothed forecasts at the smallest precipitation thresholds. This finding made sense because lighter precipitation events are much more common (and therefore better represented in the training data). Additionally, heavier precipitation events may simply be more difficult to predict since they have a greater dependence on mesoscale (or smaller) features, which have less predictability than synoptic features (since errors saturate at the small scales first; e.g., Zhang et al. 2007; Greybush et al. 2017). Overall, the result suggests that, while RFs are useful post-processing tools, they are most helpful—at least relative to other post-processing methods—for more common or routine forecasting situations. Such a result gives support to the current practice of the Weather Prediction Center to rely most on RF-based guidance for common, low-impact events (Novak 2021).

While previous studies have shown that RFs can be used to obtain skillful precipitation forecasts from CAEs (e.g., Gagne et al. 2014) and global ensembles (Herman and Schumacher 2018c), Chapter 2 was among the first to systematically compare how the same RF-based post-processing procedure benefits a CAE and similar convection-parameterizing ensemble. This was an important study because it was unknown, for example, if a convection-parameterizing ensemble could achieve CAE-caliber skill with RF-post-processing (or by how much a CAE would retain its skill advantage after RF-post-processing).

Due to the extensive biases in the convection-parameterizing SREF (e.g., Eckel and Mass 2005), I hypothesized that RF post-processing would benefit the SREF more than the HREFv2 (H1.2). I also expected RF-based SREF forecasts to have better reliability and similar resolution compared to the raw HREFv2 forecasts (H1.2). My work largely supported H1.2. I found that RF post-processing almost always increased the BSS more (relative to raw [i.e., un-post-processed] and spatially smoothed ensemble probabilities) in the SREF compared to the HREFv2. I also found that SREF RF forecasts had similar BSSs and BS resolution and better AUC and reliability compared to raw HREFv2 forecasts, owing to the RF's ability to forecast continuous probabilities. One implication of these findings is that RFs may be particularly valuable when global ensemble data is the only NWP tool available, such as for long lead time forecasts. With that said, I also found that post-processed (i.e., spatially smoothed and RF) HREFv2 forecasts had better AUC, BSS, and resolution than the SREF RF forecasts (but similar near-perfect reliability). This made sense, given that the skill of an RF is indelibly linked to the skill of the underlying dynamical ensemble (e.g., Gagne et al. 2014) and CAEs are more skillful than convection-parameterizing ensembles for precipitation (e.g., Clark et al. 2009). Thus, CAE data should be used, if available, to achieve the greatest forecast skill.

Because frequent updates to NWP models often precludes the existence of long (i.e., multi-year), stationary ensemble data archives, it is important to know how much data is sufficient for the creation of RF-based forecasts. For precipitation forecasting, I expected that approximately one season (or about 3 months) of data would be needed to provide the RFs with sufficient examples of heavy precipitation events (H1.3). Since lighter precipitation events occur more frequently, I hypothesized that less training data would be required to skillfully predict those events (H1.3), especially since Gagne et al. (2014) obtained skillful 0.1-, 0.25-, and 0.5-inch (i.e., 2.54-, 6.35-, and 12.7-mm) forecasts using only 34 days of data. My work largely supported H1.3. While all RF forecasts had positive BSSs with only 31 days of training data, I found that AUC and BSS improved dramatically when the training set was extended to 93 days (about one season). AUC and BSS increased more gradually with more training data up to a training length of 217 days (about 7 months). Interestingly, I found that all precipitation thresholds required about 93-217 days of training data to perform optimally, although the larger thresholds experienced greater benefits from increasing the training dataset length from 31 to 93 days. Overall, these results suggest that the generally limited data archives are not likely to pose a problem for operational RF implementation.

In the second research component, RFs were implemented for severe weather prediction. The primary hypotheses associated with the second component were:

H2.1: For all severe and significant severe weather hazards (including any-severe and any-significant-severe categories), RFs will have better discrimination ability, BSS, reliability, and resolution than corresponding calibrated UH2-5km-based forecasts. However, RFs will have worse discrimination ability, BSS, and resolution than corresponding (discrete and continuous) SPC human forecasts. Continuous RFs will have better reliability and resolution than discrete (i.e., binary) SPC significant severe hazard forecasts, but discrete RF forecasts will not perform better than discrete significant severe SPC forecasts.

H2.2: The RF forecasts will perform best in the seasons and locations for which severe weather climatological frequency is maximized. For tornadoes and severe hail, this is expected to be the central U.S. during the spring and summer. For severe wind, this is expected to be the eastern U.S. during the summer.

The work done in Chapter 3 was among the first to systematically compare RF-based forecasts to corresponding calibrated UH2-5km and SPC human forecasts. I hypothesized that severe-weather-predicting RFs would outperform hazard-calibrated, spatially-smoothed UH2-5km forecasts (H2.1) due to RFs' ability to consider multiple relevant variables and their expected ability to implicitly account for spatiotemporally-varying UH2-5km climatology (Sobash and Kain 2017). I hypothesized that RFs would not outperform human forecasts for most hazards—at least in terms of discrimination ability, BSS, and resolution—since human forecasters had access to additional information not considered by the RFs, including radar, satellite, and sounding observational data. Since the SPC only forecasts significant severe hazards at a single probability level (10%), I expected continuous RF forecasts to have better reliability and resolution than discrete SPC significant severe forecasts. However, I expected any RF forecast benefits to disappear when RF probabilities were discretized in accordance with the SPC probabilities (H2.1).

My work mostly supported the first part of H2.1: I found that RF severe weather forecasts had substantially greater BSSs than spatially-smoothed UH2-5km forecasts for all hazards except for significant tornado and significant severe wind. I also found that, compared to calibrated UH2-5km forecasts, RF forecasts had better resolution for every hazard, and they had better reliability for every hazard except for severe wind. These results are likely at least partially due to the UH2-5km forecasts being calibrated over the full year and full domain rather than by region and season, as Sobash and Kain (2017) suggest is optimal. However, current operational post-processing techniques tend to use a static UH2-5km threshold for a given model or ensemble member (e.g., Roberts et al. 2019). Thus, the above findings suggest that the RF method developed in this dissertation substantially outperforms one of the current operational standards for automated severe hazard guidance.

Very surprisingly, the second part of H2.1 was refuted. I found that RF hazard forecasts had greater—and sometimes substantially greater—BSSs than human SPC forecasts for most hazards and most locations and seasons, even when RF probabilities were discretized or SPC probabilities were made continuous. The third part of H2.1 was supported, however: I found that continuous—but not discrete—RF significant severe forecasts had substantially greater BSSs and better BS components compared to corresponding SPC forecasts.

Collectively, these results suggest that the RFs introduced in Chapter 3 could substantially improve existing SPC day-1 human forecasts. Indeed, the results from Chapter 3 nicely complement another study, Hill et al. (2020; published around the same time), who found that RF probabilistic severe weather forecasts based on *convection-parameterizing* predictors outperformed SPC human forecasts at 2- and 3-day but not 1-day lead times. While artificial intelligence (AI) techniques have achieved super-human performance in other domains [e.g., chess (Silver et al. 2018); go (Silver et al. 2016, 2017, 2018); and no-limit Texas Hold'em poker (Brown and Sandholm 2019)], the study in Chapter 3 and Hill et al. (2020) are among the first to suggest current RF-based techniques are capable of matching or outperforming human severe weather forecasters at day 1-3 lead times.

Based on the results from Chapter 2, the severe-weather-predicting RFs were not expected to perform uniformly for all hazards in all regions and seasons. Rather, I expected the RFs to perform best for the most frequent severe weather events (H2.2). Thus, skill was expected to be largest for tornadoes (SPC 2021d) and severe hail (SPC 2021c) in the Midwest (Fig. 3.2) in the spring and summer and for severe wind in the East (Fig. 3.2) in the summer (SPC 2021f). These expectations were consistent with Hitchens et al. (2016), who noted that SPC outlooks produced from UH2-5km were most skillful in the spring and summer. Chapter 3 partially supported H2.2: I found that severe and significant severe RF hail skill was maximized in the Midwest during the spring but that RF tornado and severe wind BSS values were maximized in the eastern U.S. during the winter. With that said, I found that, relative to SPC and UH2-5km forecasts, the RF severe wind forecasts performed best during the summertime in the eastern U.S. Thus, RF severe hail and wind forecasts may be most helpful to forecasters in the spring and summer in the Midwest (for hail) or East (for wind), as expected.

In the third research component, differently-configured severe-weather-predicting RFs were compared and analyzed to determine how RFs use CAE data to produce skillful forecasts. The hypotheses associated with the third component are:

H3.1: Greater forecast skill will result from providing an RF with individual member predictors at a single grid point.

H3.2: RFs will emphasize storm variables, but index and environment variables will also be important since simulated storms (and their attributes) do not always correspond with observed storms.

H3.3: RFs will learn to emphasize different variables for each hazard (e.g., significant hail parameter [SHIP; SPC2021b] and UH2-5km for severe hail; significant tornado parameter [STP; Thompson et al. 2012] and 0-3 km updraft helicity [UH0-3km] for tornadoes). For all hazards, RFs will learn positive—but nonlinear—relationships between many storm variables (e.g., UH2-5km, simulated reflectivity, maximum upward vertical velocity, etc.) and observed severe weather probability. Indeed, it is hypothesized that many of these variables will have an "S-shaped" relationship with severe weather probability. However, RFs are also expected to learn (and use) important relationships between multiple variables/predictors and observed severe weather.

One important question addressed by Chapter 4 is how RF systems based on CAE data should be designed. Especially for an "ensemble of opportunity" (e.g., Roberts et al. 2020), it was unclear whether single-point individual member predictors or multi-point ensemble mean predictors should be used. Given the high diversity of the HREFv2.1 (Roberts et al. 2020), it was hypothesized that the RFs would produce better forecasts by learning and exploiting the systematic biases of each member individually (H3.1). I expected that spatially smoothing the storm fields would sufficiently account for the spatial uncertainty of simulated convection, making the consideration of HREFv2.1 data at multiple spatial points unnecessary. However, H3.1 was refuted; I found that the ensemble mean RFs achieved greater forecast skill for all hazards, since ensemble mean predictors had stronger relationships with observed severe weather compared to those from any individual member. This result is important because it suggests that, even for RFs, one of the best ways to utilize ensemble information is with a simple ensemble mean. It also implies that, while ML techniques can generate probabilistic forecast guidance with just a single member, dynamical NWP ensembles are still important to run, since ensemble mean fields make better ML predictors than individual member fields. With that said, individual member RFs were only marginally worse than ensemble mean RFs, and I noted that the RFs using individual member predictors can still be beneficial by potentially learning and communicating (through interpretability techniques) deficiencies in individual ensemble members, which can aid model developers.

An important feature of CAMs and CAEs is their ability to explicitly simulate storms and associated storm-related variables. Indeed, past studies have shown the usefulness of hourly-maximum storm fields [e.g., UH2-5km (Kain et al. 2008, 2010; Sobash et al. 2011, 2016b, 2019) and simulated vertical velocity and reflectivity (Kain et al. 2010; Roberts et al. 2019)] for severe weather forecasting. Thus, I expected RFs to learn to emphasize storm fields. However, I also expected index and environment variables to be important, since simulated storms don't always correspond with observed storms (H3.2). H3.2 was mostly supported. I found that RFs run with only storm predictors achieved the greatest skill compared to RFs run with only index and only environment predictors, respectively. Further, using Tree Interpreter (TI), I found that the storm variables were among the most important predictors for all three hazards. However, with TI, I also showed that relevant index predictors (e.g., STP for tornado prediction) were also important and helped modulate the severe probability, particularly when top storm variables (e.g., UH2-5km) were marginal. Meanwhile, I found that environment variables were least important and least skillful, likely since they were obtained from temporal averaging and—individually—had a weaker association with observed severe weather compared to index variables. These findings showed the power of using complex predictors with a pre-determined association with severe weather, when available, to facilitate easier RF learning. From a severe weather forecasting perspective, these results suggested that forecasters should emphasize simulated storm fields but also consider the favorability of the simulated environment when constructing forecast hazard probabilities.

Given the high degree of skill achieved by the severe-weather-predicting RFs in Chapter 3, a main question driving the research in Chapter 4 was how, exactly, RFs learn to forecast severe weather. Given the different physical mechanisms required to produce the different hazards, I hypothesized that RFs would emphasize different, but physically-relevant, variables for each hazard (H3.3). (This hypothesis is not as trivial as it may first appear, given that UH2-5km has been effectively used to forecast all three hazards [e.g., Jirak et al. 2014; Gallo et al. 2016].) I expected RFs to learn positive relationships with many simulated storm variables that have been associated with observed severe weather [e.g., UH2-5km and UH0-3km (Sobash et al. 2016a, 2019); and maximum upward vertical velocity and simulated reflectivity (Kain et al. 2010; Roberts et al. 2019)]. Given the thresholding approach commonly used to infer simulated severe weather reports (e.g., Sobash et al. 2011, 2016b, 2019; Gallo et al. 2016; Loken et al. 2017, 2020; etc.), I expected RFs to learn an "S-shaped" relationship between many storm variables and observed severe weather probability (H3.3). I also expected RFs to learn complex relationships between observed severe weather and combinations of multiple predictors (H3.3), since RFs significantly outperformed UH2-5km-only forecasts in Chapter 3 and since Gallo et al. (2016) improved UH2-5km-based tornado forecasts by accounting for the simulated environment.

My work mostly supported H3.3. First, I found that the RFs indeed emphasized different variables for different hazards. For example, top predictors were UH0-3km and STP for tornadoes; UH2-5km and SHIP for severe hail; and UH2-5km, UH0-3km, and maximum upward velocity for severe wind. I also found, as hypothesized, that RFs learned positive relationships between observed severe weather and most storm and index variables (e.g., UH2-5km, UH0-3km, maximum vertical velocity, STP, SHIP, simulated reflectivity) for each hazard. Many of these relationships were nonlinear, but only some (e.g., UH2-5km and SHIP for severe hail, maximum vertical velocity for severe hail and wind) were S-shaped. The learned S-shaped relationships between UH2-5km and severe hail and wind probability are consistent with the use of UH2-5km thresholds to forecast severe weather (e.g., Sobash et al. 2011, 2016b, 2019; Gallo et al. 2016; Loken et al. 2020). However, the continuous nature of the RFs' learned S-shaped relationships suggests that the commonly-used thresholding process is oversimplified, since not all values above (or below) the threshold are equivalent.

Not all learned relationships were intuitive or easy to explain. For example, I found that larger values of (spatially-smoothed, daily maximum) simulated reflectivity did not substantially increase severe hail probabilities but increased severe wind probabilities more notably. Why this was the case will have to be explored in future work.

Somewhat surprisingly, I found that single variables were generally much more important than multi-variable predictor combinations, although multivariate relationships were not negligible. I found that the most important multivariate relationships involved two storm fields or one storm and one index field. Specifically, I found that RFs calibrated their probabilities by considering the precise value of certain multi-variable combinations at the same and/or neighboring grid points. For example, with only a modest value of UH2-5km forecast at the point of prediction, RFs still learned to increase severe hazard probability if UH2-5km was large at a neighboring grid point or if a relevant index variable (or another storm variable) was large at the point of prediction. This is an important finding because it suggests that the RFs learned to implicitly account for model error.

2. Summary and discussion of key lessons learned

This dissertation enhanced our understanding of how to construct RFs for next-day precipitation and severe weather prediction, how RFs compare to other top-performing post-processing techniques and human forecasts, and how RFs use ensemble data to produce skillful severe weather forecasts. Specifically, we have learned the following:

1. For next-day precipitation and severe weather forecasts, a simple but effective way of preprocessing hourly CAE data is to take a 24-h minimum/maximum/standard deviation, upscale to a coarser grid, and use grid-point-based ensemble mean fields. Temporally aggregating 24-h data helps account for model uncertainties in time, while upscaling helps account for model uncertainties in space. Grid-point-based predictors allow for skillful predictions even when simulated and observed storms do not perfectly correspond.

Previous studies (e.g., Gagne et al. 2014; Herman and Schumacher 2018c; Hill et al. 2020) have used point-based ensemble data at multiple forecast hours as RF predictors. While such a method has demonstrated skill, it assumes perfect ensemble member timing information. In contrast, temporally aggregating ensemble data uses less predictors and does not require modeled convection or environments to be perfectly simulated in time. Further, temporal maximum storm variables (e.g., UH2-5km) have repeatedly shown a strong association with observed severe weather (e.g., Kain et al. 2010; Sobash et al. 2011). Spatial upscaling further reduces the number of predictors and does not penalize slight spatial errors of ensemble members' simulated convection. Indeed, upscaling makes it computationally feasible to use grid-point-based predictors, which are advantageous because they do not require perfect correspondence between observed and simulated storms and they allow for the easy creation of 2-dimensional output probabilities.

> *2. RF-based post-processing benefits convection-parameterizing ensembles and lower precipitation thresholds more than CAEs and higher precipitation thresholds, respectively. Convection-parameterizing ensembles receive more benefit than CAEs because they have more initial biases, while the lower precipitation thresholds are climatologically more common and thus provide more training examples from which the RF can learn.*

This finding has several implications for how RFs are used operationally. First, it suggests that RFs may be most helpful to forecasters in the most common/routine forecasting situations and may (comparatively) struggle with producing calibrated guidance for rare events. Since the rarest forecasting situations (e.g., extremely high-end precipitation, EF-5 tornadoes, etc.) also tend to be the most impactful, this finding suggests that RFs (human forecasters) may provide the least (most) value in the most high-end situations. Indeed, the Weather Prediction Center (WPC) is already beginning to view the relationship between human forecaster and RF as dynamic depending on the nature of the expected event (e.g., Novak et al. 2021). With that said, RFs are still skillful even for higher-end events and more work is needed to determine exactly when RFs add most and least value to human forecasts (see the Recommendations for future research section below).

The above finding also suggests that RFs may be particularly useful in situations when CAM or CAE data is unavailable, such as when forecasting at multi-day lead times using a global ensemble (e.g., as in Hill et al. 2020). Of course, RFs still provide considerable value to CAE forecasts, especially in the domain of severe weather forecasting (see point 4 below), since severe weather hazards are not explicitly simulated by the CAE.

3. Only approximately one season of training data is necessary to obtain skillful and useful RF precipitation forecasts, even for relatively "rare" events, such as 3-inch-or-greater precipitation.

This is an important finding that has implications for the use of RFs in operations. Since dynamical models and ensembles frequently undergo updates, long multi-year data archives for any given configuration are rare. Fortunately, the results in this dissertation suggest that they are not strictly needed—at least for precipitation—since only a relatively modest amount of training data is required to obtain skillful RF forecasts.

4. Automated next-day RF severe weather forecasts are better than corresponding calibrated-UH forecasts and as good or better than SPC human forecasts for all hazards. For the significant severe hazards, most of this benefit comes from RFs' ability to forecast continuous probabilities below 10%. However, RFs still perform well even when their probabilities are discretized to match the probabilities used by the SPC and when *continuous SPC probabilities are compared against continuous RF probabilities. Severe-weather-predicting RFs perform best for wind and hail in the warm season in the central and eastern U.S. and struggle more with tornado prediction, especially in the west, where tornado reports are much rarer.*

This finding is important because it suggests that, for next-day severe weather forecasts, RFs not only outperform a top non-ML method, but also human-generated forecasts in many situations. Indeed, the work in Chapter 3 is among the first to equitably compare RF and SPC forecasts, and it shows that RFs have begun to achieve super-human performance in some next-day severe weather forecasting situations (e.g., warm season severe hail and wind prediction). The work in Chapter 3 also suggests that RFs could improve the utility of the SPC's next-day significant severe hazard forecasts by facilitating the creation of continuous probabilities—which the SPC currently does not provide. At the same time, Chapter 3 suggests that RFs have a much more difficult time with predicting tornadoes compared to severe wind and hail; thus, more sophisticated methods are likely required to achieve clearly superior RF tornado forecasts (see Recommendations for future research below).

5. Ensemble mean predictors are slightly more skillful than individual member predictors, even for an "ensemble of opportunity" whose members have different climatologies and biases. The reason is that, for a given forecast field, the ensemble mean has a clearer relationship with observed severe weather than any individual ensemble member forecast.

The implication of this finding is straightforward: to maximize RF skill, RFs should be trained using ensemble mean variables as opposed to predictors from individual ensemble members. However, the caveat is that RFs using individual member predictors can provide useful information about the importance of different members' raw forecast fields, which can potentially alert model developers to specific members' strengths and deficiencies.

6. RFs emphasize different, but physically-relevant, fields for each hazard. Overall, storm predictors are found to be most important, followed by index and then environment predictors. However, the most skillful RFs consider both storm and index predictors.

The work in Chapter 4 suggests that RFs emphasize similar variables as human forecasters to make their predictions. For tornadoes, the most important RF predictors are UH0-3km, UH2-5km, STP, and 0-1km SRH. For severe hail, the RF emphasizes SHIP, UH2-5km, UH0-3km, and upward and downward vertical velocity, and for severe wind prediction, the RF relies most on UH2-5km, UH0-3km, upward and downward vertical velocity, simulated reflectivity, and longitude. For the most part, these predictors are intuitive, as many of these variables have been independently shown to be skillful severe weather hazard predictors. The result is an important one, then, not because it necessarily identifies new severe weather predictors, but because it fosters trust that the RFs are primarily relying on ensemble data that is known to be important for severe weather prediction.

With that said, the RFs also provide some useful insights into the forecasting process. For example, they suggest that upward and downward vertical velocity are best used to enhance probability of detection of severe wind and hail, and they hint that other variables (e.g., UH2-

5km and upward and downward vertical velocity) are more useful for severe wind forecasting than simulated maximum wind speed itself. It is also interesting that, overall, RFs learn to emphasize the storm variables but also learn to assign non-negligible "weight" (or relative consideration) to index variables. This suggests that RFs place a high degree of trust in the simulated storms but also recognize that these simulated storms do not always adequately reflect reality.

> *7. Severe-weather-predicting RFs learn to use predictors in complex but intuitive ways. For each hazard, they emphasize storm and index predictors at the point of prediction but also learn to consider other relevant variables at nearby grid points. Importantly, rather than requiring UH2-5km to exceed a static threshold, RFs calibrate their probabilities by considering the precise value of UH2-5km (and other storm predictors) in the context of other variables at the point of prediction and neighboring points. For example, the same marginal value of UH2-5km can help increase RF hail probabilities if the significant hail parameter (SHIP) is large (e.g., greater than 2) or decrease RF hail probabilities if SHIP is near 0. Similarly, UH2-5km can increase hail probabilities at a point if forecast UH2-5km at that point is 0 if the UH2-5km at neighboring grid points is relatively large. This suggests that RFs, like human forecasters, learn to account for model errors when formulating their forecast probabilities.*

This finding is important because it demonstrates that RFs use ensemble data in a similar manner as human forecasters—with the understanding that the underlying dynamical model is imperfect and that the precise values of ensemble forecast variables matter. Other, non-ML automated severe weather guidance products (e.g., Sobash et al. 2011, 2016b, 2019; Gallo et al. 2016) simply don't function this way. In general, they require UH2-5km (or another storm variable, such as maximum vertical velocity) to exceed a given threshold and thus cannot distinguish between two similar values below (or above) the given threshold. This dissertation shows that not only can RFs distinguish between two similar UH2-5km values, but they can also distinguish between the same UH2-5km value in differing environments. While human forecasters may do this to a degree (e.g., by discounting a low (high) value of UH2-5km in the presence of a very (un-)favorable environment), RFs excel at accurately quantifying the impact of the precise set of variables. Although more work is needed on the subject, it is likely that this advantage of precise calibration is what gives RFs an edge over human forecasters in many severe weather forecasting situations.

3. Recommendations for future research

This dissertation described techniques for creating and analyzing next-day, random forest- (RF-) based high-impact weather forecast guidance. In doing so, it laid a foundation on which future work should build. One obvious avenue for future research would be to extend the analysis to a wider range of spatiotemporal scales, since this dissertation focused exclusively on next-day lead times and spatial scales. Indeed, research is already beginning to investigate how to use machine learning (ML) and/or deep learning to better predict high-impact weather at lead times of 90 minutes or less (e.g., Lagerquist et al. 2017, 2020), 0-3 hours (e.g., Flora et al. 2021), 1-3 days (e.g., Herman and Schumacher 2018c; Hill et al. 2020), and 1-4 weeks (e.g., Scheuerer et al. 2020). Future work should explicitly compare the benefits of RFs at each scale and interrogate how different predictors are used at the different scales. This type of analysis should provide insights into high-impact weather predictability. Relatedly, future work may wish to examine methods of “blending,” or transitioning, guidance from one scale to another, which may be beneficial for operational use. Indeed, it is conceivable that, in the not-too-distant future, multiple ML products will collectively enable the production of continuously-updating hazard probabilities from lead times of months to minutes. Such a vision would fit nicely into the Forecasting a Continuum of Environmental Threats (FACETs; Rothfusz et al. 2018) paradigm.

Another important avenue for future research is to examine the impact of including more types of predictors than just NWP forecast variables. For example, it is possible that observed radar, satellite, and sounding data will enhance RF forecast skill at next-day (and shorter) lead times. Future work should investigate the best ways of “summarizing” this information to RFs or other ML algorithms. The learned importance of observed and NWP predictors, respectively, should also be examined to enhance our understanding of high-impact weather predictability.

Of course, future work should continue exploring how to create better ML predictors from pure CAE forecast data. Since a storm’s morphology is related to the hazards it produces (e.g., Gallus et al. 2008; Smith et al. 2012; Thompson et al. 2012), it is speculated that explicit mode-related predictors could improve RF hazard prediction. Thus, it is conceivable that a future severe-weather-predicting RF could achieve greater skill by considering automated mode guidance from a second, mode-predicting ML system (e.g., Jergensen et al. 2020).

A final, but crucial, area for future research is determining when and how RFs (and other ML guidance products) add value to human forecasters. For example, in cases where RFs are used operationally, it is currently assumed that they provide most value on the lowest-impact days (e.g., Novak 2021). While this makes sense based on results from this dissertation (i.e., that RFs are most skillful for the most frequent events, which tend to be less impactful), research has not yet fully examined exactly when and how RFs provide most and least value to humans. A big part of this research will involve further testing RF (and interpretability) products in real-time testbed (e.g., Gallo et al. 2017; Clark et al. 2021) and operational environments. More in-depth, situation-specific comparisons between human and RF forecasts are also planned for the near future.

4. Research acknowledgements

Support for this dissertation's three research components (presented in Chapters 2-4) was provided by NOAA/Office of Oceanic and Atmospheric Research under NOAA-University of Oklahoma Cooperative Agreement NA11OAR4320072, U.S. Department of Commerce. The Developmental Testbed Visitor Center (DTC) Visitor Program provided additional support for Chapters 2 and 3. The DTC Visitor Program is funded by the National Oceanic and Atmospheric Administration, the National Center for Atmospheric Research, and the National Science Foundation. High-performance computing for Chapters 2-4 was conducted on the Cheyenne supercomputer (doi:10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. Some of the computing for Chapter 2 was also performed at the OU Supercomputing Center for Education and Research (OSCER) at the University of Oklahoma (OU). The Stage IV data in Chapter 2 (accessed from https://data.eol.ucar.edu/) were provided by NCAR/EOL under the sponsorship of the National Science Foundation. Two anonymous reviewers helped improve the manuscript in Chapter 3. AJC's and CDK's contributions to Chapter 3 were made as part of regular duties at the federally funded NOAA/National Severe Storms Laboratory. Montgomery Flora provided useful guidance and discussions on the Tree Interpreter method described in Chapter 4. AJC's contributions to

Chapter 4 were made as part of regular duties at the federally funded NOAA/National Severe Storms Laboratory. The statements, findings, conclusions, and recommendations presented in this dissertation do not necessarily reflect the views of NOAA or the U.S. Department of Commerce.

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, doi:https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

Adrianto, I., T. Trafalis, and V. Lakshmanan, 2009: Support vector machines for spatiotemporal tornado prediction. *Int. J. Gen. Syst.*, **38**, 759–776, doi: https://doi.org/10.1080/03081070601068629.

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

Albright, B., and S. Perfater, 2018: 2018 Flash Flood and Intense Rainfall Experiment. Weather Prediction Center Rep., 96 pp., https://www.wpc.ncep.noaa.gov/hmt/2018_FFaIR_final_report.pdf.

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

AMS, 2021: Weather and Forecasting. Accessed 25 March 2021. https://journals.ametsoc.org/view/journals/wefo/wefo-overview.xml

Baldwin, M. E., J. S. Kain, and S. Lakshmivarahan, 2005: Development of an automated classification procedure for rainfall systems. *Mon. Wea. Rev.*, **133**, 844–862.

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

Bergthorsson, P., B. R. Doos, S. Fryklund, O. Haug, and R. Lindquist, 1955: Routine forecasting with the barotropic model. *Tellus*, **7**, 272–274, doi: 10.1111/j.2153-3490.1955.tb01162.x.

Bermowitz, R. J., 1975: An application of model output statistics to forecasting quantitative precipitation. *Mon. Wea. Rev.*, **103**, 149–153, https://doi.org/10.1175/1520-0493(1975)103<0149:AAOMOS>2.0.CO;2.

Betts, A. K., 1986: A new convective adjustment scheme. Part I: Observational and theoretical basis. *Quart. J. Roy. Meteor. Soc.*, **112**, 677–691, doi:https://doi.org/10.1002/qj.49711247307.

Blake, B. T., J. R. Carley, T. I. Alcott, I. Jankov, M. E. Pyle, S. E. Perfater, and B. Albright, 2018: An adaptive approach for the calculation of ensemble gridpoint probabilities. *Wea. Forecasting*, **33**, 1063–1080.

Bolin, B., 1955: Numerical forecasting with the barotropic model. *Tellus*, **7**, 27-49, doi: 10.1111/j.2153-3490.1955.tb01139.x.

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

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

Bright, D. R., and S. L. Mullen, 2002: Short-range ensemble forecasts of precipitation during the Southwest monsoon. *Wea. Forecasting*, **17,** 1080–1100.

Brown, N., and T. Sandholm, 2019: Superhuman AI for multiplayer poker. *Science*, **365**, 885-890.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Wea. Forecasting*, **14**, 168–189, doi:10.1175/1520-0434(1999)014<0168:PPOPUT>2.0.CO;2.

Bukovsky, M. S., 2011: Masks for the Bukovsky regionalization of North America, Regional Integrated Sciences Collective, Institute for Matematics Applied to Geosciences, National Center for Atmospheric Research, Boulder, CO. Downloaded 2019-08-01. [http://www.narccap.ucar.edu/contrib/bukovsky/].

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*, https://doi.org/10.1175/WAF-D-19-0105.1.

Bushby, F. H., and M. S. Timpson, 1967: A 10-level atmospheric model and frontal rain. *Quart. J. Roy. Meteor. Soc.*, **93**, 562–564, doi: 10.1002/qj.49709339825.

Carley, J. R., B. R. J. Schwedler, M. E. Baldwin, R. J. Trapp, J. Kwiatkowski, J. Logsdon, and S. J.Weiss, 2011: A proposed model-based methodology for feature-specific prediction for high-impact weather. *Wea. Forecasting*, **26**, 243–249, doi:https://doi.org/10.1175/WAF-D-10-05008.1.

Carter, G. M., 1975: Automated prediction of surface wind from numerical model output. *Mon. Wea. Rev.*, **103,** 866–873, *https://doi.org/10.1175/1520-0493(1975)103<0866:APOSWF>2.0.CO;2*.

Charney, J. G., R. Fjortoft, and J. von Neumann, 1950: Numerical integration of the barotropic vorticity equation. *Tellus*, **2**, 237–254, doi: 10.1111/j.2153-3490.1950.tb00336.x.

Chollet, F., 2018: *Deep learning with Python*. Manning Publications Co., 361 pp.

Chollet, F. and Coauthors, 2015: Keras. GitHub. Available from https://github.com/fchollet/keras.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, A. Wimmers, J. Brunner, and W. Bellon, 2020: A Deep-Learning Model for Automated Detection of Intense Midlatitude Convection Using Geostationary Satellite Images. *Weather and Forecasting* 35, 2567-2588, https://doi.org/10.1175/WAF-D-20-0028.1.

Clark, A. J., 2017: Generation of ensemble mean precipitation forecasts from convection-allowing ensembles. *Wea. Forecasting*, **32**, 1569–1583, https://doi.org/10.1175/WAF-D-16-0199.1.

Clark, A. J., and Coauthors, 2011: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Wea. Rev.*, **139**, 1410–1418, doi:https://doi.org/10.1175/2010MWR3624.1.

Clark, A. J., and Coauthors, 2012a: An overview of the 2010 Hazardous Weather Testbed experimental forecast spring experiment. *Bull. Amer. Meteor. Soc.*, **93**, 55–74, https://doi.org/doi:10.1175/BAMS-D-11-00040.1.

Clark, A. J., 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.

Clark, A. J., J. Gao, P. Marsh, T. Smith, J. Kain, J. Correia, M. Xue, and F. Kong, 2013: Tornado pathlength forecasts from 2010 to 2011 using ensemble updraft helicity. *Wea. Forecasting*, **28**, 387–407, https://doi.org/10.1175/WAF-D-12-00038.1.

Clark, A. J., J. S. Kain, P. T. Marsh, J. Correia Jr., M. Xue, and F. Kong, 2012b: Forecasting tornado pathlengths using a three-dimensional object identification algorithm applied to convection-allowing forecasts. *Wea. Forecasting*, **27**, 1090–1113.

Clark, A. J., and Coauthors, 2018: The Community Leveraged Unified Ensemble (CLUE) in the 2016 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. Bull. Amer. Meteor. Soc., 99, 1433–1448, doi:10.1175/BAMS-D-16-0309.1.

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, doi: 10.1175/2009WAF2222222.1.

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

Clark, A. J., W. A. Gallus Jr., M. Xue, and F. Kong, 2010b: Growth of spread in convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **25**, 594–612, https://doi.org/10.1175/2009WAF2222318.1.

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.

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

Colquhoun, J. R., 1987: A decision tree method of forecasting thunderstorms, severe thunderstorms and tornadoes. *Weather and Forecasting*, **2**, 337 –345, doi: https://doi-org.ezproxy.lib.ou.edu/10.1175/1520-0434(1987)002<0337:ADTMOF>2.0.CO;2.

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, doi: 10.1175/2009WAF2222258.1.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part II: Application to convective rain systems. *Mon. Wea. Rev.*, **134,** 1785–1795.

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.

Deloncle, A., R. Berk, F. D'Andrea, and M. Ghil, 2007: Weather regime prediction using statistical learning, *J. of Atmos. Sciences*, **64**, 1619–1635, doi: https://doi-org.ezproxy.lib.ou.edu/10.1175/JAS3918.1

desJardins, M. L., K. F. Brill, and S. S. Schotz, 1991: Use of GEMPAK on UNIX workstations. *Proc. Seventh Int. Conf. on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology,* New Orleans, LA, Amer. Meteor. Soc., 449–453.

Done, J., C. A. Davis, and M. Weisman, 2004: The next generation of NWP: Explicit forecasts of convection using the Weather Research and Forecasting (WRF) Model. *Atmos. Sci. Lett.*, **5**, 110–117, doi:https://doi.org/10.1002/asl.72.

Du, J. 2011. GCIP/EOP Surface: Precipitation NCEP/EMC 4KM Gridded Data (GRIB) Stage IV Data. Version 1.0. UCAR/NCAR - Earth Observing Laboratory. https://doi.org/10.5065/D6PG1QDD. Accessed 02 Apr 2021.

Du, J., G. DiMego, B. Zhou, D. Jovic, B. Ferrier, and B. Yang, 2015: Regional ensemble forecast systems at NCEP. Preprints, *27th Conf. on Weather Analysis and Forecasting/23rd Conf. on Numerical Weather Prediction*, Chicago, IL, Amer. Meteor. Soc., 2A.5. [Available online at https://ams.confex.com/ams/27WAF23NWP/webprogram/Manuscript/Paper273421/NWP2015_NCEP_RegionalEnsembles_paper.pdf.]

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125,** 2427–2459.

Duda, J. D., and W. A. Gallus, 2010: Spring and summer Midwestern severe weather reports in supercells compared to other morphologies. *Wea. Forecasting*, **25**, 190–206, https://doi.org/10.1175/2009WAF2222338.1.

Dvorak, V. F., 1975: Tropical cyclone intensity analysis and forecasting from satellite imagery. *Mon. Wea. Rev.*, **103**, 420–430.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350, https://doi.org/10.1175/WAF843.1.

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. NOAA/NCEP/Environmental Modeling Center Office Note 442, 14 pp. [Available online at: https://www.emc.ncep.noaa.gov/officenotes/newernotes/on442.pdf].

Epstein, E. S., 1969: The role of initial uncertainties in prediction. *J. Appl. Meteor.*, **8,** 190–198.

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), http://www.ejssm.org/ojs/index.php/ejssm/article/viewArticle/33.

Evans, C., D. F. Van Dyke, and T. Lericos, 2014: How do forecasters utilize output from a convection-permitting ensemble forecast system? Case study of a high-impact precipitation event. *Wea. Forecasting*, **29**, 466–486, doi:https://doi.org/10.1175/WAF-D-13-00064.1.

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. [Available online at http://ams.confex.com/ams/SLS_WAF_NWP/techprogram/paper_47241.htm.]

Flora, M. L., C. K. Potvin, P. S. Skinner, S. Handler, and A. McGovern, 2021: Using machine learning to generate storm-scale probabilistic guidance of severe weather hazard in the Warn-on-Forecast System. *Monthly Weather Review*, **149**, 1535 –1557, doi: https://doi.org/10.1175/MWR-D-20-0194.1.

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

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

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.

Gagne, D., A. McGovern, 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.

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

Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2019: Incorporating UH occurrence time to ensemble-derived tornado probabilities. *Wea. Forecasting*, **34**, 151-164, https://doi.org/10.1175/WAF-D-18-0108.1

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.

Gallo, B. T., 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.

Gallus, W. A., N. A. Snook, and E. V. Johnson, 2008: Spring and summer severe weather reports over the Midwest as a function of convective mode: A preliminary study. *Wea. Forecasting*, **23**, 101–113, https://doi.org/10.1175/2007WAF2006120.1.

Gauss, C. F., 1809: Theoria motus corporum coelestium in sectionibus conicis solem ambientium.

Gauss, C. F., 1821: Theoria combinationis observationum erroribus minimis oboxiae (Theory of the combination of observations least subject to error).

Gensini, V., 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.

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.

Glahn, H. R., and D. A. Lowry, 1972b: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, doi:https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Goldstein, A., A. Kapelner, J. Bleich, and E. Pitkin, 2015: Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, **24**, 44-65, doi: 10.1080/10618600.2014.907095.

Good, P. I., 2006: *Resampling Methods*. Birkhauser Boston, 228 pp.

Grell, G. A., 1993: Prognostic evaluation of assumptions used by cumulus parameterizations. *Mon. Wea. Rev.*, **121**, 764–787, https://doi.org/10.1175/1520-0493(1993)121<0764:PEOAUB>2.0.CO;2.

Greybush, S. J., S. Saslo, and R. Grumm, 2017: Assessing the ensemble predictability of precipitation forecasts for the January 2016 and 2016 East Coast winter storms. *Wea. Forecasting*, **32**, 1057–1078, doi:10.1175/WAF-D-16-0153.1.

Guyer, J. L., and I. L. Jirak, 2014: The utility of storm-scale ensemble forecasts of cool season severe weather events from the SPC perspective. *27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., P3.37, https://ams.confex.com/ams/27SLS/webprogram/Paper254640.html.

Hall, T., H. E. Brooks, and C. A. Doswell III, 1999: Precipitation forecasting using a neural network. *Wea. Forecasting*, **14**, 338-345.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea.*

Forecasting, **14**, 155–167, doi:https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, https://doi.org/10.1175/MWR3237.1.

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

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, doi:https://doi.org/10.1175/2007MWR2411.1.

Han, D., L. Chan, and N. Zhu, 2007: Flood forecasting using support vector machines. *J. of Hydroinformatics*, **9**, 267–276, doi: 10.2166/hydro.2007.027.

Han, J., and H.-L. Pan, 2011: Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. *Wea. Forecasting*, **26**, 520–533, doi:https://doi.org/10.1175/WAF-D-10-05038.1.

Herman, G. R., and R. S. Schumacher, 2016: Extreme precipitation in models: An evaluation. *Wea. Forecasting*, **31**, 1853–1879, https://doi.org/10.1175/WAF-D-16-0093.1.

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

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

Herman, G. R., and R. S. Schumacher, 2018b: Flash flood verification: Pondering precipitation proxies. *J. Hydrometeorology.*, **19**, 1753–1776.

Herman, G. R., and R. S. Schumacher, 2018c: 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. *Monthly Weather Review*, 148, 2135–2161.

Hitchens, N. M., H. E. Brooks, and R. S. Schumacher, 2013: Spatial and temporal characteristics of heavy hourly rainfall in the United States. *Mon. Wea. Rev.*, **141**, 4564–4575, https://doi.org/10.1175/MWR-D-12-00297.1.

Hitchens, N. M., R. A. Sobash, and A. J. Clark, 2016: A multi-year evaluation of NSSL-WRF surrogate severe thunderstorm forecasts. *28th Conference on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc., P. 111. [Available online at https://ams.confex.com/ams/28SLS/webprogram/Paper300988.html].

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.

Hong, S.-Y., 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.

Hssina, B. A. Merbouha, H. Ezzikouri, and M. Erritali, 2014: A comparative study of decision tree ID3 and C4.5. *Int. J. Adv. Comput. Sci. Appl.*, 4, 13–19.

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, doi:https://doi.org/10.1016/0169-2070(86)90048-8.

Janjić, Z. I., 1994: The step-mountain Eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.*, **122**, 927–945, doi:https://doi.org/10.1175/1520-0493(1994)122<0927:TSMECM>2.0.CO;2.

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.

Janjić, Z. I., 2003: A nonhydrostatic model based on a new approach. *Meteorology and Atmospheric Physics*, **82**, 271-285.

Janjić, Z. I., J. P. Gerrity, Jr., and S. Nickovic, 2001: An alternative approach to modeling. *Monthly Weather Review*, **129**, 1164-1178.

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

Jergensen, G. E., A. McGovern, R. Lagerquist, and T. Smith, 2020: Classifying convective storms using machine learning. *Weather and Forecasting*, 35, 537– 559.

Jirak, I. L., 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 Conference on Numerical Weather Prediction, Denver, CO, Amer. Meteor. Soc., 14B.6, https://ams.confex.com/ams/29WAF25NWP/webprogram/Paper345640.html.

Jirak, I. L., 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.

Jirak, I. L., C. J. Melick, and S. J. Weiss, 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.

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 Conference on Severe Local Storms, Nashville, TN, Amer. Meteor. Soc., P9.137, https://www.spc.noaa.gov/publications/jirak/sseo_hwt.pdf.

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

Johnson, A., and X. Wang, 2013: Object-based evaluation of a storm scale ensemble during the 2009 NOAA Hazardous Weather Testbed Spring Experiment. *Mon. Wea. Rev.*, **141**, 1079–1098.

Kain, J. S., 2004: The Kain–Fritsch convective parameterization: An update. *J. Appl. Meteor.*, **43**, 170–181, doi:https://doi.org/10.1175/1520-0450(2004)043<0170:TKCPAU>2.0.CO;2.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing NWP. *Wea. Forecasting*, **23**, 931–952, https://doi.org/10.1175/WAF2007106.1.

Kain, J. S., S. J. Weiss, J. J. Levit, M. E. Baldwin, and D. R. Bright, 2006: Examination of convection-allowing configurations of the WRF Model for the prediction of severe convective weather: The SPC/NSSL Spring Program 2004. *Wea. Forecasting*, **21**, 167–181, doi: 10.1175/WAF906.1.

Kain, J. S., S. R. Dembek, S. J. Weiss, J. L. Case, J. J. Levit, and R. A. Sobash, 2010: Extracting unique information from high-resolution forecast models: Monitoring selected fields and phenomena every time step. *Wea. Forecasting*, **25**, 1536–1542, https://doi.org/10.1175/2010WAF2222430.1.

Kang, J.-H., M.-S. Suh, K.-O. Hong, and C. Kim, 2011: Development of updateable model output statistics (UMOS) system for air temperature over South Korea. *Asia-Pac. J. Atmos. Sci.*, **47**, 199–211, doi: https://doi.org/10.1007/s13143-011-0009-8.

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

Karstens, C. D., R. Clark III, I. L. Jirak, P. T. Marsh, R. Schneider, and S. J. Weiss, 2019: Enhancements to Storm Prediction Center Convective Outlooks. *9th Conference on Transition of Research to Operations*, Phoenix, AZ, Amer. Meteor. Soc., J7.3, https://ams.confex.com/ams/2019Annual/webprogram/Paper355037.html.

Key, J., J. Maslanik, and A. Schweiger, 1989: Classification of merged AVHRR and SMMR Arctic data with neural networks. *Photogramm. Eng. Remote Sensing*, **55**, 1331–1338.

Kishtawal, C. M., S. Basu, F. Patadia, and P. K. Thapliyal, 2003: Forecasting summer rainfall over India using genetic algorithm. *Geophysical Research Letters*, **30**, doi: 10.1029/2003GL018504.

Klein, W. H., B. M. Lewis, and I. Enger, 1959: Objective prediction of five-day mean temperatures during winter. *J. Meteor.*, **16**, 672–682, doi: https://doi.org/10.1175/1520-0469(1959)016<0672:OPOFDM>2.0.CO;2.

Klein, W. H., and H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. *Bull. Amer. Meteor. Soc.*, **55**, 1217–1227, doi:10.1175/1520-0477(1974)055<1217:FLWBMO>2.0.CO;2.

Kuligowski, R. J., and A. P. Barros, 1998: Localized precipitation forecasts from a numerical weather prediction model using artificial neural networks. *Wea. Forecasting*, **13**, 1194-1204.

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. *Monthly Weather Review* 148, 2837-2861, https://doi.org/10.1175/MWR-D-19-0372.1.

Lagerquist, R., A. McGovern, and D. Gagne, 2019: Deep learning for spatially explicit prediction of synoptic-scale fronts. *Wea. Forecasting*, 34, 1137–1160, https://doi-org.ezproxy.lib.ou.edu/10.1175/WAF-D-18-0183.1.

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

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.ezproxy.lib.ou.edu/10.1175/JTECH-D-13-00205.1.

Lakshmanan, V., K. Hondl, and R. Rabin, 2009: An efficient, general-purpose technique for identifying storm cells in geospatial images. *J. Atmos. Oceanic Technol.*, **26**, 523–537, https://doi.org/10.1175/2008JTECHA1153.1.

LeCun, Y., 1988: A theoretical framework for back-propagation. In D. Touretzky, G. Hinton, and T. Sejnowski (Eds.), *Proceedings of the 1988 connectionist models summer school* (pp. 21–28). CMU, Pittsburgh, PA: Morgan Kaufmann.

Legendre, A. M., 1805: Nouvelles méthodes pour la détermination des orbites des cometes. F. Didot.

Legg, T., and K. Mylne, 2004: Early warnings of severe weather from ensemble forecast information. *Wea. Forecasting*, **19**, 891–906, doi:https://doi.org/10.1175/1520-0434(2004)019<0891:EWOSWF>2.0.CO;2.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102,** 409–418.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *Journal of Computational Physics*, **227**, 3515-3539.

Lin, G-F., G-R. Chen, M-C. Wu, and Y-C. Chou, 2009: Effective forecasting of hourly typhoon rainfall using support vector machines. *Water Resources Research*, **45**, W08440, doi: 10.1029/2009WR007911.

Lin, Y., 2011. GCIP/EOP Surface: Precipitation NCEP/EMC 4KM Gridded Data (GRIB) Stage IV Data, version 1.0. UCAR/NCAR Earth Observing Laboratory, accessed 15 April 2019, https://data.eol.ucar.edu/dataset/21.093.

Linnainmaa, S., 1970: The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. Master's thesis, Univ. Helsinki.

Liong, S-Y, and C. Sivapragasam, 2002: Flood stage forecasting with support vector machines, **38**, 173–186.

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

Loken, E. D., A. J. Clark, M. Xue, and F. Kong, 2019b: Spread and skill in mixed- and single-physics convection-allowing ensembles. *Wea. Forecasting*, **34**, 305–330.

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating Probabilistic Next-Day Severe Weather Forecasts from Convection-Allowing Ensembles Using Random Forests. *Weather and Forecasting* 35, 1605-1631, https://doi.org/10.1175/WAF-D-19-0258.1.

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. doi: 10.1175/WAF-D-16-0200.1.

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.

Malone, T., 1955: Application of statistical methods in weather prediction. *Proc. Natl. Acad. Sci. USA*, **41**, 806–815, doi:https://doi.org/10.1073/pnas.41.11.806.

Manzato, A., 2007: Sounding-derived indices for neural network based short-term thunderstorm and rainfall forecasts. *Atmospheric Research*, **83**, 349-365.

Markowski, P. M., Straka J. M., and Rasmussen E. N., 2002: Direct surface thermodynamic observations within the rear-flank downdrafts of nontornadic and tornadic supercells. *Mon. Wea. Rev.*, **130 ,** 1692–1721.

Marsh, P. T., J. S. Kain, V. Lakshmanan, A. J. Clark, N. M. Hitchens, and J. Hardy, 2012: A method for calibrating deterministic forecasts of rare events. *Wea. Forecasting*, **27**, 531–538.

Marzban, C., 2004: The ROC curve and the area under it as performance measures. *Wea. Forecasting*, **19**, 1106–1114, https://doi.org/10.1175/825.1.

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

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

Mason, S. J., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291–303.

McCulloch, W., and W. Pitts, 1943: A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **7**, 115–133.

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

McGovern, A., C. D. Karstens, T. Smith, and R. Lagerquist, 2019a: Quasi-Operational Testing of Real-Time Storm-Longevity Prediction via Machine Learning. *Weather and Forecasting* 34, 1437-1451, https://doi.org/10.1175/WAF-D-18-0141.1.

McGovern, A., R. Lagerquist, D. John Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019b: 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.

Melhauser, C., and F. Zhang, 2012: Practical and intrinsic predictability of severe and convective weather at the mesoscales. *J. Atmos. Sci.*, **69**, 3350–3371, https://doi.org/10.1175/JAS-D-11-0315.1.

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.

Mills, G. A., and J. R. Colquhoun, 1998: Objective prediction of severe thunderstorm environments: Preliminary results linking a decision tree with an operational regional NWP model. *Weather and Forecasting*, **13**, 1078–1092, doi: https://doi-org.ezproxy.lib.ou.edu/10.1175/1520-0434(1998)013<1078:OPOSTE>2.0.CO;2.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

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, doi:https://doi.org/10.1023/B:BOUN.0000020164.04146.98.

Nakanishi, M., and H. Niino, 2006: An improved Mellor–Yamada Level-3 model: Its numerical stability and application to a regional prediction of advection fog. *Bound.-Layer Meteor.*, **119**, 397–407, doi:https://doi.org/10.1007/s10546-005-9030-8.

NCEI, 2019: U.S. billion-dollar weather and climate disasters: Overview. NOAA/NCEI, https://www.ncdc.noaa.gov/billions/.

NCEI, 2021: U.S. Billion-Dollar Weather and Climate Disasters. NOAA/NCEI, https://www.ncdc.noaa.gov/billions/, DOI: 10.25921/stkw-7w73

Nielsen, E. R., and R. S. Schumacher, 2018: Dynamical insights into extreme short-term precipitation associated with supercells and mesovortices. *Journal of the Atmospheric Sciences*, **75**, 2983–3009.

NOAA, 2020: NOAA Research and Development Vision Areas: 2020-2026. https://nrc.noaa.gov/Council-Products/Research-Plans.

Novak, D., 2021: The Forecaster – NWP Partnership. *20th Conf. on Artificial Intelligence*, Virtual. Amer. Meteor. Soc., J2.6, https://ams.confex.com/ams/101ANNUAL/meetingapp.cgi/Person/204493.

Olah, C., A. Mordvintsev, and L. Schubert, 2017: Feature visualization. Distill, https://distill.pub/2017/feature-visualization/.

Ortiz-Garcia, E. G., S. Salcedo-Sanz, and C. Casanova-Mateo, 2014: Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data. *Atmospheric Research*, **139**, 128-136.

Palmer, T., 2017: The primacy of doubt: Evolution of numerical weather prediction from determinism to probability. *Journal of Advances in Modeling Earth Systems*, **9**, 730-734.

Parker, D. B., 1985: *Learning-logic. Technical report TR-47.* Center for Comp. Research in Economics and Management Sci., MIT.

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, http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Quinlan, J. R., 1979: Discovering rules by induction from large collections of examples. In D. Michie (Ed.), *Expert systems in the micro electronic age*. Edinburgh University Press.

Quinlan, J. R., 1983: Learning efficient classification procedures and their application to chess endgames. In R.S. Michalski, J.G. Carbonell & T.M. Mitchell, (Eds.), *Machine learning: An artificial intelligence approach*. Palo Alto: Tioga Publishing Company.

Quinlan, J. R., 1986: Induction of decision trees. *Mach. Learn.*, **1**, 81–106.

Quinlan, J. R., 1993: *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, Inc.

Rajendra, P., K. V. N. Murthy, A. Subbarao, and R. Boadh, 2019: Use of ANN models in the prediction of meteorological data. *Model. Earth Syst. Environ.*, **4**, 1051–1058, https://doi-org.ezproxy.lib.ou.edu/10.1007/s40808-019-00590-2.

Rasmussen, E. N., and D. O. Blanchard, 1998: A baseline climatology of sounding-derived supercell and tornado forecast parameters. *Wea. Forecasting*, **13**, 1148–1164.

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? *Weather and Forecasting*, 35, 2293–2316.

Roberts, B., I. Jirak, A. Clark, S. Weiss, and J. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensembles. *Bull. Amer. Meteor. Soc.*, **100**, 1245–1258, https://doi-org.ezproxy.lib.ou.edu/10.1175/BAMS-D-18-0041.1.

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

Roebber, P. J., D. M. Schultz, B. A. Colle, and D. J. Stensrud, 2004: Toward improved prediction: High-resolution and ensemble modeling systems in operations. *Wea. Forecasting*, **19**, 936–949, doi: 10.1175/1520-0434(2004)019<0936:TIPHAE>2.0.CO;2.

Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev.*, **142**, 4519–4541, https://doi.org/10.1175/MWR-D-14-00100.1.

Rothfusz, L. P., R. Schneider, D. Novak, K. Klockow-McClain, A. E. Gerard, C. Karstens, G. J. Stumpf, and T. M. Smith, 2018: FACETs: A Proposed Next-Generation Paradigm for High-Impact Weather Forecasting. *Bulletin of the American Meteorological Society* 99, 2025-2043, https://doi.org/10.1175/BAMS-D-16-0100.1.

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

Scheuerer, M., M. B. Switanek, R. P. Worsnop, and T. M. Hamill, 2020: Probabilistic subseasonal precipitation forecasts over California. *Monthly Weather Review*, 148, 3489–3506.

Schmeits, M. J., K. J. Kok, and D. H. P. Vogelezang, 2005: Probabilistic forecasting of (severe) thunderstorms in the Netherlands using model output statistics. *Wea. Forecasting*, **20**, 134–148, https://doi.org/10.1175/WAF840.1.

Schmidhuber, J., 2015: Deep learning in neural networks: An overview. *Neural Networks*, **61**, 85–117.

Schoen, J., and W. S. Ashley, 2011: A climatology of fatal convective wind events by storm type. *Wea. Forecasting*, **26**, 109–121, https://doi.org/10.1175/2010WAF2222428.1.

Schultz, M. G., C. Betancourt, B. Gong, F. Kleinert, M. Langguth, L. H. Leufen, A. Mozaffari, and S. Stadtler, 2021: Can deep learning beat numerical weather prediction? *Phil. Trans. R. Soc. A.*, **379**, doi: http://doi.org/10.1098/rsta.2020.0097.

Schumacher, R. S., and R. H. Johnson, 2006: Characteristics of U.S. extreme rain events during 1999–2003. *Wea. Forecasting*, **21**, 69–85, https://doi.org/10.1175/WAF900.1.

Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145**, 3397–3418, https://doi.org/10.1175/MWR-D-16-0400.1.

Schwartz, C. S., and Coauthors, 2010: Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Wea. Forecasting*, **25**, 263–280, doi:https://doi.org/10.1175/2009WAF2222267.1.

Schwartz, C., G. Romine, K. Fossell, R. Sobash, and M. Weisman, 2017: Toward 1-km ensemble forecasts over large domains. Mon. Wea. Rev. doi: 10.1175/MWR-D-16-0410.1.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2015: NCAR's experimental real-time convection-allowing ensemble prediction system. Wea. Forecasting, 30, 1645– 1654.

Schwartz, C. S., G. S. Romine, R. A. Sobash, K. R. Fossell, and M. L. Weisman, 2019: NCAR's Real-Time Convection-Allowing Ensemble Project. *Bulletin of the American Meteorological Society* **100**:2, 321-343.

Schwartz, C. S., Z. Liu, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, doi:https://doi.org/10.1175/WAF-D-13-00145.1.

Selvaraju, R., M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 2017: Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. Conf. on Computer Vision*, Venice, Italy, IEEE, https://doi-org.ezproxy.lib.ou.edu/10.1109/ICCV.2017.74.

Silver, D., and Coauthors, 2016: Mastering the game of go with deep neural networks and tree search. *Nature*, **529**, 484–489, https://doi.org/10.1038/nature16961.

Silver, D., and Coauthors, 2017: Mastering the game of go without human knowledge. *Nature*, **550**, 354–359, https://doi.org/10.1038/nature24270.

Silver, D., and Coauthors, 2018: A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, **362**, 1140-1144, doi: 10.1126/science.aar6404

Simonyan, K., A. Vedaldi, and A. Zisserman, 2014: Deep inside convolutional networks: Visualizing image classification models and saliency maps. arXiv, https://arxiv-org.ezproxy.lib.ou.edu/abs/1312.6034.

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.

Skinner, P. S., L. J. Wicker, D. M. Wheatley, and K. H. Knopfmeier, 2016: Application of two spatial verification methods to ensemble forecasts of low-level rotation. *Wea. Forecasting*, **31**, 713–735, https://doi.org/10.1175/WAF-D-15-0129.1.

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

Sobash, R. A., and J. S. Kain, 2017: Seasonal variations in severe weather forecast skill in an experimental convection-allowing model. *Wea. Forecasting*, **32**, 1885–1902, https://doi.org/10.1175/WAF-D-17-0043.1.

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

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

Sobash, R. A., C. S. Schwartz, G. S. Romine, 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.ezproxy.lib.ou.edu/10.1175/WAF-D-19-0044.1.

Sobash, R. A., G. S. Romine, and C. S. Schwartz, 2020: A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived form a convection-allowing model. *Weather and Forecasting*, 35, 1981–2000.

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, doi:https://doi.org/10.1175/WAF-D-10-05046.1.

SPC, 2019a: Storm Prediction Center WCM page: Severe weather database files (1950–2017). Accessed 16 December 2019, https://www.spc.noaa.gov/wcm/.

SPC, 2019b: Severe weather event summaries: NWS local storm reports. Accessed 16 December 2019, https://www.spc.noaa.gov/climo/online/.

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

SPC, 2021b: Significant Hail Parameter. Accessed 26 March 2021, https://www.spc.noaa.gov/exper/mesoanalysis/help/help_sigh.html.

SPC, 2021c: Storm Prediction Center Hail Probabilities (1982-2011). Accessed 1 April 2021, https://www.spc.noaa.gov/new/SVRclimo/climo.php?parm=allHail.

SPC, 2021d: Storm Prediction Center Tornado Probabilities (1982-2011). Accessed 1 April 2021, https://www.spc.noaa.gov/new/SVRclimo/climo.php?parm=allTorn.

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

SPC, 2021f: Storm Prediction Center Wind Probabilities (1982-2011). Accessed 1 April 2021, https://www.spc.noaa.gov/new/SVRclimo/climo.php?parm=allWind.

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.*, **127,** 433–446.

Stevenson, S. N., and R. S. Schumacher, 2014: A 10-year survey of extreme rainfall events in the central and eastern United States using gridded multisensor precipitation analyses. *Mon. Wea. Rev.*, **142**, 3147–3162, https://doi.org/10.1175/MWR-D-13-00345.1.

Szpiro, G. G., 1997: Forecasting chaotic time series with genetic algorithms, *Phys. Rev. E.*, **55**, 2557–2568.

Tang, B.H., V. A. Gensini, and C. R. Homeyer, 2019: Trends in United States large hail environments and observations. *npj Clim 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.ezproxy.lib.ou.edu/10.1175/2008MWR2387.1.

Thompson, G., R. M. Rasmussen, and K. Manning, 2004: Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: Description and sensitivity analysis. *Mon. Wea. Rev.*, **132**, 519–542, doi:https://doi.org/10.1175/1520-0493(2004)132<0519:EFOWPU>2.0.CO;2.

Thompson, R. L., 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.

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. Preprints, *21st Conf. on Severe Local Storms,* San Antonio, TX, Amer. Meteor. Soc., J3.2. [Available online at https://ams.confex.com/ams/SLS_WAF_NWP/techprogram/paper_46942.htm.]

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

Torn, R. D., and G. S. Romine, 2015: Sensitivity of central Oklahoma convection forecasts to upstream potential vorticity anomalies during two strongly forced cases during MPEX. *Mon. Wea. Rev.*, **143**, 4064–4087, https://doi.org/10.1175/MWR-D-15-0085.1.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations, Bull. Am. Meteorol. Soc., **74**, 2317–2330.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, **8**, 378–398.

Trafalis, T.B., H. Ince, and M. Richman, 2003: Tornado detection with support vector machines. In *Computational science – ICCS 2003*, Edited by: P. M. Sloot, D. Abramson, A. Bogdanov, J. J. Dongarra, A. Zomaya, and Y. Gorbachev, 289–298. Berlin/Heidelberg: Springer-Verlag.

Trafalis, T. B., B. Santosa, and M. Richman, 2004: Bayesian neural networks for tornado detection. *WSEAS transactions on systems*, **3**, 3211–3216.

Trafalis, T. B., B. Santosa, and M. Richman, 2005: Learning networks for tornado forecasting: A Bayesian perspective. *WIT transaction on information and communication technologies*, **35**, 5–14.

Trier, S. B., G. S. Romine, D. A. Ahijevych, R. J. Trapp, R. S. Schumacher, M. C. Coniglio, and D. J.Stensrud, 2015: Mesoscale thermodynamic influences on convection initiation near a surface dryline in a convection-permitting ensemble. *Mon. Wea. Rev.*, **143**, 3726–3753, https://doi.org/10.1175/MWR-D-15-0133.1.

Vapnik, V., and C. Cortes, 1995: Support-vector networks. *Machine Learning*, **20**, 273–297.

Wagstaff, K., and J. Lee, 2018: Interpretable discovery in large image data sets. arXiv, https://arxiv-org.ezproxy.lib.ou.edu/abs/1806.08340.

Wandishin, M. S., S. L. Mullen, D. J. Stensrud, and H. E. Brooks, 2001: Evaluation of a short-range multimodel ensemble system. *Mon. Wea. Rev.*, **129,** 729–747.

Weisman, M. L., C. Davis, W. Wang, K. W. Manning, and J. B. Klemp, 2008: Experiences with 0–36-h explicit convective forecasts with the WRF-ARW Model. *Wea. Forecasting*, **23**, 407–437, https://doi.org/10.1175/2007WAF2007005.1.

Weisman, M. L., W. C. Skamarock, and J. B. Klemp, 1997: The resolution dependence of explicitly modeled convective systems. *Mon. Wea. Rev.*, **125**, 527–548, doi:https://doi.org/10.1175/1520-0493(1997)125<0527:TRDOEM>2.0.CO;2.

Werbos, P. J., 1981: Applications of advances in nonlinear sensitivity analysis. *Proceedings of the 10th IFIP conference, 31.8-4.9, NYC* (pp. 762–770).

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.

Wilks, D. S., 2001: A skill score based on economic value for probability forecasts. *Meteorological Applications*, **8**, 209–219.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences: Third Edition*. Elsevier Inc., 676 pp.

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

Wong, K. Y., C. L. Yip, and P. W. Li, 2008: Automatic tropical cyclone eye fix using genetic algorithm. *Expert systems with applications*, **34**, 643-656.

Xue, M., and Coauthors, 2007: CAPS real-time storm-scale ensemble and high-resolution forecasts as part of the NOAA Hazardous Weather Testbed 2007 Spring Experiment. Preprints, *22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction,* Salt Lake City, UT, Amer. Meteor. Soc., 3B. [Available online at http://ams.confex.com/ams/pdfpapers/124587.pdf].

Zhang, F., Bei N., Rotunno R., Snyder C., and Epifanio C., 2007: Mesoscale predictability of moist baroclinic waves: Convection permitting experiments and multistage error growth dynamics. *J. Atmos. Sci.*, **64**, 3579–3594.