UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

IMPACTS OF MULTISCALE PREDICTORS ON RANDOM FOREST BASED PROBABILISTIC FORECASTS OF SEVERE WEATHER HAZARDS

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By
Daniel Kubalek
Norman Oklahoma
2024

IMPACTS OF MULTISCALE PREDICTORS ON RANDOM FOREST BASED PROBABILISTIC FORECASTS OF SEVERE WEATHER HAZARDS

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Aaron Johnson, Chair

Dr. Xuguang Wang

Dr. Zachary Lebo

Acknowledgements

I would like to first and foremost thank my advisor Dr. Aaron Johnson, for everything you have helped me with throughout the thesis and my masters. From your guidance to mentoring I am forever grateful. To my co-advisor Dr. Xuguang Wang, for her guidance and pushing me to new heights of expectations for myself and future work for which I am thankful. To Dr. Zachary Lebo, for graciously accepting a position on my committee and for suggestions to help bring this thesis together.

This project could not have been possible without funding. I would like to give credit to the NOAA JTTI grant NA20OAR4590358 for making the opportunity to conduct research possible, as well as giving me the opportunity to pursue my masters degree at the University of Oklahoma.

I would also like to thank the University of Oklahoma Super Computing Center for Education and Research (OSCER) for providing the computational resources and support needed to conduct research. Lastly, I would like to thank my friends and family for supporting me throughout my masters and this thesis.

Contents

List of Figures

List of Tables

Abstract

Machine learning (ML) algorithms utilized for post-processing of convection-allowing model/ensemble (CAM/CAE) output has been a major area of research to handle limitations with CAM/CAE forecasts. ML has been used to correct systematic biases, relate observed variables to numerical output, and synthesize extremely large data into probabilistic forecasts. In particular, numerous studies have shown random forests (RFs) to be successful in severe weather forecasting applications utilizing predictors from global scale and/or CAE output. However, predictors used in the RF models are typically fixed and treated independently when training the RF models. This can consequently leave out important information about the large-scale flow pattern that is necessary for assessing severe weather risk. This thesis develops a method for manifesting multiscale flow-dependence into RF models through direct incorporation of CAE-based predictors that are pre-processed at increasing spatial length scales. The different length scales account for different scales of motion with the goal to improve probabilistic forecast skill for a variety of severe weather hazards for next-day (12-12 UTC) - or 24hr and 4hr (20-00 UTC) forecasts. In order to verify the impacts of the multiscale predictors on the skill of the RF models, a control (CTLRF) and experimental (EXPRF) set of RF models were created. The CTLRF models were trained with only predictors pre-processed to 80 kilometers (km) and the EXPRF models were trained with predictors pre-processed to 80 km in addition to larger, spatially smoothed 80 km predictors. Both models were verified against the storm prediction center (SPC) reports quantitatively and qualitatively. Results show that the EXPRF models had higher brier skill score's (BSS) than the CTLRF models for all sub-significant severe weather hazards for both forecast periods, but significantly higher BSS's when forecasting any severe weather hazard (24hr and 4hr), wind (24hr), hail (24hr and 4hr), and significant winds (24hr). The EXPRF forecasts generally had the best resolution component of the brier score (BS), of which some severe weather hazards were significantly higher than CTLRF forecasts. Furthermore, both models generally had small calibration error. However, the CTLRF 24hr and 4hr wind forecasts had significantly lower calibration error compared to EXPRF. In general, neither model's probabilistic forecasts were consistently more reliable than the other. Predictor contributions determined via tree interpreter (TI) showed when severe weather did not occur, on average, the meso-γ scale storm-attribute predictors contributed more to forecast skill than the meso-β scale storm-attribute predictors for 24hr forecasts. Whereas the opposite was true for the 4hr forecasts. When severe weather did occur, on average, the meso-β scale storm-attribute predictors contributed the most to skill in general. Meanwhile, the meso-γ scale environmental predictors dominated environment-related contributions to forecast skill, but in general, most multiscale predictors still contributed to skill. Through case-studies, it was found that the meso-β and meso-α scale storm-attribute predictors accounts for spatial uncertainty of simulated storms similar to neighborhood-based CAM forecasts. Meanwhile the environment predictors, in particular the convective environment predictors, had greater sensitivity to smoothing and sometimes did not benefit from losing sharp gradients and local extremes that can be associated with synoptically predictable features.

Introduction

Severe weather forecasting has evolved substantially over the years due in part to more sophisticated forecasting methods and computational resources. A large component of this evolution is made up of convection allowing models or CAMs. Unlike global models with coarser grids, CAMs have fine enough horizontal resolution (normally $\leq\sim$ 4 km) to allow the primary mesoscale circulations that facilitate the morphology, evolution, timing, initiation, etc. of convection all while avoiding the need for convective-parameterization (Weisman et al., 1997; Kain et al., 2006; Done et al., 2004). CAMs have undergone expansive improvements over the years for a myriad of reasons, such as improved initialization through advanced convective scale data assimilation (Johnson et al., 2015, 2022; Johnson and Wang, 2017; Wang and Wang, 2017, 2020, 2021, 2023a,b; Degelia et al., 2019; Chipilski et al., 2020; Gasperoni et al., 2022; Chandramouli et al., 2022; Yang and Wang, 2023b,a), improved model physics (Flora et al., 2018; Beck et al., 2016), etc. Another important advancement of CAMs is the generation of convection allowing ensembles or CAEs. Due to limited predictability of forecasting the details of convection, errors can grow rapidly in CAM based deterministic forecasts (Loken et al., 2017) that can translate to poor probabilistic forecasts. CAEs can estimate forecast uncertainties using different initial conditions (ICs) (Johnson et al., 2014; Johnson and Wang, 2024, 2016) and varied model physics (Roebber et al., 2004; Kain et al., 2006, 2008; Johnson et al., 2011a,b; Duda et al., 2014, 2016, 2017; Gasperoni et al., 2020; Johnson and Wang, 2020). As a result, a range of possible outcomes are accounted for which can lead to better probabilistic forecasts that can be derived from a deterministic forecast. CAEs play an important role in operational severe weather forecasting (Benjamin et al., 2016; Dowell et al., 2022; Roberts et al., 2019)

Despite their advantages, CAMs/CAEs also have inherent limitations. Although they have the ability to simulate key mesoscale structures of convection that can produce severe weather, their severe weather hazards are not fully resolved. For the application interested in severe weather, tornadoes, max-size hail cores, and max-wind gusts - or microbursts - these are still not resolved at convection-allowing resolution (e.g., Loken et al., 2020; Clark and Loken, 2022). The CAMs/CAEs can also contain bias (e.g., Herman and Schumacher, 2018b; Loken et al., 2019; Jasper Velthoen and Jongbloed, 2023). A full ensemble member suite can be systematically biased such as convection initiates too late, too far east, and/or magnitude of precipitation can be too low or high, etc. (Weisman et al., 2008; Davis et al., 2006; Gagne et al., 2014). The spread of ensemble members can also be biased, typically under dispersed, which can result in poorly calibrated forecast probabilities or missed outlier events (Gebhardt et al., 2011; Romine et al., 2014; Novak et al., 2008). There is also large data volume to synthesize (e.g., Roebber et al., 2004). Even for small ensembles (e.g. 10-20 members), it can become challenging for forecasters to parse through the data to interpret and contextualize numerous forecast variables.

To address these limitations, usually one or more post-processing technique is often utilized to derive the probabilities from the ensemble and to calibrate the probabilistic forecasts. For example, neighborhood-based methods (e.g., Schwartz and Sobash, 2017; Blake et al., 2018) and object-based probability methods (e.g., Johnson and Wang, 2012; Johnson et al.,

2013; Johnson and Wang, 2020; Wilkins et al., 2021) have been used to produce convective scale probability forecasts. There are numerous techniques that have been leveraged for this, especially statistically driven techniques to calibrate or correct CAEs probability forecasts (e.g., Johnson and Wang, 2012; Jasper Velthoen and Jongbloed, 2023). Recently, as one of the statistically driven methods, machine learning (ML) has been used to calibrate CAE forecasts (e.g., McGovern et al., 2017). As stated in Loken et al. (2020), ML algorithms ingest historical data and find patterns in the data. These learned patterns can be utilized to relate CAM/CAE resolved variables to severe weather reports (Loken et al., 2020), correct systematic biases (Loken et al., 2019; Zarei et al., 2021; Baez-Villanueva et al., 2020; Jasper Velthoen and Jongbloed, 2023) and synthesize ensemble output that can be post-processed as simple probabilistic forecasts (e.g., Gagne et al., 2014; Clark and Loken, 2022; Herman and Schumacher, 2018b; Hill et al., 2020). In addition to this, ML models are often nonlinear which makes them quite suitable for relating CAEs and severe weather hazards. Some are also human-readable, facilitating the ability to investigate what relationships the model has identified in relation to an event being forecasted (e.g., McGovern et al., 2017).

Among all ML methods, the random forest (RF) algorithm has become a popular ML method for post-processing numerical weather prediction (NWP) output and has had success in severe weather forecasting applications, using both CAM and global ensemble variables (e.g., Clark and Loken, 2022; Hill et al., 2020; Loken et al., 2019, 2020) as predictors; as well as for convective related hazards such as severe hail, damaging straight-line winds, heavy precipitation, etc. (Gagne et al., 2014, 2017; McGovern et al., 2011; Herman and Schumacher, 2018b; Medina et al., 2019; Yao et al., 2020). For example, Loken et al. (2020) obtained next-day (12-12 UTC) severe weather probabilistic forecasts utilizing RFs to relate CAE forecast variables to severe weather reports. This technique was used to challenge severe weather forecasts issued from the calibrated Storm-Scale Ensemble of Opportunity (SSEO) 2–5 km updraft helicity (UH) forecast and storm prediction center (SPC) convective outlooks. Their results showed that the RF-based forecasts had consistently higher brier skill scores (BSSs) for all severe hazards (any severe, wind, hail, tornadoes, and significant counterparts) regardless of the evaluation domain. The most skillful RF forecasts, when compared against the UH and SPC forecasts, were for wind and hail during the spring and summer months (March-August). In addition, despite their relatively higher BSSs across all severe hazards, they did find that the RF-based forecasts for all the severe hazards had the best resolution, but not necessarily the best reliability. Still for most severe hazards, the RF-based forecasts outperformed SPC and UH forecasts. RFs are also attractive due to their physical interpretability as demonstrated in Loken et al. (2022). Using a python-based module called tree interpreter (TI), they studied contributions made from all predictors for the same hazards and found that the RF models emphasized different predictors for different hazards in a physically meaningful way. For example, maximum updraft helicity (UH) 0-3 km was found to contribute more for tornado probabilities than for wind or hail.

Although RFs have shown skill in many facets of meteorology, in particular high-impact severe weather forecasting applications (Gagne et al., 2017; Hill et al., 2020; McGovern et al., 2017; Loken et al., 2020; Clark and Loken, 2022; Zeng et al., 2022; Medina et al., 2019), there is always a need for improvements of these models and one of the problems lies within the

spatial mapping. In Liu et al. (2020); Snellman (1982) they explain that features on all resolved spatial scales contribute to severe weather risk. In addition, forecasting severe weather requires information of the large-scale flow (Johns and Doswell III, 1992; Ostby, 1999) to assess mesoscale evolution that ultimately governs convection. Furthermore, relationships between CAM forecast variables and observed severe weather are likely to be flow-dependent (e.g., Loken et al., 2020). However, since the RF models are only using predictors that are pre-processed to fixed and/or small grids such as (30 km; Clark and Loken, 2022), (20km; Loken et al., 2019), (80 km; Loken et al., 2020), and treated independently when training the RF models, the RF models could be missing useful information about the large scale flow and/or features. To address this issue, Shearer et al. (2023) separately trained RF models for different classifications of large-scale flow pattern within the same season. However, past studies have not yet accounted for the large-scale flow dependence of severe risk by directly including CAE-based RF predictors that are based on the multiple scales of motion. This study is aiming to answer if and how the incorporation of predictors that represent a broad range of spatial scales improves RF model probability forecast performance and how these predictors impact probabilistic forecast performance. In particular, this is accomplished through direct inclusion of multiscale predictors. The goals of the study are to (1) determine if the inclusion of multiscale predictors improves the probability forecast skill for a variety of severe weather hazards and (2) contribute to physical understanding for how the multiscale predictors affect the RF-based probability forecast performance.

The rest of this thesis is organized as the methodology detailing the datasets used for pre-processing for the RF models and the experiment design for this study, followed by the results with analysis of a couple of representative case studies, and closing out with the discussion of the results and conclusions of this study with suggestions for future work.

Methodology

Datasets

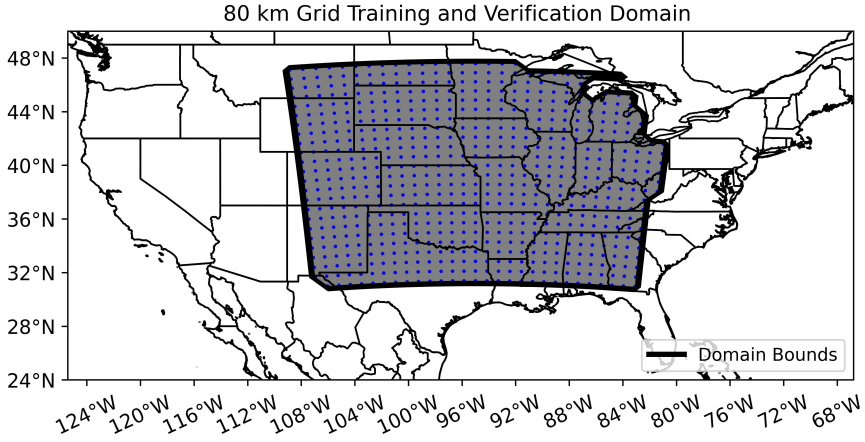

Figure 1: Training/verification domain (grey shading) and 80 km grid points (blue dots)

For this thesis, NWP output used for the RF-based forecasts came from the University of Oklahoma (OU) Multiscale data Assimilation and Predictability (MAP) Lab 10-member CAE produced during the 2018 and 2019 NOAA Hazardous Weather Testbed Spring Forecast Experiments (Johnson et al., 2023). All OU MAP Lab CAE member forecasts are made on a 3 km grid over a contiguous United States (CONUS) domain with 1620 x 1120 grid points with a temporal resolution of 1hr. Furthermore, a regional masking was applied to designate the training and verification domain. This domain spans approximately the central most region of the CONUS yielding approximately 24 x 30 gridpoints (Fig. 1). The training

Month	2018	2019	Total
April	30	29-30	3
May	1-4, 7-11, 14-18, 21-25, 28-31	1-3,6-10, 14-18, 20-24, 27-31	46
June	1	-	1
Total	25	25	50

Table 1: OU MAP Lab Produced CAE Initialized Days. 25 days from 2018 and 25 days from 2019. Only 2018 has a day in June. All forecast initialization's were at 0000 UTC

and verification domain are chosen to focus on an interior region with reliable severe weather reports (e.g., excluding oceans and international boundaries) and sufficient distance from the computational domain boundary to enable smoothing over large distances. Forecasts were initialized at 0000 UTC and ran for 36 hours with forecast lead time 12-36 hours (i.e., 12-12 UTC). Since not all days during these periods had severe weather forecasted, only 50 cases were available (Table 1).

Severe weather observations used for verification and training of the RF models, were pulled from the filtered SPC storm report database. The reports were then remapped to an 80 km grid and were encoded as a 1 (a report) or 0 (no report) depending on if there was a report within the 80 km grid box during the forecast period. This is similar to SPC verification that uses a 40 km radius from a given point. The storm reports consist of severe wind ($\geq$ 50 kt or 58 mph), significant winds ($\geq$ 65 kt or 75 mph), severe hail (maximum hailstone diameter $\geq$ 1.00 in), significant hail (maximum hailstone diameter $\geq$ 2.00 in), tornadoes, and significant tornadoes ($\geq$ EF2).

Report climatology (Fig. 2a-d) within the forecast-domain showed that any severe weather reports had the largest spatial distribution with a majority of reports from the central/southern plains and the midwest. Wind and hail reports made up most of the severe weather reports with greater spatial distribution from wind reports over hail reports. Although there were 8 severe weather hazards forecasted in this study, only four were chosen to show since the remaining hazards had similar distributions, but much less in frequency.

Predictor Fields

45 model forecast variables were made available to the RF models for training (Table 2). The model variables were first re-gridded to 80 km before being used for training of the RF

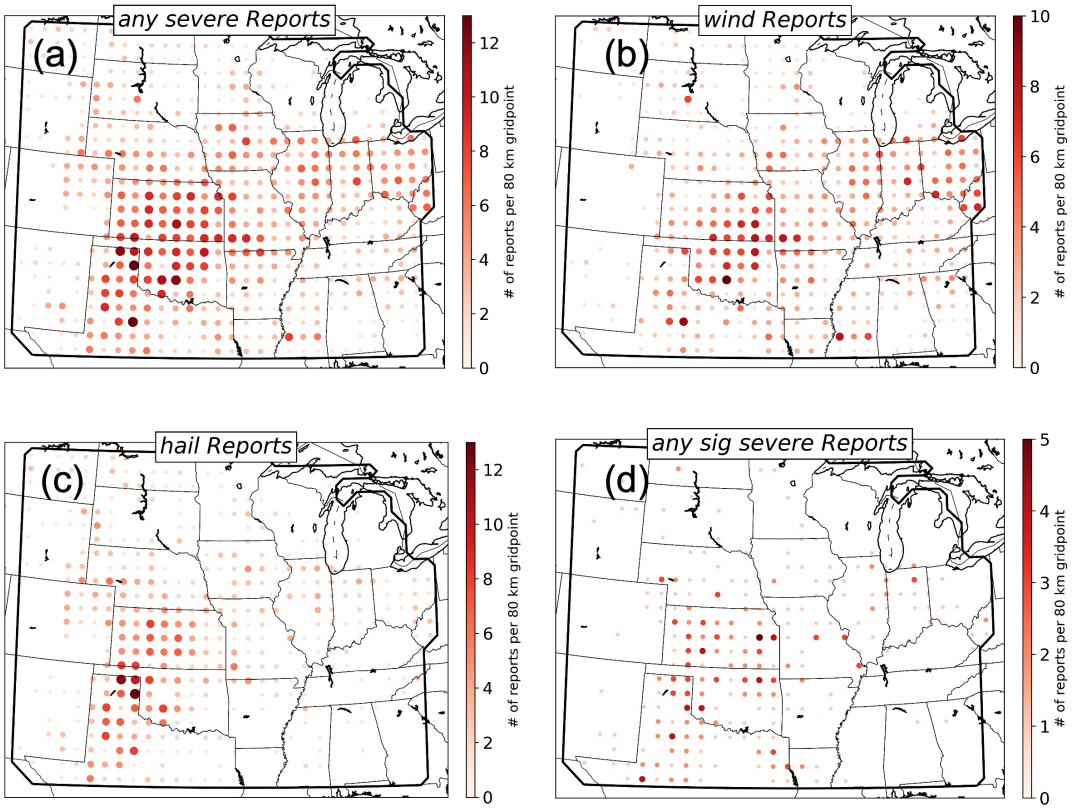

Figure 2: Spatial frequency and distribution of severe weather reports for 80 km points, for the 4 largest report samples. Size and color are based on size with an interval of 0.75 for for visual aid. (a) any severe, (b) wind, (c) hail, and (d) any sig. severe. Forecast domain (black line) for reference

models. There are four groups of fields: storm-attribute, environmental, convective-related, and other (encompassing latitude and longitude). For simplicity in the physical quantification of the predictors, the convective-related predictors are grouped as environmental predictors yielding only three groups for comparison.

Random Forests (RFs)

RFs are made up of an ensemble of decision trees and can either be used for regression or classification (Breiman, 2001). For this study, RFs are used as a classification method following past studies (e.g., Hill et al., 2020; Herman and Schumacher, 2018b; Zeng et al., 2022; Medina et al., 2019; Ahijevych et al., 2016). This simplifies the relationship of the resolved CAE variables to the severe weather reports. Furthermore, RFs were used for this thesis since they are parallelizable, do not require standardized input, require less hyperparameters to tune, and have physical interpretability (Loken et al., 2019). These are what makes them particularly attractive for this study. There have been numerous successful studies implementing RFs as a means to post-process NWP outputs to produce probabilistic forecasts (Clark and Loken, 2022; Gagne et al., 2014; Loken et al., 2019, 2020; McGovern

80 km RF Predictors			
Storm-Attribute Fields[5]	Environment-related Fields[3]	Convective-related Fields	Other[3]
Maximum Updraft Helicity 2-5km (uh25)	2m - Temperature (T2m)	180mb ML CAPE[1] (mlcape180)	Latitude (lat)
Maximum Updraft Helicity 0-3km (uh03)	2m - Dewpoint Temperature (Td2m)	180mb ML CIN[2] (mlcin180)	Longitude (lon)
Maximum dBZ 10c (dbz10c)	10m - u-wind (u10)	90mb ML CAPE[1] (mlcape90)	-
Maximum dBZ 1km (dbz1km)	10m - v-wind (v10)	90mb ML CIN[2] (mlcin90)	-
Maximum Hourly Precip (hrprecip)	mean sea level pressure (mslp)	225mb MUCAPE[1] (mucape255)	-
Maximum Windspeed 10m (wspd10m)	precipitable water (pw)	225mb MUCIN[2] (mucin255)	-
Column Hail (colhail)	850 mb - u-wind (u850)	surface-based LCL (sblcl)[3]	-
-	850 mb - v-wind (v850)	SBCAPE[1] (sbcape)	-
-	850 mb - Heights (z850)	SBCIN[2] (sbcin)	-
-	850 mb - Temperature (T850)	0-1km Helicity[3] (hlcy01)	-
-	850 mb - Dewpoint Temperature (Td850)	0-3km Helicity[3] (hlcy03)	-
-	700 mb - u-wind (u700)	-	-
-	700 mb - v-wind (v700)	-	-
-	700 mb - Heights (z700)	-	-
-	700 mb - Temperature (T700)	-	-
-	700 mb - Dewpoint Temperature (Td700)	-	-
-	500 mb - u-wind (u500)	-	-
-	500 mb - v-wind (v500)	-	-
-	500 mb - Heights[4] (z500)	-	-
-	500 mb - Temperature (T500)	-	-
-	500 mb - Dewpoint Temperature (Td500)	-	-
-	250 mb - u-wind (u250)	-	-
-	250 mb - v-wind (v250)	-	-
-	250 mb - Heights (z250)	-	-

Table 2: 80 km predictors available to train RF models. Superscripts denote ensemble and temporal aggregation type. For example, a predictor type of 1 is a mean-max, where the first aggregate denotes ensemble aggregation and second denotes temporal aggregation. Predictor Types: 1 - mean-max single time, 2 - mean-min single time, 3 - mean-mean single time, 4 - 500mb Height special case, 5 - max-max hourly max

et al., 2017). They are quite powerful in the sense they can ingest extremely large amounts of data and find important patterns that can summarize information from CAEs quickly and effectively. Since RFs can be trained using the resolved CAE variables as predictors and severe weather reports as targets, relationships can be identified between them acting to essentially resolve even smaller scale events. RFs require pre-processing of the global-scale or CAE variables in order to be quick and effective. There are two steps that are employed to reduce dimensionality of the variables, a temporal and a spatial step. Predictors are also selected in some manner to use for training the model.

Building blocks of the RF algorithm begins with decision trees. A single decision tree recursively splits a dataset based off the most optimized split of a node. Optimized is based off a feature (i.e., predictor) and feature value that minimizes the impurity of a sample. Note

every feature is considered in determining the optimal split. The splitting continues until all leaf nodes are pure (i.e., only containing samples that are all true or false). However, decision trees alone are deterministic and optimized to a given training data set thus heavily prone to overfitting. RFs alleviate this issue with an ensemble of decision trees. The RF algorithm works by growing numerous unique decision tress. From which, stochasticity makes them unique. This is manifested in the decision trees in two ways. The subset of training samples are determined by bootstrap resampling and the splitting of the nodes are determined by random subsets of variables. Once all the decision trees have been split and, either all terminal nodes are pure, or a certain criteria is met, a democratic processes follows. All decision trees will have their own final classification for the training set and majority vote rules. For example, if there are 500 decision trees and 300 of those classify something as a 0, then the classification issued by the RF model will be a 0. Expanding upon this for the probabilistic forecasts, each grid point uses the RF classifier that takes in a new sample and runs it through every tree in the forest. Instead of a pure classification of a 0 or 1 being classified at that grid point, a probability is given that is based off the mean fraction of training samples that are associated with a given class at the relevant leaf node across all decision trees. Furthermore, RFs are also attractive due to their relative insensitivity to parameters in terms of model performance (Herman and Schumacher, 2018a). For this study, all RFs had a forest size of 1000 trees, maximum tree depth of 10, a minimum of 30 samples per leaf, square root of the total number of features for maximum number of features considered at each branch, max samples 0.25 of the total for building each tree, and splitting criterion set using entropy. All the hyperparameters were consistent for all RF's. These hyperparameters were based off sensitivity tests performed in Herman and Schumacher (2018a) with the most deviant parameter being forest size from previous studies. However, this does not impact skill. This was discussed in Herman and Schumacher (2018a), where forest size skill increases with larger forests and asymptote's at some level, but beyond this the trees just become redundant and computationally expensive. The RFs and RF probabilities were created using random forest classifiers from the Python module Scikit-Learn (Pedregosa et al., 2011). Discussion of the pre-processing of predictors, experiment design, train/test split, predictor selection, and differences among the experiments are explained further in the next few sections.

Pre-processing of Data

a.) 80 km Predictors

For all 10 members of the CAE, the following forecast variables were pre-processed as follows. In order to conserve the environment-related and convective-related forecast variables, they were remapped from the 3 km grid to the 80 km grid using nearest neighbor-averaging (i.e. remapping) as explained in Accadia et al. (2003). For the storm-attribute output variables, they used a maximum output value from each 3 km grid point within the 80 km grid. This effectively created 80 km predictors available to the RF as shown in Table 2. Due to the high dimensionality of the CAE output, multiple steps are used to further reduce dimensionality of the dataset to make the RF computation feasible during training (Loken et al., 2020). This is accomplished by aggregating all the ensemble members and

the temporal dimension. For the environment-related variables, the temporal mean of the ensemble mean was used as this was the representative value over the entire forecast period. For the convective-related variables, the temporal maximum (or minimum) was taken of the ensemble mean. This was representative of the ensemble prediction at time that was most favorable for severe weather. Lastly, the storm-attribute predictors used the temporal maximum of the ensemble maximum representative as the "worse-case" or most severe storm within the ensemble during the forecast period. Though more of the ensemble distribution or temporal distribution could've been utilized, the simplicity of the aggregations used here reflect at least one or more aggregation other studies have used (e.g., Loken et al., 2017, 2019, 2020; Clark and Loken, 2022). Furthermore, to reduce complications of how different aggregations could impact the RF skill for each predictor, thus deviating from the scope of the smoothing of the predictors, these aggregations were representative enough for the corresponding predictors.

b.) Multiscale Predictors

Obtaining the multiscale predictors required one more step increasing the spatial scales of the original 80 km predictors via spatial smoothing. To start the smoothing process, first ±0, 1, 3, and 9 80 km gridboxes were created around every 80 km gridbox to define a larger spatial box. Note ±0 is the un-smoothed 80 km gridbox. For example, a ±1 80 km gridbox would yield a 240 km × 240 km box including the original 80 km gridpoint. Next, following similar procedure for creating a neighborhood around a given point using the geometry of a circle as in (Schwartz and Sobash, 2017), smoothing radii (i.e. radius of influence) were applied to all newly defined gridboxes from the previous step corresponding to the number of 80 km gridboxes that makes up the larger spatial box: $r = 0, 1, 3, 9 \times 80$ km where r is the number of grid points away from a given 80 km gridbox point. To clarify, the smoothing radius is equivalent to $r \times 80$ km. In order to effectively represent the spatial smoothing of each predictor, an "effective diameter" is defined that is representative of the number of 80 km gridboxes that are included within the smoothing radii. For example (Fig. 3), for a smoothing of ±1 80 km gridbox, the width of 80 km gridboxes is 240 km whose gridpoints are within the smoothing radius, thus resulting in a smoothed predictor considered to be at the 240 km scale. With this definition, the categories of spatial scales are then defined as 80 km, 240 km, 560 km, and 1520 km. This procedure is calculated for all model output variables, length-scales, and gridpoints within the domain,

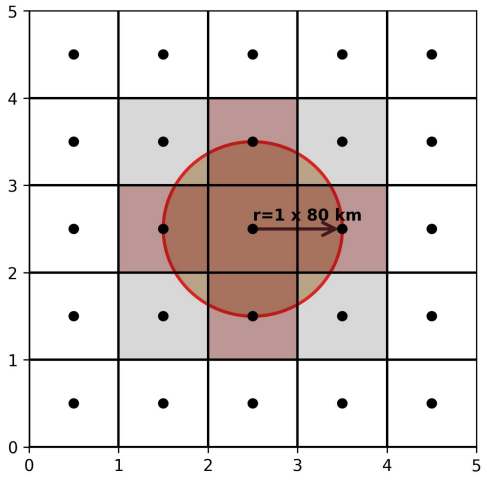

Figure 3: Schematic example of smoothing procedure for effective diameter = 240 km. Black dots representative of 80 km gridpoints. Gridboxes representative of 80 km grid boxes. Grey filled boxes denote ±1. Smoothing radii r is equivalent to number of grid boxes to smooth over × 80 km. Maroon filled boxes denote actual gridboxes included in averaging making up effective diameter

effectively creating predictors that utilize increasingly larger spatial scales.

Characteristic RF models

A RF model is created for each severe weather hazard (8) and forecast period (2). In addition to test and understand what impacts the multiscale predictors had on the forecast skill of a RF model, two "characteristic" RF models were also developed. Thus, 32 total RF models were created. Characteristic is defined here as the nature of which predictor scheme a given RF model is trained on. The first characteristic model, is the Control RF model (CTLRF) which is trained with only the 80 km predictors. The second characteristic model, is the experimental RF model (EXPRF) which is trained with 80 km, 240 km, 560 km, and 1520 km predictors.

Training and Testing

Training and testing of the RF models started out with 50 total cases from Table 1. Of those 50, 30 were randomly selected for training to break any potential highly correlated days that could follow subsequent severe weather events. The remaining 20 cases were used for testing and verification of the RF models. Both CTLRF and EXPRF models are each trained for forecasting any severe weather, wind, hail, and tornadoes as well as their significant severe counterparts for a 24hr and 4hr period.

At this point, a distinction between this thesis and others' work is the method of predictor selection for training. Typically, predictors are chosen in a subjective way that matches current knowledge of relationships between the predictand and predictors. However, for this study the nature of the physical relationships of multiscale predictors to the predictand's are not necessarily known. The same variables, but at different scales as predictors are available which could lead to redundancy. There is also a need to build a simple yet skillful RF model for physical interpretation. For those reasons, use for an objective predictor selection method was warranted.

Forward predictor selection (FPS)

An objective method for predictor selection called Forward predictor selection (FPS) (see McGovern et al., 2019; Jasper Velthoen and Jongbloed, 2023) was used for this study. It is an objective method that iteratively builds the RF models from a null RF model to an optimized RF model. This is used since it's been shown that including too many variables often leads to a decrease in statistical efficiency and degrades model interpretability (Jasper Velthoen and Jongbloed, 2023). Furthermore it reduces noisy and/or redundant predictors (Hall et al., 2011; Ahijevych et al., 2016) and has been shown to attain the most optimal set of predictors (Jasper Velthoen and Jongbloed, 2023). Thus, FPS can accomplish the task of building simple yet skillful RF models. The procedure starts with a null RF model (e.g. 24-hr CTLRF forecasting any severe weather hazard). Then a corresponding number of RF models to available predictors are trained. Some verification metric is calculated, in this case BSS, for all RF models trained. The RF model trained with the predictor that yields the highest

BSS is the first predictor selected for the null model. After designating the first predictor for the null model, this becomes the "single-predictor" model from which is then trained with each remaining predictor (in addition to the predictor that was selected) once more. Now, in the case of the predictors available to the CTLRF models, 45 RF models are retrained with the retained predictor in the previous iteration. The predictor appended to the RF model yielding the highest BSS is the next predictor selected now making a two-predictor model. These iterations continue until some criteria is met. For this study, this criteria was the difference in BSS between the "current" iteration and the previous iteration being ≤ 0.001, indicating a very marginal addition of skill or a drop in skill. Once this criteria is met, the RF model is complete. This procedure is employed for all RF models created for each severe weather hazard and forecast period. Another goal of using this method was to reduce the computational cost associated with the EXPRF models using 174 possible predictors that would likely be highly correlated since the additional predictors are smoothed variations of the forty-five 80 km predictors.

RF Verification

In order to verify the performance of the RF models, summary metrics such as brier skill score (BSS), brier score (BS) and the components of BS (Wilks, 2019): reliability (REL), resolution (RES), and area under the relative operating charactersitic (ROC) curve (AUC) as well as graphical devices such as performance diagrams (Roebber, 2009), attribute diagrams (Hsu and Murphy, 1986; Wilks, 2019), and ROC curve diagrams were utilized (Wilks, 2019). The brier score is a summary metric for the magnitude of probabilistic forecast error for a given model (Loken et al., 2020). It is negatively oriented such that a 0 indicates perfect forecasts and 1 indicates no skill. The BS can be algebraically decomposed into three components (Wilks, 2019): reliability, resolution, and uncertainty.

$$BS = \frac{1}{N}\sum_{i=1}^{I}(p_i - o_i)^2 = \frac{1}{N}\sum_{k=1}^{K} N_k(p_k - \overline{o}_k)^2 - \frac{1}{N}\sum_{k=1}^{K} N_k(\overline{o}_k - \overline{o})^2 + \overline{o}(1-\overline{o}) \tag{1}$$

where N is the total number of forecasts/observation pairs, p_i is the forecast probability at point i, o_i is the binary observation (0 or 1) at point i, K is the number of forecast probability bins, N_k is the number of forecasts in the k^{th} probability bin, p_k is the probability for the k^{th} bin, $\overline{o}_k$ is the average relative frequency of observations (0 or 1) in probability bin k, and finally $\overline{o}$ is the average observation frequency for all forecasts (i.e., climatology). The first term is the reliability, second term resolution, and last term uncertainty. Reliability can be thought of as the error between forecast probabilities and observed relative frequencies, resolution is the mean squared difference between observed relative frequency and climatology, and uncertainty is essentially the brier score of climatological forecasts (Wilks, 2019). Reliability is a good summary metric to assess forecast skill since it explains how well-calibrated a forecast model is. It is negatively-oriented so the larger (smaller) reliability is the poorer (better) the calibration of the model is. Resolution is also another good summary metric in that it explains how well a forecast model is able to discern differences in probabilities of events. It is positively oriented so higher (lower) resolution means greater (lower) differentiation of forecast probabilities and events. Lastly, uncertainty reflects the climatological

frequency of the event. Although, the brier score clues into how "accurate" a forecast model is, there is a need to compare forecast skill of the model to that of some reference forecast skill.

The BSS, defined in Eq. (2), is the fractional decrease in probabilistic forecast error relative to some reference forecast (Wilks, 2019).

$$BSS = \frac{BS - BS_{ref}}{0 - BS_{ref}} = 1 - \frac{BS}{BS_{ref}} \tag{2}$$

Typically, this reference forecast is climatology. BSS is positively oriented where a 1 means perfect forecasts. Unlike brier score, BSS can have negative values where values ≤ 0 indicates no skill. This metric provides a valuable assessment for summarizing how well one forecast model is relative to some reference forecast model. Furthermore, since trivial negatives can impact brier score, the BSS results are considered much more heavily (Loken et al., 2022; Wilks, 2019). In order to discern significant differences in forecast skill between the two characteristic RF models, permutation resampling of the BS differences (Johnson et al., 2013) was used to assess significance of skill between the RF models. Resampling verification statistics are important for this case since independent sampling cannot be assumed.

$$BSS_{\text{diff}} = 1 - \frac{BS_{\text{exp}}}{BS_{\text{ctl}}} \tag{3}$$

In order to make a fair comparison of the EXPRF and CTLRF models, the difference in BS that will be tested for significance is defined as the ratio of the EXPRF probabilistic forecast error and CTLRF probabilistic forecast error as shown in Eq. (3). This method is applied to test significance with 80% or more confidence (i.e. $p_{value} \geq 0.2$) of the difference in forecast skill for all severe weather hazards and periods listed previously. In addition, the differences in reliability and resolution were also tested for significance using a similar method as that used for testing the BSS differences.

AUC is a summary metric for the ROC curve diagrams (Wilks, 2019). It is positively oriented. AUC measures how well probability forecasts are able to discriminate a dichotomous event. It is typically calculated using a trapezoidal approximation for calculating the area under the ROC curve. 0.5 or less is considered an unskillful forecast whereas > 0.7 is considered the lower limit of skillful forecasts (Buizza et al., 1999). Since the ROC curve plots the probability of detection (POD) versus the probability of false detection (POFD) for some probability threshold, the closer the data points are to the top-left corner the larger the AUC. However, as discussed in (Loken et al., 2020) it is not uncommon for forecast models to have AUC's ≥ 0.9 due to POFD often times being reduced due to correct negatives when forecasting severe weather events (especially when the events are rare).

Attribute diagrams are used for assessing quality of samples of probability forecasts for depicting REL and RES (Hsu and Murphy, 1986). It plots the conditional relative frequency of observations given each forecast probability against forecast probability. This was used as a way to show how well the models are calibrated and how well models forecasts are able to discern differences in forecast probabilities of events.

Performance diagrams (Roebber, 2009) are a graphical device that compliment ROC curve diagrams. They plot the success ratio (SR) versus POD for some probability threshold making it much more sensitive to incorrect forecasts when an event didn't occur. They also include frequency bias spikes from the origin of the plot and contours indicating critical success index (CSI). Values are optimized at 1 so values closer to the top-right corner yields higher performance. Since POD is related to resolution and FAR is related to reliability, lower or higher FAR or POD, respectively, would impact the BS which in turn would impact the BSS. Thus, it is a very useful verification tool for further understanding BSS.

RF Physical Interpretation/TI Module

In order to perform physical diagnostics of the RF models for determining how the multi-scale predictors impact the RF models, a similar method to that used in (Loken et al., 2022) was employed using a python-based module called TreeInterpreter (TI). This module facilitated the analysis for discerning predictor importance of the RF models as well as understanding what the RF models have learned.

Tree Interpreter (TI) is a python-based model that takes a given testing sample and runs it through each decision tree in the forest, recording at each node how a predictor impacts the training sample purity (or training climatology). From which the sum of the impacts (or contributions) are made over all nodes in the forest and reports the mean contribution of a predictor. More information and specifics can be found in (Loken et al., 2022). In summary, TI features the ability to decompose the final RF probability into components of bias and predictor contribution

$$P_r(RF) = \text{Bias} + \sum_{i=1}^{N} P_i \tag{4}$$

where $P_r(RF)$ is the final RF probability, Bias is the training sample climatology, P_i mean probability contribution from i^{th} predictor, and N is total number of predictors in RF model. In addition to this decomposition, the components are also stratified based off class. For this case, TI stratifies the decomposed final RF probability into the binary "no" reports (o_0) and "yes" reports (o_1). Note that since the classes are binary, the magnitudes will be the same for both classifications but one will be a positive contribution and the other will be negative and vice versa. Due to this, in order to assess predictor contribution to forecast skill the entire testing sample was separated into two sub-samples, one with no reports and the other with reports. Contribution to forecast skill was considered when there were negative average contributions from a predictor (i.e. decreasing probabilities) for the class o_0 in the no report sample. Likewise, contribution to forecast skill was considered when there were positive average contributions from a predictor (i.e. increasing probabilities) for the class o_1 in the yes report sample. Furthermore, considering that each RF model had different predictors selected based off FPS, the predictors average contributions were grouped together based off predictor-type (i.e. environment, storm-attribute, or lat/lon) and smoothing spatial scale allowing for comparison of contributions to forecast probability despite the differences in individual predictors selected. Note that the convective-environment predictors were grouped

with the the environment predictors for simplicity. The groups were simply the sum of the average contributions from the predictors within that group. This is because a given final prediction is the bias and sum of the individual predictors' average contributions. It wouldn't make logical sense to discuss the grouped contributions as an average as this wouldn't equate to the final RF probability as if it was just from the predictors themselves.

To summarize all this information, stacked bar plots were chosen. Though it is not clear which individual predictors are shown and groups can sometimes have only one predictor, it summarizes the most important contribution information from the multiscale contributions against the 80 km predictor contribution well. Lastly, since the CTLRF and EXPRF models were trained with different predictors that were chosen by the FPS, a direct comparison of contributions from the CTLRF and EXPRF models was not performed, but inferences were made based off the skill differences and how the 80 km predictors that were selected for the EXPRF model contributed against the multiscale predictors selected.

Results

RF Predictors Selected

Predictor selection for the RF models was based off the FPS method discussed in Forward predictor selection (FPS). Each characteristic RF model (i.e. CTLRF and EXPRF) was built with predictors that yielded the highest BSS. For both periods and both CTLRF

FPS Random Forests CTLRF 12-12 and 20-00 UTC Models

Severe Hazard	Predictors (24 hr)	Predictors (4 hr)
Any Severe	uh25, sbcape, v500, wspd10m, T500, dbz1km, mslp	uh25, dbz10c, sbcin, v850, mslp, u10m, T250
Wind	colhail, wspd10m, T500, dbz1km, T2m, uh03, v10	colhail, v850, z500, T250, sbcin, Z250, lon
Hail	uh25, wspd10m, mlcape90, lon, lat, Td2m	uh25, mlcape180, lon, sbcin, u10, dbz10c, mslp
Tornado	colhail, v500, u500, dbz1km, wspd10m, mslp, sbcape	colhail, v500, u500, z700, sbcape, lat
Any Sig.Severe	uh25, v500, mlcape255, u500, mslp, T700, lon, Td500	T850, T700, dbz1km, u250, Td2m
Sig. Wind	uh25, lat, lon, mslp	sbcin, z500, lat, uh25, mslp
Sig. Hail	uh25, lon, mlcape90	mucape255, Td2m, dbz10c, Td850
Sig. Tornado	z850, v500, lon	u500, mucin255, u700, u250, sbcin, lon, hlyc1km, T850

Table 3: 24hr (left) and 4hr (right) CTLRF predictors selected from 45, 80 km predictors after each iteration during forward predictor selection (FPS). Each predictor was attained based off the iterative model with the highest brier skill score possessing said predictor.

and EXPRF models, the FPS selected predictors corresponding to environment, convective, and storm-attribute fields. For all of the 24hr sub-significant severe weather hazards, the

FPS Random Forests EXPRF				
Hazard	**r = 80 km**	**r = 240 km**	**r = 560 km**	**r = 1,520 km**
Any (24 hr)	sbcape, wspd10m, mslp, T500	uh25, colhail, wspd10m	v10	v10, mlcin90
Wind (24 hr)	colhail, T2m, sblcl	wsdp10m, T500, dbz1km	uh03	sbcape
Hail (24 hr)	uh25, mlcape90, dbz10c, sbcape, u500, lon	Td2m, wspd10m	pw	mucape255mb
Tornado (24 hr)	lon	dbz1km, v250	uh25, Td700	v10m, Td500
Any sig. (24 hr)	lon	uh25, mslp	hlcy3km, T700, u500	Td500, T250, mlcape90mb, Td700
Sig. Wind (24 hr)	mslp, lon, lat	uh25, hrprecip, sblcl	-	-
Sig. Hail (24 hr)	uh25, lon	-	-	dbz1km, v10m
Sig.Tornado (24 hr)	-	-	v10	z850, z500
Any (4 hr)	mlcape90, mucin255, u10	dbz10c, wspd10m, colhail	-	v10m, mslp
Wind (4 hr)	-	dbz10c, colhail	v10, T2m	T850, dbz10c, mslp
Hail (4 hr)	uh25, mlcape180, sbcin, maxdbz10c, u10m, z850	dbz10c	Td2m	-
Tornado (4 hr)	maxspd10m, lat	-	u10	v700, z700
Any Sig. (4 hr)	mlcin180, mucape255, lon	u500	mlcin180	mucin255, u850
Sig. Wind (4 hr)	sbcin, z500	-	mlcin90	-
Sig. Hail (4 hr)	mucape255, Td2m, dbz1km, Td850	-	-	u10
Sig. Tornado (4 hr)	-	u10	-	sbcin, T850, mlcin90

Table 4: 24hr and 4hr EXPRF predictors selected from 174, 80 km and multiscale predictors after each iteration during forward predictor selection (FPS). Each predictor was attained based off the iterative model with the highest brier skill score possessing said predictor.

EXPRF models utilized predictors representative of all larger spatial scales. In addition to this, all other severe hazards used variations of the larger-scale predictors (Table 4). It was also common for most of the sub-significant severe weather hazards (i.e., 24hr and 4hr any severe, wind, and 24hr tornado) to utilize two or more larger-scale storm-attribute predictor ranging between smoothing levels of 240 km and 560 km. Most frequent of which, was the 2-5 km maximum updraft helicity (maxuh25). However, it was more common for the fixed 80 km convective environment predictors to be selected (e.g., surface based CAPE, 90mb mixed layer CAPE, 255mb most unstable CIN, etc.) over the larger-scale variations, especially for hail forecasts. Also unique to all the hail forecasts, was the substantial use of the fixed 80 km convective and storm-attribute predictors over the larger-scale variations. For the more synoptic influenced predictors (e.g., u and v winds, temperature (T), and dewpoint temperature (Td) of varying pressure levels) the larger-scale variations were selected more often.

There also appeared to be more predictors selected for the EXPRF model that would provide greater predictability of the afternoon severe weather (20-00 UTC) parameter space (SBCAPE, MUCAPE255mb, SBCIN, etc.) than were selected for the CTLRF model (Table 3 right column and Table 4 bottom 8 rows)

CTLRF Verification

Qualitative Overview of Forecasts

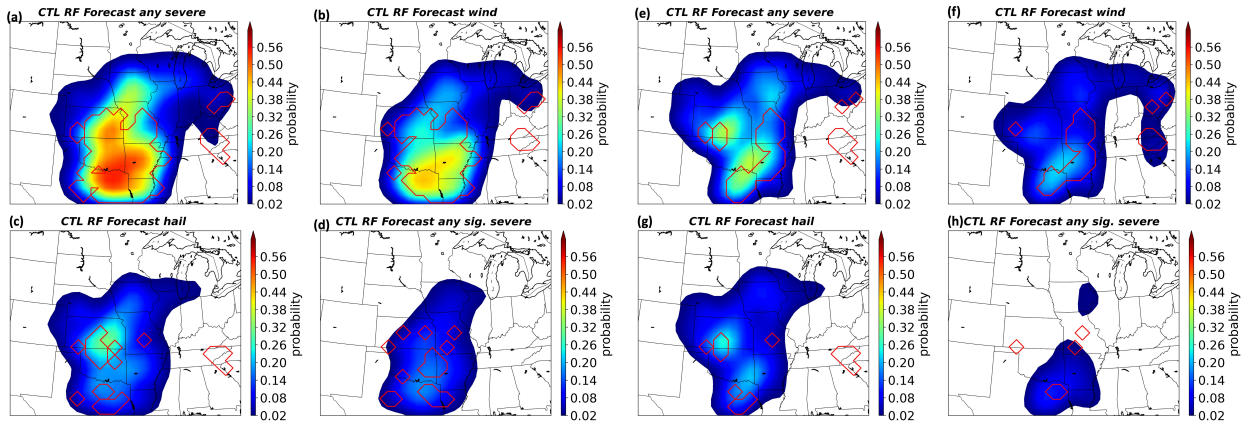

Figure 4: CTLRF forecasts for May 18th, 2019. Filled contours are continuous probabilistic forecast values contoured at 0.2% intervals between 0 and $\geq$ 60%. Red contour lines are region of 80 km storm reports. Left set of 4 panels are for 24hr forecasts for: (a) any severe, (b) wind, (c) hail forecasts, and (d) any sig. severe. Right set of panels for 4hr forecasts for: (e) any severe, (f) wind, (g) hail, and (h) any sig. severe

A case study from May 18th, 2019 was used to assess a baseline of qualitative skill of the CTLRF model in both forecast periods (Fig. 4). Based off radar images and the SPC convective outlook regarding to this forecast period (not shown), the vast majority of the severe weather impacts were due to a dual squall line event. One squall line evolved from southeastern OK and moved across LA while the other moved across OK into southwest and central MO over the 24hr period. Examining the CTLRF model forecasts for 24hr any severe weather hazard, severe wind, severe hail, and any sig. severe (Fig. 4a-d) forecasts were qualitatively skillful. Meanwhile in the 4hr period, there were some discrepancies that promoted lower qualitative skill relative to the 24hr forecasts.

In particular, in the 24hr any severe and wind forecast, the majority of the probabilities were within the report regions, especially the highest probabilities across northeast TX into central AR (Fig. 4a,b). The 24hr hail probabilities were highest in a region of severe weather reports. In addition to appreciable probabilities throughout OK and northern TX, where scattered hail reports occurred (Fig. 4c). Meanwhile, though probabilities were low for 24hr any sig. severe, the trend continued with appreciable forecast probabilities corresponding to regions of reports. In the 4hr period, generally across all 3 forecast hazards (Fig. 4d-f),

higher to the highest probabilities were embedded within regions of severe weather reports. Furthermore, there was a greater tendency for over-forecasting, especially for any severe hazard where the probability maxima in southeast OK lied outside of the main report axis from northeastern TX stretching into central MO (Fig. 4e).

Although the 24hr forecasts were skillful, there were still some forecast errors to be addressed. Among all three 24hr forecasts for any severe, wind, and hail; all had consistently relatively higher probabilities across northern MO, into central IA outside of the report regions (Fig. 4a-d). In addition, equal maximum hail probabilities were forecasted both inside and outside of the region of hail reports in KS, most substantially in northeastern KS (Fig. 4c) and for both 24hr any severe and wind, there was a substantial reduction in probabilities across a large report swath in central MO (Fig. 4a,b). It was also true that CTLRF forecast had one area forecasted with highest wind potential (Fig. 4b; highest forecast probability swath).

Quantitative Overview of Forecasts

To establish a baseline for a quantitative investigation between the CTLRF and EXPRF

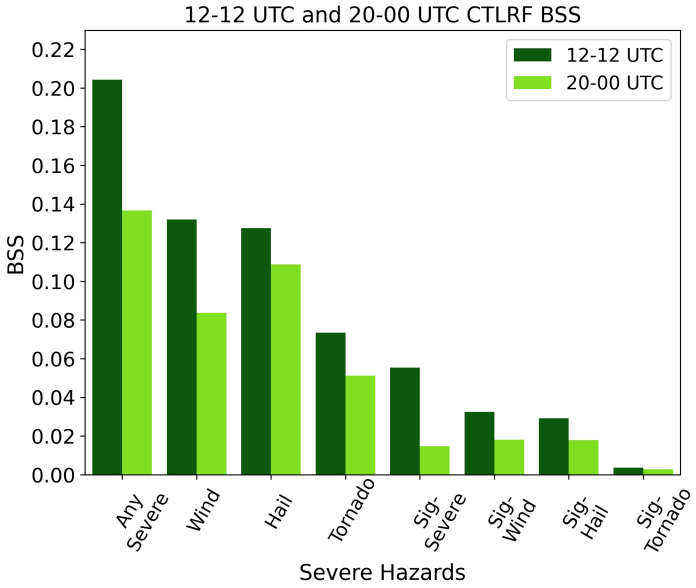

Figure 5: Forecast domain-wide BSS for CTLRF-based probabilistic forecasts for 24hr forecasts (dark green) and 4hr forecasts (light green)

models, Fig. 5 shows the BSS for the 24 hour and 4 hour CTLRF models forecasting all 8 severe weather hazards. Note the BSS shown here was calculated relative to the climatology of the severe weather reports for each hazard, respectively from Eq. (2). In the 24hr period, the CTLRF model forecasting any severe weather hazard boasts the highest skill relative to all other hazards and periods (results similar to that in Loken et al., 2020). As expected for the remaining hazards, skill decreased corresponding to sample size of the reports (e.g., there were more wind reports than there were hail and tornado reports; refer to Fig. 2).

For the 4hr period, once again CTLRF forecasts for any severe weather hazard were the most skillful. However, 4hr hail forecasts were more skillful than 4hr wind contrary to the 24hr hail. For the 24hr and 4hr significant severe weather hazards, the findings for the sub-significant severe hazards translated into the 4 hour period with lower skill, especially where sample size became quite small. Between the 24hr and 4hr CTLRF forecasts, the 24hr CTLRF models were more skillful forecasting all 8 severe weather hazards than the 4hr models, demonstrated in the case study (see Fig. 4)

ROC curve diagrams give a graphical representation of the RF model's ability to classify different events at different probability thresholds. In Fig. 6, it was clear that all RF models

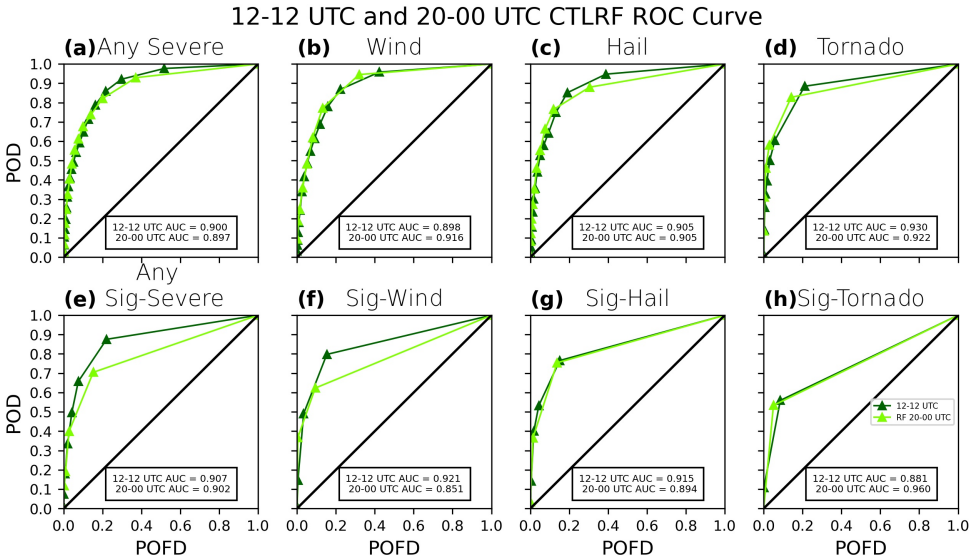

Figure 6: ROC curve plots for probability threshold every 3% with 24hr and 4hr AUC values annotated adjacent to x-axis for a.) any severe hazard, b.) wind, c.) hail, d.) tornado, e.) any significant severe, f.) significant wind, g.) significant hail, h.) significant wind, and i.) significant tornado, 24hr forecasts (dark green line, with dark green triangles marking each probability threshold) and 4h forecasts (same as 24hr except light green)

for both 24hr and 4hr periods were skillful in that they had an AUC > 0.7. However, for the sub-significant severe weather hazards, the calculated AUC's all lie above 0.9. This is common for severe weather due to numerous trivial negatives (i.e. rare events). Since the ROC curve utilizes POFD (which is a function of true negatives), a large number of true negatives can heavily influence the AUC for these RF models (Loken et al., 2020). Despite this issue, the ROC curve still demonstrated the RF models event detection from non-events (i.e. higher POD rates than POFD rates; e.g.,Fig. 6a,d,g,etc.). Similar to what was seen in the BSS results, AUC also agreed with the CTLRF model skill decreasing with severe weather hazard sample size. However, there were some deviations from this big-picture trend. In the 24hr vs 4hr tornado forecasts, the 4hr CTLRF model was better able to discern tornado events than the 24hr. Possibly an artifact of the lower sample size in the 4hr period inflating AUC. Just as in the BSS results, except for a couple nuances, the 24hr CTLRF models were

also more skillful than the 4hr CTLRF models via AUC verification (Fig. 6). The AUC's for the 24hr and 4hr were also comparable to that of previous studies with most forecast hazards attaining AUC's greater than 0.9.(Fig. 6)

EXPRF vs. CTLRF Verification

Quantitative

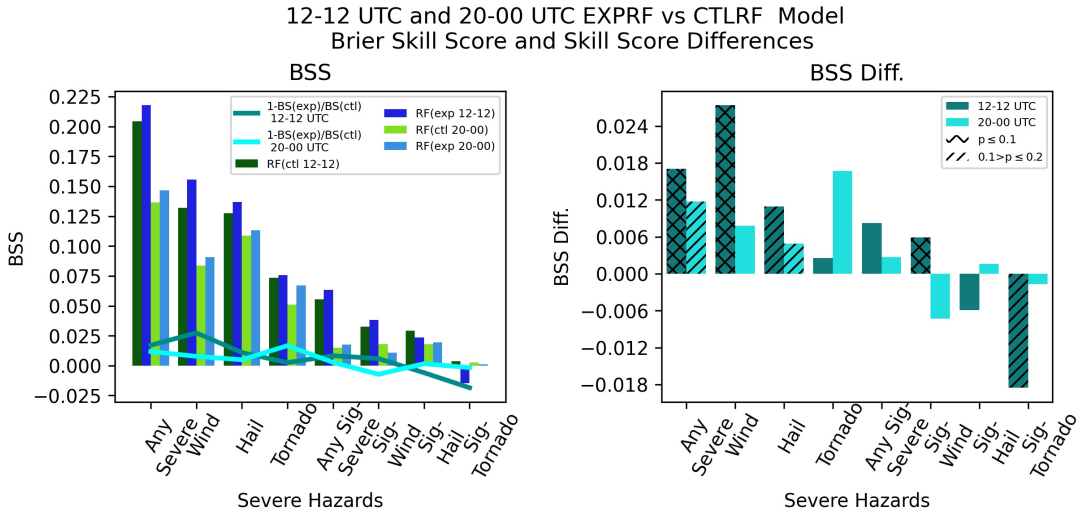

Figure 7: Forecast domain-wide BSS (left) for 24hr CTLRF-based probabilities (dark green bars) and 24hr EXPRF-based probabilities; 4hr CTLRF (light green) and 4hr EXPRF (light blue). Two lines plotting the difference of BSS calculated using Eq. (3) for 24hr (dark cyan) and 4hr (cyan). Difference of BS (right) calculated using Eq. (3) for 24hr (dark cyan) and 4hr (cyan) probabilistic forecasts. x-hatch mark denotes significance with a p-value ≤ 0.1. Forward slash hatch denotes p-values greater than 0.1, but less than 0.2

For the 24hr forecast period, EXPRF had a higher BSS than CTLRF forecasting any severe, wind, hail, tornado, any sig. severe, and sig. wind. In the 4hr period, this was also true for any severe, wind, hail, tornado, any sig. severe, and sig, hail. CTLRF had a higher BSS than EXPRF when forecasting 4hr sig. wind, 24hr sig. hail and both 24hr and 4hr sig. tornado. These results were based off Eq. (2) for each severe weather hazard and period, respectively. In order to give a fair comparison of the model performance against each other (Wilks, 2019), the BSS's were calculated using Eq. (3) treating the reference BS as the CTLRF BS. When the BSS of the EXPRF model was calculated with respect to CTLRF model, the results were in agreement with what was found when calculated relative to respective climatology, for the 24hr and 4hr forecast period (Fig. 7 left; line plots).

The EXPRF probabilistic forecasts with the most statistically significant differences from the CTLRF probabilistic forecasts (i.e. $> 90\%$ confidence) were for 24hr any severe, wind, and significant wind (Fig. 7 right). The EXPRF forecasts were also significant at a minimum level of 80% significance for 4hr any severe and hail. Although CTLRF forecasts for

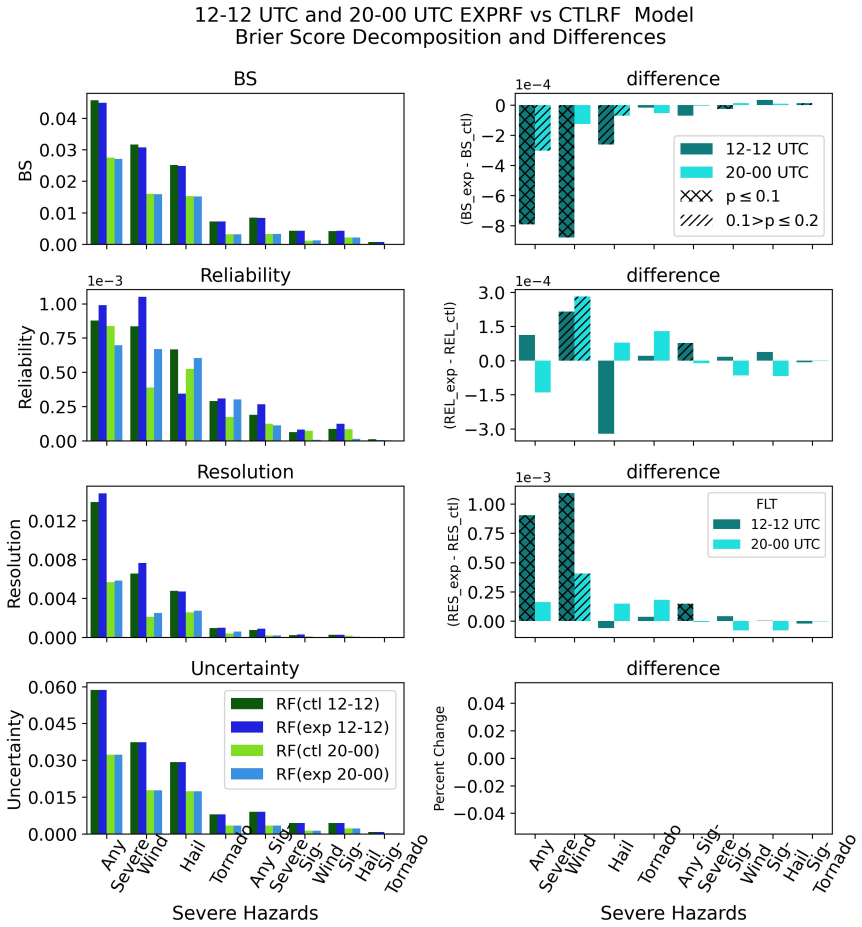

Figure 8: Decomposition of brier score components among all 8 severe hazards. (left) each row corresponds to (top to bottom) brier score, reliability, resolution for 24hr and 4hr CTLRF and EXPRF probabilities (same color scheme as in BSS figure). (Right) is the difference between EXPRF and CTLRF (brier score, reliability, resolution, and uncertainty). X-hatch marks denote significance difference at a p-value $\leq$0.1 and forward-slash indicate a p-value larger than 0.1, but smaller than 0.2. Last row, last column difference in climatology for the same forecast period and hazard is zero.

24hr significant tornado were also statistically significant at this level, this was likely due to bias in forecast probabilities over samples that were almost all "no report", and therefore of negligible practical significance. Meanwhile, though the EXPRF forecast skill differences in the 24hr period for tornadoes and any sig. severe weather hazard were positive, they were not statistically significant suggesting larger sample sizes maybe needed to see a significant difference. Additionally, EXPRF forecast skill differences in the 4hr period for wind and hail were also positive, but EXPRF did not have significant advantage with these forecasts. Overall, except for the vanishingly rare sig. tornadoes, all statistically significant differences favor EXPRF over CTLRF.

While the BSS in Fig. 7 showed for which severe weather hazards and periods EXPRF

was more skillful overall, the BS components aided in distinguishing whether these differences in skill were due to differences in calibration error or resolution. The EXPRF forecasts had statistically significant higher resolution for 24hr: any severe, wind, and any significant severe hazard; 4hr windFig. 7 right column; 3rd row). This was an important metric to favor the EXPRF model considering the calibration error was already small so not much further improvement was needed, thus impacts on BS differences were greater with resolution than were reliability. Although most of the significant differences in BSS between the EXPRF and CTLRF forecasts were due to resolution, 24hr hail was mostly due to substantially lower calibration error (Fig. 7left and right column; 2nd row). Beyond any significant severe hazard, the rest of the significant severe weather hazards resolution for both CTLRF and EXPRF models were very small. Furthermore, when comparing the reliability and resolution of the infrequent and significant hazards (e.g., sig. hail, sig. tornadoes) to their uncertainty, it was quite clear that there was marginal to no advantage from either model (Fig. 8 left column; last row), especially from significant tornadoes.

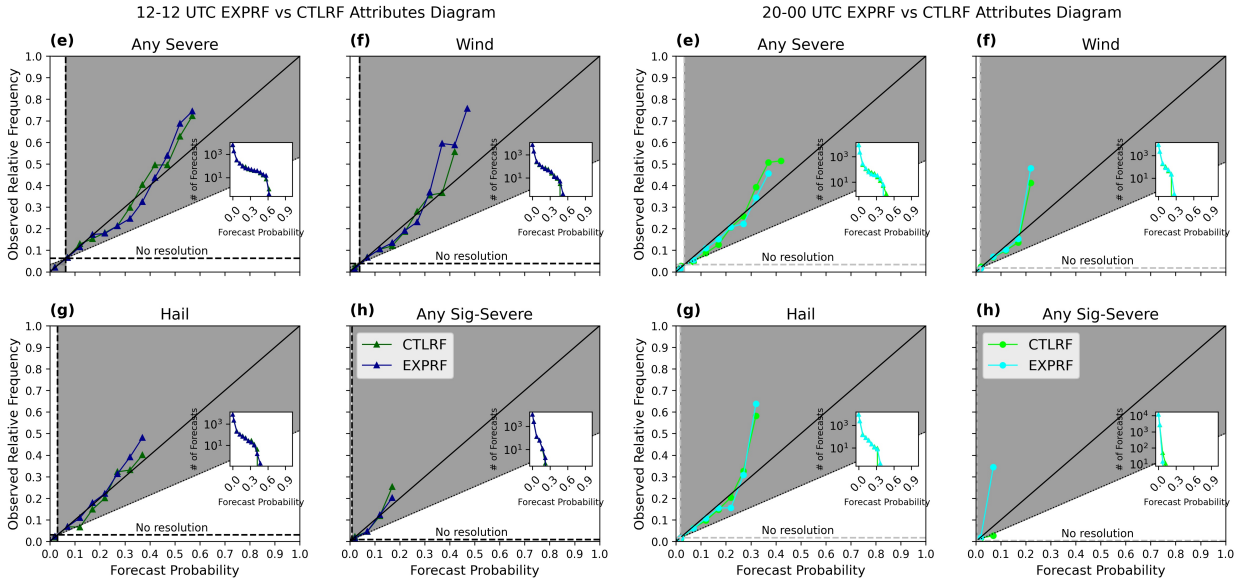

Figure 9: Attribute diagrams for (a) any severe weather, (b) wind, (c) hail, and (d) any sig. severe. 24hr CTLRF (dark green; triangles correspond to probability bin), 24hr EXPRF (dark blue), 4hr CTLRF (light green; circles correspond to probability bin), 4hr EXPRF (light blue). Probability bins are [1%,[5-10%),[10-15%),...,100%] with plotted points representing midpoint of probability bins. 5% intervals provided most smoothing all while persevering important trends. Horizontal dashed line is climatological frequency, vertical dashed line is climatological forecast probability, line that intersects these lines is the no skill line. All points bounded by these values (grey area) indicative larger resolution than magnitude of reliability, thus positive skill to BSS

From an attributes diagram perspective neither model was consistently more reliable. It seemed this was mostly due to both EXPRF and CTLRF having an over-forecasting bias in the middle range of forecast probabilities and an under-forecasting bias at higher forecast

probabilities (e.g., Fig. 9a and b) which contributed to the higher resolution, but poorer calibration. Thus, the calibration error was susceptible of going either way for both models. For instance, EXPRF and CTLRF probabilities for 24hr any severe hazard (Fig. 9a) showed well-calibrated forecasts below 20% forecast probabilities, but both over-forecasted between the 20% and 30% forecast probability bins and under-forecasted forecast probabilities over 30%. For 4hr any severe hazard (Fig. 9b) both EXPRF and CTLRF forecasts under 20% probability were marginally over-forecasted then under-forecasted for forecast probabilities over 30%. Yet in the 24hr period for any severe weather, CTLRF corrected within the mid-range of forecast probabilities while EXPRF was over-forecasting and then in the 4hr period for any severe weather, both EXPRF and CTLRF were over-forecasting most of the forecast probability range, but EXPRF was just slightly over-forecasting less. So the similarities discussed between the EXPRF and CTLRF attribute diagrams were consistent with the skill differences primarily coming from resolution.

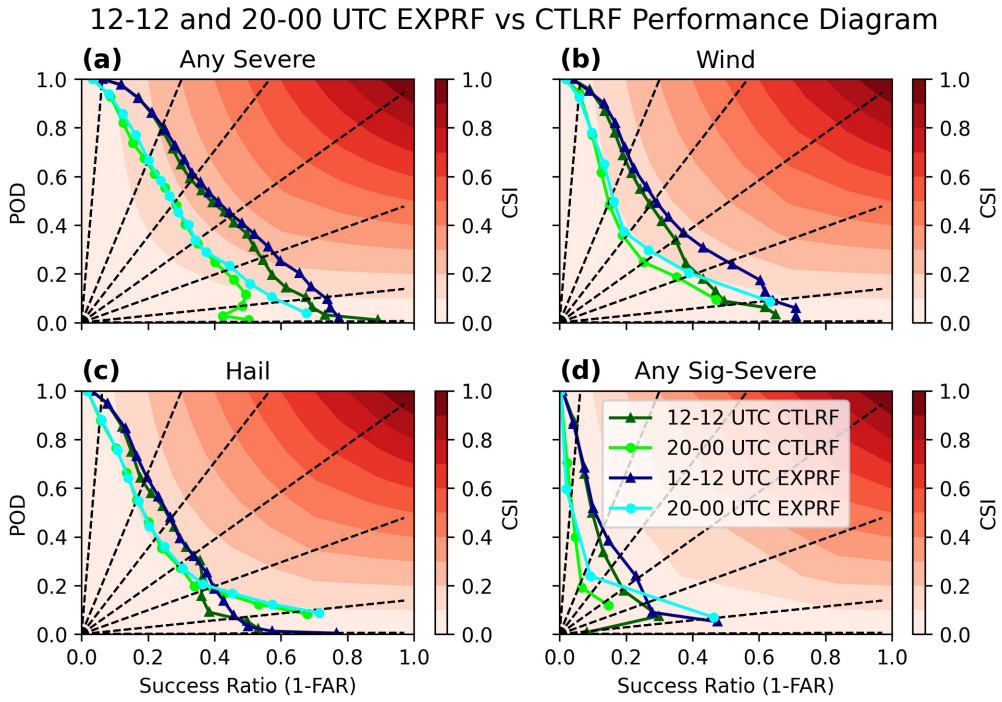

Figure 10: Performance diagrams for (a) any severe weather, (b) wind, (c) hail, and (d) any sig. severe. 24hr CTLRF (dark green; triangles correspond to probability bin), 24hr EXPRF (dark blue), 4hr CTLRF (light green; circles correspond to probability bin), 4hr EXPRF (light blue). Probability bins are [1%,[3-6%),[6-9%),...,99%] with plotted points representing midpoint of probability bins. Black dashed spikes from origin are constant lines of bias; from top-left to bottom-right: 16, 3, 2, 1, 0.5, 0.1, 0.0

Performance diagrams provided an additional perspective on the skill differences between EXPRF and CTLRF forecast performance. Compared to ROC curves the POFD is replaced with SR therefore does not account for true negatives (trivial in severe weather forecasting), which gives a different perspective of forecast skill and detectability over ROC curve

diagrams. The diagrams showed most advantages for 24hr EXPRF any severe weather forecasts, with pronounced advantages in both 24hr and 4hr wind forecasts; but less pronounced for both forecast periods for hail (Fig. 10a-c). For any sig. severe weather, there were some advantages of EXPRF over CTLRF, but the curves were noisier corresponding to smaller sample size (Fig. 10d).

When investigating the 24hr and 4hr EXPRF and CTLRF any severe weather probabilistic forecast performance, EXPRF outperformed CTLRF forecasts at higher probabilities due to forecasts having higher POD and lower FAR and consistently stayed closer to a higher CSI level indicating overall better detection rates than CTLRF. Additionally, even at nearly the same CSI level, EXPRF forecasts for most of the remaining forecast probabilities had slightly higher POD and lower FAR which aided in the EXPRF 24hr any severe weather forecasts having the most advantage overall. EXPRF wind and hail performance (Fig. 10b and c) were the most different from each other with substantially greater detection rates in the 24hr period, due to higher CSI, at the same forecast probability levels for high wind probabilities with appreciable differences in the 4hr. While for hail, there weren't nearly as pronounced differences in both 24hr and 4hr forecasts. Interestingly, the 4hr EXPRF hail forecasts still slightly edged CTLRF all while both attained higher POD and lower FAR than both 24hr RF forecasts.

Case-Studies

Case 1: May 10, 2018

May 10th, 2018 demonstrated some features that were representative of the quantitative results showed previously. EXPRF had higher any severe forecast probabilities that corresponded to severe weather reports over NE (Fig. 11a), but also had slightly higher probabilities than CTLRF down along western KS into northern TX. Just as seen in the performance diagram at the higher probabilities, EXPRF forecasts had higher POD relative to the CTLRF (Fig. 10a). In the case of 24hr wind, there was a greater demonstration of EXPRF increasing and decreasing probabilities appropriately (Fig. 11b) with higher wind probabilities in regions where wind reports occurred and broad swaths of decreased probabilities where reports did not occur. However, just as seen in the reliability diagram (Fig. 9b), the tendency to slightly overforecast at moderate probabilities relative to CTLRF was demonstrated by the increased probabilities in eastern NE and western IA where no reports occurred. Lastly, the 24hr EXPRF hail forecast decreased probabilities in a region where reports occurred, but the large decrease in probabilities across western KS into the panhandle of OK and parts of NE that did not receive reports, outweighed the decreased probabilities within the reports region (Fig. 11c) thus, attributing to slightly higher BSS relative to CTLRF. This partially reflected the well-calibrated lower probability EXPRF forecasts that were seen in (Fig. 9c). Overall, this case reflected the positive BSS difference in favor of EXPRF for any severe, wind, and hail (0.0484, 0.00225, 0.0441) for a modest severe weather event.

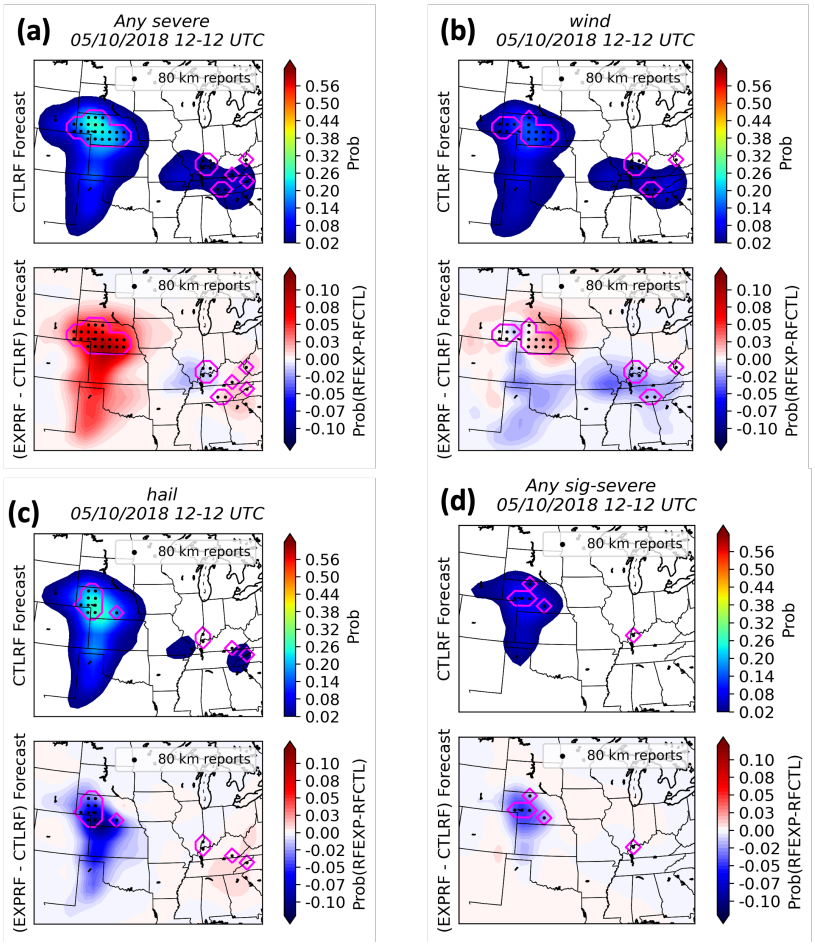

Figure 11: May 10th, 2018 12-12 UTC CTLRF probabilistic forecast map vs. difference of EXPRF and CTLRF probabilistic forecast map. a.) Any severe weather probabilistic forecast. Top panel: CTLRF probabilistic forecast. Filled contours are probabilities every 2%, black dots are 80 km gridded reports, and magenta contours outline 80 km reports for readability purposes. Bottom panel: Probabilistic forecast difference (EXPRF - CTLRF) where red filled contours are positive differences and blue filled contours are negative differences plotted every 0.03 units between -10 and 10 units Panel b.) same as a.) except for 24hr wind. Panel c.) same as a.) except for 24hr hail. Panel d.) same as a.) except for 24hr any sig. severe

Case 2: May 18, 2019

In contrast to the first case, this case offers some insight to understand how RF skill for any severe, wind, hail, and any sig. severe reflect a larger magnitude of the BSS results as seen in Fig. 7 with exception to any severe hazard, corresponding to a robust severe weather event. The 24hr EXPRF model had improvement in forecast skill for all four significantly skilled forecast hazards in the 24 hour period (any severe, wind, hail, and any sig.severe; 3 shown in the figures). Although 24hr any. sig severe was not considered statistically significant, there was an appreciable increase in skill from the EXPRF forecast (i.e. relative to CTLRF about 0.0300 units) as supported in the overall BSS results (Fig. 7).

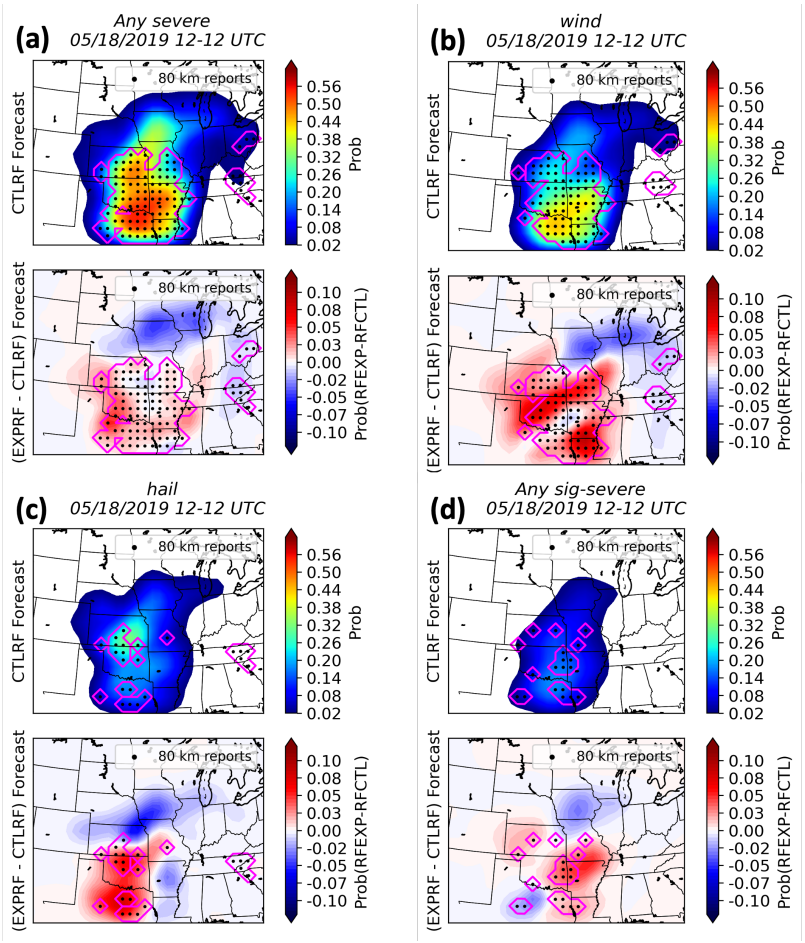

Figure 12: May 18th, 2019 12-12 UTC CTLRF probabilistic forecast map vs. difference of EXPRF and CTLRF probabilistic forecast map. (a) Any severe weather probabilistic forecast. Top panel: CTLRF probabilistic forecast. Filled contours are probabilities every 2%, black dots are 80 km gridded reports, and magenta contours outline 80 km reports for readability purposes. Bottom panel: Probabilistic forecast difference (EXPRF - CTLRF) where red filled contours are positive differences and blue filled contours are negative differences plotted every 0.03 units between -10 and 10 units Panel b.) same as a.) except for 24hr wind. Panel c.) same as a.) except for 24hr hail. Panel d.) same as a.) except for 24hr any sig. severe

Substantial forecast improvements over the CTLRF forecasts were seen for 24hr any severe, wind, hail, and any significant severe weather with a BSS difference of 0.0307, 0.0686, 0.0441, and 0.0300 relative to the CTLRF, respectively. Namely, these forecasts showed improvement due to much higher forecast probabilities within regions of reports, and decreased probabilities across regions that did not have reports. This also reflected the higher POD and lower FAR relative to the CTLRF forecasts that was demonstrated in the performance diagrams (Fig. 10b-d)

In particular, the 24hr wind forecast demonstrated two maximum region of probabilities for severe winds, rather than just one as seen in the CTLRF forecast(Fig. 12b), all while agreeing with the CTLRF forecasts on moderate probabilities across northern AR into northeastern TX. This was representative of the higher resolution that was previously seen in Fig. 8c and much higher CSI in Fig. 10b at moderate-high probabilities. In the 24hr EXPRF hail forecasts (Fig. 12c), there was continued indication of higher resolution owing to higher probabilities in the event of greater hail reports (compared to that of May 10th, 2018) which also resulted in the lower reliability since there were several cases of over-forecasting relative to CTLRF in-between the regions of severe reports.

RF Interpretability

Predictor Importance

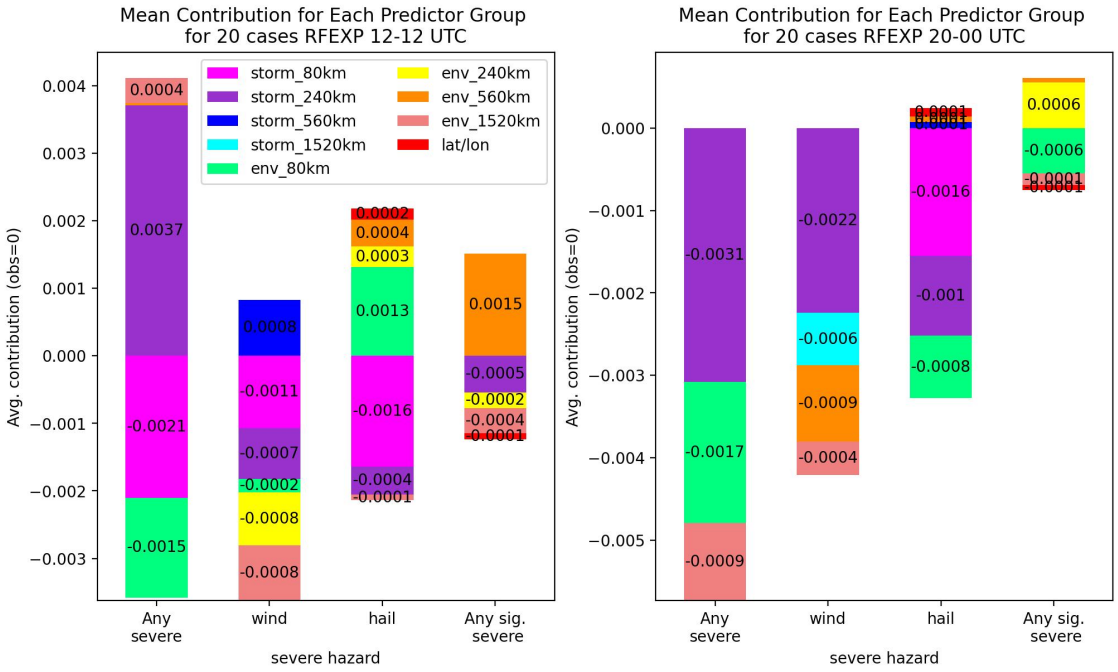

Figure 13: Aggregate of average contributions from all 20 test cases from all predictors sorted into groups based off meteorological characteristics and spatial length scale. Shown for no severe weather reports sample. Blocks are color coded by predictor characteristic and spatial length scale from smallest spatial scales to largest 12-12 UTC forecast period (left), 20-00 UTC forecast period (right). Values within blocks are sum of average contributions from individual predictors within respective group.

As explained in the RF Physical Interpretation/TI Module section, predictors were grouped based off predictor type and spatial smoothing. An extension of this was made to make physical connections of the scales of motion the smoothing was representative of. Thus, from herein the "effective diameters" are represented as synoptic (1520 km), meso-α (560 km), meso-β (240 km), or meso-γ (80 km) scale features related to storm-attributes or

storm-environments. Importance was measured by the average contributions of the predictor groups for a given probabilistic forecast with particular emphasis on the differences between the multiscale and meso-γ predictors. In the following discussion, predictors were considered to be contributing skill to the forecast probabilities when they had a positive average contribution in samples with a severe report, or a negative average contribution in samples without a severe report. For both figures (Figs. 13 and 14) it was important to note the y-axis scale relative to each other noting that when there were severe reports, contributions were much higher. This observation holds true for both case studies because the "bias" terms is already near-zero and the forecast probability is zero-bounded. The purpose of this section is to better understand the specific multi-scale features contributing the skill of the EXPRF model.

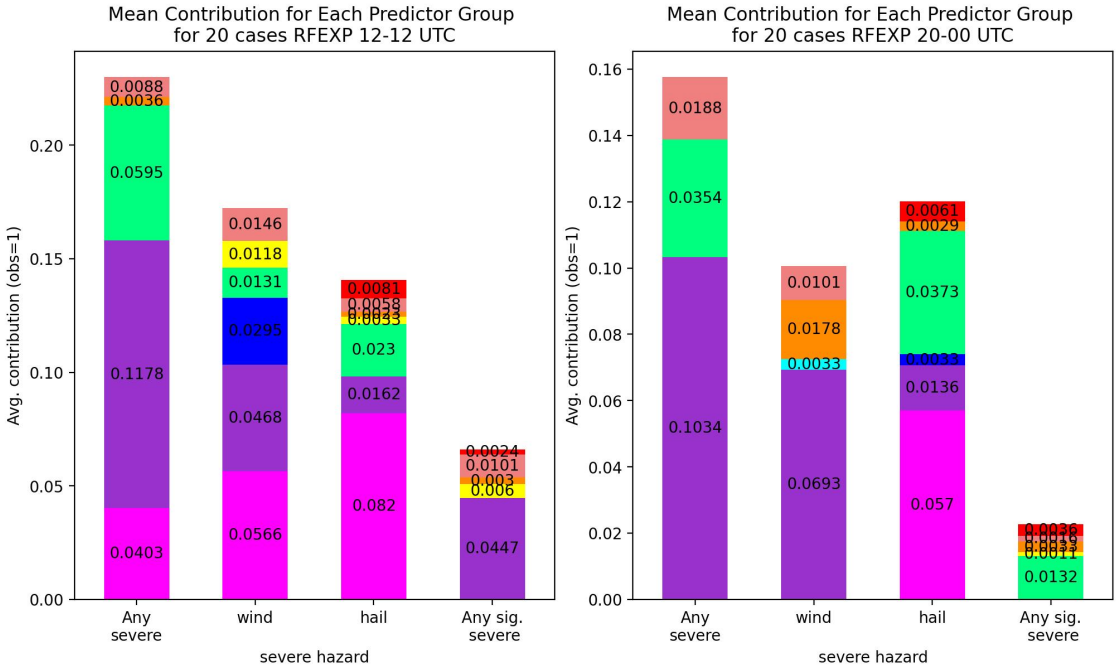

Figure 14: Aggregate of average contributions from all 20 test cases from all predictors sorted into groups based off meteorological characteristics and spatial length scale. Shown for severe weather reports sample. Blocks are color coded by predictor characteristic and spatial length scale from smallest spatial scales to largest 12-12 UTC forecast period (left), 20-00 UTC forecast period (right). Values within blocks are sum of average contributions from individual predictors within respective group.

Overall at locations without a severe report, the storm-attribute predictors' contributions to skill (i.e. reducing probability on average) mainly came from meso-γ predictors for any severe, wind, and hail in the 24hr period (Fig. 13 left) and for hail in the 4hr period (Fig. 13 right). larger-scale predictors, in particular meso-β scale storm-attribute and synoptic scale environment for 24hr any-severe, actually increased forecast probability on average at locations with out a severe report. The meso-β scale storm-attribute predictors had a more dominant contribution to skill for any severe hazard and wind in the 4hr period

(Fig. 13right), while environment predictors mainly contributed to the skill for any sig. severe in both the 24hr and 4hr periods (Fig. 13 left, right).

The overall magnitudes of average contributions were much larger for locations with a report (Fig. 14 left, right) compared to those locations with no severe reports (Fig. 13 left, right) in part due to the already near-zero base rate (i.e., bias term) of the severe weather hazards. All predictor groups contributed to skill in samples with a severe report (Fig. 14). For several severe weather hazards (i.e. 24hr any severe; 4hr any severe, wind) the meso-β scale storm-attribute predictors contributed to skill much more than the meso-γ storm-attribute predictors (Fig. 14). For the other severe weather hazards (i.e. 24hr wind, hail; 4hr hail) the larger-scale meso-β scale storm-attribute predictors did contribute positively, but not as much as the meso-γ storm-attribute predictors. For the environment predictors there were much greater contributions from the meso-γ predictors than the larger-scale predictors, although the larger-scale environment predictors (e.g., synoptic) still contributed to further increase forecast probabilities where severe weather was reported.

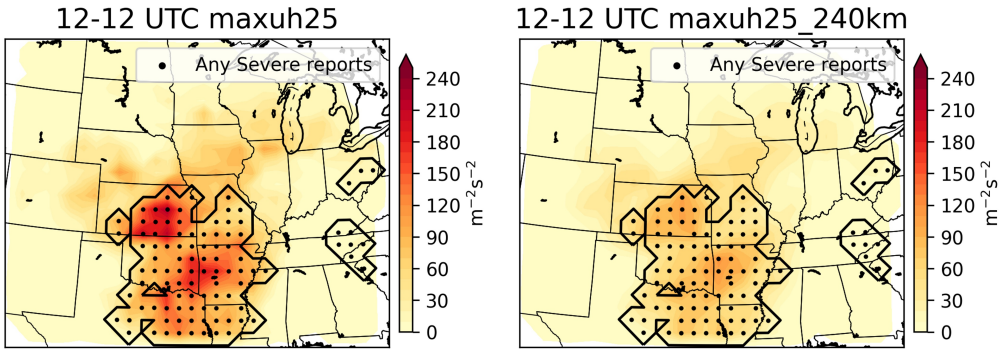

Figure 15: Example of 2-5km maximum updraft helicity with no smoothing (left) and with 240 km smoothing (right). Notice the substantial drop in magnitude, but key regions of max helicity are retained. Report contour (black line) and 80 km reports.

The greater reliance on the meso-β scale storm-attribute predictors for at least several variables, can be understood in terms of the location uncertainty of explicitly simulated storms. In the case of the 80 km maxuh25 field (Fig. 15 left) there were substantial maxima's within the report regions that would obviously benefit the forecast, but compared to the smoothed maxuh25 field there was a benefit such that the key regions of higher maxuh25 were retained, spread out further (greater representation of accounting spatial uncertainty), and noisier local maxima were smoothed out. In contrast to the storm-attribute predictors, environment predictors may have small scale structures (e.g sharp gradients, local maxima) that are related to more predictable synoptic scale features that may not be as beneficial to smooth out. For instance, in Fig. 16a-d, it becomes more obvious as to why the multi-scale smoothed environment predictors would be less likely to contribute to severe weather probabilities especially at the 1520 km smoothing level. Particularly, the sbcape meso-β and meso-α scale smoothing retained the shape of the instability plume, there was a loss of local gradients and local maxima.

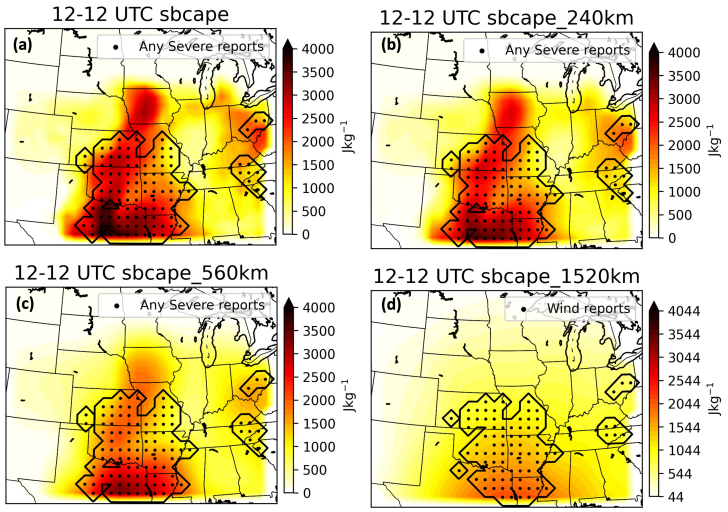

Figure 16: Example of increasing spatial average of surface based CAPE with no smoothing (a) 240 km smoothing (b), 560 km smoothing (c), and 1520 km smoothing (d). General shape of instability plume remains, but local features and gradients are loss, especially by 1520 km smoothing. Report contour (black line) and 80 km reports.

Representative Case-Studies

Two case studies were selected to provide some physical understanding of the physical features corresponding to different TI contributions from different severe weather hazards and of different caliber of severe weather events.

Just as in the aggregate results, May 10th, 2018 had positive probability contributions from the meso-β scale storm-attribute predictors for 24hr any severe hazard when there were no reports (Fig. 17). Also, when there were severe reports, the meso-β scale storm-attribute predictors contributed more on average to increasing skill than the meso-γ storm-attribute predictors for 24hr any severe and wind, while the meso-γ storm-attribute predictors contributed on average, more than the meso-β storm-attribute predictors for 24hr hail (Fig. 18). Furthermore there were positive contributions coming from all multiscale predictors for 24hr wind (Fig. 18) as seen in the aggregate when there were severe wind reports (Fig. 14).

Recall from Fig. 11 that the EXPRF advantage for 24hr any severe hazard on the May 10th, 2018 case resulted from an increase in forecast probability in the region of severe reports in NE that outweighed increases in forecast probabilities down into northwestern TX. The smoothing of the storm-attribute predictors provided a better focus on the region of severe weather reports. In particular, the maxuh25 meso-β scale storm-attribute predictor retained the more favorable region for severe weather after smoothing (Fig. 19a) as well as the columnhail meso-β scale storm-attribute predictor (Fig. 19c), though less in magnitude (Fig. 19f). The smoothed maxuh25 field most closely reflected the increased forecast probabilities for 24hr any severe weather across NE where severe reports occurred (Fig. 19b). Furthermore, the maxspd10m meso-α storm-attribute predictor (Fig. 19b) contributed the

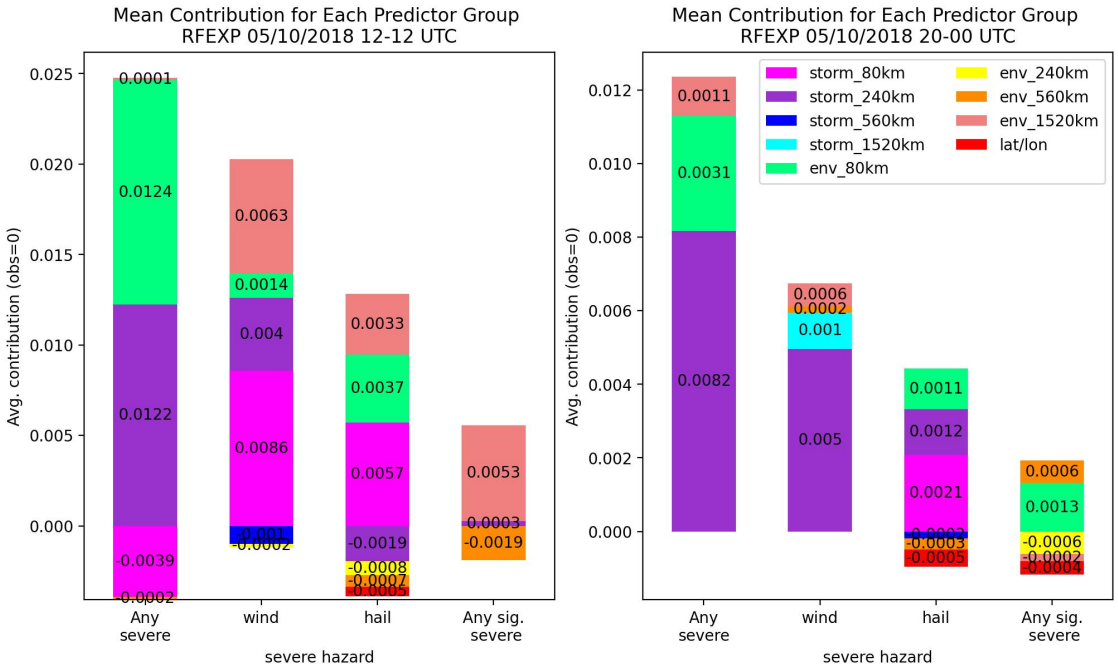

Figure 17: May 10th, 2018 case of average contributions from all predictors sorted into groups based off meteorological characteristics and spatial length scale. Shown for severe weather reports sample. Blocks are color coded by predictor characteristic and spatial length scale from smallest spatial scales to largest 12-12 UTC forecast period (left), 20-00 UTC forecast period (right). Values within blocks are sum of average contributions from individual predictors within respective group.

most positive forecast probabilities in the reports region (Fig. 19e), but the magnitude was largest for the maxuh25 meso-β scale storm-attribute predictor. This was consistent with what was demonstrated in Fig. 15

Recall from Fig. 11 that EXPRF advantage for 24hr wind on May 10th, 2018 resulted from an increase in probability within the region of reports with several swaths of decreased probabilities that, in combination, outweighed the increased wind probabilities outside of the report regions and the deceased probabilities in the isolated cases (Fig. 11). As described in the 24hr wind storm-attribute contributions, there were a couple of the multiscale storm-attribute predictors contributing on average to forecast skill. A slight difference in this case from the aggregate, was that the meso-γ storm-attribute predictors were marginally contributing (though positive) whereas the meso-β and meso-α scale storm-attribute predictors, on average, were dominating forecast skill. After checking forecast skill differences it was important to understand these differences as this case had a higher BSS in favor of the EXPRF 24hr wind all while showing primary contributions to forecast skill from multiscale storm-attribute predictors.

Similar to the 24hr any severe case, the smoothed maxuh03 (Fig. 20a) used in 24hr wind, retained the key regions of relatively higher maxuh03. Per contributions, the meso-α scale

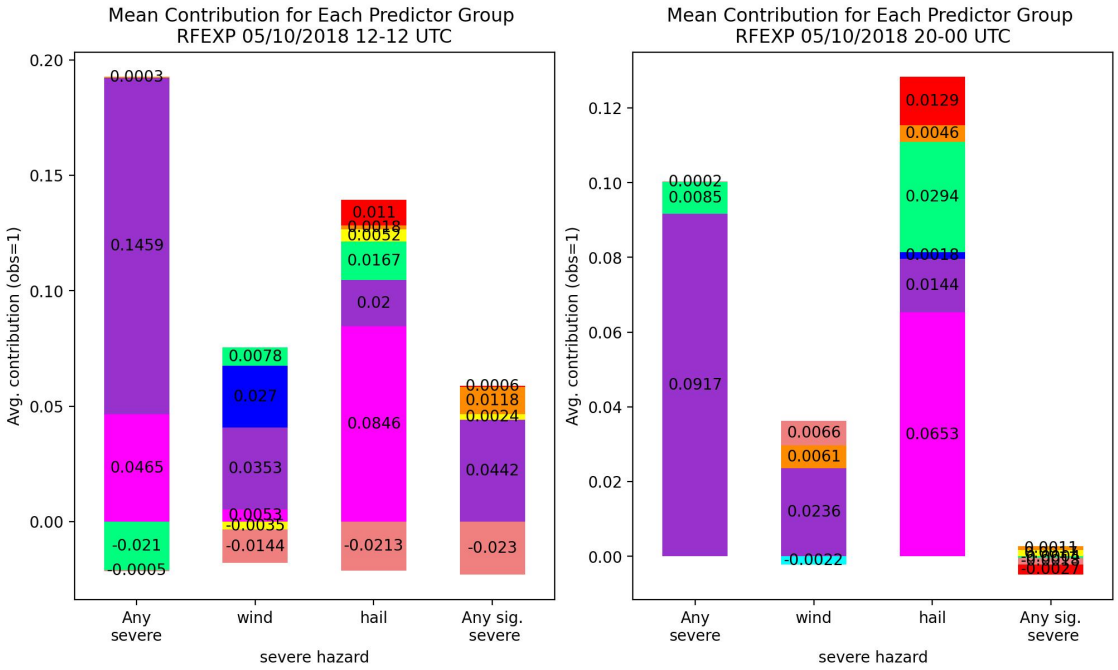

Figure 18: May 10th, 2018 case of average contributions from all predictors sorted into groups based off meteorological characteristics and spatial length scale. Shown for severe weather reports sample. Blocks are color coded by predictor characteristic and spatial length scale from smallest spatial scales to largest 12-12 UTC forecast period (left), 20-00 UTC forecast period (right). Values within blocks are sum of average contributions from individual predictors within respective group.

maxuh03 contributed on average the most to forecast skill even at the higher smoothing level (Fig. 20d). In addition, the meso-β scale maxdbz1km maxima were associated with positive contributions and even seemed to only contribute positively when values exceeded approximately 56 dBZ (Fig. 20b and e). In addition to the aspects of the overall results that could be understood using the May 10th, 2018 case, the May 18th, 2019 case demonstrated how the large scale environment predictors also contributed to skill even while the storm-attributes were dominant in 24hr any severe (Fig. 21 left). Consistent with what has been the case for the aggregate, May 10th, 2018, and now May 18th, 2019 results, the meso-β scale maxuh25 was a favorable predictor for retaining the key regions of simulated storms all while reducing noisier and unimportant maxuh25 values in a large severe weather event as such (Fig. 22). Considering this was such a large event, it made sense that there would be greater contributions from the synoptic scale environment predictors. In addition, the smoothing

The synoptic driven 90mb mlcin (Fig. 22b), inferred from the large-scale warm air advection that was evident in the synotpic scale v10m wind (Fig. 22f) seemed to contribute in a way that corresponded to the large-scale flow pattern (suggested from the synoptic scale v10m wind Fig. 22c). Though it was not obvious why there was a large area of positive contributions across the entire reports region into an isolated region of reports (Fig. 22e),

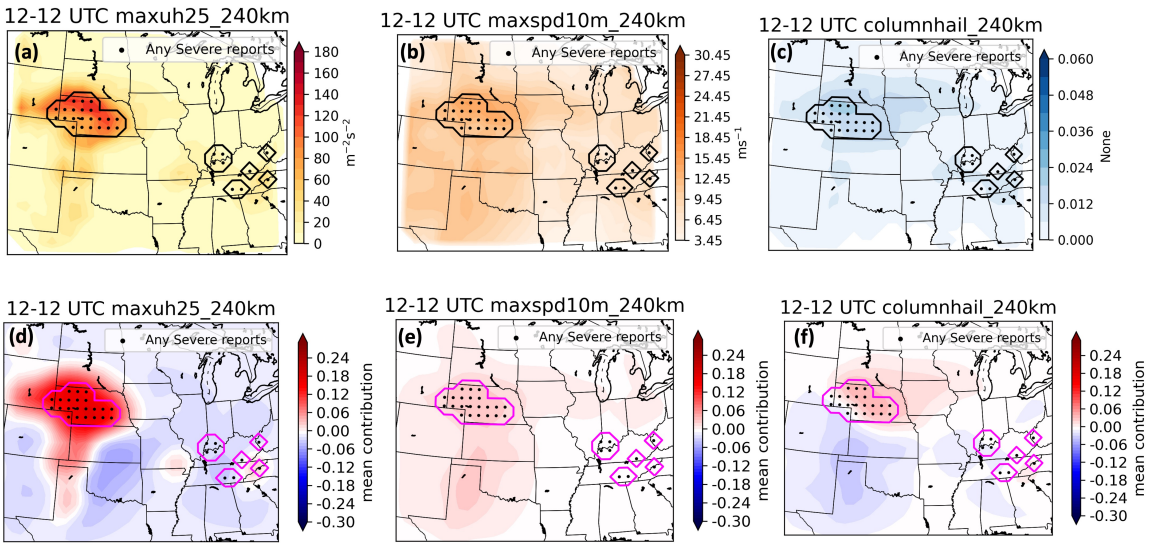

Figure 19: Top 3 contributing storm-attribute predictors when locations had severe weather May 10th, 2018. From pre-processed 10 member ensemble following steps for pre-processing as described in methods. (a) 240 km smoothed 2-5 maximum updraft helicity, (b) maximum 10m wind speed, (c) 240km smoothed columnhail, (d) corresponding contributions to (a), (e) corresponding to (b), and (f) corresponding to (c). Contribution levels every 0.01 units from -0.3,0.3

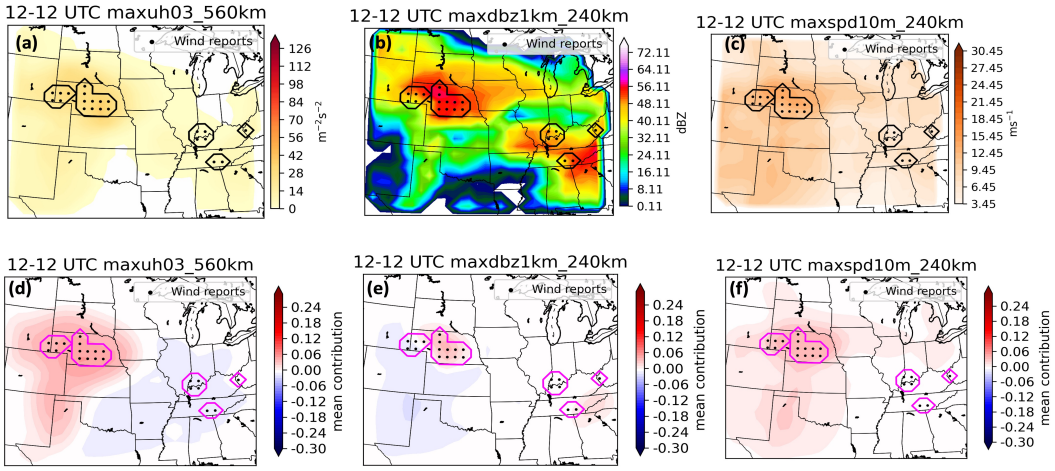

Figure 20: Top 3 contributing storm-attribute predictors when locations had severe wind reports May 10th, 2018. From pre-processed 10 member ensemble following steps for pre-processing as described in methods. (a) 560 km smoothed 0-3 maximum updraft helicity, (b) maximum 1km dBZ, (c) maximum 10m windspeed, (d) corresponding contributions to (a), (e) corresponding to (b), and (f) corresponding to (c). Contribution levels every 0.01 units from -0.3,0.3

this seemed to be a case of predictor interaction considering the positive contributions from the synoptic scale 90mb mlcin seems to be directed in similar flow to that of the synoptic

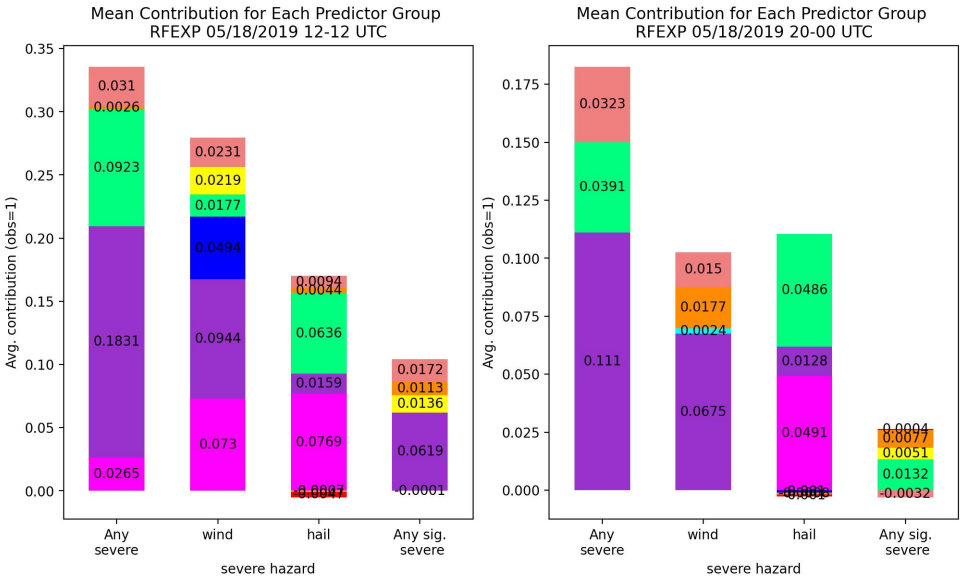

Figure 21: May 18th, 2019 case of average contributions from all predictors sorted into groups based off meteorological characteristics and spatial length scale. Shown for severe weather reports sample. Blocks are color coded by predictor characteristic and spatial length scale from smallest spatial scales to largest 12-12 UTC forecast period (left) 20-00 UTC forecast period (right). Values within blocks are sum of average contributions from individual predictors within respective group.

v10m wind based off contributions (Fig. 22 e,f). Though, this would need further evaluation.

Summary and Discussion

When evaluating the impacts of multiscale predictors on RF-based probabilistic forecasts, it was evident that the EXPRF model often produced skillful forecasts relative to the CTLRF. In several other case studies (not shown), the EXPRF forecasts had a tendency to increase probabilities over the same areas CTLRF had issued appreciable probabilities, but it was this consistent increase in forecast probabilities over regions associated with severe reports that gave the EXPRF models a performance edge for several severe weather hazards in terms of detection rates. Furthermore, the May 18th, 2019 case was the best example of the EXPRF handling well a major severe weather event. What was not shown were the raw severe weather reports demonstrating the two squall-lines mentioned. This was a rather marked achievement for the EXPRF model to have forecasted both swaths of severe wind reports (asserted from two regions of maximized wind probabilities and co-located wind reports) that occurred across southeast OK into LA and southwest into central MO where the CTLRF model only had one forecasted area; the more obvious squall line that moved across southeast OK into LA. In addition, there was substantial support for the importance of including predictors that directly incorporate flow-dependence as demonstrated by FPS selection of the multiple larger scale predictors and the consistent physical implications of the meso-β and meso-α scale storm-attribute predictor (i.e., maxuh25, maxuh03, columnhail,

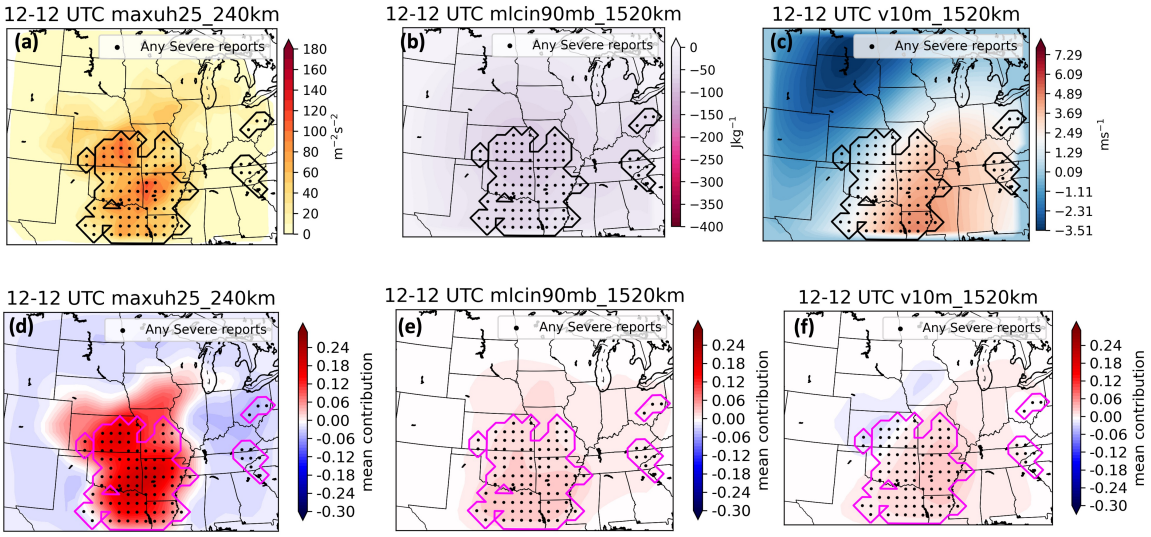

Figure 22: Top contributing storm-attribute predictors and Top 2 contributing environment predictors when locations had severe weather May 18th, 2019. From pre-processed 10 member ensemble following steps for pre-processing as described in methods. (a) 240 km smoothed 2-5 maximum updraft helicity, (b) 1520 km smoothed 90mb mixed layer CIN, (c) 1520km smoothed v10m wind, (d) corresponding contributions to (a), (e) corresponding to (b), and (f) corresponding to (c). Contribution levels every 0.01 units from -0.3,0.3

maxdbz1km, etc.) often times contributing on average, the most skill for a given forecast and successfully reducing noisy regions and removing unimportant features that might otherwise be considered important. Although, despite the advantage of the multiscale storm-attribute predictors, in the case of hail forecasts, it seemed important to retain un-smoothed values (which makes physical sense in terms of forecasting the most intense simulated storms). This was true for environment predictors, in particular convective environment predictors, which seemed to be more sensitive to the smoothing. This was demonstrated for most hazards showing non-smoothed meso-γ scale convective environment predictors contributing more than the smoothed convective environment predictors. This was again most true for hail forecasts, but several other severe weather hazards demonstrated the need for the meso-γ convective environment predictors. Overall, it was still shown that multiscale predictors included into the random forest with the 80 km predictors, were beneficial and generally improved over the RF's without the multiscale predictors.

Conclusion

This thesis utilized two characteristically different random forest models (i.e. CTLRF and EXPRF) to forecast 1200-1200 UTC next-day (24hr) and 2000-0000 short-term UTC (4hr) for 8 individual severe weather hazards (any severe, wind, hail, tornadoes, and significant counterparts). The predictors for both models came from the OU Map Lab produced 10-member CONUS-domain convection-allowing ensemble forecasts during the 2018 and 2019 HWT SFE.

Typically, the RF predictors are created in a couple of steps − spatial pre-processing and temporal pre-processing. The spatial pre-processing step was the most critical part that separated previous studies from this study. RF predictors are usually pre-processed to some fixed grid (e.g., 80 km) during the spatial pre-processing step to reduce dimensionality. However, since it has been shown that features on all spatial scales contribute to severe weather risk (e.g., Snellman, 1982), these spatial pre-processing methods to fixed grids can potentially leave out important information about the large scale flow pattern in the RF predictors. In order to test this, the EXPRF models were created via predictors that directly included the larger scale of motions via a neighborhood averaging of the 80 km forecast variables at increasingly larger spatial length scales. These scales effectively represented the meso-γ (80 km), meso-β (240 km), meso-α (560 km), and synoptic scales (1520 km). This larger-scale pre-processing step was intended to account for the multiscale flow-dependence of severe weather. To further simplify the RF models, in an objective way and to reduce redundancy and/or noise associated with the multiscale predictors, a forward predictor selection (FPS) method was utilized. This objectively selected which predictors were to be used for training the RF models for the all eight severe weather hazards for the two different forecast periods. After the spatial pre-processing step, all the predictors ensembles were aggregated and then temporally aggregated. One of the study were to test whether the inclusion of the multiscale predictors improved RF based probabilistic severe weather forecasts. Thus, the creation of RF models with fixed 80 km predictors (i.e., CTLRF) and multiscale, flow-dependent predictors (i.e., EXPRF) were tested against CTLRF to directly assess the impacts of the multiscale predictors on RF-based probabilistic forecast skill. Following the demonstrated impacts, it was necessary to then determine a physical understanding as to how the multiscale predictors contributed to the RF-based probabilistic forecast skill through use of TI.

The EXPRF and CTLRF models created were compared against each other using several verification methods. It was found that for nearly all of the 24hr and 4hr forecasted severe weather hazards, the EXPRF model had a higher BSS than the CTLRF model. Most of the 24hr sub-significant severe weather forecasts were significantly more skillful with EXPRF than CTLRF. In general, both models had small calibration errors, but the CTLRF model had greater reliability forecasting wind for both periods. Most of the EXPRF forecasts had higher resolution (about an order of magnitude larger) than the CTLRF of which was significant for 24hr any severe and wind forecasts corresponding to two of the most significantly skilled forecast hazards. In addition, performance diagrams which showed the significantly more skilled EXPRF forecasts for both 24hr and 4hr, having generally higher CSI than the CTLRF forecasts, especially for 24hr wind at higher forecast probabilities.

While forecast skill was generally improved overall for EXPRF relative to CTLRF as stated previously, another goal of this study was to better understand how the multiscale predictors can contribute to the RF-based forecast skill. To achieve this goal, the python-based module tree interpreter was used to investigate the contributions of specific predictors to the performance of the EXPRF model. Over all the cases, it was found that on average when severe weather did not occur, the meso-γ scale storm-attribute predictors contributed more to skill than the meso-β scale storm-attribute predictors for 24hr forecasts, but opposite

was true for 4hr forecasts. When severe weather did occur, the meso-β scale storm-attribute predictors contributed the most to skill in general. However, the meso-γ scale environment predictors dominated environment-related contributions to forecast skill. Lastly, nearly all of the multiscale storm-attribute and environment predictors on average were contributing to forecast skill.

Two case studies were utilized to demonstrate the physical implications of the multiscale predictors. It was found that the most attributable reason for the substantial contributions of the meso-β storm-attribute predictor in general, was due to accounting for location uncertainty of explicitly simulated storms similar to neighborhood-based CAM forecasts. The meso-γ storm-attribute predictors were consistently contributing higher forecast probabilities. On the contrary, for more environment based predictors with important implications from sharp gradients (e.g. surface based CAPE, 2m-Temperature), the smoothing of those predictors constituted the loss of valuable information of intensity and locations of said gradients that can be associated with synoptically predictable features. In combination, forecasting severe hazards with multiscale predictors does provide an advantage, especially for explicitly resolved storms.

These results are promising in that the RF-based probabilistic forecasts of severe weather are statistically significantly improved by the inclusion of multiscale predictors for several severe weather hazard types/periods. Thus, it is confirmed that there are aspects of the large scale flow pattern that are able to further improve the RF-based forecast if accounted for in the RF predictors. However, this study was limited in some aspects. Sample size was somewhat limited, the larger-scale dependence was represented by a simple averaging radius, and predictors that were more representative of the different severe weather hazard types could have been utilized (e.g. greater number of low-level shear predictors for tornadoes). Future studies may benefit from using larger sample size to further optimize the spatial scales across different lead times and severe hazard types, and yield more robust results related to the rarer severe hazard types (e.g., tornado, and sig. severe). While the TI did provide some physical justification for the benefit of the multiscale predictors, future work should also continue to focus on the contributions from predictor-interactions when evaluating physical impacts of multiscale predictors.

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. *Weather and Forecasting*, **18 (6)**, 918 – 932, https://doi.org/10.1175/1520-0434(2003)018⟨0918:SOPFSS⟩2.0.CO;2.

Ahijevych, D., J. Pinto, J. K. Williams, and M. Steiner, 2016: Probabilistic forecasts of mesoscale convective system initiation using the random forest data mining technique. *Weather And Forecasting*, **31 (1)**, 581 – 599, https://doi.org/10.1175/WAF-D-15-0113.1.

Baez-Villanueva, O. M., and Coauthors, 2020: Rf-mep: A novel random forest method for merging gridded precipitation products and ground-based measurements. *Remote Sensing of Environment*, **239**, 111 606, https://doi.org/https://doi.org/10.1016/j.rse.2019.111606.

Beck, J., F. Bouttier, L. Wiegand, C. Gebhardt, C. Eagle, and N. Roberts, 2016: Development and verification of two convection-allowing multi-model ensembles over western europe. *QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY*, **142 (700, PT A)**, 2808–2826, https://doi.org/10.1002/qj.2870.

Benjamin, S. G., and Coauthors, 2016: A north american hourly assimilation and model forecast cycle: The rapid refresh. *Monthly Weather Review*, **144 (4)**, 1669 – 1694, https://doi.org/10.1175/MWR-D-15-0242.1.

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. *Weather and Forecasting*, **33 (4)**, 1063 – 1080, https://doi.org/10.1175/WAF-D-18-0035.1.

Breiman, L., 2001: Random forests. *Machine Learning*, **45 (1)**, 5–32, https://doi.org/10.1023/A:1010933404324.

Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli, 1999: Probabilistic predictions of precipitation using the ecmwf ensemble prediction system. *Weather and Forecasting*, **14 (2)**, 168 – 189, https://doi.org/10.1175/1520-0434(1999)014⟨0168:PPOPUT⟩2.0.CO;2.

Chandramouli, K., X. Wang, A. Johnson, and J. Otkin, 2022: Online nonlinear bias correction in ensemble kalman filter to assimilate goes-r all-sky radiances for the analysis and prediction of rapidly developing supercells. *Journal of Advances in Modeling Earth Systems*, **14 (3)**, e2021MS002 711, https://doi.org/https://doi.org/10.1029/2021MS002711, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021MS002711.

Chipilski, H. G., X. Wang, and D. B. Parsons, 2020: Impact of assimilating pecan profilers on the prediction of bore-driven nocturnal convection: A multiscale forecast evaluation for the 6 july 2015 case study. *Monthly Weather Review*, **148 (3)**, 1147 – 1175, https://doi.org/10.1175/MWR-D-19-0171.1.

Clark, A. J., and E. D. Loken, 2022: Machine learning–derived severe weather probabilities from a warn-on-forecast system. *Weather and Forecasting*, **37 (10)**, 1721 – 1740, https://doi.org/10.1175/WAF-D-22-0056.1.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. part ii: Application to convective rain systems. *Monthly Weather Review*, **134 (7)**, 1785 – 1795, https://doi.org/10.1175/MWR3146.1.

Degelia, S. K., X. Wang, and D. J. Stensrud, 2019: An evaluation of the impact of assimilating aeri retrievals, kinematic profilers, rawinsondes, and surface observations on a forecast of a nocturnal convection initiation event during the pecan field campaign. *Monthly Weather Review*, **147 (8)**, 2739 – 2764, https://doi.org/10.1175/MWR-D-18-0423.1.

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. *Atmospheric Science Letters*, **5 (6)**, 110–117, https://doi.org/https://doi.org/10.1002/asl.72, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/asl.72.

Dowell, D. C., and Coauthors, 2022: The high-resolution rapid refresh (hrrr): An hourly updating convection-allowing forecast model. part i: Motivation and system description. *Weather and Forecasting*, **37 (8)**, 1371 – 1395, https://doi.org/10.1175/WAF-D-21-0151.1.

Duda, J. D., X. Wang, F. Kong, and M. Xue, 2014: Using varied microphysics to account for uncertainty in warm-season qpf in a convection-allowing ensemble. *Monthly Weather Review*, **142 (6)**, 2198 – 2219, https://doi.org/10.1175/MWR-D-13-00297.1.

Duda, J. D., X. Wang, F. Kong, M. Xue, and J. Berner, 2016: Impact of a stochastic kinetic energy backscatter scheme on warm season convection-allowing ensemble forecasts. *Monthly Weather Review*, **144 (5)**, 1887 – 1908, https://doi.org/10.1175/MWR-D-15-0092.1.

Duda, J. D., X. Wang, and M. Xue, 2017: Sensitivity of convection-allowing forecasts to land surface model perturbations and implications for ensemble design. *Monthly Weather Review*, **145 (5)**, 2001 – 2025, https://doi.org/10.1175/MWR-D-16-0349.1.

Flora, M. L., C. K. Potvin, and L. J. Wicker, 2018: Practical predictability of supercells: Exploring ensemble forecast sensitivity to initial condition spread. *MONTHLY WEATHER REVIEW*, **146 (8)**, 2361–2379, https://doi.org/10.1175/MWR-D-17-0374.1.

Gagne, D. J., A. McGovern, S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. *Weather and Forecasting*, **32 (5)**, 1819 – 1840, https://doi.org/10.1175/WAF-D-17-0010.1.

Gagne, D. J., A. McGovern, and M. Xue, 2014: Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts. *Weather and Forecasting*, **29 (4)**, 1024 – 1043, https://doi.org/10.1175/WAF-D-13-00108.1.

Gasperoni, N. A., X. Wang, and Y. Wang, 2020: A comparison of methods to sample model errors for convection-allowing ensemble forecasts in the setting of multiscale initial conditions produced by the gsi-based envar assimilation system. *Monthly Weather Review*, **148 (3)**, 1177 – 1203, https://doi.org/10.1175/MWR-D-19-0124.1.

Gasperoni, N. A., X. Wang, and Y. Wang, 2022: Using a cost-effective approach to increase background ensemble member size within the gsi-based envar system for improved radar analyses and forecasts of convective systems. *Monthly Weather Review*, **150 (3)**, 667 – 689, https://doi.org/10.1175/MWR-D-21-0148.1.

Gebhardt, C., S. Theis, M. Paulat, and Z. Ben Bouallègue, 2011: Uncertainties in cosmo-de precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmospheric Research*, **100 (2)**, 168–177, https://doi.org/https://doi.org/10.1016/j.atmosres.2010.12.008.

Hall, T. J., C. N. Mutchler, G. J. Bloy, R. N. Thessin, S. K. Gaffney, and J. J. Lareau, 2011: Performance of observation-based prediction algorithms for very short-range, probabilistic clear-sky condition forecasting. *Journal of Applied Meteorology and Climatology*, **50 (1)**, 3 – 19, https://doi.org/10.1175/2010JAMC2529.1.

Herman, G. R., and R. S. Schumacher, 2018a: Money doesn't grow on trees, but forecasts do: Forecasting extreme precipitation with random forests. *Monthly Weather Review*, **146 (5)**, 1571 – 1600, https://doi.org/10.1175/MWR-D-17-0250.1.

Herman, G. R., and R. S. Schumacher, 2018b: "dendrology" in numerical weather prediction: What random forests and logistic regression tell us about forecasting extreme precipitation. *Monthly Weather Review*, **146 (6)**, 1785 – 1812, https://doi.org/10.1175/MWR-D-17-0307.1.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Monthly Weather Review*, **148 (5)**, 2135 – 2161, https://doi.org/10.1175/MWR-D-19-0344.1.

Hsu, W., and A. H. Murphy, 1986: The attributes diagram a geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting*, **2 (3)**, 285–293, https://doi.org/https://doi.org/10.1016/0169-2070(86)90048-8.

Jasper Velthoen, J.-J. C., and G. Jongbloed, 2023: Forward variable selection for random forest models. *Journal of Applied Statistics*, **50 (13)**, 2836–2856, https://doi.org/10.1080/02664763.2022.2095362, https://doi.org/10.1080/02664763.2022.2095362.

Johns, R. H., and C. A. Doswell III, 1992: Severe local storms forecasting. *Wea. Forecasting*, **7 (4)**, 588–612.

Johnson, A., F. Han, Y. Wang, and X. Wang, 2023: Scale-dependent verification of the ou map convection allowing ensemble initialized with multi-scale and large-scale perturbations during the 2019 noaa hazardous weather testbed spring forecasting experiment. *Atmosphere*, **14 (2)**, https://doi.org/10.3390/atmos14020255.

Johnson, A., and X. Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Monthly Weather Review*, **140 (9)**, 3054 – 3077, https://doi.org/10.1175/MWR-D-11-00356.1.

Johnson, A., and X. Wang, 2016: A study of multiscale initial condition perturbation methods for convection-permitting ensemble forecasts. *Monthly Weather Review*, **144 (7)**, 2579 – 2604, https://doi.org/10.1175/MWR-D-16-0056.1.

Johnson, A., and X. Wang, 2017: Design and implementation of a gsi-based convection-allowing ensemble data assimilation and forecast system for the pecan field experiment. part i: Optimal configurations for nocturnal convection prediction using retrospective cases. *Weather and Forecasting*, **32 (1)**, 289 – 315, https://doi.org/10.1175/WAF-D-16-0102.1.

Johnson, A., and X. Wang, 2020: Interactions between physics diversity and multiscale initial condition perturbations for storm-scale ensemble forecasting. *Monthly Weather Review*, **148 (8)**, 3549 – 3565, https://doi.org/10.1175/MWR-D-20-0112.1.

Johnson, A., and X. Wang, 2024: Impacts of initial condition perturbation blending in 10- and 40-member convection-allowing ensemble forecasts. *Monthly Weather Review*, https://doi.org/10.1175/MWR-D-23-0188.1.

Johnson, A., X. Wang, J. R. Carley, L. J. Wicker, and C. Karstens, 2015: A comparison of multiscale gsi-based enkf and 3dvar data assimilation using radar and conventional observations for midlatitude convective-scale precipitation forecasts. *Monthly Weather Review*, **143 (8)**, 3087 – 3108, https://doi.org/10.1175/MWR-D-14-00345.1.

Johnson, A., X. Wang, and T. Jones, 2022: Impacts of assimilating goes-16 abi channels 9 and 10 clear air and cloudy radiance observations with additive inflation and adaptive observation error in gsi-enkf for a case of rapidly evolving severe supercells. *Journal of Geophysical Research: Atmospheres*, **127 (11)**, e2021JD036 157, https://doi.org/https://doi.org/10.1029/2021JD036157, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2021JD036157.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2011a: Hierarchical cluster analysis of a convection-allowing ensemble during the hazardous weather testbed 2009 spring experiment. part i: Development of the object-oriented cluster analysis method for precipitation fields. *Monthly Weather Review*, **139 (12)**, 3673 – 3693, https://doi.org/10.1175/MWR-D-11-00015.1.

Johnson, A., X. Wang, F. Kong, and M. Xue, 2013: Object-based evaluation of the impact of horizontal grid spacing on convection-allowing forecasts. *Monthly Weather Review*, **141 (10)**, 3413 – 3425, https://doi.org/10.1175/MWR-D-13-00027.1.

Johnson, A., X. Wang, M. Xue, and F. Kong, 2011b: Hierarchical cluster analysis of a convection-allowing ensemble during the hazardous weather testbed 2009 spring experiment. part ii: Ensemble clustering over the whole experiment period. *Monthly Weather Review*, **139 (12)**, 3694 – 3710, https://doi.org/10.1175/MWR-D-11-00016.1.

Johnson, A., and Coauthors, 2014: Multiscale characteristics and evolution of perturbations for warm season convection-allowing precipitation forecasts: Dependence on background flow and method of perturbation. *Monthly Weather Review*, **142 (3)**, 1053 – 1073, https://doi.org/10.1175/MWR-D-13-00204.1.

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. *Weather and Forecasting*, **21 (2)**, 167 – 181, https://doi.org/10.1175/WAF906.1.

Kain, J. S., and Coauthors, 2008: Some practical considerations regarding horizontal resolution in the first generation of operational convection-allowing nwp. *Weather and Forecasting*, **23 (5)**, 931 – 952, https://doi.org/10.1175/WAF2007106.1.

Liu, X., K. Zhou, Y. Lan, X. Mao, and R. J. Trapp, 2020: On the construction principle of conceptual models for severe convective weather forecasting operations in china. *Weather and Forecasting*, **35 (1)**, 299 – 308, https://doi.org/10.1175/WAF-D-19-0026.1.

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 (4)**, 1605 – 1631, https://doi.org/10.1175/WAF-D-19-0258.1.

Loken, E. D., A. J. Clark, and A. McGovern, 2022: Comparing and interpreting differently designed random forests for next-day severe weather hazard prediction. *Weather and Forecasting*, **37 (6)**, 871 – 899, https://doi.org/10.1175/WAF-D-21-0138.1.

Loken, E. D., A. J. Clark, A. McGovern, M. Flora, and K. Knopfmeier, 2019: Postprocessing next-day ensemble probabilistic precipitation forecasts using random forests. *Weather and Forecasting*, **34 (6)**, 2017 – 2044, https://doi.org/10.1175/WAF-D-19-0109.1.

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. *Weather and Forecasting*, **32 (4)**, 1403 – 1421, https://doi.org/10.1175/WAF-D-16-0200.1.

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. *Bulletin of the American Meteorological Society*, **98 (10)**, 2073 – 2090, https://doi.org/10.1175/BAMS-D-16-0123.1.

McGovern, A., D. John Gagne II, N. Troutman, R. A. Brown, J. Basara, and J. K. Williams, 2011: Using spatiotemporal relational random forests to improve our understanding of severe weather processes. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, **4 (4)**, 407–429, https://doi.org/https://doi.org/10.1002/sam.10128, https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.10128.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bulletin of the American Meteorological Society*, **100 (11)**, 2175 – 2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Medina, B. L., L. D. Carey, C. G. Amiot, R. M. Mecikalski, W. P. Roeder, T. M. McNamara, and R. J. Blakeslee, 2019: A random forest method to forecast downbursts based on dual-polarization radar signatures. *Remote Sensing*, **11 (7)**, https://doi.org/10.3390/rs11070826.

Novak, D. R., D. R. Bright, and M. J. Brennan, 2008: Operational forecaster uncertainty needs and future roles. *Weather and Forecasting*, **23 (6)**, 1069 – 1084, https://doi.org/10.1175/2008WAF2222142.1.

Ostby, F. P., 1999: Improved accuracy in severe storm forecasting by the severe local storms unit during the last 25 years: Then versus now. *Weather and Forecasting*, **14 (4)**, 526 – 543, https://doi.org/10.1175/1520-0434(1999)014⟨0526:IAISSF⟩2.0.CO;2.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, **12 (85)**, 2825–2830.

Roberts, B., I. L. Jirak, A. J. Clark, S. J. Weiss, and J. S. Kain, 2019: Postprocessing and visualization techniques for convection-allowing ensembles. *Bulletin of the American Meteorological Society*, **100 (7)**, 1245 – 1258, https://doi.org/10.1175/BAMS-D-18-0041.1.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Weather and Forecasting*, **24 (2)**, 601 – 608, 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. *Weather and Forecasting*, **19 (5)**, 936 – 949, https://doi.org/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. *Monthly Weather Review*, **142 (12)**, 4519 – 4541, https://doi.org/10.1175/MWR-D-14-00100.1.

Schwartz, C. S., and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Monthly Weather Review*, **145 (9)**, 3397 – 3418, https://doi.org/10.1175/MWR-D-16-0400.1.

Snellman, L., 1982: Impact of afos on operational forecasting. *Ninth Conf. on Weather Forecasting and Analysis*.

Wang, Y., and X. Wang, 2017: Direct assimilation of radar reflectivity without tangent linear and adjoint of the nonlinear observation operator in the gsi-based envar system: Methodology and experiment with the 8 may 2003 oklahoma city tornadic supercell. *Monthly Weather Review*, **145 (4)**, 1447 – 1471, https://doi.org/10.1175/MWR-D-16-0231.1.

Wang, Y., and X. Wang, 2020: Prediction of tornado-like vortex (tlv) embedded in the 8 may 2003 oklahoma city tornadic supercell initialized from the subkilometer grid spacing analysis produced by the dual-resolution gsi-based envar data assimilation system. *Monthly Weather Review*, **148 (7)**, 2909 – 2934, https://doi.org/10.1175/MWR-D-19-0179.1.

Wang, Y., and X. Wang, 2021: Rapid update with envar direct radar reflectivity data assimilation for the noaa regional convection-allowing nmmb model over the conus: System description and initial experiment results. *Atmosphere*, **12 (10)**, https://doi.org/10.3390/atmos12101286.

Wang, Y., and X. Wang, 2023a: A multivariate additive inflation approach to improve storm-scale ensemble-based data assimilation and forecasts: Methodology and experiment with a tornadic supercell. *Journal of Advances in Modeling Earth Systems*, **15 (1)**, e2022MS003 307, https://doi.org/https://doi.org/10.1029/2022MS003307, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003307.

Wang, Y., and X. Wang, 2023b: Simultaneous multiscale data assimilation using scale- and variable-dependent localization in envar for convection allowing analyses and forecasts: Methodology and experiments for a tornadic supercell. *Journal of Advances in Modeling Earth Systems*, **15 (5)**, e2022MS003 430, https://doi.org/https://doi.org/10.1029/2022MS003430, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2022MS003430.

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. *Weather and Forecasting*, **23 (3)**, 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. *Monthly Weather Review*, **125 (4)**, 527 – 548, https://doi.org/10.1175/1520-0493(1997)125⟨0527:TRDOEM⟩2.0.CO;2.

Wilkins, A., A. Johnson, X. Wang, N. A. Gasperoni, and Y. Wang, 2021: Multi-scale object-based probabilistic forecast evaluation of wrf-based cam ensemble configurations. *Atmosphere*, **12 (12)**, https://doi.org/10.3390/atmos12121630.

Wilks, D. S., Ed., 2019: *Preface to the Third Edition*. Fourth edition ed., Elsevier, xvii pp., https://doi.org/https://doi.org/10.1016/B978-0-12-815823-4.09983-1, URL https://www.sciencedirect.com/science/article/pii/B9780128158234099831.

Yang, Y., and X. Wang, 2023a: A comparison of 3denvar and 4denvar for convective-scale direct radar reflectivity data assimilation in the context of a filter and a smoother. *Monthly Weather Review*, **152 (1)**, 59 – 78, https://doi.org/10.1175/MWR-D-23-0082.1.

Yang, Y., and X. Wang, 2023b: Impact of radar reflectivity data assimilation frequency on convection-allowing forecasts of diverse cases over the continental united states. *Monthly Weather Review*, **151 (2)**, 341 – 362, https://doi.org/10.1175/MWR-D-22-0095.1.

Yao, H., X. Li, H. Pang, L. Sheng, and W. Wang, 2020: Application of random forest algorithm in hail forecasting over shandong peninsula. *Atmospheric Research*, **244**, 105 093, https://doi.org/https://doi.org/10.1016/j.atmosres.2020.105093.

Zarei, M., M. Najarchi, and R. Mastouri, 2021: Bias correction of global ensemble precipitation forecasts by random forest method. *Earth Science Informatics*, **14 (2)**, 677–689, https://doi.org/10.1007/s12145-021-00577-7.

Zeng, Q., Z. Qing, M. Zhu, F. Zhang, H. Wang, Y. Liu, Z. Shi, and Q. Yu, 2022: Application of random forest algorithm on tornado detection. *Remote Sensing*, **14 (19)**, https://doi.org/10.3390/rs14194909.