UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

# Storm-scale Ensemble-based Severe Weather Guidance: Development of an Object-based Verification Framework and Applications of Machine Learning

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

MONTGOMERY LEE FLORA
Norman, Oklahoma
2020

# STORM-SCALE ENSEMBLE-BASED SEVERE WEATHER GUIDANCE: DEVELOPMENT OF AN OBJECT-BASED VERIFICATION FRAMEWORK AND APPLICATIONS OF MACHINE LEARNING

A DISSERTATION APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Corey Potvin, Chair

Dr. Cameron Homeyer, Co-Chair

Dr. Amy McGovern

Dr. Xuguang Wang

Dr. Steven Cavallo

Dr. Andrew Fagg

## Dedication

I dedicate this dissertation to God, my wife Kristina, my son Noah, and the rest of loving family for supporting me as I pursued my dreams.

## Acknowledgements

I have pursued this doctorate degree since my junior year of high school and many individuals have a contributed to my success. I reflect on those that have embraced my goals and never once discouraged me from pursuing them. I want to thank God for placing wonderful people in my life and I wish to briefly acknowledge some of them here (I apologize if there is any I have missed!).

To my caring and loving wife, Kristi, you have been strong by my side for over 8 years and I cannot express in words how much your love and support has guided me through graduate school. I know it has been difficult moving so far away from our family and friends in Indiana, but our family is growing and we are learning to make Oklahoma our home. I hope in the years to come we can continue to grow together, encourage one another and become the best versions of ourselves.

To my beautiful son, Noah, you have changed my life in ways I could never have imagined. After experiencing my love for you, I better understand the love that God feels for his children. When I was my younger self, I thought earning this degree would be my life's greatest accomplishment, but I was sorely mistaken. Other than your future siblings, nothing I will ever create can compare to you.

To my advisor, Corey, you have been a tremendous mentor and guiding force through my REU internship, graduate school and into my early career. Your approaches to research are methodical, well-thought out, and speak highly of your character. It is these traits I aspire to and hope to further develop as a scientist.

To my family (Mom, Judy, Shell, Janus, Gail, Darrel, Tim, just to name a few), your love and support has not gone unnoticed. Many of you encouraged me to pursue this degree, and I could not have done it without your support. It is incredibly difficult to accomplish great things on your own, and the difference between being successful and unsuccessful can often come down to strength of the relationships in our lives. *"The love of family is life's greatest blessing."*

To my wonderful friend, Shawn, your friendship has meant so much to me. It's difficult to create new friendships, especially after moving so far away from home. Healthy relationships are such an important part of life and being a good scientist. Being able to freely share ideas with you truly helped me to be a better scientist.

To Amy McGovern and Patrick Skinner, I also want to acknowledge your part in my success. It was Amy McGovern's graduate course that introduced me to the exciting field of machine learning. Her overall demeanor and resolve as a professor made the learning process enjoyable, and her commitment to best practices in machine learning has pushed me to be a better scientist. As for Patrick Skinner, many of the crucial components of this dissertation came from my conversations with him. Patrick has also been a very approachable person and his often optimistic take on things was a very refreshing experience. I know my success as a Ph.D. student and a machine learning scientist would have greatly suffered without the role these two wonderful mentors played.

## Table of Contents

## List of Figures

# Abstract

A goal of the National Oceanic and Atmospheric Administration (NOAA) Warn-on-Forecast (WoF) project is to provide rapidly updating probabilistic guidance to human forecasters for short-term (e.g., 0-3 h) severe weather forecasts. Several case studies have shown that experimental WoF systems (WoFS) can produce accurate short-term probabilistic guidance for hazards such as tornadoes, hail, and heavy rainfall. However, without an appropriate probabilistic verification method for WoFS-style forecasts (which provide guidance for individual thunderstorms), a robust evaluation of WoFS performance has been lacking. In this dissertation, I develop a novel object-based verification method for short-term, storm-scale probabilistic forecasts and apply it to WoFS probabilistic mesocyclone guidance and further adapted to evaluate machine learning-based calibrations of WoFS severe weather probabilistic guidance.

The probabilistic mesocyclone guidance was generated by calculating grid-scale ensemble probabilities from WoFS forecasts of updraft helicity (UH) in layers 2-5 km (mid-level) and 0-2 km (low-level) above ground level (AGL) aggregated over 60-min periods. The resulting ensemble probability swaths are associated with individual thunderstorms and treated as objects. Each ensemble track object is assigned a single representative probability value. A mesocyclone probability object, conceptually, is a region bounded by the ensemble forecast envelope of a mesocyclone track for a thunderstorm over 1 hour. The mesocyclone probability objects were matched against rotation track objects in Multi-Radar Multi-Sensor data using the *total interest score*, but with the maximum displacement varied between 0, 9, 15, and 30 km. Forecast accuracy and reliability were assessed at four different forecast lead time periods: 0-60 min, 30-90 min, 60-120 min, and 90-150 min. In the 0-60 minute forecast period, the low-level UH probabilistic forecasts had a POD, FAR, and CSI of 0.46, 0.45, and 0.31, respectively, with a probability threshold of 22.2% (the threshold of maximum CSI). In the 90-150 minute forecast period, the POD and CSI dropped to 0.39 and 0.27 while FAR remained relatively unchanged. Forecast probabilities >60% over-predicted the likelihood of observed mesocyclones in the 0-60 min period; however, reliability improved when allowing larger maximum displacements for object matching and at longer lead times.

To evaluate the ability of machine learning (ML) models to calibrate WoFS severe weather guidance, the probability object-based method was generalized for identifying any ensemble storm track (based on individual ensemble updraft tracks rather than mesocyclone tracks). Using these ensemble storm tracks, three sets of predictors were extracted from the WoFS forecasts: intra-storm state variables, near-storm environment variables, and morphological attributes of the ensemble storm tracks. Random forests, gradient-boosted trees, and logistic regression algorithms were then trained to predict which WoFS 30-min ensemble storm tracks will produce a tornado, severe hail, and/or severe wind report. To provide a baseline against which to test the ML models performance, I extracted the probability of mid-level UH exceeding a threshold (tuned per severe weather hazard) from each ensemble storm track. The three ML algorithms discriminated well for all three hazards and produced far more reliable probabilities than the UH-based predictions. Using state-of-the-art ML interpretability methods, I found that the ML models learned sound physical relationships and the appropriate responses to the ensemble statistics. Intra-storm predictors were found to be more important than environmental predictors for all three ML models, but environmental predictors made positive contributions to severe weather likelihood in situations where the WoFS fails to analyze ongoing convection. Overall, the results suggest that ML-based calibrations of dynamical ensemble output can improve short term, storm-scale severe weather probabilistic guidance.

# Chapter 1: Introduction

"*Whatever may be the progress of science, never will observers who are trustworthy, and careful of their reputation, venture to foretell the state of the weather*"

- Francois Arago, 19th century French Mathematician

"*But who wants to be foretold the weather? It is bad enough when it comes, without our having the misery of knowing about it before hand*"

- Jerome K. Jerome, Three Men in a Boat

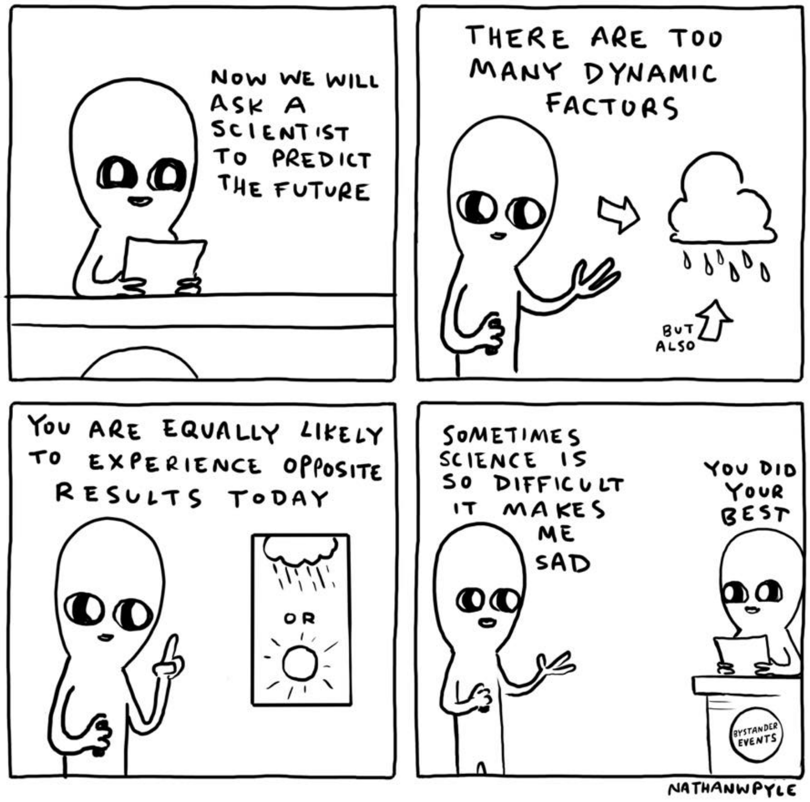

NOW WE WILL ASK A SCIENTIST TO PREDICT THE FUTURE
THERE ARE TOO MANY DYNAMIC FACTORS
BUT ALSO
YOU ARE EQUALLY LIKELY TO EXPERIENCE OPPOSITE RESULTS TODAY
OR
SOMETIMES SCIENCE IS SO DIFFICULT IT MAKES ME SAD
YOU DID YOUR BEST
BYSTANDER EVENTS
NATHANWPYLE

Figure 1.1: The difficulties of predicting the weather.

Forecasting severe convective thunderstorms and their associated hazards (e.g., wind gusts, torrential rain, hail and sometimes tornadoes) is a crucial task since they present a serious threat to human lives and property. From 1980 to 2020, severe storms have caused the highest number of billion-dollar disaster events as compared to other disasters such as tropical cyclones, drought, or flooding with an average event cost of $ 2.1 billion (NCEI 2020). In 2020, there have been 10+ billion-dollar severe storm events (NCEI 2020). Accurately forecasting the location and timing of severe convective hazards, however, remains a challenge for human forecasters. In the current framework known as "warn-on-detection," the National Weather Service (NWS) issues hazardous weather warnings based on radar observations, spotters reports or when an impending hazard is deemed imminent by the forecasters knowledge of the storm environment (e.g., Coleman et al. 2011; Brotzge and Donner 2013). The "warn-on-detection" paradigm is limited as the observational network resolution is often too coarse to capture important storm-scale processes and performance at longer lead times (e.g., beyond 30-60 min) remains highly in question. For example, although considerable effort has been made to distinguish tornadic environments from non-tornadic ones, tornado warning lead times have remained relatively static since 1986 (e.g., Stensrud et al. 2013; Brooks and Correia 2018, see Figure 1.2).

In recent years, observation platforms, numerical weather prediction (NWP) models, data assimilation algorithms, and computational resources have progressed considerably. It is becoming increasingly possible to incorporate satellite, radar, and in situ observations via an ensemble-based data assimilation method (e.g., Ensemble Kalman Filter) in real-time. This allows for the generation of more realistic initial conditions for NWP models, which have proven to be helpful in providing severe weather warning guidance (Roebber et al. 2004; Stensrud et al. 2009, 2013). Thus, researchers have been exploring a transition from warn-on-detection to "warn-on-forecast" (WoF), where numerical guidance plays a more crucial role in the severe weather warning process by significantly extending warning lead times (Stensrud et al. 2009, 2013).

Several case studies have showed that experimental WoF systems (WoFS) can produce accurate short-term probabilistic guidance for hazards such as tornadoes (Snook et al. 2012; Yussouf et al. 2013a,b; Wheatley et al. 2015; Yussouf et al. 2015; Jones et al. 2016), hail

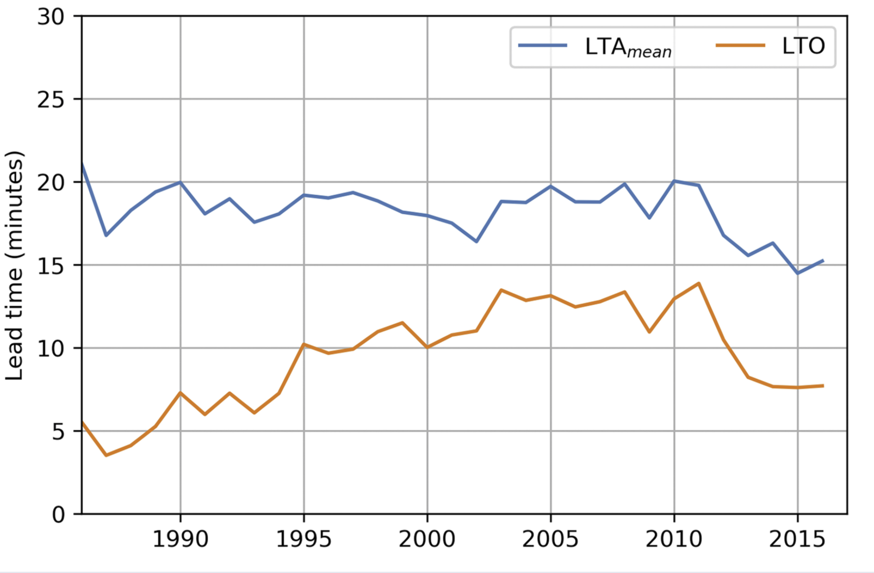


Figure 1.2: From Brooks and Correia (2018), their Figure 3. Average Lead time in advance (LTA$_{mean}$) and official lead time (LTO) for tornado warnings. LTA$_{mean}$ considers only those warnings issued prior to occurrence of a tornado in the warned area while LTO assigns a leadtime of 0 for any tornado that does not have a warning issued before the tornado occurs. See Brooks and Correia (2018) for more details.

(Snook et al. 2016; Labriola et al. 2017, 2019), and heavy rainfall (Yussouf et al. 2016; Lawson et al. 2018a). With continual development of WoFS, however, it is critical to objectively assess the quality of its forecasts, the impact of system configuration changes (e.g., improvements in data assimilation or increasing grid resolution) and inclusion of post-processing techniques (e.g., machine learning calibration) on probabilistic forecast performance. Recently, object-based frameworks have become increasingly common for the verification of convection-allowing model (CAM) forecasts of various severe weather hazards (e.g., Gallus 2010; Johnson et al. 2013; Clark et al. 2014; Cai and Dumais 2015; Stratman and Brewster 2017; Skinner et al. 2018; Jones et al. 2018; Adams-Selin et al. 2019). Object-based verification can easily diagnose or intuitively account for displacement errors between a forecast and observations, and it provides object properties (e.g., orientation, aspect ratio, area) as additional forecast attributes for evaluation (Davis et al. 2006; Ahijevych et al. 2009). Skinner et al. 2018 (hereafter S18) established the first WoFS baseline for the performance of deterministic thunderstorm and mesocyclone predictions. Using an object-based framework, they determined that deterministic forecasts provided for both thunderstorms and mesocyclones across 32 spring cases were skillful overall based on contingency table metrics such as probability of detection and false alarm ratio (defined in Section 2.3). However, a limitation of the work was that no assessment of the accuracy and reliability of the WoFS probabilistic guidance was performed. As an extension of S18, this dissertation develops a novel object-based verification method for storm-scale probabilistic guidance and first applies it to WoFS mesocyclone guidance (see Chapter 4) and then further adapts it for any ensemble storm track for calibrating the WoFS severe weather guidance using machine learning (ML).

Objective verification of probabilistic mesocyclone forecasts from convection-allowing ensembles has thus far been performed in the next-day (6-36 hr) paradigm using grid-based frameworks with neighborhood post-processing (e.g., Gallo et al. 2016, 2018, 2019; Sobash et al. 2016a; Dawson et al. 2017). For next-day forecasts, there are multiple reasons for utilizing neighborhood post-processing. First, at these forecast lead times, intrinsic predictability limits restrict skillful forecasts to broader mesoscale regions rather than the scales representative of individual convective storms (Lorenz 1969). Second, a well-documented flaw of grid-based verification in high resolution forecasts is the infamous "double penalty," where a small spatial displacement between the forecast and an observation leads to both a missed observation and false alarm forecast (Ebert 2008). The result is an unduly negative evaluation of a forecast's predictive skill since, operationally, small spatial displacements are tolerable. Post-processing techniques such as neighborhooding, filtering, or upscaling (i.e., coarsening the verification grid) applied to both forecasts and observations can relax the condition of an exact match and instead assess the scale at which forecasts have the best performance (for a comprehensive discussion on such techniques see Gilleland et al. [2009; 2010] and Schwartz and Sobash [2017]).

A difference between WoF-style and next-day ensemble forecasts is that WoF should provide forecast guidance for individual thunderstorms (Stensrud et al. 2009, 2013). Grid-based verification of WoF guidance can quantify errors associated with the numerical model or data assimilation technique. However, the neighborhooding/filtering/upscaling techniques used by grid-based verification smooth spatial scales associated with convective storms. Therefore, this dissertation I developed a complementary verification technique for WoF guidance that keeps storm-scale forecast information, but allows for operationally tolerable spatial displacements.

Using an object-based framework, we can conceive of forecast probability swaths associated with individual thunderstorms as "probabilistic" forecast objects[1] with a single, representative probability value. Conceptually, we assign a probability of event[2] occurrence within a storm-scale region bounded by the forecast envelope of the event location. The prescribed probability value predicts the likelihood of a storm producing an event rather than the likelihood of an event affecting any point; this distinction and the advantages of event-based probabilistic forecasts are further discussed in Section 4.2. Object-based verification emulates initial forecaster interpretations of WoFS guidance, where forecasters key in on coherent areas of interest in the WoFS model output rather than using the forecast information in a strictly point-by-point basis (Wilson et al. 2019).

Using this object-based approach, one can also objectively assess the potential skill of applying ML-based calibrations to the WoFS forecasts, which has recently become a popular approach for calibrating severe weather probabilistic guidance (e.g., Gagne et al. 2017; Lagerquist et al. 2017; McGovern et al. 2017; Cintineo et al. 2014, 2018; Burke et al. 2019; McGovern et al. 2019b; Hill et al. 2020; Lagerquist et al. 2020; Cintineo et al. 2020; Loken et al. 2020; Sobash et al. 2020; Steinkruger et al. 2020). A key advantage of ML models is their ability to leverage multiple input predictors and learn complex relationships to produce skillful, calibrated probabilistic guidance. An additional advantage for real-time operational

[1] Probability objects will also be referred to as ensemble storm tracks throughout

[2] The event considered in this dissertation is a mesocyclone; however, the technique applies to any storm-generated hazard, as will be shown for the ML-derived probabilities

settings is that once an ML model has been trained, making predictions on new data is computationally quick ($\lll$ 1 s per example). Further discussion on the history of ML in severe weather forecasting can be found in Section 2.2.

In this dissertation I trained gradient-boosted classification trees (Friedman 2002; Chen and Guestrin 2016), random forests (Breiman 2001a), and logistic regression models on WoFS forecasts from the 2017-2019 Hazardous Weather Testbed Spring Forecasting Experiments (HWT-SFE; Gallo et al. 2017) to determine which storms predicted by the WoFS will produce a tornado, severe hail, and/or severe wind report. Besides evaluating the ML performance, this dissertation explores a suite of state-of-the-art ML interpretability methods. ML models unfortunately have the reputation of being seen as "black boxes" where the perception is that the end-user cannot understand the internal workings of the model (McGovern et al. 2019b). Some ML systems, in low-risk situations (e.g., Netflix recommending movies for a user) do not require interpretability, but in high-risk situations (e.g., severe weather forecasting) where missing an event or issuing a false alarm can be costly, decision making must be more deliberate and requires knowing why a model came to its prediction. In the latter situations, robust verification of a complex, end-to-end automated ML system is nearly impossible as one cannot possibly account for a complete list of failure modes (Doshi-Velez and Kim 2017). Therefore, human forecasters will continue to play a role in automated guidance (known as the human in the loop paradigm) and research has shown that the combination of human forecasters and automated guidance has outperformed solely automated guidance for severe weather forecasting (Karstens et al. 2018). Thus, to build human forecasters' trust in ML predictions and maximize the use of automated guidance requires explaining the "why" of an ML model's prediction in understandable terms and creating real-time visualizations of these methods (Hoffman et al. 2017; Karstens et al. 2018).

The following is a summary of my contributions to atmospheric and data science, which are published in Flora et al. (2019) and Flora et al. (2020). I developed a novel object identification method to identify "ensemble storm tracks" from storm-scale probabilistic guidance.

Using this method, I produced the first verification of WoFS-style probabilistic guidance, which is described in Chapter 4. Additionally, I used the novel object identification method to generate severe weather probabilistic guidance from the WoFS using ML (described in Chapter 5). These models for the three severe weather hazards (tornadoes, severe hail, and severe wind) were found to be more skillful and reliable than a competitive baseline generated from the raw WoFS output. To verify the results, I also built on previous research to derive new verification metrics associated with the performance diagram (Roebber 2009). These metrics normalize for the climatological event frequency and allow for the comparison between different datasets. Lastly, I implemented several state-of-the-art interpretation methods to explore and identify relationships learned by the ML models. To achieve this, I developed a full python package known as Model Interpretability in Python (MintPy; Flora and Handler 2020).

The outline of this dissertation is as follows. Chapter 2 discusses how past research has used environmental soundings, CAM-based predictions, and ML methods for severe weather hazard predictions and the verification methods used herein. The WoFS forecast and verification datasets are briefly described in Chapter 3. Chapter 4 describes the initial development of a novel object-based method and its application to verifying WoFS probabilistic mesocyclone guidance, which was published in Flora et al. (2019). The ensemble object identification method was improved upon in subsequent research, which is described in Chapter 5. Chapter 5 also discusses the ML methods used herein, including the predictor engineering, the three ML models mentioned above, and the model tuning and evaluation methods. The ML interpretability methods used are described in Chapter 6. Results from Chapters 4 - 6 are presented in Chapter 7. Conclusions, limitations of the different studies, and avenues for future work are presented in Chapter 8. Additional figures and analysis are provided in 3 appendix chapters.

# Chapter 2: Literature Review

This chapter briefly discusses three main topics:

- The environmental predictors and the development and/or verification of convection-allowing model (CAM) forecasts for the three severe weather hazards
- The history of applied machine learning (ML) research for severe weather hazard prediction
- Important concepts/metrics/diagrams for the verification of rare event probabilistic forecasts of binary outcomes

By understanding the processes associated with severe weather hazards and analyzing past efforts made by researchers to predict them, we can make a better choice of predictors to extract from the WoFS output.

## 2.1 Severe Weather Hazards

### 2.1.1 Severe Hail

**Environmental Parameter-based Prediction**

Forecasting hail severity is challenging given the complexity of hail formation, our limited understanding of the association between storm environments and a given hail size being produced, and the limitation of current microphysical parameterizations to explicitly predict hail size. The difficulty is also compounded by the relatively small sample size of reliable hail observations to calibrate/verify existing methods (an issue for all severe weather hazards) and the regional variability in the parameter space. Of the different convective modes, supercell thunderstorms are the most prolific producer of severe hail [$\geq$1 in (2.5 cm)] and significant severe hail ($\geq$2 in [5 cm]; Duda and Gallus 2010; Smith et al. 2012). The large, quasi-steady state, rotating updraft of a supercell is often sufficiently strong and exists long enough to sustain and grow hailstones in their most efficient hail-formation layer (above the freezing level). Therefore, effective parameters for predicting hail size/severity are those often associated with supercells (e.g., atmospheric instability, 0-6-km wind shear, and 0- to 3-km wind shear and storm-relative helicity) as other properties such as the melting/freezing height and super-cooled water content in the efficient hail-formation layer are often unobserved or poorly sampled by the current sounding network. The strength of the updraft is key as it must balance the downward fall speed of the hailstones. If the updraft is too strong, it can eject hailstones before significant growth occurs, but if the updraft is too weak, hailstone fallout will occur. The environmental wind profile is also critical, as strong storm-relative winds help inject hail embryos into the updraft and deep-layer shear can increase the horizontal extent of the updraft, making it more conducive for hail growth (Dennis and Kumjian 2017).

Given that significant hail growth requires a strong updraft, several studies have attempted to find a relationship between atmospheric instability and hail severity, but they found mixed results (e.g., Huntrieser et al. 1997; Edwards and Thompson 1998; Groenemeijer and Delden 2007; Johnson and Sugden 2014; Tuovinen et al. 2015; Pucik et al. 2015). Edwards and Thompson (1998) found that CAPE was a poor discriminator between different hail sizes, but significant severe hail (≥2 in) did not occur with modified CAPE [the ratio of CAPE to convective cloud depth] less than 1300 $m^2$ $s^{-2}$. Jewell and Brimelow (2009) and Johnson and Sugden (2014) also showed that CAPE exhibited little-to-no skill in discriminating severe hail from non-severe hail events. In contrast, in Europe, Huntrieser et al. (1997) found that greater mid-level instability was associated with thunderstorms producing hail damage (no explicit prediction of hail size). Using a 28-yr dataset from the Netherlands (over 60 K soundings), Groenemeijer and Delden (2007) also found that CAPE distinguished environments with large hail producing thunderstorms from non-hail producing thunderstorms.

CAPE may not always be the best predictor of updraft strength, especially in the most efficient hail-formation layer, which explains the mixed results. However, several studies have shown that moderate CAPE coupled with a high shear environment (e.g., prototypical supercell environments) is associated with large hail (assuming storm initiation; Johnson and Sugden 2014; Tuovinen et al. 2015; Pucik et al. 2015; Dennis and Kumjian 2017; Kumjian et al. 2019). For supercells, the vertical perturbation pressure gradient force associated with the environmental wind shear enhances the longevity of the updraft, which can increase the residence time of a hailstone in the hail-growth region (Dennis and Kumjian 2017; Kumjian and Lombardo 2020). Johnson and Sugden (2014) found that larger hail sizes were associated with higher storm-relative helicity (SRH) and stronger storm-relative winds above 6 km, which is consistent with recent modeling results (Dennis and Kumjian 2017; Gagne II et al. 2019) and smaller hail tends to be associated with weaker shear environments (Kumjian et al. 2019). In a recent study, Kunz et al. (2017) found that both 0-6 km wind shear and 0-3 km SRH are important quantities for large hail (≥2 in), but only in combination with longer-duration storms. Dennis and Kumjian (2017) found significant changes in hail production when environmental wind shear was altered in high-resolution supercell simulations. By increasing the deep-layer zonal shear, the storm's updraft was elongated in the same direction, which increased the favorable region of hail growth and hailstone residence times within the updraft. However, increasing low-level meridional wind shear reduced hail mass by separating the favorable embryo source region and hydrometeors to serve as embryos.

Other potential predictors of hail severity are based on the depth (or the minimum height) of the optimal hail growth layer above the freezing level (Edwards and Thompson 1998; Johnson and Sugden 2014). Moisture content below the freezing level or in the boundary layer also has an influence on hydrometeor density and the growth rates of larger hail (Allen et al. 2015; Johnson and Sugden 2014). Grant and van den Heever (2014) analyzed the impact of varying mid-level moisture content on different simulated supercell structures ("classic" vs. "low precipition" supercells) and their respective hail productions. Changing the mid-level moisture content altered the storm-relative winds and led to different hail growth mechanisms for the different supercell structures. Classic supercells had higher riming rates on the western side of the updraft while riming rates in low-precipitation supercells were higher on the north/northeast side of the updraft. Studies have also suggested that the lifting condensation level may be a useful predictor of hail size (Pucik et al. 2015; Groenemeijer and Delden 2007).

Researchers have developed composite parameters to forecast hail severity, but a robust evaluation of their performance is lacking. For example, forecasts using the significant hail parameter (a combination of CAPE, mixing ratio of a parcel, environmental mid-level lapse rate, 500-hPa temperature, and deep-layer shear) have yet to be rigorously evaluated in the literature. Johnson and Sugden (2014) tested the significant hail parameter, but found it did not differentiate well for hail size compared to other methods. Instead, they derived the large hail parameter (LHP), which includes properties of the vertical wind profile, most unstable CAPE, mid-level lapse rate, and hail growth zone thickness and found that it could better discriminate between ≥2 in hail and smaller hail sizes as compared to simpler CAPE-deep-layer wind shear products.

### CAM-based Prediction

Since estimating hail size from environmental predictors can have varying degrees of success, researchers have explored using CAM model surrogates (e.g., updraft helicity; Gagne et al. 2017, Adams-Selin et al. 2019; Burke et al. 2019), explicit hail size prediction from the microphysics parameterization (e.g., Mansell et al. 2010; Milbrandt and Morrison 2013; Morrison and Milbrandt 2015), or from an additional model coupled to the NWP model (e.g., HAILCAST, Brimelow et al. 2002; Jewell and Brimelow 2009; Adams-Selin and Ziegler 2016) to predict severe hail. Though horizontal grid spacing used in most operational CAM ensembles is too coarse to resolve severe weather hazards, storm surrogates such as updraft helicity have showed skill for next-day and short-time [e.g, $O(1\text{ h})$] severe weather prediction (e.g., Sobash et al. 2011, 2016b; Snook et al. 2012; Yussouf et al. 2013a,b; Wheatley et al. 2015; Yussouf et al. 2015; Jones et al. 2016; Skinner et al. 2016, 2018; Jones et al. 2019;

Flora et al. 2019; Yussouf et al. 2020). Unfortunately, few studies have verified the skill of CAM severe storm surrogates to isolate the hail-specific threat. Typically, hail-specific surrogates from CAM output have only be verified as a baseline product for other methods (e.g., machine learning-based products, WRF-HAILCAST; Gagne et al. 2017, Adams-Selin et al. 2019; Burke et al. 2019). In those studies, updraft helicity was found to be a successful predictor of severe hail, but it is a limited product since it does not account for non-rotating, severe thunderstorms and only leverages a small portion of the CAM model output.

Besides CAM severe weather surrogates, we can estimate hail size from the predicted microphysical state variables (Snook et al. 2016; Labriola et al. 2017, 2019, 2020). These studies have found explicit hail forecasts to be marginally successful, but the methods have several limitations: our understanding of microphysical processes is lacking (and therefore the processes are poorly modelled), properly capturing the hail size distribution often requires higher-moment models (Milbrandt and Yau 2006), and the thresholds for determining severe hail from a hail size distribution are defined ad hoc. Moreover, these studies tend to be research-oriented (e.g., 500-m resolution, double- or triple-moment microphysics schemes) and it is unclear whether their results will translate in real-time settings with more operationally-relevant schemes.

Another option for explicit hail prediction is using a coupled model such as HAILCAST (Brimelow et al. 2002). The original HAILCAST model was a stand-alone hail growth model that relied on sounding-derived predictors and an approximation for updraft longevity to estimate maximum hail size at the surface. Unlike microphysics schemes that predict the total amount of hail over an extensive region, HAILCAST predicts the growth of just a few hailstones at each grid point to determine how large a hailstone can grow given a vertical profile. This method also has limitations: it is a single column model that cannot advect hail horizontally (important to hail production), and also uses poorly understood microphysical processes. The recent implementation of HAILCAST, known as WRF-HAILCAST (Adams-Selin and Ziegler 2016), uses NWP model predicted variables and coupled microphysics parameterization to predict the maximum hail size at the surface. Recent verification of WRF-HAILCAST in the NOAA/Hazardous Weather Testbed Spring Forecasting Experiments found that it was comparable in skill to storm surrogate fields (e.g., updraft helicity) and human forecasters when predicting >2 in hail (Adams-Selin et al. 2019).

### 2.1.2 Severe Wind

#### Environmental Parameter-based Prediction

Severe convective wind gusts (also referred to as non-tornadic, damaging straight-line winds), unlike large hail and tornadoes, can occur in a wide variety of environments. Severe convective wind gusts can be attributed to either long-lived convective windstorms, short, local downbursts, or a combination of both. To compound the difficulty, similar environmental conditions may cause a quasi-stationary mesoscale convective system (MCS) or a rapidly moving one, depending on the orientation of the prevailing flow to the storm outflow (Corfidi 2003). One primary mechanism for severe convective winds is a strong downdraft (known as a downburst) driven by precipitation-cooled air (and precipitation loading) and steep lapse rates, which allows the downdraft air to remain negatively buoyant as it warms upon descent. As the downdraft hits the surface and spreads out horizontally, the surface wind can be intense and cause damage.

Downbursts are one mechanism for producing severe convective winds, but there are additional processes that increase the strength of the downdraft and corresponding horizontal momentum relative to specific convective modes. For example, dynamic pressure forces can also drive supercell downdrafts (such as the rear-flank or occlusion downdrafts) (Wakimoto 2001). As the low-level mesocyclone intensifies, the pressure is lowered locally, and the dynamically-induced pressure gradient draws the air down from above. Nonlinear dynamic pressure perturbation forces in the region between the low-level mesocyclone and anti-mesocyclone can also cause momentum surges in the rear-flank downdraft (Skinner et al.

2015). MCSs can enhance severe winds at the surface through cold pool dynamics and/or a descending rear-inflow jet (Houze Jr. 2004). As the outflows from the many updrafts merge into a single cold pool, the horizontal pressure gradients associated with the cold pool can cause severe winds in the absence of any intense downbursts. The internal dynamics of organized systems can also contribute to local enhancements in surface winds when a rear-inflow jet descends to the surface (Weisman 1993). Lastly, an MCS's cold pool can lead to convective redevelopment in environments already capable of producing severe convective winds.

MCSs and supercells are the primary producers of severe convective wind reports (Smith et al. 2013; more so MCSs), so the parameters often associated with these convective modes are proxies for severe wind potential. A study by Doswell and Evans (2003) found that proximity soundings for strongly forced bow echoes and supercells were almost identical. Thus, it is not surprising that Coniglio et al. (2010) found that long-lived MCSs thrive on higher CAPE and vertical wind shear similar to supercells, but Evans and Doswell (2001) found that CAPE and vertical wind shear do not separate derechos—a long-lived MCS producing widespread, damaging wind—from non-severe MCSs. In some situations, severe wind gusts can occur in high shear, low CAPE windstorms with strong horizontal pressure gradients and synoptic-scale forcing where wind gusts are amplified by convection (Evans and Doswell 2001; Clark et al. 2009; Gatzen 2011; Pucik et al. 2015), though these situations are more relegated to cold season thunderstorms. Local downbursts may even form with both small CAPE and weak shear, in cases where the boundary layer is deep and dry (Wakimoto 1985). The boundary layer dryness enhances evaporative cooling and promotes negative buoyancy in the downdraft. To estimate the strength of the downdraft, forecasters use downdraft CAPE (DCAPE; Gilmore and Wicker 1998). In a systematic evaluation of severe convective wind environments, Kuchera and Parker (2006) found that the combination of DCAPE and ground-relative wind in a storm's inflow layer was the most successful predictor of convective severe wind gusts.

#### CAM-based Predictions

Researchers have primarily focused on using CAM forecasts to predict tornadoes and severe hail with little work done to develop or verify techniques for diagnosing severe winds. One issue is that CAMs cannot fully resolve the convective processes necessary for the correct representation of near-surface convective wind gusts (Bryan et al. 2003), so forecasters are required to use a threshold lower than 50 kts to separate severe from non-severe winds in CAM forecast output. For example, Hepper et al. (2016) used a 30 kts threshold for Storm-Scale Ensemble of Opportunity forecasts, which was found to be too low to generate meaningful guidance for severe wind likelihood since it could not discriminate between a high-end derecho event and a low-end non-severe MCS event. Jirak et al. (2014) found that of the three severe weather hazards, the Short-Range Ensemble Forecasts performed worst at predicting severe wind likelihood. They found the result unsurprising, as a variety of convective modes and environments can produce storms with damaging wind gusts. Severe wind reports are also notorious for being of suspect quality (perhaps more so than the other two severe weather hazards; Trapp et al. 2006), which limits reliably assessing the performance of severe wind guidance.

### 2.1.3 Tornadoes

#### Environmental Parameter-based Prediction

Tornadoes, especially ≥EF2, are almost exclusively associated with supercells (Duda and Gallus 2010; Smith et al. 2012). However, distinguishing between tornadic and non-tornadic supercell storms has a long, storied history and remains an active area of research. Tornadoes are favorable in supercells because of their internal dynamics (e.g., the low-level mesocyclone, dynamic pressure perturbation forces). A necessary precursor for tornadogenesis is the development of a low-level mesocyclone (LLM). The LLM forms from the tilting of the storm-generated low-level horizontal vorticity associated with horizontal buoyancy gradients produced by the forward-flank downdraft (FFD). The development of the LLM provides the low-level updraft necessary for stretching near-surface vertical vorticity into the cloud base. The existence of a LLM, though, is not a sufficient precursor for tornadogenesis as observational studies like Trapp et al. (2005) have found that only 40% of LLMs are associated with tornadoes.

Besides the traditional parameters associated with supercells (e.g., CAPE, deep-layer shear), we know from proximity sounding analysis that tornadic supercells are favorable in environments with lower lifting condensation level (LCL) heights and strong low-level storm-relative helicity (SRH) and wind shear, respectively (e.g., Brooks et al. 1994; Rasmussen and Blanchard 1998; Markowski et al. 2003; Thompson et al. 2003; Anderson-Frey et al. 2017; Coffer et al. 2019; Coniglio and Parker 2020). Lower LCL heights are important for two reasons:

1. lower LCL heights mean a lower cloud base for the stretching vertical vorticity column to attach to,

2. lower LCL heights can limit the rear-flank downdraft (RFD) strength (and its potential to undercut the LLM and cause tornadogenesis failure) because of weaker evaporative cooling (since the air is closer to saturation; Markowski et al. 2002).

By limiting the storm outflow strength, the mid-level mesocyclone and LLM can stay in alignment making for favorable dynamic updraft forcing to stretch and intensify low-level rotation (Brown and Nowotarski 2019; Homeyer et al. 2020). Strong low-level SRH is important as stream-wise vorticity coupled with the baroclinically-induced horizontal vorticity from the evaporatively cooled downdraft can increase the strength and longevity of the LLM (Davies-Jones 1984; Davies-Jones and Brooks 1993; Markowski and Richardson 2013; Mashiko 2016b; Coffer and Parker 2016). In recent modeling studies, tornadogenesis is well correlated with LLM strength (e.g., Mashiko 2016a,b; Roberts et al. 2016, 2020; Yokota et al.

2018). In the most extensive study of supercell environments, Coniglio and Parker (2020) found that tornadic environments also have smaller 0-3-km temperature lapse rates because of weaker/shallower capping inversions and larger 0-3-km CAPE.

### CAM-based Predictions

Of the three severe weather hazards, tornadoes have received the most attention from the operational CAM research community (Sobash et al. 2011, 2016a; Clark et al. 2012, 2013; Gallo et al. 2016, 2017, 2018, 2019; Sobash et al. 2019). Early studies by Clark et al. (2012, 2013) found that daily accumulated updraft helicity (UH) swaths were positively correlated with total tornado path length. Sobash et al. (2016a) found that next-day forecasts of strong low-level rotation occurred in environments consistent with proximity sounding based tornadic environments. Gallo et al. (2016, 2017, 2018, 2019) combined next-day CAM ensemble forecasts of UH with the significant tornado parameter (STP; Thompson et al. 2003), environmental information, and climatological tornado frequencies, respectively, to produce skillful and relatively reliable probabilistic tornado guidance.

Several studies have also examined the ability of a WoF-type system to assimilate observed tornadic supercells and provide 0-1 h probabilistic numerical forecasts of low-level vertical vorticity and/or UH (Dawson et al. 2012; Yussouf et al. 2013b,a, 2015, 2016; Potvin and Wicker 2013; Wheatley et al. 2015; Skinner et al. 2018; Flora et al. 2019). Dawson et al. (2012) and Potvin and Wicker (2013) conducted experiments with horizontally homogeneous ICs and generated probabilistic forecasts of low-level rotation of supercell storms. Both studies concluded that short-range probabilistic forecasts of low-level rotation could be achieved with reasonable accuracy. Yussouf et al. (2013a, 2015) showed the capability of the WoF-type system to provide relatively accurate estimates of intense LLM tracks that align well with the locations of radar-derived rotation tracks associated with the observed tornadic storm. In particular, Yussouf et al. (2013b) found that an improved representation of mesoscale heterogeneity in the near-storm environment produced more accurate ensemble

Kalman Filter analyses of tornadic supercell thunderstorms and improved probabilistic forecasts of low-level rotation. As for Yussouf et al. (2016), they concluded that both low-level rotation and rainfall probabilistic forecasting are possible with a WoF-type system. Wheatley et al. (2015) found that a WoF-type system could produce areas of intense low-level rotation approximately 30 minutes before the first observed tornado in cases of supercells and MCSs. Including the clear-sky satellite data in the data assimilation, Jones et al. (2016) found that it reduced anomalous cloud cover and improved thermodynamics conditions leading to higher probabilistic forecasts of strong low-level rotation that corresponds well with observed tornado tracks.

## 2.2 Applications of Machine Learning in Severe Weather Prediction

The previous sections discussed sounding- and CAM-based prediction for the three severe weather hazards. However, the sounding network and operational CAM horizontal grid spacing are often too coarse to resolve smaller-scale processes and information that would be valuable to forecasters. This section highlights an additional effort to improve severe weather prediction by machine learning (ML) methods which rely on a data-driven process to develop a prediction system.

Using ML methods to produce probabilistic severe weather forecasts dates as far back as the early 1970s (Alaka et al. 1973; Reap 1974; Klein and Glahn 1974; Reap and Foster 1979; Charba 1979). These early studies used forward stepwise[1] multiple linear regression (linear regression with multiple predictors), a process made popular in meteorology by model output statistics (MOS; Glahn and Lowry 1972). In Charba (1979), the goal was to predict any severe weather hazard (severe wind gusts >50 kts, tornado or hail >0.75 in) in a 4-hour

[1] also known as screening and/or forward selection in the literature; the forward stepwise method (Glahn and Lowry 1972) refers to the predictor selection process. In this method, the first predictor is the one most correlated with the target variable. Then the next predictor is the one that leads to the greatest reduction of variance when coupled with the first predictor. The selection process continues until some stopping criterion is met.

window (2-6 h lead time) over an 85 x 85 nautical mile square area. The predictors included hourly observed surface conditions, numerical weather prediction (NWP) output from the Limited Area Fine Mesh model, and radar data for 37 predictors total. The prediction produced a positive Brier skill score (defined in Section 2.3.3) and a bias near 1.0. As a complement to Charba (1979), Reap and Foster (1979) focused on severe weather prediction at longer lead times (e.g., 12-36 h), which also produced fairly reliable results.

Although early studies showed promise, ML approaches to severe weather prediction were not widely adopted until the mid-1990s, which coincided with the development of the Weather Surveillance Radar-1988 Doppler (WSR-88D) network. With the WSR-88D network, meteorologists could collect large amounts of observational data, including reflectivity and radial velocities. With these large datasets, the focus of applied ML research in severe weather prediction shifted to nowcasting (<1 h lead times) approaches (Kitzmiller et al. 1995; Billet et al. 1997; Marzban and Stumpf 1996, 1998; Alexiuk et al. 1999; Marzban and Witt 2001). Kitzmiller et al. (1995) developed the Severe Weather Potential algorithm, which used linear regression to predict the likelihood of a storm cell producing any severe weather hazard within the next 20 minutes. The input predictors included multiple variations of vertically integrated liquid (VIL) and the horizontal areal extent of the storm cell. Using linear and logistic regression (defined in Section 5.2.1), Billet et al. (1997) derived equations from a combination of VIL, freezing level, and low-level storm inflow to predict hail diameter and probability of severe hail (size ≥0.75 in), respectively, which was the first method to predict a specific hazard rather than "any severe." Although predicting hail size was found to be of limited use, logistic regression produced a fairly reliable probability of severe hail.

Until the mid-1990s, linear regression-based algorithms were the common approach in meteorology, but with the development of techniques like back-propagation (Rumelhart et al. 1985) there was a renewed interest in neural networks. Marzban and Stumpf (1996) is the earliest example of a neural network-based severe weather prediction system. They trained a neural network to predict whether a circulation detected by the National Severe

Storm Laboratory (NSSL) mesocyclone detection algorithm (MDA; Stumpf et al. 1998) would produce a tornado in the next 20 minutes. They found the method outperformed the pre-existing rule-based algorithm for classifying MDA-identified circulations. In a follow-up paper, Marzban and Stumpf (1998) applied a neural network to the NSSL MDA-identified circulations with the goal of predicting the probability of all damaging winds (both straight-line and tornadic) using only radar-derived predictors. They found that including hidden nodes improved performance, but the effect of balancing event and non-event examples in the training dataset was error metric dependent. Alexiuk et al. (1999, 2000) used a variety of ML algorithms (decision trees, Fuzzy $K$-means clustering, neural networks, $K$ nearest neighbors, learning vector quantization) to classify storm cells into one of four classes: tornado, hail, severe wind, and heavy rain. Alexiuk et al. (1999) found that fuzzy $K$-means clustering produced the best results and tornado events were much more easily discriminated from hail events than either heavy rain or severe wind events. Building upon that work, Alexiuk et al. (2000) used principal component analysis (PCA) to reduce the dimensionality of the data. However, PCA led to a decrease in performance in all cases. To complement the neural network developed for tornado/wind prediction (e.g., Marzban and Stumpf 1996, 1998), Marzban and Witt (2001) developed a neural network for explicit hail size prediction and one for different, nominal categories (small, medium, large) for objects identified by the NSSL hail-detection algorithm. The neural network outperformed the NSSL hail-detection algorithm at predicting hail size, while the probabilistic neural network produced highly reliable and discriminatory probabilities for the smallest and largest hail categories, but struggled for mid sized hail.

In the early 2000s, ML-based severe weather forecasting became increasingly focused on tornado prediction and improving upon the operational NSSL MDA (Trafalis et al. 2003, 2005; Lakshmanan et al. 2005; Trafalis et al. 2007; Adrianto et al. 2009; Trafalis et al. 2013). Given the computational limitations in operational settings, many studies only used radar data-derived predictors. Researchers also began exploring support vector machines

(SVMs; Cortes and Vapnik 1995), which had become popular at the turn of the century. In Trafalis et al. (2003, 2005), they found Bayesian neural networks and SVMs performed significantly better than traditional neural networks for tornado forecasting. Lakshmanan et al. (2005) and Adrianto et al. (2009) used fuzzy logic and SVMs, respectively, to produce a 30-min gridded tornado probability. Trafalis et al. (2013) built on the work of Marzban and Stumpf (1996), with the goal of finding a better solution to the rare-event problem. They applied three ML models–logistic regression, SVMs, and random forests–to radar data and reanalysis-derived near-storm environment (NSE) data and found that the best predictors were related to deep-layer shear, relative humidity, DCAPE, and low-level rotation.

In the last decade, studies have incorporated distinct datasets beyond radar data (e.g., satellite observations, surface data, NWP model output, etc) as predictors and implemented previously untested methods such as random forests, gradient-boosted trees, and convolutional neural networks (Lopez et al. 2007; Gagne et al. 2012; Manzato 2013; Cintineo et al. 2014; Lagerquist et al. 2017; Cintineo et al. 2018; Czernecki et al. 2019; Lagerquist et al. 2020; Cintineo et al. 2020; Yao et al. 2020; Steinkruger et al. 2020). Lopez et al. (2007) developed a short-term hail occurrence forecast from sounding-derived indices using logistic regression. Spatiotemporal relational random forests (SRRFs; McGovern et al. 2013) were used to predict the tornado probability of radar-observed supercells (Gagne et al. 2013) and next-day severe hail from CAM ensemble output (Gagne et al. 2012). Manzato (2013) used an ensemble of neural networks to predict hail occurrence and size using sounding-derived indices. The ProbSevere model (Cintineo et al. 2014, 2018) is a naïve Bayesian classifier and reliably predicts severe weather likelihood up to a lead time of 90 min. In a newer version, ProbSevere v2.0, the system can now produce probabilistic guidance for individual severe weather hazards (tornadoes, hail >1 in., and/or wind gusts >50 kts; Cintineo et al. 2020) and recently became an operational product. In an idealized framework, Steinkruger et al. (2020) explored using ML methods to produce automated tornado warning guidance and found promising results. Using 4 different algorithms –random forest, neural networks, gradient-boosted trees, and logistic regression– Lagerquist et al. (2017) produced skillful probabilistic severe wind predictions for radar-observed storms using radar data and NSE variables from NWP model output as predictors. Czernecki et al. (2019) trained a random forest on radar reflectivity, lightning detection data, and sounding-indices derived from reanalysis data to predict large hail. The model produced fairly skillful results and was largely driven by the radar reflectivity and composite indices such as the significant hail parameter and large hail parameter. Recently, using a convolution neural network (CNN; LeCun et al. 1990), a deep learning technique, Lagerquist et al. (2020) produced a next-hour tornado prediction system comparable to the ProbSevere system. Yao et al. (2020) using a 15-yr dataset, trained a random forest to predict 0-6 h hail occurrence. They found that the random forest focused on thermal predictors such as the lifted index, Showalter stability index, and total index.

Recently, studies have investigated ML-based severe weather forecasting at longer lead times (24-36-h; e.g., Gagne et al. 2017; Burke et al. 2019; Hill et al. 2020; Loken et al. 2020; Sobash et al. 2020) because of the growing archive of CAM forecasts. ML models such as random forests (Breiman 2001a) have produced competitive next-day hail predictions (Gagne et al. 2017; Burke et al. 2019), reliable next-day severe weather hazard guidance (Loken et al. 2020), and even outperformed the Storm Prediction Center (SPC) Day 2 and 3 outlooks (Hill et al. 2020). Neural networks have also shown success in predicting next-day severe weather and were more skillful than an UH baseline (Sobash et al. 2020).

## 2.3 Verification of Probabilistic Forecasts of Binary Outcomes

What is a good forecast? To answer this question, Murphy (1993) identified 3 “types” of goodness:

1. Consistency: the correspondence between forecasters’ judgments and their forecasts

2. Quality: correspondence between the forecasts and the matching observations

3. Value: the benefit realized by the end user’s use of the forecast

Traditionally, it is difficult to assess consistency and value, and therefore this dissertation will focus solely on forecast quality. To assess the forecast quality of forecast probabilities of binary outcomes requires discussing 3 important verification diagrams (and their accompanying scalar metrics). These diagrams include the receiver operating characteristic (ROC; Metz 1978) diagram, the performance diagram (Roebber 2009), and the attribute diagram (Hsu and Murphy 1986). Though additional diagrams and verification metrics exist, these three verification diagrams summarize how well the forecast probabilities can discriminate between event and non-event (ROC diagram), how correctly the probabilities can predict events (performance diagram), and how reliable the probabilities are (attribute diagram).

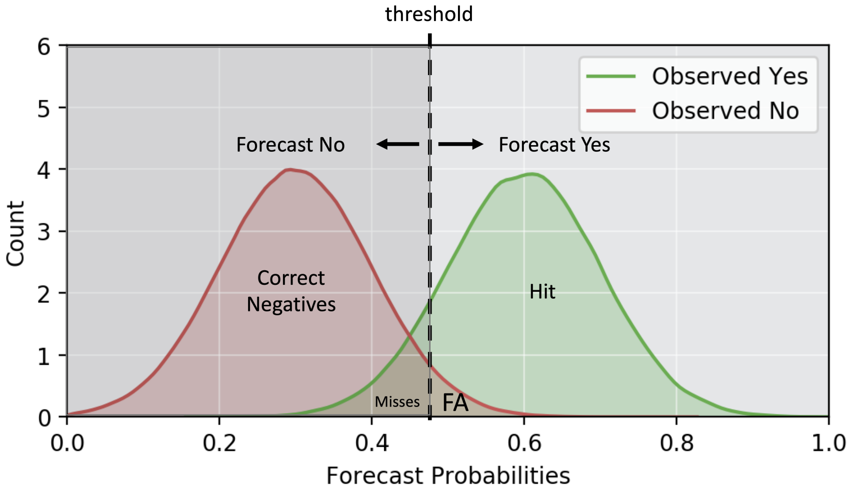


Figure 2.1: Distribution of forecast probabilities conditioned on being matched to an observed yes (green) or observed no (red). Forecast probabilities are converted to yes/no forecasts based on some threshold (e.g., 45% in this example). The regions of the two distributions are annotated by their corresponding contingency table term. FA is short for false alarms.

Before discussing these diagrams, it is important to define some key terms and provide illustrations that will help facilitate our understanding of the following verification metrics.

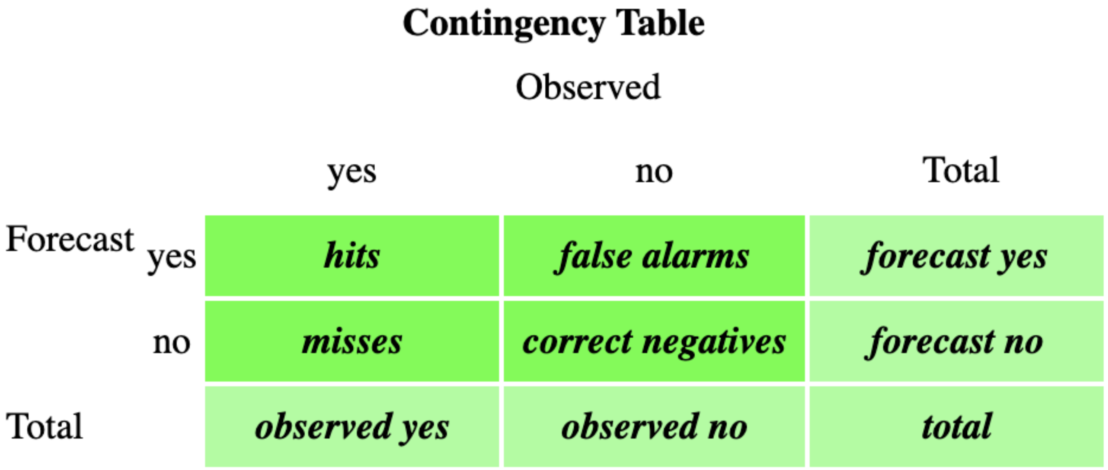
For binary outcomes, forecast probabilities are either associated with an event (observed yes) or a non-event (observed no; Figure 2.1). The forecast probabilities can then be converted to yes/no forecasts based on some threshold. We can then build a contingency table from the binarized forecast probabilities and binary outcomes, respectively (Figure 2.2). The four combinations in the contingency table (which can be seen in Figure 2.1) are:

1. Hits (h): forecast for event to occur and the event occurred

2. False Alarms (f): forecast for event to occur, but the event did not occur

3. Misses (m): forecast for event to not occur, but the event did occur

4. Correct Negatives (c): forecast for event to not occur and the event did not occur



**Contingency Table**

Observed

| | | yes | no | Total |
|---|---|---|---|---|
| Forecast | yes | ***hits*** | ***false alarms*** | ***forecast yes*** |
| | no | ***misses*** | ***correct negatives*** | ***forecast no*** |
| Total | | ***observed yes*** | ***observed no*** | ***total*** |

Figure 2.2: Example of a contingency table, which includes four components: hits, false alarms, misses, and correct negatives. See text for definitions of these terms. The figure comes from https://www.cawcr.gov.au/projects/verification/.

These terms for the contingency table components are the nomenclature in meteorology, but the generic names are true positives (hits), false positives (false alarms), false negatives (misses), and true negatives (correct negatives), respectively (true/false refers to the forecast while positive/negative refers to the binary outcome). One can compute multiple metrics

Table 2.1: Common verification metrics associated with the components of the contingency table (non-exhaustive list). The terms $h, m, f, c$ refer to hits, misses, false alarms, and correct negatives, respectively.

| Metrics | Formulas |
|---|---|
| Probability of Detection (POD) | $\frac{h}{h+m}$ |
| Probability of False Detection (POFD) | $\frac{f}{f+c}$ |
| Success Ratio (SR) | $\frac{h}{h+f}$ |
| Critical Success Index (CSI) | $\frac{h}{h+m+f}$ |
| False Alarm Ratio (FAR) | $\frac{f}{h+f}$ |
| Frequency Bias (BIAS) | $\frac{h+f}{h+m}$ |

Table 2.2: Aliases for the contingency metrics in Table 2.1.

| Metric | Aliases |
|---|---|
| Probability of Detection (POD) | Sensitivity, Recall, Hit Rate, True Positive Rate |
| Probability of False Detection (POFD) | Fall-out or False Positive Rate |
| Success Ratio (SR) | Precision |
| Critical Success Index (CSI) | Threat Score |
| False Alarm Ratio | False Discovery Rate |

from the contingency table components (Table 2.1). Again, the names are nomenclature in meteorology, but are often referred to by their generic names in other disciplines (Table 2.2).

### 2.3.1 The ROC Diagram

The ROC diagram plots POD against POFD for a series of different probability thresholds (Figure 2.3). The POD is the probability that if an event occurs that it will be forecasted correctly and POFD is the probability that if an no event occurs that it will be forecasted incorrectly. Therefore, a forecast system that can maximize POD while minimizing POFD can discriminate well between events and non-events. To summarize the ROC curve as a single metric, one can compute the area under the ROC curve (AUC). The AUC (which

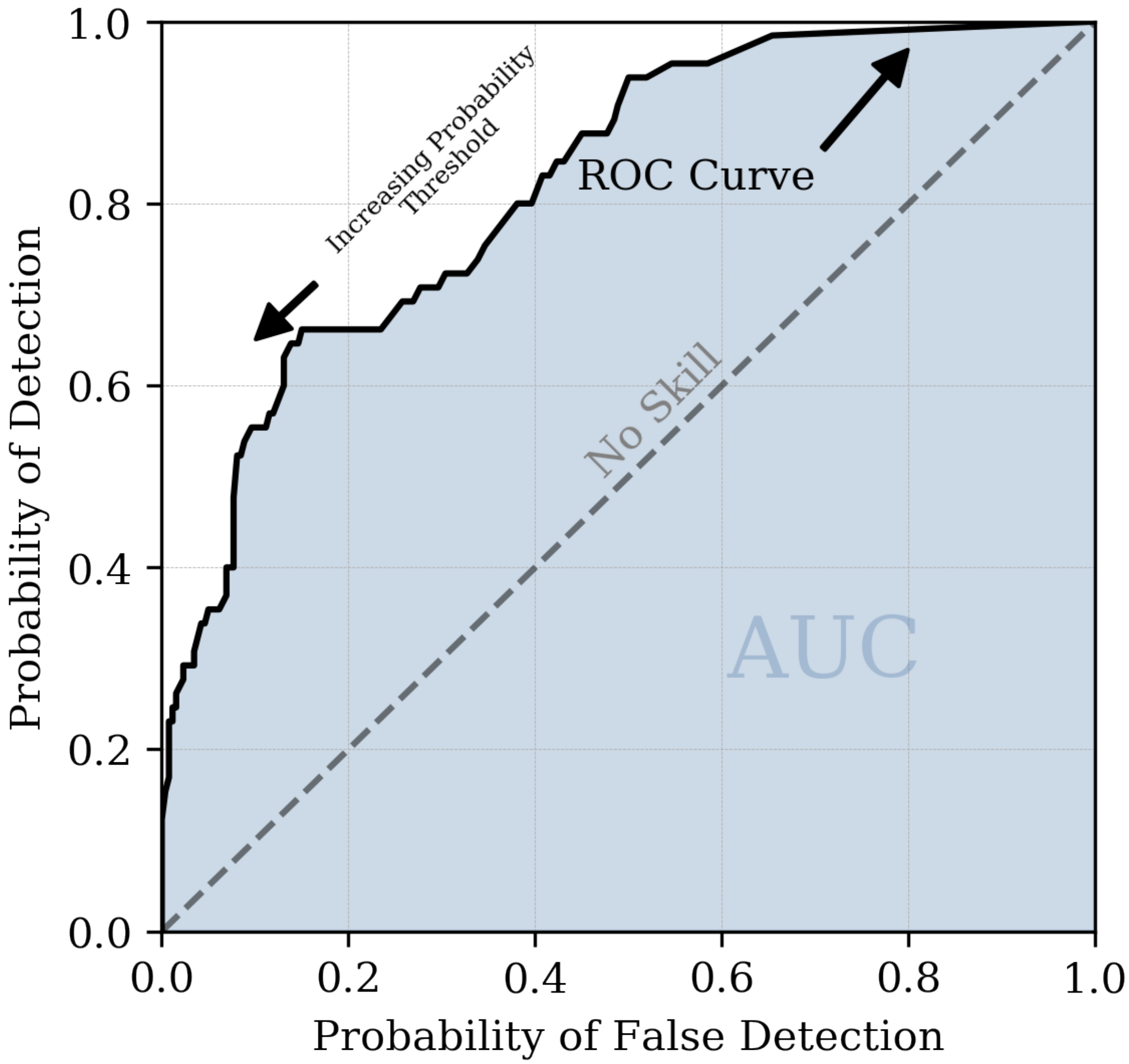


Figure 2.3: An example of a ROC diagram. The ROC curve is derived by computing POD and POFD for forecast probabilities based on a series of probability thresholds (where the increasing probability threshold is from the upper right hand to the lower left hand). The curve is summarized by the area under the curve (shown in blue shading). An AUC = 0.5, indicated by the dashed diagonal lines, represents a no skill system.

is a special case of the Mann-Whitney U-test; Neuhäuser 2011) can be interpreted as the probability that when given a random pair of event and non-event examples, our model will correctly rank them. A classifier that can perfectly discriminate between events and non-events will produce an AUC = 1 while a classifier that predicts randomly (has no skill) will produce an AUC = 0.5. The AUC has two important properties:

1. Scale-Invariant
2. Skew-Invariant

The first property of AUC says that it is insensitive to the absolute value (scale) of the forecast probabilities, since it only considers how well the forecast probabilities are ranked. For example, dividing or multiplying all the forecast probabilities by a constant term will not alter their rank. Thus, it is important to couple AUC with metrics that are penalized by poor calibration (e.g., Brier skill score; defined in Section 2.3.3). The second property of AUC says that it is insensitive to the ratio of events to non-events in the dataset (often referred to as its skew) as it weights events and non-events equally. Therefore, AUC by itself is not an appropriate metric for evaluating rare event forecasting. For example, AUC may provide an overly optimistic assessment of discrimination in applications where less importance is placed on correctly predicting non-events. For severe weather prediction, correct negatives are conditionally important because it is only desirable to accurately predict non-events in environments that favor severe weather (to reduce false alarms).

### 2.3.2 The Performance Diagram

The performance diagram[2] plots the SR against the POD for a series of different probability thresholds and assesses the ability of the forecast probabilities to correctly predict an event while ignoring correct negatives (Roebber 2009; Figure 2.4). The SR is the probability that when an event is forecasted an event will occur. Therefore, a perfect forecast system

[2] Commonly known as the precision-recall diagram (Manning and Schtze 1999)

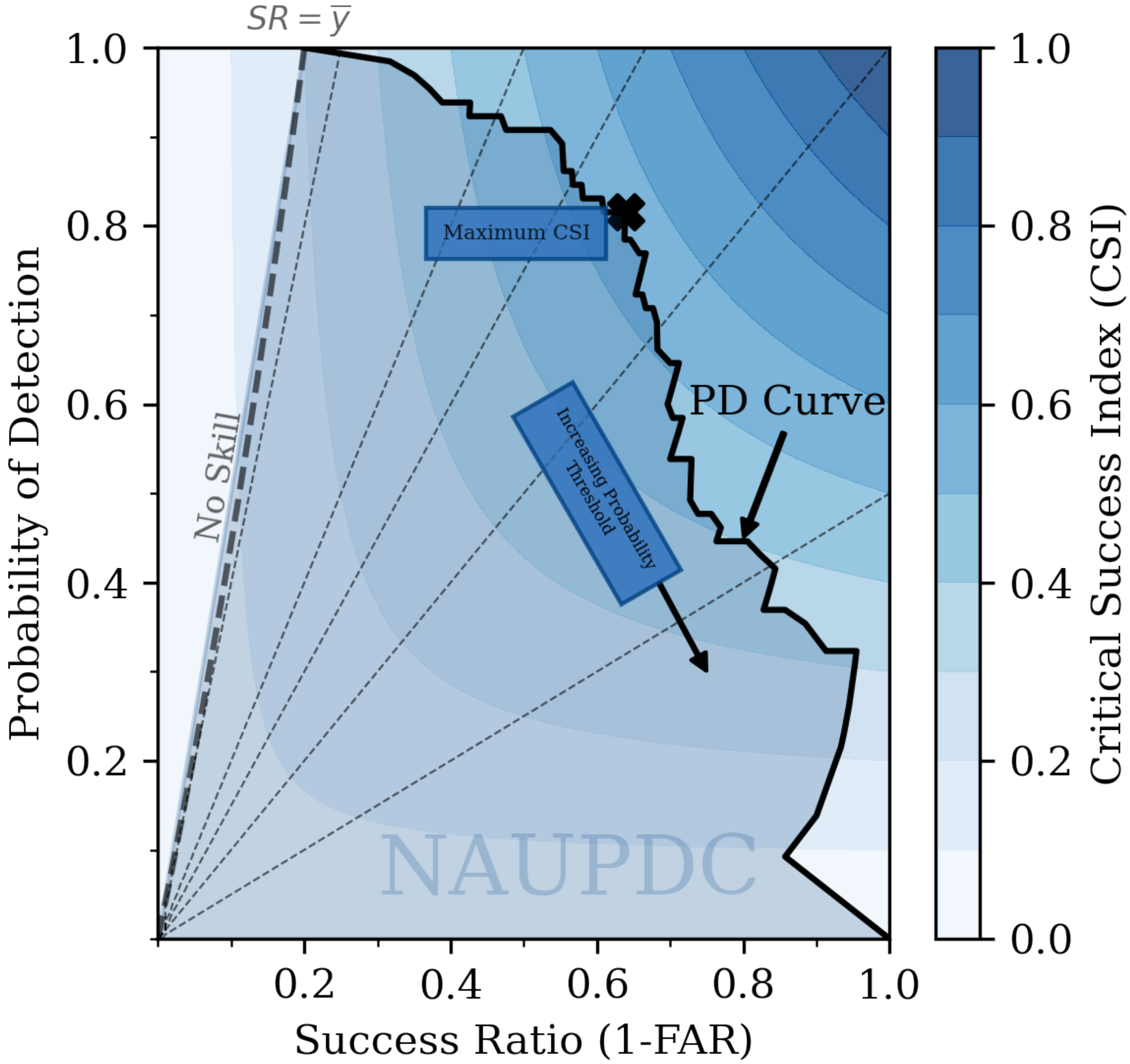


Figure 2.4: An example of a performance diagram (PD). The filled contours are the critical success index while the black dashed diagonal lines emanating from the origin are the frequency bias. The PD curve is derived by computing POD and SR for forecast probabilities based on a series of probability thresholds (where the increasing probability threshold is from the upper left hand to the lower right hand). A no-skill system produces a PD curve along the gray dashed line which is dependent on the climatological event frequency of the dataset ($\overline{y}$ where $y$ is the binary target variable). The curve can summarized based on the area between the no-skill curve and PD curve, which is known as the normalized area under the PD curve (NAUPDC). Another important feature is the location of maximum critical success index (CSI).

should maximize SR and POD. The performance diagram is complementary to the ROC curve, especially for imbalanced prediction problems (like severe weather forecasting) where it is more important to correctly predict events than non-events (Davis and Goadrich 2006). CSI and frequency bias are functionally related to POD and SR and are also displayed on the performance diagram. A probabilistic forecast is considered to have perfect performance when the CSI and frequency bias are equal to 1 (corresponding to the upper right corner) for some probability threshold. However, for probabilistic forecasts of rare events, a maximum CSI of 1 is practically unachievable (Hitchens et al. 2013) and the maximum CSI tends to be associated with a frequency bias >1 (Baldwin and Kain 2006).

Similar to the ROC Diagram, one can compute the area under the performance diagram curve (AUPDC[3]). Rather than computing the area through integration, which can be too optimistic, it is more robust to compute AUPDC from the weighted average SR[4] (Boyd et al. 2012):

$$\mathrm{AUPDC} = \sum_{k=1}^{K} (\mathrm{POD}_k - \mathrm{POD}_{k-1})\mathrm{SR}_k, \tag{2.1}$$

where $K$ is the number of probability thresholds used to calculate POD and SR. Unlike AUC, AUPDC is skew-dependent and changing the ratio of events to non-events will alter the minimum possible SR defined in Boyd et al. (2012) as:

$$SR_{min} = \frac{c\mathrm{POD}}{1 - c + c\mathrm{POD}}, \tag{2.2}$$

where $c$ is the climatological event frequency of the dataset (number of events divided by the total number of examples). If a curve lies along $SR_{min}$, the prediction system is considered to have no skill. Therefore, one can normalize AUDPC by the minimum possible AUPDC (Boyd et al. 2012), which facilitates comparing the model skill on datasets with different climatological event frequencies for a given hazard or comparing model performance for

[3] Also known as the area under the precision-recall curve, which is often acronymized as AUPRC or AUCPR

[4] Known better by the term "average precision" where precision is synonymous with success ratio

different hazards with different climatological event frequencies. The minimum AUPDC is:

$$\text{AUPDC}_{min} = \frac{1}{pos} \sum_{i=1}^{pos} \frac{i}{i + neg}, \tag{2.3}$$

where *pos* and *neg* are the number of event and non-event examples in the verification dataset, respectively (Boyd et al. 2012). The normalized AUPDC (NAUPDC) is defined as:

$$\text{NAUPDC} = \frac{\text{AUPDC} - \text{AUPDC}_{min}}{1 - \text{AUPDC}_{min}}, \tag{2.4}$$

Regardless of climatological event frequency, the best possible classifier will have an NAUPDC of 1 and the worst possible classifier will have an NAUPDC of 0. Moreover, based on Theorem 1 (which is an original work and provided in Appendix A), we can normalize the maximum CSI by the climatological event frequency ($c$) using the following equation (hereafter referred to as NCSI):

$$NCSI = \frac{\text{CSI}_{max} - c}{1 - c} \tag{2.5}$$

### 2.3.3 The Attributes Diagram

The attribute diagram (also known as the reliability diagram) shows forecast probabilities against their conditional event frequencies (Figure 2.5). Thus, the plot for a perfectly reliable forecast system will lie along the one-to-one line. The conditional event frequency, however, can be sensitive to the bin interval, especially for smaller datasets. To address uncertainty in the conditional event frequency, one can compute the “consistency bars” from Bröcker and Smith (2007), which allows for immediate interpretation of the confidence of the reliability of a prediction system. Reliability is then assessed as the extent to which the conditional event frequencies fall within the consistency bars rather than based on their distance from the diagonal. In addition to the reliability curve, the attribute diagram also displays a histogram of forecast probabilities in each bin (to measure sharpness), a no-skill curve, and

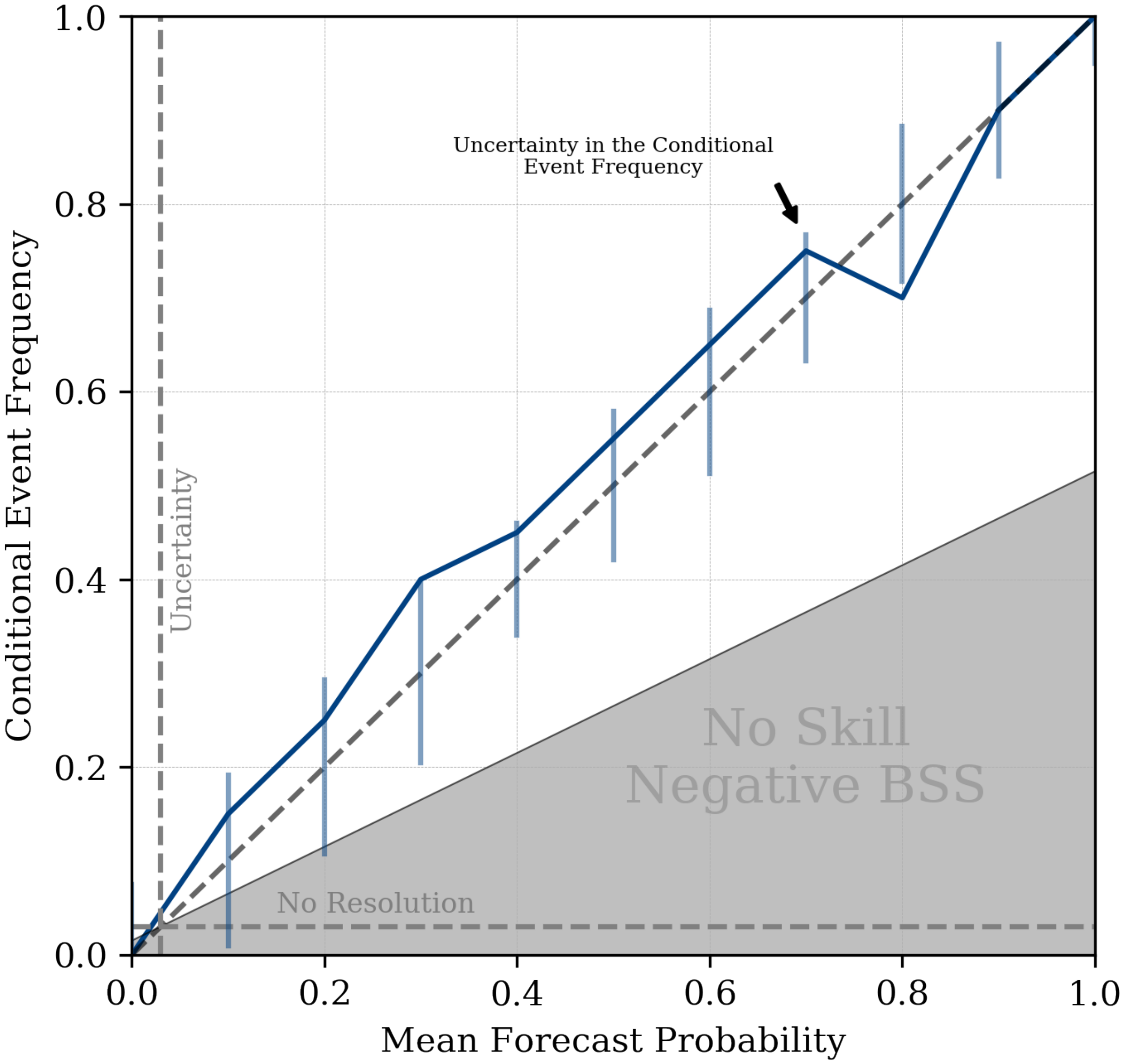


Figure 2.5: An example of an attribute diagram. Forecast probabilities are separated into equally spaced bins from which the mean forecast probability and conditional event frequency are computed, which results in a reliability curve (shown in blue). The dashed diagonal curve references a perfectly reliable system. However, given the sensitivity of conditional event frequency to the bin interval size, error bars on the conditional event frequency are shown as the vertical light blue lines. The gray shaded regions delineates positive from negative Brier skill score. The dashed gray horizontal and vertical lines are the no resolution and uncertainty of the verification dataset, respectively.

uncertainty/no resolution curves (both related to the climatological event frequency). The no-skill curve, defined as (x=0, y=0.5$\overline{y}$) to (x=1, 0.5[1+$\overline{y}$]), where $\overline{y}$ is the climatological event frequency, delineates regions of positive and negative Brier skill score (BSS), a common metric associated with the attribute diagram (Hsu and Murphy 1986). The Brier score (BS) is defined as:

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \tag{2.6}$$

where $p$ is the forecast probabilities, $o$ is the binary outcome/target variable, and $N$ is the number of examples in the verification dataset. By binning the forecast probabilities into $K$ bins (similar to the attribute diagram), the BS can be decomposed into three terms:

$$BS = \frac{1}{N}\sum_{k=1}^{K} n_k(p_k - \overline{o_k})^2 - \frac{1}{N}\sum_{k=1}^{K} n_k(\overline{o_k} - \overline{o})^2 + \overline{o}(1-\overline{o}), \tag{2.7}$$

where $n_k$ is the number of samples in the $kth$ bin. The three terms of equation 2.11 are known as the reliability (REL), resolution (RES), and uncertainty (UNC) terms. The reliability measures the weighted average difference between forecast probabilities and the conditional event frequencies and will be zero for a perfectly reliable forecast. The resolution term measures the weighted average difference between the conditional event frequencies and the climatological frequency and should be 1 for a perfectly reliable forecast. The final term is the uncertainty in the observations and does not reflect forecast quality. To convert the BS into a skill score, it has to be measured with respect to some baseline forecast, which in most cases is the climatological frequency. The BSS is defined as:

$$BSS = \frac{\mathrm{BS} - \mathrm{BS}_{ref}}{\mathrm{BS}_{ref}} = \frac{\mathrm{RES} - \mathrm{REL}}{\mathrm{UNC}}, \tag{2.8}$$

where the reference BSS ($BSS_{ref}$) is the score associated with a forecast is that is always the climatological event frequency. Like other skill scores, the BSS ranges from $(-\infty, 1]$, with higher values considered better. A positive BSS (RES > REL) means that the model is better than climatology, but it can be difficult to compare BSS from two different datasets since it is heavily impacted by the skew.

# Chapter 3: Data

## 3.1 Forecast and Verification Data

### 3.1.1 Description of the Forecast Dataset

The WoFS is a rapidly-updating ensemble data assimilation and prediction system. WoFS consists of a 36-member multiphysics ensemble (see S18, their table 1) that uses the Advanced Research Weather Research and Forecasting model (WRF-ARW; Skamarock et al. 2008) with 3-km horizontal grid spacing. WoFS is initialized with initial and lateral boundary conditions provided by the experimental 3-km High-Resolution Rapid Refresh Ensemble (HRRRE; Dowell et al. 2016) on a 750 x 750 km domain re-centered daily over the region of greatest severe weather potential. Radar, satellite (i.e., GOES-16 cloud water path), and Oklahoma Mesonet (when available) observations are assimilated every 15 min with conventional observations assimilated hourly using the ensemble adjustment Kalman filter (Anderson 2001) included in the Data Assimilation Research Testbed (DART) software. After five 15-min assimilation cycles (i.e., starting at 1900 UTC), 18-member forecasts (a subset of the 36 analysis members) are issued every 30 min and provide forecast output every 5 min for up to 6 hours of lead time.

The evaluation of the WoFS probabilistic mesocyclone guidance uses all available cases (63) generated during the 2017 and 2018 Hazardous Weather Testbed Spring Forecasting Experiments (HWT-SFE; Gallo et al. 2017) and 2018 Hydrometeorology Testbed Flash Flood and Intense Rainfall experiment (HMT-FFaIR; Barthold et al. 2015; Albright and Perfater 2018). The WoFS configuration described above was used during the 2017 and 2018 HWT-SFEs, but during the 2018 HMT-FFaIR the domain was enlarged to 900 x 900 km, the Community Gridpoint Statistical Interpolation based Ensemble Kalman Square Root Filter (GSI-EnKF; DTC 2017a,b) was used as the data assimilation scheme, and forecasts were initialized every hour between 1800-0400 UTC. The changes to the domain size and forecast length introduced during the 2018 HMT-FFaIR experiments were designed to focus on heavy rainfall forecasts at longer lead times. Overall model performance between both configurations was similar (not shown). Although forecast periods varied, to ensure that cases were weighted equally, only forecasts initialized at the top of the hour between 1900 - 0300 UTC were considered for our evaluations.

To evaluate the skill and reliability of WoFS probabilistic mesocyclone guidance, 60-min forecasts of updraft helicity (UH) in the 2—5 and 0–2km layers above ground level (AGL) are examined in this dissertation. Assessing UH in the two different layers can help determine if WoFS probabilistic mesocyclone guidance accurately distinguishes between supercells with and without low-level mesocyclones, which can be used as a proxy for tornado occurrence (e.g., 40% of low-level mesocyclones are associated with tornadoes; Trapp et al. 2005). To examine the decrease in skill of the WoFS probabilistic model guidance with forecast lead time, the following four 60-min forecast periods were used: 0-60 min, 30-90 min, 60-120 min, and 90-150 min.

The ML-calibration of WoFS's severe weather probabilistic guidance uses 81 cases (provided in Table B.1 in Appendix B) generated during the 2017-2019 HWT-SFEs. During these experiments, WoFS domains were frequently centered over the Great Plains and mid-Atlantic with less focus on the Southeast and Midwest (Figure 3.1). This is not surprising as severe weather is most common over the Great Plains during the spring (severe weather has a less pronounced springtime maximum over the mid-Atlantic) and becomes more common elsewhere during the summer or cool season (SPC 2020). Overall, the dataset sufficiently samples environments relevant for springtime severe weather forecasting, but the trained ML algorithms may not be appropriate for year-round use.

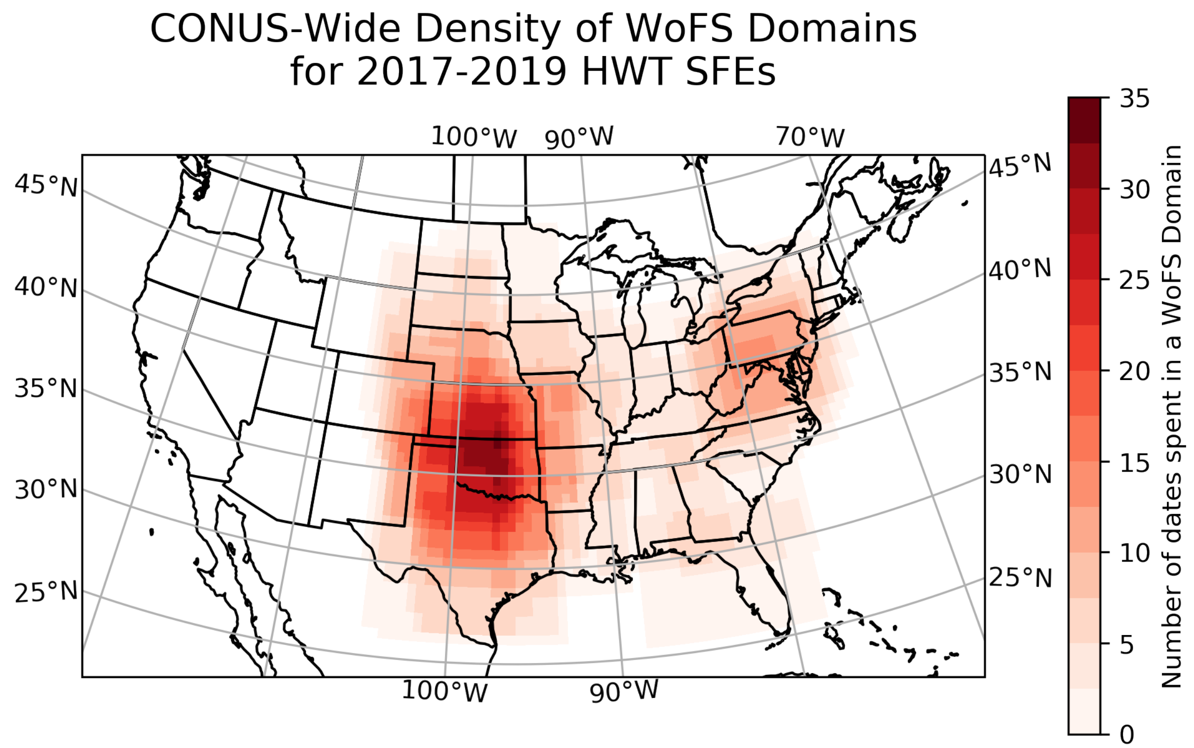

CONUS-Wide Density of WoFS Domains
for 2017-2019 HWT SFEs
100°W
90°W
70°W
45°N
40°N
35°N
30°N
25°N
Number of dates spent in a WoFS Domain
0
5
10
15
20
25
30
35

Figure 3.1: Map of the number of times a 0.5 x 0.5 degree region was in a WoFS domain during the 2017-2019 HWT-SFEs.

To be consistent with recent WoFS verification studies (e.g., Skinner et al. 2018) and typical National Weather Service (NWS) warning lead times (Brooks and Correia 2018), the WoFS forecast data were aggregated into 30-min periods up to a lead time[1] of 150 min (e.g., 0-30, 5-35, ..., 120-150 min). Given the rapid model error growth on the spatiotemporal scales represented in WoFS forecasts, the whole dataset was split in two based on the forecast lead time, whereby forecasts beginning in the first hour (i.e., 0-30, 5-35, ..., 60-90 min) are in one dataset (referred to as FIRST HOUR hereafter) and forecasts beginning in the second hour are in a second dataset (i.e., 65-95, 70-100, ..., 120-150 min; referred to as SECOND HOUR hereafter). The choice to split at the hour mark is ad-hoc, but it is informed by my previous

[1] It takes approximately 20—25 minutes to produce and disseminate the first two forecast hours of WoFS guidance to real-time users, so the effective lead time is reduced from values calculated from forecast initialization

research of storm-scale predictability (Flora et al. 2018). The different lead times within the FIRST HOUR and SECOND HOUR are uniformly distributed. Splitting the dataset in this way allows the ML models to learn from the different forecast error characteristics in the two datasets (e.g., larger ensemble spread in SECOND HOUR than in FIRST HOUR), which should improve the models' skill. The predictability of individual storm-scale features greatly diminishes beyond 150 min lead times (Flora et al. 2018), and therefore forecasts at those lead times are not considered in this dissertation.

### 3.1.2 Description of the Verification Dataset

The verification dataset for the WoFS probabilistic mesocyclone guidance is derived from radar-derived rotation tracks rather than local storm reports, similar to several recent studies (e.g., Skinner et al. 2016; Dawson et al. 2017, S18). Although radar-derived rotation tracks are imperfect, they avoid some limitations of using local storms reports, which suffer from poor estimates of intensity (Trapp et al. 2006; Verbout et al. 2006), non-meteorological bias (Brooks et al. 2003; Doswell et al. 2005) and under-sampling in rural areas (e.g., Potvin et al. 2019). Low- and mid-level (0-2 and 2-5 km AGL, respectively) radar-derived rotation tracks are generated from the maximum range-corrected NSSL Multi-Radar Multi-Sensor (MRMS) cyclonic azimuthal wind shear data (Smith and Elmore 2004; Miller et al. 2013; Smith et al. 2016; Mahalik et al. 2019) in each layer calculated every 5 min over the WoFS domain. Following quality control and interpolation onto the WoFS grid (fully described in S18), I aggregated these azimuthal wind shear data to produce 60-min rotation tracks. In S18, radar data in regions too close or too far (i.e., less than 5 km or greater than 150 km) from the nearest WSR-88D site were ignored to mitigate range-related impacts. However, in this dissertation, radar data outside the 150 km radius or inside the 5 km radius are included in both the forecast and verification dataset. Re-calculation of verification scores presented in S18 showed minimal sensitivity to including these data.

# Chapter 4: Object-based Framework for Ensemble-based Probabilistic Guidance

## 4.1 Deterministic-based and Verification Object Identification

The goal of the mesocyclone object identification is to isolate strong mid- and low-level rotation which may be associated with severe weather (e.g., winds >50 kts, hail >1.0 in, or a tornado) in both the forecast and verification dataset. In S18, single thresholds based on the 99.95th percentile value in the forecast and verification dataset were used for object identification. However, there are known limitations to the single threshold method. Object identification in a single threshold method will be sensitive to small changes in the size and intensity of objects near the threshold. Without using an excessively high threshold, the single threshold method can perform poorly at separating distinct, overlapping features. A candidate object identification method well-suited to mitigate these issues is the enhanced watershed algorithm, which identifies local maxima and then grows objects pixel-by-pixel from a quantized version of the original field until they reach a specified area or intensity criteria (Lakshmanan et al. 2009). Objects are restricted from growing into regions less than the minimum threshold (e.g., mid-level UH <40 $m^2$ $s^{-1}$) and once an object is identified, a larger region surrounding the objects is demarcated as a no-grow region for additional objects ensuring separation (i.e., the *foothills* region in Lakshmanan et al. 2009).

The enhanced watershed algorithm available in the open-source Hagelslag Python package (Gagne et al. 2016), which is a Python implementation of Lakshmanan et al. (2009) was used. The parameters for the Hagelslag enhanced watershed algorithm (Table 4.1) were tuned to improve the identification of both MCS and supercell rotation tracks, but there are sensitivities to these parameters.

Given that objects identified by the enhanced watershed algorithm are restricted from growing into regions less than the minimum threshold, a higher minimum threshold can shrink objects or potentially separate tracks where the intensity fluctuates below the mini-

Table 4.1: Parameters of the Hagelslag watershed algorithm for all identified objects. The minimum and maximum intensity thresholds (`min_thresh` and `max_thresh`, respectively) for the azimuthal wind shear reflect that of the rescaled values. A larger saliency criterion (`size_threshold_pixels`) than past studies (e.g., Sobash et al. 2016a) was required to prevent tracks from being broken into multiple objects. For more details on the parameters, the open-source Hagelslag Python package is available at https://github.com/djgagne/hagelslag.

| | Azi. Wind Shear | LL UH | ML UH | Ens. Probabilities |
|---|---|---|---|---|
| `min_thresh` | $0.003 * 10^4$ $s^{-1}$ | 10 $m^2$ $s^{-2}$ | 40 $m^2$ $s^{-2}$ | 0 |
| `max_thresh` | $0.008 * 10^4$ $s^{-1}$ | 50 $m^2$ $s^{-2}$ | 250 $m^2$ $s^{-2}$ | 75 |
| `data_increment` | 2 | 5 | 5 | 10 |
| `size_threshold_pixels` | 200 | 200 | 200 | 200 |

mum threshold (a limitation of the single threshold method as well). However, lowering the minimum threshold identifies weaker rotation tracks where the intensity inside the object is similar to the minimum threshold. To address this concern, I applied the image processing concept of hysteresis (Jain 1989; Lakshmanan et al. 2009) where objects are identified at a lower threshold, but must contain pixels above a second, higher threshold. Essentially, the lower minimum threshold is used to prevent shrinkage and/or separation of identified objects, but the additional threshold removes objects with weaker intensity. Rather than using the maximum intensity inside an object for the second threshold which can be unrepresentative and isolated to a single point, the 75th percentile value was used; a value representative of a quarter of the pixels within an object. The choice of a 75th percentile value threshold for mid- and low-level azimuthal wind shear was varied between 0.003-0.005 $s^{-1}$ with the identified objects matched against local storm reports to determine a representative value for "severe" rotation. Although increasing the intensity value improved matches against the local storm reports, there were diminishing returns in bulk verification metrics as increasing the threshold removed too many objects. I also did not strive for a perfect match owing to the under-reporting bias noted above. A 75th percentile threshold of 0.0035 $s^{-1}$ was found to best balance these identification criteria for both mid- and low-level azimuthal shear.

Object identification thresholds for mid- and low-level UH swaths were determined by trying to produce a similar number of forecast objects as observed objects. Sobash et al.

(2016b) and Sobash and Kain (2017) motivated this method as they maximized forecast fraction skill score when the number of severe surrogate probabilistic forecasts was equivalent to the number of severe reports. The thresholds for low (mid)-level UH objects found to produce a forecast object count similar to the observed object count are 20 $\mathrm{m^2s^{-2}}$ (80 $\mathrm{m^2s^{-2}}$). Although these values were not hyper-tuned, they still reflect the current WoFS dataset and may be defined sub-optimally. I found that decreasing these values and thereby increasing the number of forecast objects improved the contingency table metrics (increased CSI), but degraded reliability. Similar to Sobash et al. (2016b) and Sobash and Kain (2017), I found that matching the forecast object count to the observed object count was an good trade-off between the contingency table metrics and reliability.

Another sensitivity to the watershed method is that a larger area threshold (or saliency criterion as denoted in Lakshmanan et al. 2009) is required to prevent separation and shrinkage. However, in the current implementation of Hagelslag, the separation of local maxima is a function of the area threshold. Thus, when using a larger area threshold, it is possible that it identifies only a single rotation track amongst a cluster of two or more tracks. To allow for identification of additional nearby tracks, I introduced a new criterion that sets the minimum separation of local maxima. Through tuning, I found that 30 km was sufficient to separate near-by storms. If the threshold was much lower then too many local maxima were identified.

After identification, a series of quality control measures were applied. First, forecast and observed objects that did not meet a 90 $\mathrm{km^2}$ minimum area threshold were removed. Next, forecast and observed objects with a minimum distance less than 12 km were merged into a single object and objects with a duration less than 15 minutes were removed. Finally, the 75th percentile value threshold (i.e., the hysteresis threshold) was applied to remove weaker rotation tracks identified by the watershed method.

**Ensemble Probability-based Object Identification**

Forecast probability swaths associated with individual thunderstorms can be conceived of as individual "probabilistic" forecast objects with a prescribed single, representative probability value. The parameters for the Hagelslag enhanced watershed algorithm for identifying probability objects are provided in Table 4.1. The parameters for identifying probability objects were tuned for both MCS and supercell cases, but they cannot distinguish between closely spaced rotation objects. The poorer performance in these cases is because of the sensitivity of the enhanced watershed algorithm to the scale of the phenomena to be identified (noted in Lakshmanan et al. 2009) and absence of universal parameters that cover all relevant spatial scales.

After object identification of the probability swaths, the maximum grid point probability within an object was assigned to each grid point. Ideally, the likelihood of a mesocyclone occurring within a storm is the total number of ensemble members producing a mesocyclone divided by the ensemble size, which is typically equal to the maximum probability within the object. However, sometimes, UH forecast objects amongst the ensemble members may not overlap at a single grid point (particularly at later lead times). In these cases, the maximum number of ensemble members forecasting a mesocyclone at a point will be less than the total number of ensemble members forecasting a mesocyclone within a storm. In these instances, the maximum probability within the object will underestimate the ensemble probability of a mesocyclone occurring within a storm.

## 4.2 Object-based Verification of Probabilistic Guidance

### 4.2.1 Generating the Grid-Scale Ensemble Probability of Event Occurrence

Schwartz and Sobash (2017) discussed multiple methods for generating forecast probabilities from CAM ensembles. To generate grid-scale ensemble probabilities, $f_{ij}$ forecasts for $i =$

$1, \ldots, M$ grid points and $j = 1, \ldots, N$ ensemble members are converted to binary using an event threshold $q$ (e.g., rainfall $>$ 1 in) to produce $N$ binary probability fields (BP):

$$BP(q)_{ij} = \begin{cases} 1 & \text{if } f_{ij} \geq q; \text{and} \\ 0 & \text{if } f_{ij} < q \end{cases}, \tag{4.1}$$

where the binary probability fields are a function of the event threshold. The ensemble probability (EP) at the $i$th grid point is then calculated as an ensemble-average of the binary probability fields:

$$EP(q)_i = \frac{1}{N} \sum_{j=1}^{N} BP(q)_{ij}. \tag{4.2}$$

In this dissertation, a similar definition is adopted, but the binary probability field of event occurrence at the $i$th grid point for the $j$th member ($BP_{ij}$) is defined using the deterministic forecast objects

$$BP_{ij} = \begin{cases} 1 & \text{if } i \in S_j; \text{and} \\ 0 & \text{if } i \notin S_j. \end{cases}, \tag{4.3}$$

where $S_j$ is the set of grid points within a deterministic forecast objects for the $j$th ensemble member. Calculating the ensemble probability from the quality-controlled deterministic forecast objects, rather than using an event threshold (e.g., on the raw time-aggregated UH forecasts), helps ensure that the probability swaths are associated with coherent forecast tracks. For this dissertation, no additional alterations (e.g., upscaling, smoothing, filtering, neighborhooding) are made to the ensemble probabilities of event occurrence.

### 4.2.2 Grid-Based Verification of WoFS Mesocyclone Probabilistic Guidance

Forecast probability accuracy and reliability are traditionally evaluated in a grid-based framework where forecast probabilities and observations are verified on the native grid (e.g., 3-km grid for our dissertation) or upscaled and evaluated on a coarser grid. The reliability of the 0-60 minute low-level UH probabilistic guidance on the native 3-km grid is given in Figure 4.1d with an example forecast shown in Figure 4.1a. The grid-scale forecast probabilities exhibit the sharpness and spatial scales of individual thunderstorms, but greatly over-predict the likelihood of a mesocyclone impacting a point (similarly for mid-level UH; not shown). The large over-prediction bias of the WoFS probabilistic guidance on the native 3-km grid shows considerable under-dispersion. Quantifying and attributing the under-dispersion in the WoFS is beyond this dissertation.

Traditionally, correcting for under-dispersion requires applying neighborhood maxing (replacing the value at a gridpoint with the maximum value within a radius of that point) and spatial smoothing to the forecast probabilities, which can improve reliability. To improve the reliability of the forecast probabilities on the native 3-km grid without altering the observations requires substantial spatial smoothing ($\sigma = 300km$), which is unsurprising as a point in the WoFS domain had 0.02% chance of being within observed low-level rotation over the 63 cases. For a rare event, reliable grid-scale forecast probabilities (especially on high resolution grids) will be low, near the climatological frequency, especially as predictability decreases (Murphy 1991). This smoothing can limit the usefulness of WoFS probabilistic guidance to human forecasters for hazards associated with individual thunderstorms between the watch and warning time scales. This is because one can misinterpret the smoothed probabilities as each thunderstorm having a low likelihood of producing an event rather than an event impacting *any particular point* as having a low likelihood; Ebert et al. (2011) pointed out this ambiguity for heavy rainfall forecasting.

It is possible to keep higher probabilities (e.g., >50%) using neighborhood maxing in combination with smoothing, but again at the cost of spatial resolution, as shown in Figures 4.1b,c,e,f. In Figures 4.1b,e (Figures 4.1c,f), the neighborhood maximum ensemble probability (NMEP; Schwartz and Sobash 2017) is calculated within a 3x3 (5x5) grid point neighborhood and smoothed with a 6-km (12-km) Gaussian filter a while 3x3 (5x5) grid point

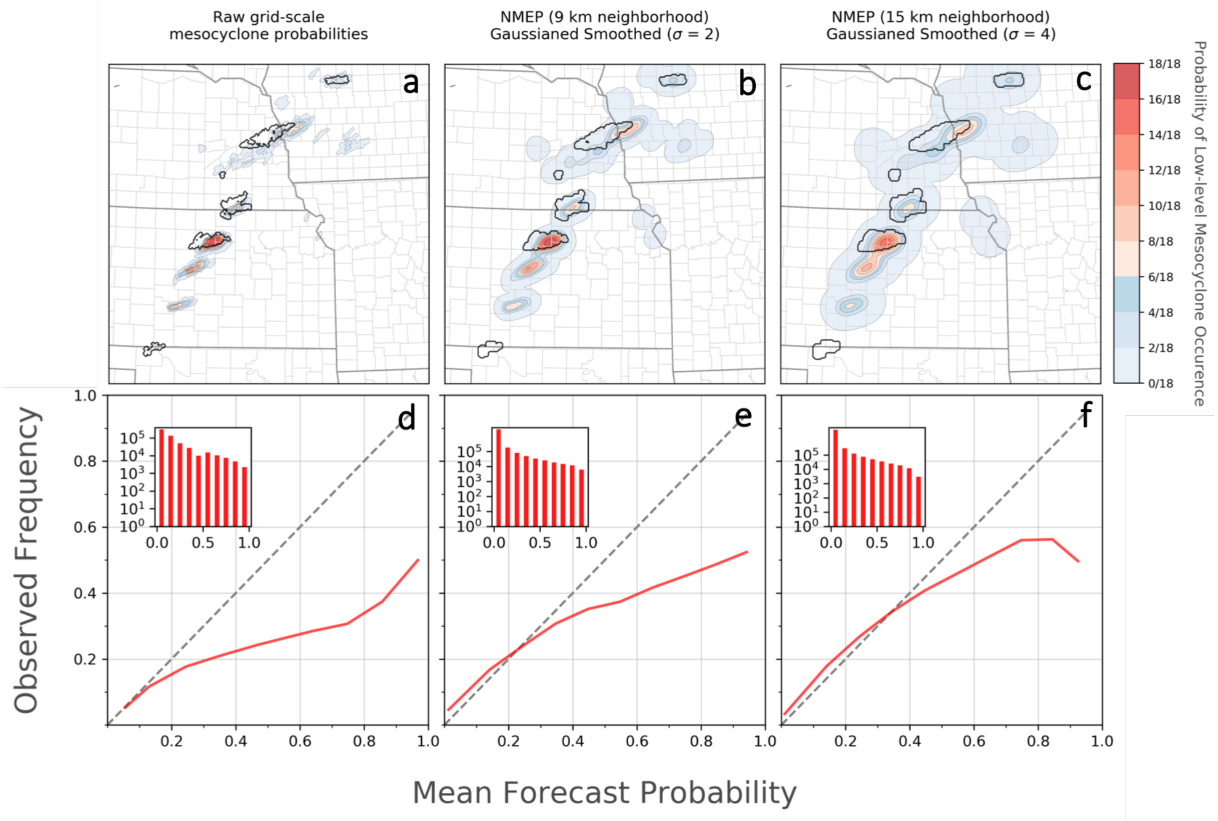
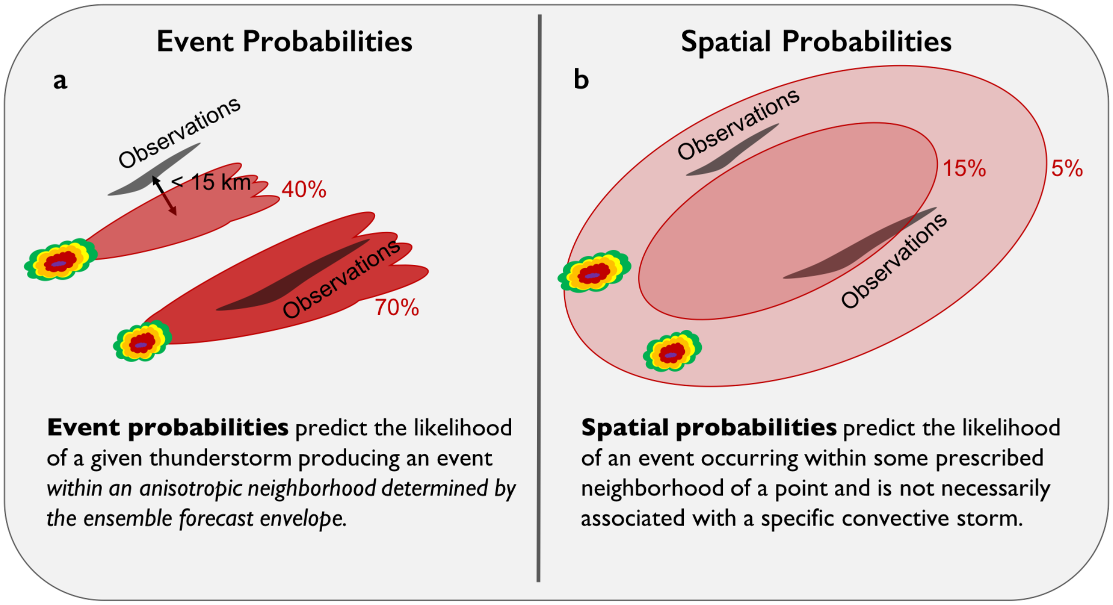


Figure 4.1: Top row: 0-60 minute probabilistic forecast of low-level mesocyclone occurrence initialized at 2300 UTC on 01 May 2018 with a) forecast probabilities and observations on the native 3-km grid and no post-processing, b) NMEP in 3x3 grid point neighborhood with Gaussian smoothing ($\sigma = 2$) and 3x3 grid point maximum value filter applied to the observations, c) NMEP in 5x5 grid point neighborhood with Gaussian smoothing ($\sigma = 4$) and 5x5 grid point maximum value filter applied to the observations. Observed hour-long low-level rotation tracks are outlined with black contours. Bottom row: reliability diagrams for the 0-60 minute WoFS low-level updraft helicity probabilities calculated for all 63 cases and evaluated in a grid-based framework. The three panels (d-f) correspond to probabilities calculated in the manner described for (a-c).

maximum filter was applied to the observations. These neighborhoods are much smaller than those used for next-day convection-allowing ensembles (e.g., 40 km smoothing and maximum value radii are typical for next-day verification). Although improved reliability and higher probabilities are present in both cases (more so in Figure 4.1f), much of the thunderstorm-scale forecast information has been filtered out. For example, the high probabilities associated with four distinct supercells in Kansas are strongly damped or aggregated into broad, coarser regions of forecast probabilities (cf. Figure 4.1a with Figure 4.1b or Figure 4.1c). Ultimately, the forecast probabilities are unreliable on the native 3-km grid owing to underdispersion and improving reliability through post-processing techniques obscures storm-scale information.

### 4.2.3 Distinction between grid- and object-based verification of probabilities

Figure 4.1a suggests WoFS, which uses rapidly cycled data assimilation to produce accurate storm-scale initial conditions, can produce highly confident short-term forecasts of a rare event. To retain unsmoothed, high forecast probabilities valid at finer spatial scales, I am distinguishing *spatial* probabilities and *event* probabilities, which is illustrated in Figure 4.2. Event probabilities predict the likelihood of a storm producing an event *within a neighborhood determined by the ensemble forecast envelope* while spatial probabilities predict the likelihood of an event occurring *within some prescribed neighborhood* of a point and are not necessarily associated with a specific convective storm. Therefore, one can measure the consistency of probabilistic forecasts in complementary event- or spatial-based frameworks (e.g., I assessed the consistency of the spatial probabilities in Section 4.2b). The event probability framework is tolerant of small spatial displacements between ensemble member forecasts of a mesocyclone, but is conditional on the predicted mesocyclones developing within the same parent thunderstorm. This changes the interpretation of the forecast probabilities from the likelihood of an event occurring within a prescribed radius of a point to the likelihood a particular storm will produce an event. The ensemble-determined footprint is flow-dependent and can grow in time as forecast uncertainty increases while using a static neighborhood in traditional methods measures forecast quality at the same spatial scales for each available lead time. Event-based verification permits the consistency of WoFS's probabilistic guidance for rare events to be assessed.



**Event probabilities** predict the likelihood of a given thunderstorm producing an event *within an anisotropic neighborhood determined by the ensemble forecast envelope.*

**Spatial probabilities** predict the likelihood of an event occurring within some prescribed neighborhood of a point and is not necessarily associated with a specific convective storm.

Figure 4.2: Illustration of distinction between spatial and event reliability of probabilistic forecasts. Event reliability (a) measures the consistency of probabilistic forecasts associated with an individual thunderstorm within an anisotropic neighborhood determined by the forecast ensemble envelope (forecast probabilities [shown in red] are the likelihood of the event occurring). Spatial reliability (b) measures the consistency of probabilistic forecasts of an event occurring within some prescribed neighborhood of a point and are not associated with a specific convective storm (forecast probabilities [shown in red] are the likelihood of the event impacting a particular point).

### 4.2.4 Verification of probability swaths in an object-based framework

I focus on two questions for evaluating WoFS probabilistic guidance:

1. Are probabilistic mesocyclone forecasts for individual thunderstorms skillful?

2. Are probabilistic mesocyclone forecasts for individual thunderstorms reliable?

To answer the first question, I apply object matching between the probability and observed rotation tracks objects. Object matching allows for calculation of verification metrics based on traditional contingency table statistics (i.e., hits, misses, and false alarms), which are intuitive and easily interpreted. Traditionally, matched forecast objects are classified as "hits," unmatched forecast objects as "false alarms," and unmatched verification objects as "misses." However, probability forecast objects generated from multiple predicted UH swaths (e.g., broad MCS probability objects) may overlap with several observed mesocyclones, especially at later lead times. In these situations, the number of "hits" in a single forecast will vary depending on whether matched forecast or observed objects are counted. Based on the contingency table, the total number of possible "hits" is the number of observed objects. Thus, when "hits" were classified as matched forecast objects, the number of hits was reduced within the contingency table, resulting in lower probabilistic forecast skill (roughly a 0.1 drop in CSI; not shown).

To remain consistent in the contingency table, if "hits" are classified as matched forecast objects, then in situations with multiple observed objects overlapping a single forecast object, we would consider all but one observed object as a "miss". As this situation arises within probability swath objects associated with MCSs or nearby cellular convection, I classify "hits" as the number of observed rotation track objects that are matched to forecast probability objects.

The verification metrics for the WoFS probabilistic guidance was limited to those that consider only hits, misses, and false alarms, which can be visualized using a performance diagram (Roebber 2009) and attribute diagram (Hsu and Murphy 1986). These metrics do not address the impact of correct negatives, which is a known limitation of the current object matching methods (Davis et al. 2009). We can label probability forecast objects as "no" forecasts through a probability threshold, but they remain a poor sample of the "true" number of correct negatives for rare-event forecasting, given that most of the forecast domain is not within any object. The necessity of ignoring correct negatives prevents the use of traditional probabilistic forecast verification metrics such as Brier skill score (BSS), the receiver operating curve (ROC) and area under the ROC (AUC).

To address the second question on assessing the reliability of the probabilistic mesocyclone forecast, we can use the event reliability definition from Figure 4.2. Similar to grid-based reliability, the probabilities associated with an object can be binned and compared against the observed frequency. For this dissertation, I define the observed frequency as the number of matched probability objects divided by the total (matched and unmatched) number of probability objects in a probability bin. Unlike the contingency table metrics, probability objects are binned on every other discrete ensemble probability ([1/9, 2/9,..., 9/9]) as large variations in number of samples exist when binning on each discrete probability.

The object matching in S18 used a simplified version of the total interest score (Davis et al. 2006a; see equation 1 in S18) that included only the minimum spatial displacement and centroid and timing displacements. I do not consider the timing displacement factor for the 60-min forecast periods used in this dissertation. A match must exceed a minimum total interest score of 0.2, which reduces the matching distance. To explore the sensitivity of forecast skill and reliability to matching distance, the maximum distance for both centroid and minimum displacement used in the total interest score is varied from 0, 9, 15, and 30 km and is hereafter referred to as the matching neighborhood.

The method for generating grid-scale probabilities and identifying probability swaths as objects is summarized in Figure 4.3. First, forecast rotation track objects are identified and quality controlled from the raw UH field for all ensemble members (Figure 4.3a; Section 3.1). The grid-scale ensemble probability of mesocyclone occurrence is then calculated from the forecast rotation track objects (Figure 4.3b; Section 4.2), and probability swath objects are identified using the enhanced watershed algorithm with the maximum probability value assigned to the swath object (Figure 4.3c; Section 4.2). A fuller discussion on the ensemble object identification method and additional procedural details are provided in the following chapter.

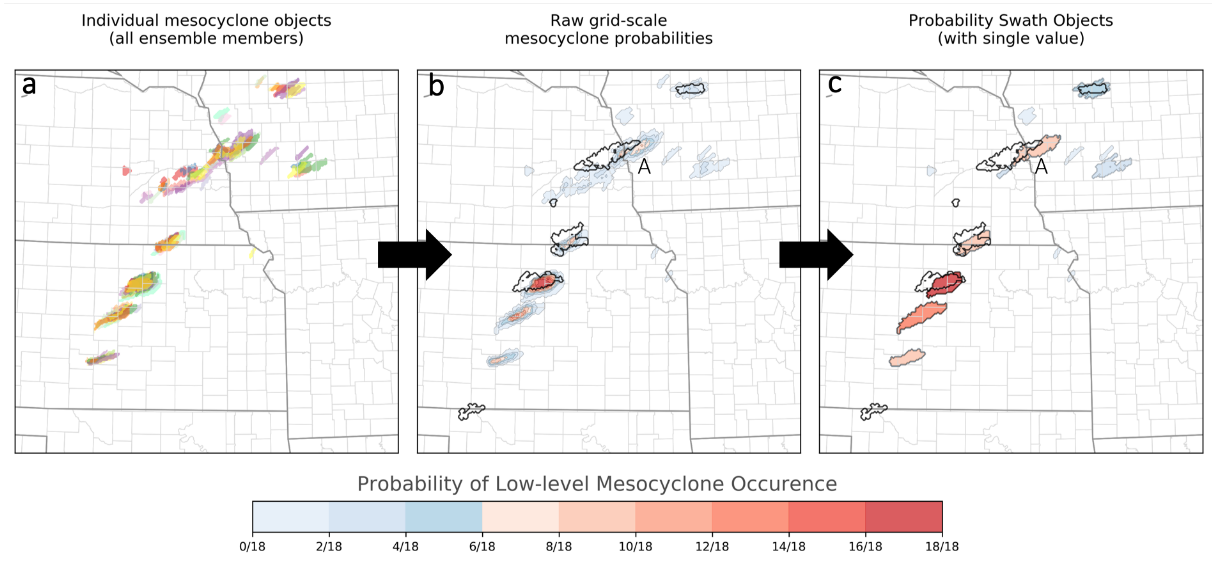

Individual mesocyclone objects
(all ensemble members)
Raw grid-scale
mesocyclone probabilities
Probability Swath Objects
(with single value)
a
b
A
c
A
Probability of Low-level Mesocyclone Occurence
0/18
2/18
4/18
6/18
8/18
10/18
12/18
14/18
16/18
18/18

Figure 4.3: Illustration of transforming individual ensemble member mesocyclone objects into probabilistic mesocyclone objects with a single, representative probability value. a) Paintball plot of forecast mesocyclone objects identified from raw updraft helicity aggregated over 60 minutes, then quality controlled as described in Section 4.1. b) Raw, grid-scale ensemble probability of low-level mesocyclone occurrence. c) Probability objects are identified using the enhanced watershed algorithm and assigned the maximum probability occurring in the object (shown as the filled color). The technique is demonstrated using a 0-60 min probabilistic forecast of low-level mesocyclone occurrence initialized at 2300 UTC on 01 May 2018. Observed hour-long low-level rotation tracks are outlined with black contours. The large probability swath near **A** denotes a potential limitation of the watershed algorithm where objects can be shrunk compared to the raw probability field.

# Chapter 5: ML-Based Calibration of WoFS Severe Weather Guidance

In this section, the generic procedure for identifying ensemble storm tracks is described, which builds upon the probability object method described in the previous section for identifying WoFS mesocyclone tracks. Rather than WoFS mesocyclone tracks, the method is applied to an ensemble of storm location based on overlapping 30-min updraft tracks. This section also discusses the data preprocessing procedures for ML and the ML models and methods used herein.

## 5.1 Data Pre-Processing Procedures

### 5.1.1 Ensemble storm track identification and labelling

In past ML studies using CAM ensemble output, object-based methods have been used to extract data from individual ensemble members rather than from the ensemble as a whole (e.g., Gagne et al. 2017, Burke et al. 2019). However, there are limitations to extracting data from the individual ensemble members. First, applying an ML model to individual member forecasts requires an additional procedure for combining the separate predictions into a single ensemble forecast. Second, learning on the individual member forecasts neglects important ensemble attributes like the ensemble mean, which, on average, is a better prediction than any single deterministic forecast, and the ensemble spread (e.g., standard deviation), which can be a useful measure of forecast uncertainty. Therefore, I extract ensemble information using the ensemble storm track method developed herein.

The steps of the ensemble storm track identification method are provided in the flow chart shown in Figure 5.1 with accompanying illustrations shown in Figure 5.2.

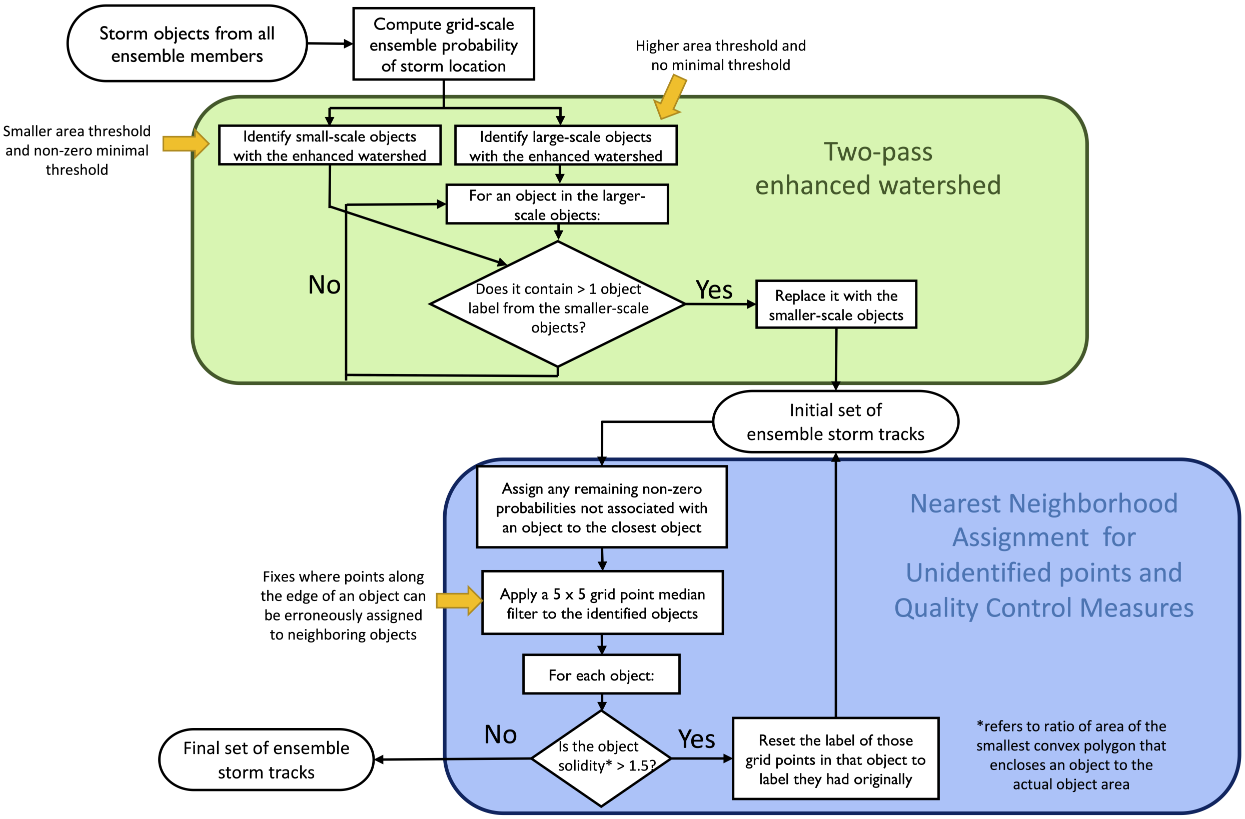

Storm objects from all ensemble members
Compute grid-scale ensemble probability of storm location
Higher area threshold and no minimal threshold
Smaller area threshold and non-zero minimal threshold
Identify small-scale objects with the enhanced watershed
Identify large-scale objects with the enhanced watershed
Two-pass enhanced watershed
For an object in the larger-scale objects:
No
Does it contain > 1 object label from the smaller-scale objects?
Yes
Replace it with the smaller-scale objects
Initial set of ensemble storm tracks
Assign any remaining non-zero probabilities not associated with an object to the closest object
Nearest Neighborhood Assignment for Unidentified points and Quality Control Measures
Fixes where points along the edge of an object can be erroneously assigned to neighboring objects
Apply a 5 x 5 grid point median filter to the identified objects
For each object:
No
Is the object solidity* > 1.5?
Yes
Reset the label of those grid points in that object to label they had originally
*refers to ratio of area of the smallest convex polygon that encloses an object to the actual object area
Final set of ensemble storm tracks

Figure 5.1: Flowchart for the ensemble storm track identification algorithm.

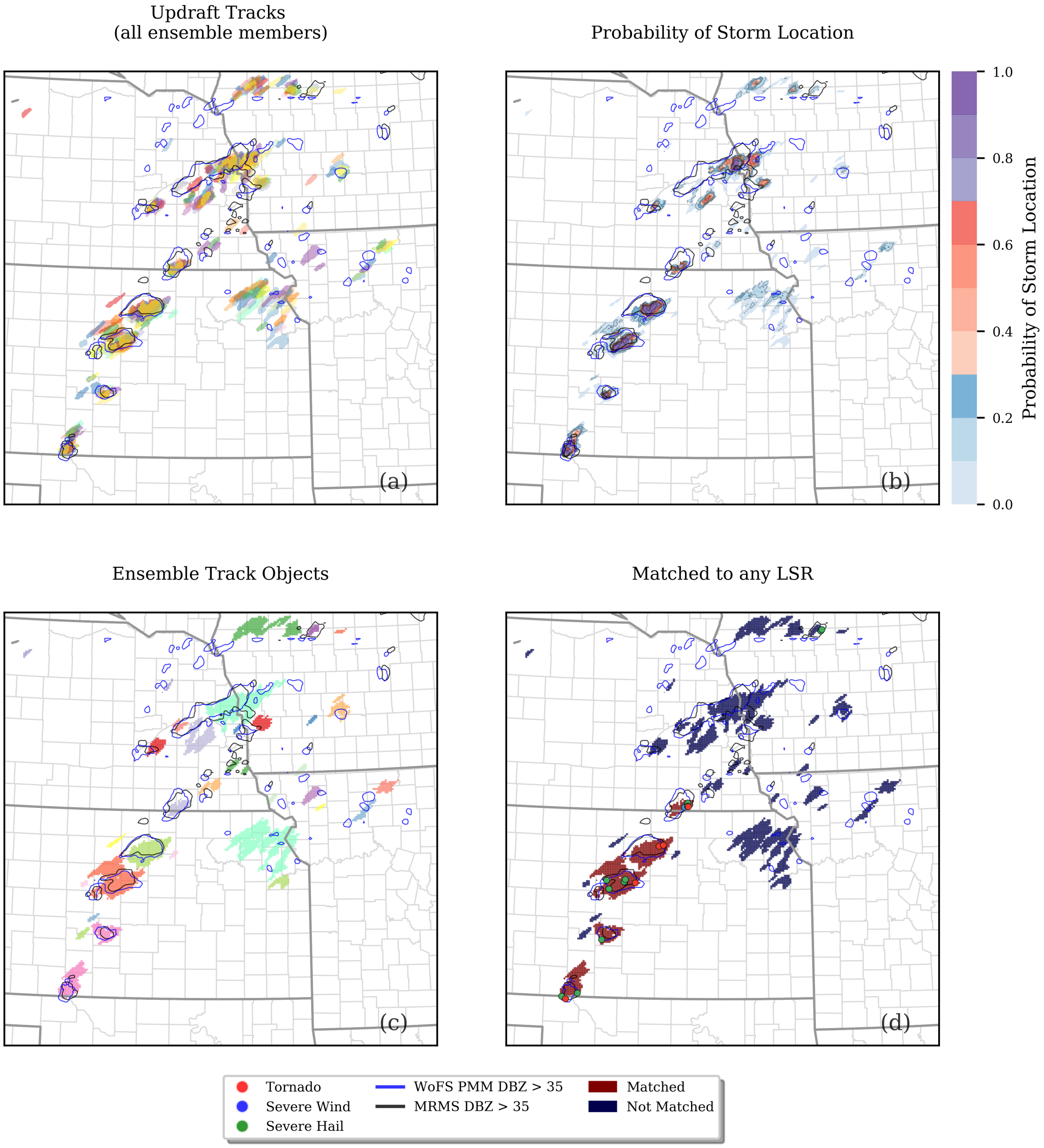


Figure 5.2: Illustration of transforming individual ensemble member updraft tracks into ensemble storm tracks. a) Paintball plot of updraft tracks identified from 30-min-maximum column-max vertical velocity, then quality controlled as described in Section 2b.1. b) Grid-scale ensemble probability of storm location is computed from the objects in (a). c) ensemble storm track objects are identified using the algorithm outlined in Section 2b.1. d) ensemble storm track objects containing a tornado (red dot), severe hail (green dot), or severe wind (blue dot) shown in red (not matched shown in blue). The technique is demonstrated using a 0-30 min forecast initialized at 2330 UTC on 01 May 2018. For context, the 35-dBZ contour of the WoFS probability matched mean (blue) and Multi-Radar Multi-System (MRMS; black) composite reflectivity at forecast initialization time, respectively, are overlaid in each panel.

First, per ensemble member, storms tracks are identified by taking peak column-maximum vertical velocity values composited over 30-min periods and thresholding them at 10 m s$^{-1}$ (Figure 5.2a). After identification, storm tracks not meeting a 108 km$^2$ (12 grid cells) minimum area threshold are removed since such storms tend to be too small and/or short-lived to be likely to produce severe weather and were found to degrade the ensemble storm track identification by producing too many objects. The ensemble probability of storm location ($EP$; Figure 5.2b) at grid point $i$ (based on $N$ ensemble members) is calculated using equation 4.2 and 4.3, but $S_j$ is defined by the updraft tracks rather than the updraft helicity tracks. The ensemble storm track objects (Figure 5.2c) are then identified from the $EP$ field with the following procedure (see Figure 5.1):

1. Identify large-scale objects by applying the enhanced watershed algorithm (Lakshmanan et al. 2009; Gagne et al. 2016) with a large area threshold (3600 km$^2$ in this study) and no minimum threshold.

2. Identify smaller-scale objects by applying the enhanced watershed algorithm with a smaller area threshold (2700 km$^2$ in this study) and some minimum threshold. I choose a threshold of 5.5% (one of 18 ensemble members) as setting the threshold higher than this causes excessive object break-up.

3. If a larger-scale object contains multiple smaller-scale objects then replace it with the smaller-scale objects.

4. Assign any remaining non-zero probabilities not associated with an object to the closest object.

5. Apply a 5 x 5 gird point median filter to each grid point with non-zero probability (assigns it the object label that occurs most frequently within a 2–grid-point radius). This is necessary to quality control the previous step where points along the edge of an object can be erroneously assigned to neighboring objects.

6. For objects with a solidity [ratio of object area to convex area (area of the smallest convex polygon that encloses the region)] greater than a given threshold (e.g, 1.5 in this study), reset the label of those grid point in that object to label they had originally. This quality control will "reset" an object if the previous steps produced an object with poor solidity.

7. Repeat steps 4-7 until no further changes occur.

This two-pass procedure coupled with the nearest neighborhood assignment allows the enhanced watershed to grow objects to a greater size while maintaining object separation.

After I identify the ensemble storm tracks, I classify each according to whether it contains a tornado, severe hail, and/or severe wind storm report (Figure 5.2d). To account for potential reporting time errors, reports were considered within ± 15 min of either side of the 30 min forecast period (a 60 min window). The choice of 15 min attempts to capture potential human reporting errors, but is only defined ad-hoc. Sometimes, an observed storm may produce severe weather, but there is no corresponding forecast storm in the WoFS guidance. This does not undermine the goal of the ML prediction system, which is to predict which WoFS storms will become severe. However, our inability to account for missed storm reports where the WoFS cannot predict the occurrence of a storm in a particular area highlights an important trade-off between the event-based prediction framework that I developed in this dissertation and the more traditional grid-based framework (which allows such misses to be included in the verification, but produces overly smooth forecasts). Last, I recognize that local storm reports are error-prone (e.g., Brooks et al. 2003; Doswell et al. 2005; Trapp et al. 2006; Verbout et al. 2006; Cintineo et al. 2012; Potvin et al. 2019), but they are the best database for individual severe weather hazards, they have been frequently used in past ML studies (e.g., Cintineo et al. 2014, 2018, Gagne et al. 2017, McGovern et al. 2017; Burke et al. 2019; Hill et al. 2020; Lagerquist et al. 2020; Sobash et al. 2020; Steinkruger et al. 2020), and are used in official evaluations of NWS warnings and SPC watches and outlooks.

### 5.1.2 Predictor Engineering

Figure 5.3 depicts the data preprocessing and predictor engineering procedure.

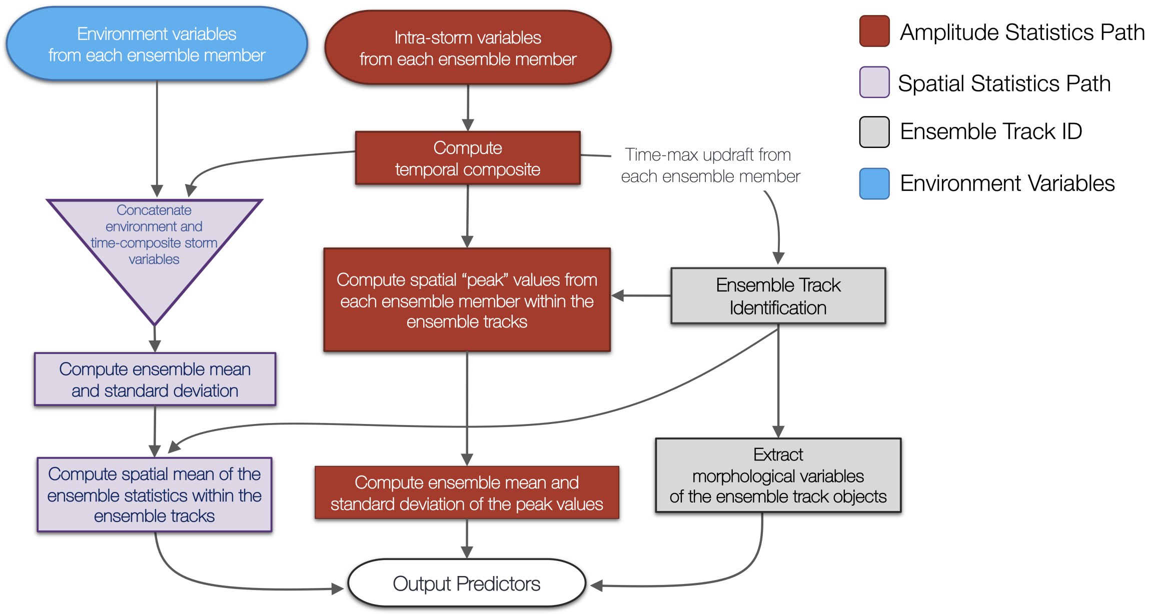

Environment variables from each ensemble member
Intra-storm variables from each ensemble member
🟥 Amplitude Statistics Path
🟪 Spatial Statistics Path
⬜ Ensemble Track ID
🟦 Environment Variables
Compute temporal composite
Time-max updraft from each ensemble member
Concatenate environment and time-composite storm variables
Compute spatial "peak" values from each ensemble member within the ensemble tracks
Ensemble Track Identification
Compute ensemble mean and standard deviation
Compute spatial mean of the ensemble statistics within the ensemble tracks
Compute ensemble mean and standard deviation of the peak values
Extract morphological variables of the ensemble track objects
Output Predictors

Figure 5.3: Flow chart of the data preprocessing and predictor engineering used in this dissertation. The three components are the ensemble storm track object identification (shown in grey), the amplitude statistics (shown in red), and the spatial statistics [shown in purple (a combination of red and blue)]. Environmental variable input is shown in blue.

First, per ensemble member, the 30-min maximum (minimum) was calculated for the positively-oriented[1] (negatively-oriented[2] ; denoted by $*$) intra-storm variables while the environment variables were taken from the beginning of a the valid forecast period to sample the pre-storm region (see Table 5.1 for the input variables). Predictors subsequently generated from these fields are of two modes: spatial statistics (shown as the purple path in

[1] Positively-oriented variables are variables where increasing magnitude is associated with larger positive values

[2] negatively-oriented variables are variables where increasing magnitude is associated with larger negative values

Table 5.1: Input variables from the WoFS. The asterisk (*) refers to negatively-oriented variables. CAPE is convective available potential energy, CIN is convective inhibition, and LCL is the lifting condensation level. Mid-level lapse rate is computed over the 500-700 hPa layer and low-level lapse rate is computed over the 0-3 km layer. HAILCAST refers to maximum hail diameter from WRF-HAILCAST (Adams-Selin and Ziegler 2016; Adams-Selin et al. 2019). The cold pool buoyancy ($B$) is defined as $B = g\frac{\overline{\theta}_{e,z=0}}{\theta'_{e,z=0}}$ where $g$ is the acceleration due to gravity, $\overline{\theta}_{e,z=0}$ is the lowest model level average equivalent potential temperature, and $\theta'_{e,z=0}$ ($= \theta_{e,z=0} - \overline{\theta}_{e,z=0}$) is the perturbation equivalent potential temperature of the lowest model level. Values in the parentheses indicate those variables are extracted from different vertical levels and/or layers.

| Intra-storm | Environment | Object Properties |
|---|---|---|
| Updraft Helicity (0-2 km, 2-5 km) | Storm-Relative Helicity (0-1 km, 0-3 km) | Area |
| Cloud Top Temperature* | 75 mb Mixed-layer CAPE | Eccentricity |
| 0-2 km Avg. Vertical Vorticity | 75 mb Mixed-layer CIN | Orientation |
| Composite Reflectivity | 75 mb Mixed-Layer LCL | Minor axis length |
| 1-3 km Maximum Reflectivity | 75 mb Mixed-Layer Equivalent Potential Temperature | Major axis length |
| 3-5 km Maximum Reflectivity | U Shear (0-6 km, 0-1 km) | Extent |
| 80-m wind speed | V Shear (0-6 km, 0-1 km) | Initialization Time |
| 10-500 m Bulk Wind Shear | 10-m U | |
| 10-m Divergence* | 10-m V | |
| Column-maximum Updraft | Mid-Level Lapse Rate | |
| Column-minimum Downdraft* | Low-level Lapse Rate | |
| Low-level updraft (1 km AGL) | Temperature (850, 700, 500 mb) | |
| HAILCAST | Dewpoint Temperature (850, 700, 500 mb) | |
| Cold Pool Buoyancy* | Geopotential Height (850, 700 500 mb) | |

Figure 5.3) or amplitude statistics (shown as the red path in Figure 5.3). For the spatial statistics, I compute the ensemble mean and standard deviation at each grid point within the ensemble storm track, then spatially average them over the storm track. I am only computing the spatial average (and not e.g., the standard deviation within the storm track) to limit the number of predictors in favor of model interpretability over model complexity.

I only compute amplitude statistics for the time-composite intra-storm variables. For the positively oriented (negatively oriented) intra-storm state variables, the spatial 90th (10th) percentile value (from grid points within an ensemble storm track) is computed from each ensemble member to produce an ensemble distribution of “peak” values. The 90th (10th) percentile is used as the “peak value” rather than maximum (minimum) since the maximum (minimum) value may be valid at only a single grid point, and therefore potentially unrepresentative. However, it is unknown whether taking the spatial maximum/minimum value may produce more skillful information to the ML models and should be explored in future studies. The ensemble mean and standard deviation are subsequently computed from each set of peak values to capture the expected amplitudes of storm features and the uncertainty therein. Reversing this procedure (i.e., computing the ensemble mean and standard deviation at each grid point and then finding the peak value) would have caused useful fine-scale details in the WoFS forecasts to be lost because of storm phase differences among ensemble members.

Lastly, I calculated a handful of properties describing the ensemble storm track object morphology. These include area, eccentricity, major and minor axis length, and orientation. Altogether, there are 30 amplitude statistics, 76 spatial statistics, and 7 object properties for a total of 113 predictors.

## 5.2 Machine Learning Methods

### 5.2.1 Machine Learning Models

#### Logistic Regression

A linear regression model is a linear combination of learned weights ($\beta_i$), predictors ($x_i$) and a single bias term ($\beta_0$) :

$$z = \beta_0 + \sum_{i=1}^{N} \beta_i x_i, \tag{5.1}$$

where $N$ is the number of predictors. For logistic regression, a logit transformation is applied to the output of the linear regression model:

$$p = \frac{1}{1 + \exp(-z)}, \tag{5.2}$$

where $p$ are the model predictions [values between (0,1)]. The weights are learned by minimizing the binary cross-entropy (also known as the log-loss; Kuhn and Johnson 2013) between the true binary labels ($y$) and model predictions with two additional terms for regularization (known together as the elastic net penalty; Kuhn and Johnson 2013):

$$C \sum_{k=0}^{K} \Big[ y_k \log_2(p_k) + (1 - y_k) \log_2(p_k) \Big] + \frac{1-\alpha}{2} \sum_{k=0}^{K} \beta_k^2 + \alpha \sum_{k=0}^{K} |\beta_k| \tag{5.3}$$

where $K$ is the number of training examples, C $\{= \frac{1}{\lambda}$ where $\lambda \in [0, \infty)\}$ is the inverse of the regularization parameter (adjusts the strength of the regularization terms relative to the log-loss), and $\alpha \in [0, 1]$ is a mixing parameter that adjusts the relative strength of the two regularization terms. The second term is known as the “ridge” penalty or $L2$ error and it penalizes the model from heavily favoring predictors by encouraging the model to keep weights small. The last term is known as the “lasso” (least absolute shrinkage and selection operator) penalty or $L1$ error and it allows weights to be zeroed out, thereby removing predictors from the model. Since logistic regression explicitly combines predictors (unlike the tree-based methods) and the scale of the predictors can vary considerably, the training and testing predictors are normalized by the training dataset mean and standard deviation for each predictor.

### Tree-based Methods

Tree-based methods are among the most common ML algorithms. A single classification tree recursively partitions a predictor space into a set of subregions using a series of decision nodes where the splitting criterion favors increasing the “purity” (consisting of only one class) of these regions (Hastie et al. 2001). To prevent overfitting (restricting the subregions from becoming too narrowly defined) decision trees can be “pruned,” for example, by requiring a maximum depth or removing final nodes (known as leaf nodes) below a minimum sample size. A classification random forest builds an ensemble of weakly correlated classification trees and merges their predictions to improve accuracy and stability over any individual decision tree (Breiman 2001a). Random forests achieve the increased performance over a single decision tree by relying on two sources of randomness, which decreases the variance of the learned model. The first source of randomness is that each tree is only trained on a bootstrap resample of the original training examples. A single decision tree tends to be sensitive to the training dataset such that a small change can result in a significantly different tree structure. Thus, training on a random subsample of the training dataset for each tree results reduces the odds of overfitting. The second source of randomness is that only a small, random subset of predictors are used per split. Instead of searching for the most important predictor while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. The random forest prediction is the ensemble average of the event frequencies (from those examples in the leaf node) predicted by each individual classification tree (all trees are weighed equally).

In contrast, an ensemble of decision trees can be combined using the statistical method known as gradient boosting where predictions are not made independently, but sequentially (Friedman 2002). The first tree is trained on the true targets and then each additional tree is trained on the error residual of the previous tree. In this dissertation, the error residual is based on the log-loss function used in equation 5.3. Conceptually, trees are added one at a time with each successive tree structure adjusted based on the results of the previous iteration. Similar to random forests, the decision trees of a gradient boosted model can also be trained on random samples of the training dataset or a random subset of predictors per node (known as stochastic gradient boosting; Kuhn and Johnson 2013). Unlike random forests, however, the maximum depth of the decision trees in a gradient-boosted model are typically between 4-8 as the goal is to produce weaker predictive models. A weak learner is one that classifies the data but with a high error rate. The final prediction of a gradient-boosted forest is the weighted sum of the predictions from the separate classification trees.

### Isotonic Regression

ML models may correctly rank predictions (predict the most probable class), yet produce highly uncalibrated probabilistic output, especially when trained on resampled data. Isotonic regression is a non-parametric method for finding a non-decreasing (monotonic) approximation of a function and is commonly used for calibrating ML predictions (Niculescu-Mizil and Caruana 2005). Past studies in weather-based studies have found success using isotonic regression-based calibrations (Lagerquist et al. 2017; McGovern et al. 2019a; Burke et al. 2019). To compute calibrated probability estimates, isotonic regression seeks the best fit of the data that are consistent with the classifier's ranking. First, pairs of $(p_i, y_i)$ are sorted based on $p_i$ where $p$ is the base classifier's uncalibrated predictions and $y$ is the true binary labels. Starting with $y_1$, the algorithm moves to the right until it encounters a ranking violation $(y_i > y_{i+1}; 0 > 1)$. Pairs $(y_i, y_{i+1})$ with ranking violations are replaced by their average and potentially averaged with previous points to maintain the monotonicity constraint. This process is repeated until all pairs are evaluated. The outcome is a model that relates a base classifier's prediction to a calibrated conditional event frequency (through the averaging of the rank violations). To prevent introducing bias, the isotonic regression is typically trained on the predictions and labels of the base model on a validation dataset. Rather than training on an independent validation dataset, I use the cross-validation approach from Platt (1999) where the base model is fit on each training fold and used to make predictions on the corresponding validation fold. The calibration model (e.g., isotonic regression) is then trained on the concatenation of the predictions from the different cross-validation folds. The base model can then be refit to the whole training dataset while the calibration model is effectively fit on the whole training dataset without biasing the predictions.

#### Models used

In this dissertation, the random forest and logistic regression models are those available in the sci-kit learn python package (Pedregosa et al. 2011). The gradient-boosted classification trees model comes from the open-source eXtreme Gradient Boosted (XGBoost) python package (Chen and Guestrin 2016). The gradient-boosted classification tree model will be referred to as the XGBoost model herein. The calibration model used is the isotonic regression model available in the sci-kit learn package (Pedregosa et al. 2011).

### 5.2.2 Developing a Baseline Prediction from the WoFS

The baseline prediction is the ensemble probability of mid-level UH exceeding a threshold, given the prior success of this diagnostic in predicting severe weather and its frequent use as a baseline in other severe-weather-based ML studies (e.g., Gagne et al. 2017; Loken et al. 2020; Sobash et al. 2020). The ensemble probabilities are computed using equation 4.1 where $f$ is updraft helicity and $q$ is the UH threshold. I then set the event probability for a storm to the maximum ensemble probability within the ensemble storm track, similar to the method used in Flora et al. (2019). To tune the threshold for each severe weather hazard, I tested the mid-level UH probabilities on the 5 validation folds (described above) and computed the cross-validation average performance for multiple metrics (Figure 5.4). Changing the UH threshold reveals there is a tradeoff between the ranking-based and calibration-based metrics. Increasing the threshold improves reliability, but decreases the ability of the probabilities to discriminate between events and non-events. The appropriate threshold was selected subjectively with the maximizing NAUPDC weighed more than the other metrics since the calibration-metrics are sensitive to climatological event frequency. For FIRST HOUR tor-

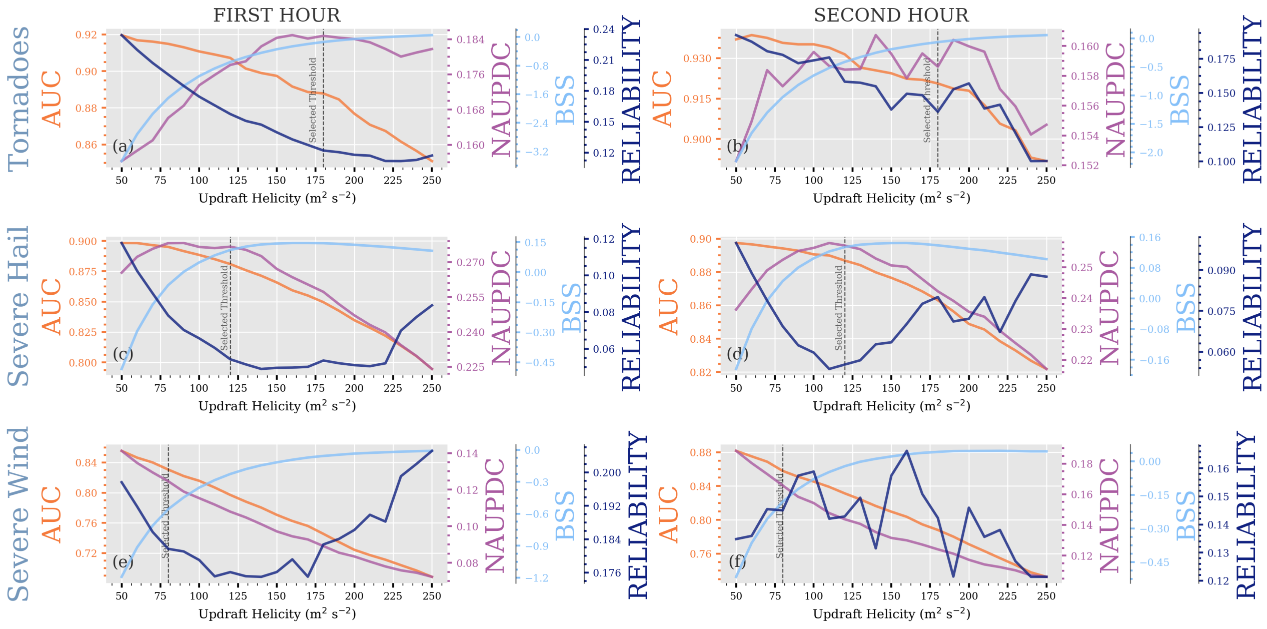


Figure 5.4: Cross-validation average (within the training dataset) performance of the baseline updraft helicity probabilities as a function of a varying threshold for predicting tornadoes (top row), severe hail (middle row), and severe wind (bottom row). Panels on the left (right) are valid for FIRST HOUR (SECOND HOUR). Metrics include AUC (red), Normalized AUPDC (NAUPDC; blue), Brier skill score (BSS; green), and the reliability component of the BSS (RELIABILITY; purple). The vertical dashed line labelled **Selected Threshold** indicates the updraft helicity threshold which optimizes certain metrics or limits tradeoffs between the various metrics (see text for details).

nado prediction, I selected a threshold of UH >180 m$^2$ s$^{-2}$ since a higher threshold degrades the ranking-based metrics although reliability continues to improve (Figure 5.4a). A similar argument can be made for the 120 m$^2$ s$^{-2}$ threshold selected for severe hail (Figure 5.4b). For severe wind (Figure 5.4e), there is no apparent optimal threshold, suggesting that UH is not the most appropriate predictor of severe wind likelihood. As a compromise, I choose a threshold of UH >80 m$^2$ s$^{-2}$ with the minimum UH threshold used to identify mid-level mesocyclones (see section 4.1). The results are similar in the SECOND HOUR dataset and therefore I kept the optimal threshold the same for simplicity (Figure 5.4b, d, f).

### 5.2.3 Model Tuning and Evaluation

To assess expected model performance, both the FIRST HOUR and SECOND HOUR datasets were split into 64 dates for training and 17 dates for testing. Rather than randomly separating the dates, I ensured that the ratio of dates with at least one event to the total number of dates was maintained for both the training and testing partitions. For example, if 40 of the 81 dates had a tornado (50%), then this ratio was approximately maintained in both the training and testing dataset. Although not perfect, this simple approach helps ensure that the testing dataset is more representative of the training dataset, which limits bias in the assessment of model performance. The number of examples in each training and testing dataset per hazard is provided in Table 5.2.

Table 5.2: Numbers of examples in the training and testing datasets for the different severe weather hazards and lead time intervals.

| | | Training | Testing |
|---|---|---|---|
| FIRST HOUR | | | |
| | Tornado | 346,341 | 82,750 |
| | Severe Hail | 349,508 | 79,583 |
| | Severe Wind | 330,840 | 98,251 |
| SECOND HOUR | | | |
| | Tornado | 262,878 | 82,483 |
| | Severe Hail | 258,270 | 87,091 |
| | Severe Wind | 258,991 | 86,370 |

Bayesian hyperparameter optimization (hyperopt; Bergstra et al. 2013) was used to identify the optimal hyperparameters for each model using 5-fold cross validation over the training dataset. The hyperopt python package is based on a random search method but implements a Bayesian approach where performance on previous iterations helps determine the optimal parameters. For this dissertation, I am using the AUPDC (defined in section 2.3.2) as our optimization metric. The default stopping criterion in hyperopt is a user-set maximum number of evaluation rounds, so I implemented an early stopping criterion where a 1%

improvement in performance must occur within a set number of rounds or else optimizing stops, which improves computational efficiency (I found that requiring said improvement at least every 10 rounds was sufficient). The hyperparameters and values used for each model are presented in Table 5.3. For those hyperparameters not listed I used the default values in version 0.22 of the scikit-learn software (Pedregosa et al. 2011) and version 0.82 of the XGBoost software (Chen and Guestrin 2016). The optimal hyperparameter values for each model and severe weather hazard for the FIRST HOUR and SECOND HOUR dataset are provided in Table 5.4 and Table 5.5, respectively.

Table 5.3: Hyperparameter values attempted for each model in the hyperparameter optimization.

| | Hyperparameter | Values |
|---|---|---|
| Random Forest | | |
| | Num. of Trees | 100, 250, 300, 500, 750, 1000, 1250, 1500 |
| | Maximum Depth | 5, 10, 15, 20, 30, 40, None |
| | Minimum Leaf Node Sample Size | 1, 5, 10, 15, 25, 50 |
| XGBoost | | |
| | Num. of Trees | 100, 250, 300, 500, 750, 1000, 1250, 1500 |
| | Minimum loss reduction ($\gamma$) | 0, 0.001, 0.01, 0.3, 0.5, 1 |
| | Maximum Depth | 2,4,7,10 |
| | Learning Rate ($\eta$) | $10^{-1}$, $10^{-2}$, $10^{-3}$, $10^{-4}$ |
| | Minimum Child Weight | 1, 5, 10, 15, 25 |
| | Ratio of predictors randomly selected per tree | 0.7, 0.8, 1.0 |
| | Subsample ratio of the examples | 0.5, 0.6, 0.7, 1.0 |
| | $L_1$ weight | 0, 0.5, 1, 10, 15 |
| | $L_2$ weight | 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1, 1.0 |
| Logistic Regression | | |
| | C | 0.0001, 0.001, 0.01, 0.1, 1.0 |
| | $\rho$ (l1_ratio) | 0.0001, 0.001, 0.01, 0.5, 1.0 |

For the final assessment, I evaluated the ML models and UH-based baselines on the independent testing datasets (severe weather hazard dependent). All metrics are bootstrap resampled (N=1000) to produce confidence intervals for significance testing. For an unbiased measure of variance, the bootstrapping method requires independent samples, but our testing samples come from overlapping forecast ranges (e.g., 0-30, 5-35, 10-40, etc) and therefore are not independent. The ensemble objects are not tracked in time and therefore I cannot

Table 5.4: Optimal hyperparameter values for each model and severe weather hazard for the FIRST HOUR dataset.

| | Hypermeter | Tornadoes | Severe hail | Severe Wind |
|---|---|---|---|---|
| Random Forest | | | | |
| | Num. of Trees | 100 | 1500 | 250 |
| | Maximum Depth | 40 | 40 | 20 |
| | Minimum Leaf Node Sample Size | 10 | 1 | 1 |
| XGBoost | | | | |
| | Num. of Trees | 300 | 250 | 300 |
| | Minimum loss reduction ($\gamma$) | 0.5 | 0 | 0 |
| | Maximum Depth | 10 | 10 | 7 |
| | Learning Rate ($\eta$) | 0.1 | 0.1 | 0.1 |
| | Minimum Child Weight | 1 | 1 | 15 |
| | Ratio of predictors randomly selected per tree | 0.7 | 0.8 | 0.8 |
| | Subsample ratio of the examples | 1.0 | 0.6 | 1.0 |
| | $L_1$ weight ($\alpha$) | 0.5 | 1 | 1 |
| | $L_2$ weight ($\lambda$) | 0.001 | 0.0005 | 0.1 |
| Logistic Regression | | | | |
| | C | 0.1 | 0.01 | 0.01 |
| | $\rho$ (l1_ratio) | 0.0001 | 0.01 | 0.001 |

Table 5.5: Same as in Table 5.4, but the SECOND HOUR dataset.

| | Hypermeter | Tornadoes | Severe hail | Severe Wind |
|---|---|---|---|---|
| Random Forest | | | | |
| | Num. of Trees | 1250 | 1250 | 250 |
| | Maximum Depth | 20 | 20 | 40 |
| | Minimum Leaf Node Sample Size | 50 | 5 | 5 |
| XGBoost | | | | |
| | Num. of Trees | 250 | 500 | 300 |
| | Minimum loss reduction ($\gamma$) | 0 | 0 | 1.0 |
| | Maximum Depth | 10 | 10 | 10 |
| | Learning Rate ($\eta$) | 0.1 | 0.1 | 0.1 |
| | Minimum Child Weight | 10 | 5 | 25 |
| | Ratio of predictors randomly selected per tree | 0.7 | 1.0 | 0.8 |
| | Subsample ratio of the examples | 0.7 | 1.0 | 0.7 |
| | $L_1$ weight | 1 | 0.5 | 10 |
| | $L_2$ weight | 0.01 | 0.1 | 1.0 |
| Logistic Regression | | | | |
| | C | 0.01 | 0.01 | 0.01 |
| | $\rho$ (l1_ratio) | 0.001 | 1.0 | 1.0 |

compute serial correlations on the full dataset, but based on a manual analysis of a small subset, I found that serial correlations for some predictors were not negligible (e.g., r=0.2), but small enough that the confidence intervals should not markedly underestimate the true uncertainty of the various verification scores.

# Chapter 6: Interpretation Methods for ML Models

The following chapter briefly describes model-agnostic ML interpretability methods for traditional ML models (for more details see [Molnar 2019a]). Molnar (2019a) identifies five scopes of ML interpretability methods, which can be summarized into three categories:

- Algorithm Transparency: *How does the algorithm create the model?*
- Global Interpretability: *How does the trained model as a whole make predictions?*
- Local Interpretability: *Why did the model make a certain prediction for an instance? Why did the model make specific predictions for a group of instances?*

Typically, when referring to model interpretability, one is referring to the last two categories whereas algorithm transparency reflects our understanding of the inner workings of a given algorithm and not a specific model or prediction. Global approaches include measuring predictor importance (e.g., Breiman 2001b; Lakshmanan et al. 2015; section 6.2) and/or ascertaining the expected functional relationship between a predictor and a ML model's prediction (e.g., Friedman 2001; Apley and Zhu 2016; Section 6.3 and Section 6.4). For the local approach, one can summarize the individual contributions of predictors for particular forecasting situations. For example, comparing predictor contributions in situations where the ML model performs well and against examples when it performs poorly (see Section 6.5).

## 6.1 Removing Redundant Information

Often at the heart of model interpretability is the tradeoff between interpretability and model complexity. Increasing model complexity can improve model performance, but often at the expense of model interpretability. One method for improving model interpretability without greatly affecting model performance is removing redundant information (often in the form of collinearities in the data). Computing multiple statistics for a four-dimensional variable, as described in Section 5.1, can provide useful information, but can also increase the amount of redundant information. While tree-based ML algorithms are fairly immune to redundant information, colinear predictors can produce instability for logistic regression (Kuhn and Johnson 2013). For example, if two predictors are highly correlated, the coefficients learned by the logistic regression can be of opposite signs, which rarely affects model performance, but can inhibit model interpretability.

There are good reasons to avoid training on data with highly correlated predictors. First, including additional predictors (especially if they may contain redundant information) may not justify the increase in model complexity. For example, McGovern et al. (2019a) found that for operational real-time settings, including additional predictors was unwarranted as it increased preprocessing time with little boost in model performance. Removing redundant predictors can also reduce over-fitting by limiting opportunities for ML models to learn noise. Lastly, predictors with redundant information can inhibit or muddle model interpretation methods (McGovern et al. 2019b; Molnar 2019a).

For this dissertation, predictors with the most correlated relationships are removed using the simple heuristic method from Kuhn and Johnson (2013). The appeal of this method is that it does not involve correlations with the target variable, so it can be performed prior to cross-validation without introducing bias or cross-contamination between training and testing sets (Hastie et al. 2001).

This removal process is as follows:

1. Calculate the linear correlation matrix of the predictors from the training dataset

2. Isolate the set of predictors with absolute correlations greater than some given threshold

3. Determine the pair of predictors with the largest absolute correlation (call them predictors A and B)

4. Compute the average correlation between A and the other variables. Do the same for predictor B

5. If A has a larger average correlation, remove it; otherwise, remove predictor B

6. Repeat steps 3-5 until all correlated pairs above the given threshold have been evaluated

Using a linear correlation threshold of 0.9, this procedure removed between 25-30 predictors out of the 113 for the different training datasets (e.g., per hazard for the FIRST HOUR and SECOND HOUR datasets). The temporal maximum and standard deviation were highly correlated for multiple intra-storm variables. This is expected, as increasing peak storm intensity is often associated with a large spread in time because of the storm evolution. The ensemble mean and standard deviation were also often highly correlated. Similar to the argument above, we often associate an increase in ensemble mean with an increase in ensemble spread. Lastly, the ensemble statistics of the "peak" values for some intra-storm variables (e.g., low-level updraft, 2-5 km UH, 10-500 m bulk wind shear) were highly correlated with the spatial average value extracted from within the ensemble track object.

In all cases, the performance of the machine learning algorithms did not substantially degrade when removing highly correlated predictors (see Appendix C). Ultimately, I favored reducing model complexity and increasing model interpretability over the marginal prediction skill increase obtained by including redundant predictors.

## 6.2 Predictor Importance

Ranking predictors based on their contribution to the model (also known as assessing their importance) is a crucial component of model interpretation. In the literature, there are multiple methods for ranking predictors:

1. Univariate relationship with the target variable

2. Expected contribution to the magnitude of a model's prediction

3. Expected contribution to the model's performance

The first method does not include the model itself and is typically based on correlations with the target variable, but can also include methods like the Kullback-Leibler J measure (Lakshmanan et al. 2015). The random forest variable importance method (Breiman 2001a), analyzing the logistic regression coefficients, SHAP dependence (Lundberg and Lee 2017), and accumulated local effect/partial dependence (defined in following sections) variance (Greenwell et al. 2018) are examples of method (2) while methods such as permutation importance and sequential backward and forward selection (McGovern et al. 2019b) are examples of method (3). In general, the first two methods (1 and 2) can be defined as measures of the predictor "relevance" while predictor "importance" is formally defined with respect to model performance (van der Laan 2006).

The different methods can produce synonymous rankings, but its not guaranteed (Marzban et al. 1999). As demonstrated in Lakshmanan et al. (2015), a predictor can have high univariate skill with respect to the target variable, but the multivariate relationship learned between other predictors in a ML model led to a bigger improvement in model performance. For example, UH often has a high univariate skill with respect to severe weather, but when provided with other predictors, the ML models in this dissertation, with few exceptions, favored other predictors over UH. Predictors with larger contributions can have an ambiguous effect on model performance. For example, in Section 7.3.2, we will see that 0-2 km UH is a top contributor to the examples matched to an tornado, but similarly contributes to false alarms and therefore would rank higher using method (2) than method (3).

The most popular method for assessing predictor importance is the permutation importance. The permutation importance method was first in introduced in Breiman (2001a), but was improved in Lakshmanan et al. (2015). Recent papers have referred to the methods in Breiman (2001a) and Lakshmanan et al. (2015) as the single-pass and multiple-pass permutation importance, respectively (e.g., McGovern et al. 2019b; Jergensen et al. 2020). Permutation importance is measured as the change in model error when values for a predictor are shuffled (permuted). If the error is relatively unchanged once a predictor is shuffled, then it is considered unimportant. The single-pass method only shuffles each predictor once and then ranks them accordingly. The multiple-pass method, however, keeps the most important predictor permuted and then re-shuffles each predictor again to determine the second most important predictor (and so on). Lakshmanan et al. (2015) developed this technique to emulate sequential backward selection (McGovern et al. 2019b) where rather than shuffling values for a given predictor, the predictor is removed from the dataset and the ML model is refit to the reduced dataset. However, retraining the model without some predictor does not demonstrate the importance of that predictor with respect to the original model, but rather highlights some characteristic of the dataset.

A limitation of the permutation importance method [holds true for all permutation-based interpretability methods (e.g., partial dependence, SHAP, etc.)] is the assumption that predictors are independent. If two predictors are strongly correlated then it can reduce their respective importance, as the ML model will treat the two predictors as being interchangeable. An advantage of the multiple-pass method is that by keeping predictors shuffled, it should break up any correlated predictors. However, if the predictors are physically correlated (e.g., updraft speed and hail size), then permuting the data can create unphysical relationships (e.g., zero updraft speed and >2 in hail), which leads to prediction instability. Therefore, it is common to compute permutation importance through several bootstrap iterations. For each iteration, the training dataset is bootstrap resampled and the loss of performance is assessed. The ranking is then assessed by the mean loss of performance from the bootstrap samples. For this dissertation, I bootstrapped the permutation importance results with N=100.

### 6.2.1 Training or Testing Dataset?

According to Molnar (2019a), favoring the training or testing dataset for predictor importance remains an open question (see their section 5.5.2). Lakshmanan et al. (2015), however, cautioned against using an independent dataset and argued for only using the training dataset. The goal of measuring predictor importance is quantifying how the model relies on each predictor and not attempting to estimate how well the model generalizes to unseen data. If the ML model learned a pattern in the training dataset that it is potentially underrepresented in the independent dataset it can bias the predictor ranking. For example, if a ML model learned that higher values of 0-3 km SRH ($> 300$ m$^2$ s$^{-2}$) significantly increased tornado likelihood, but the independent dataset had limited samples of environments with higher 0-3 km SRH, then 0-3 km SRH would appear to be an unimportant predictor. Therefore, the predictor importance in this dissertation is assessed using the training dataset. As for the remaining ML interpretability methods, they are only meant to be computed on the training dataset (except the SHAP values, which can be computed on both training and testing).

## 6.3 Partial Dependence

To complement the predictor importance, it is also crucial to understand why particular predictors are important and what their expected contribution is to the ML model prediction. A common approach for visualizing the effect of a predictor on an ML model is the partial dependence (PD) plot (Friedman 2001; McGovern et al. 2019b; Jergensen et al. 2020). The PD of predictor $x_j$ is defined as:

$$PD(x_j) = \sum_{i=1}^{N} f(\mathbf{X}^{(i)}_{\setminus x_j}, x_j = x_{j,v}) \text{ for } x_{j,v} \in x_{j,0}, x_{j,1}, ..., x_{j,V}, \tag{6.1}$$

where $f$ is the ML model, $N$ is the number of training examples, $\mathbf{X}^{(i)}_{\setminus x_j}$ is set of predictors excluding $x_j$ for the $i$th training example, and $x_{j,0}, x_{j,1}, ..., x_{j,V}$ is the set of unique values of predictor $x_j$ where the PD is evaluated. The idea is to set $x_j$ for all training examples to some value and average the resulting predictions repeating the process for multiple values of $x_j$ to produce a curve. We can then compute the "centered" partial dependence by subtracting out the average partial dependence value so the mean effect is zero:

$$PD(x_j) = \tilde{PD}(x_j) - \sum_{p=1}^{P} \tilde{PD}(x_{j,p}), \tag{6.2}$$

where $\tilde{PD}$ is the uncentered PD. The magnitude (whether positive or negative) of the centered PD shows the marginal contribution of the predictor on the predicted outcome of a ML model.

There are two main limitations of PD: it assumes that predictors are independent (a permutation-based method) and it only represents the marginal effect (the effect is averaged over the whole training dataset), which can hide heterogeneous effects and be susceptible to correlated predictors (Molnar 2019b). Because of these reasons, I do not consider PD in this dissertation, but recognize that it is a common tool for ML interpretability. Instead, I have opted for the method discussed in the following section which does not assume predictor independence and is based on conditional expectations (immune to correlated predictors).

## 6.4 Accumulated Local Effects

Though PD curves are easy to calculate and simple to understand, they assume predictors are independent (correlated features can distort the PD curve; Molnar et al. 2020) and the marginal effect can hide heterogeneous effects (Molnar 2019a). An alternative to PD is a recently developed method known as accumulated local effects (ALE; Apley and Zhu 2016). The theoretical definition of the uncentered ALE for predictor $x_j$ is :

$$ALE(x_j) = \int_{z_{0,j}}^{x_j} \mathbb{E}\left[\frac{\partial f(\mathbf{X})}{\partial X_j}\Big| X_j = z_j\right] dz_j \tag{6.3}$$

where $f$ is the ML model, $\mathbf{X}$ is the set of all predictors, and $z_j$ are values of $x_j$. For a given predictor, ALE computes the expected change in prediction over a series of conditional distributions and then accumulates (integrates) them to return the expected functional relationship of that predictor to the ML model. By computing the average change in prediction over a series of small windows, it isolates the effect of the predictor from the effects of all other predictors. Isolating the effect in this way makes ALE more immune to correlations unlike PD. Performing the calculations over a series of conditional distributions rather than the marginal distribution also avoids the pitfall of PD which can suffer from unlikely or nonphysical combination of predictor values, which introduces bias.

To estimate ALE, we bin the values of predictor $x_j$ (usually by percentile to ensure an equal number of exampes in each bin) and use the following formula:

$$ALE(x_j) = \sum_{k=1}^{K} \frac{1}{n_j(k)} \sum_{i:x_j^{(i)} \in N_j(k)} \left[ f(x_j = z_{k,j}, x_{\setminus j}^{(i)}) - f(x_j = z_{k-1,j}, x_{\setminus j}^{(i)}) \right] \tag{6.4}$$

where $K$ is the number of bins, $n_j(k)$ are the number of training examples in the $k$-th bin, $N_j(k)$ denotes the $k$-th bin interval $\{x_j \in (z_{k-1,j}, z_{k,j}]\}$. Equation 6.4 assumes a linear approximation of equation 6.3, so the bin intervals must be sufficiently small. Molnar (2019b) found that using >20-30 bins was sufficient to approximate the true ALE curve (30 bins was used in this dissertation). To clarify equation 6.4, Figure 6.1 shows a simple example of the ALE calculation from Molnar (2019a). To approximate the gradient, for those training examples in a given bin, we set the predictor $x_j$ to both the left and right side of the bin interval, compute the resulting predictions, and then take the average difference over those examples. We then take an accumulated sum over the average effect in each bin interval. Similar to PD, we subtract the average ALE so that the mean effect is zero. We can also explore feature interactions using ALE.[1] It is possible to compute the ALE for two predictors to show how they interact. The equations for 2D ALE are overwhelming and not presented here, so I refer the reader to Molnar (2019b) (see their Section 6.2). Conceptually, the 2D ALE plots estimate the additional contribution to the model due to the interaction between any two predictors. In addition to removing the average 2D ALE to adjust for the mean effect, the first-order ALE from both features are also removed to solely highlight the interaction

[1] Feature interactions can also be explored with PD, but those methods are not discussed here

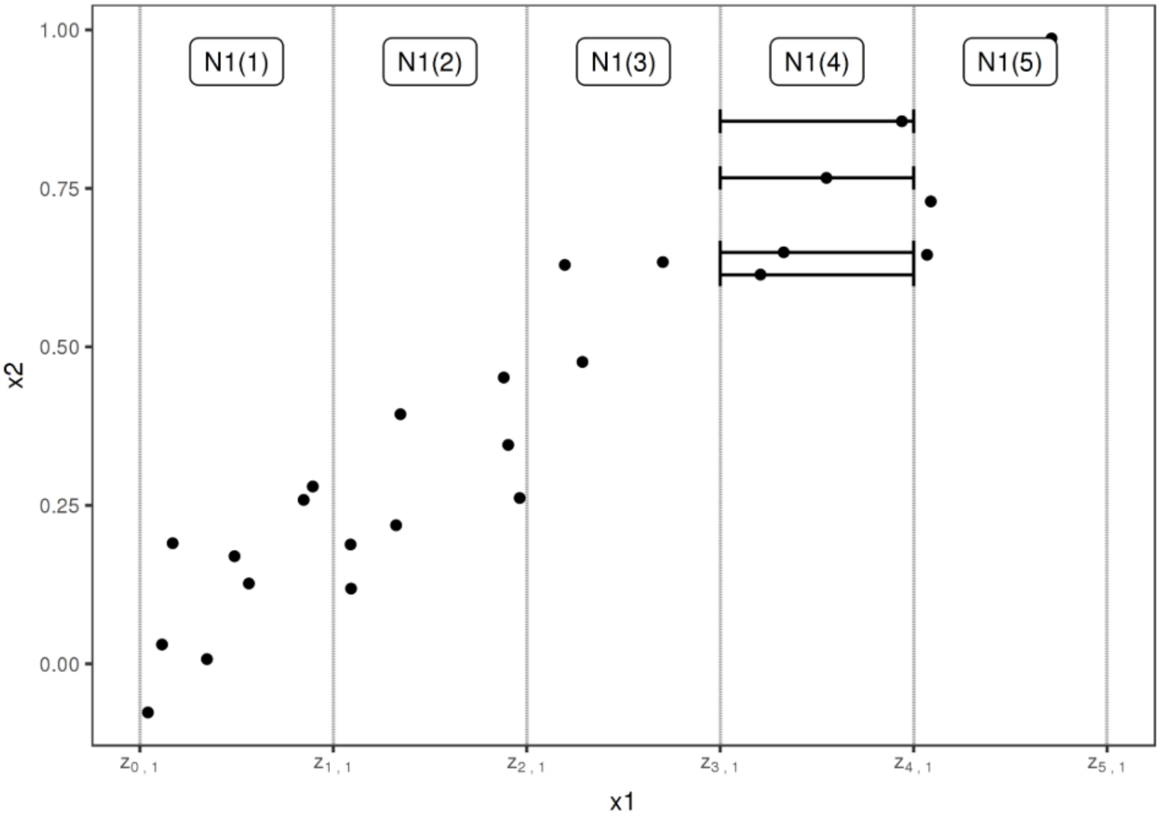


Figure 6.1: From Molnar (2019a) their Figure 5.12. Calculation of ALE for predictor $x_1$, which is correlated with $x_2$. First, we divide the predictor into intervals (vertical lines). For the data examples (points) in a interval, we calculate the difference in the prediction when we replace the predictor with the upper and lower limit of the interval (horizontal lines). These differences are later accumulated and centered, resulting in the ALE curve.

between the two predictors. For a visual interpretation, Figure 6.2 demonstrates how the 2D ALE computation is performed.

From Molnar et al. (2019), any high-dimensional prediction function (i.e., an ML model) can be decomposed as a sum of components with increasing dimensionality:

$$f(x) = \overbrace{f_0}^{\text{Intercept}} + \overbrace{\sum_{j=1}^{P} f_j(x_j)}^{\text{1st order effects}} + \overbrace{\sum_{j<k}^{P} f_{jk}(x_j, x_k)}^{\text{2nd order effects}} + ... + \overbrace{f_{1,...,P}(x_{1,...,P})}^{\text{P-th order effects}}, \tag{6.5}$$

where $P$ is the number of predictors. Using equation 6.5, we can approximate an ML model as its average model prediction plus the sum total of the first-order ALE for each predictor:

$$f(x) = \frac{1}{N}\sum_{i=1}^{N} f(x^{(i)}) + \sum_{j=1}^{P} ALE_j(x_j) + I(x), \tag{6.6}$$

where $N$ is the number of training examples and $I(x)$ is a measure of the InterAction Strength (IAS; Molnar et al. 2019) amongst the predictors (any and all second-order and higher interaction effects). We can define the IAS as an approximation error of the first-order ALE with respect to the original model predictions:

$$IAS = \frac{\sum_{i=1}^{N}\left(f[x^{(i)}] - f_{ALE1st}[x^{(i)}]\right)^2}{\sum_{i=1}^{N}\left(f[x^{(i)}] - f_0\right)^2}, \tag{6.7}$$

where $f_{ALE1st} = f_0 + ALE_1(x_1) + ... + ALE_P(x_P)$. If IAS $= 0$, then a ML model is perfectly approximated by the first-order ALE model and has no predictor interactions.

## 6.5 SHAP

PD and ALE provide the expected contributions computed over the whole training dataset, but for specific example(s), how do we explain the contributions of each predictor to the final prediction(s)? Shapley values (Shapley 1953), which have roots in game theory, have become the most promising method for explaining ML predictions of individual

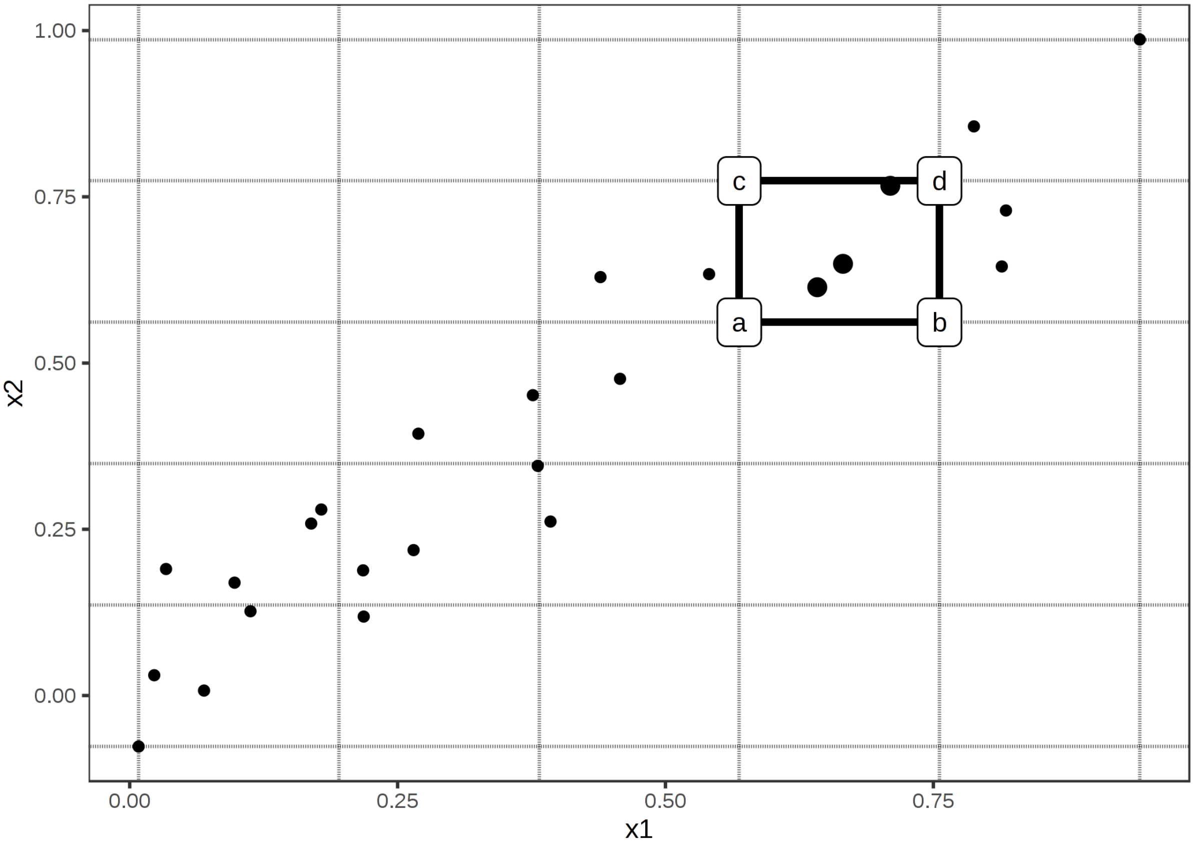


Figure 6.2: From Molnar (2019a) their Figure 5.13. Calculation of 2D-ALE. We place a grid over the two features. In each grid cell we calculate the 2nd-order differences for all [examples] within. We first replace values of $x_1$ and $x_2$ with the values from the cell corners. If $a$, $b$, $c$ and $d$ represent the corner-predictions of a manipulated [example] (as labeled in the graphic), then the 2nd-order difference is $(d - c) - (b - a)$. The mean 2nd-order difference in each cell is accumulated over the grid and centered. The first-order ALE effect is only computed and subtracted from the final computation to isolate only second-order effects.

examples. The Shapley value for predictor $x_j$ is the weighted average difference in model prediction (its contribution) when it is included and not included in some subset of predictors $S$ for all possible subsets of predictors not including predictor $x_j$ ($S \subseteq F_{\setminus\{x_j\}}$), or

$$\phi_{x_j} = \frac{1}{F!} \sum_{S \subseteq F_{\setminus\{x_j\}}} |S|!(|F| - |S| - 1)![f_{S\cup\{x_j\}}(x_{S\cup\{x_j\}}) - f_S(x_S)], \tag{6.8}$$

where $F \in \mathcal{R}^P$ is the set of all predictors, $f$ is the model, and $x_S$ is the input predictors in set $S$. The weight, $|S|!(|F|-|S|-1)!$, is based on all possible permutations of $S$ and the remaining possible ways predictors can be added to $S$ (indicating that the order in which predictors are added to the set matters). Intuitively, predictor subsets consisting of few predictors or subsets with almost all predictors will be the most informative about the effect of adding another predictor and should have a greater weight. From a game theory perspective, when players are cooperating in a coalition, Shapley values are the fairest possible payouts to the players depending on their contribution to the total payout for some game. In terms of ML, we can think of the players as the predictors and the payouts as their contributions to the final prediction (the total payout). By fairness, I am referring to the following three axioms that must be satisfied:

- Local Accuracy (additivity): The sum of the contributions (Shapley values) from each predictor plus the base rate (average predictions from the model) must equal the final prediction.

- Consistency (monotonicity): If a ML model changes such that the marginal contribution of a predictor increases or stays the same, the Shapley values must also increase or stay the same, respectively.

- Missingness: Predictors missing for some subset $S$ must have a contribution of zero to the model.

Young (1985) found that Shapley values are the only set of values that satisfy these three principles (among other properties as well).

Computing the exact Shapley values, however, is practically impossible as it requires creating the $P!$ possible subsets of predictors. Recently, Lundberg and Lee (2017) developed a computationally feasible, model-agnostic method known as KernelSHAP for approximating Shapley values. Before discussing the KernelSHAP method, however, it will be important to discuss a key term used throughout. As noted in equation 6.8, to compute Shapley values requires creating subsets of predictors, but ML models cannot have missing features. Therefore, $S$ is replaced with a binary vector of length $P$ or $S' \in \{0,1\}^P$ (the prime notation is indicating that $S'$ is a simplified version of $S$) where the 0's and 1's indicate which predictors are "present" or not in a given subset, respectively. This binary vector is referred to as the coalition vector in Molnar (2019a). Based on that definition, the KernelSHAP method is as follows:

- Produce $K$ versions of the coalition vector $S'_k \in \{0,1\}^P,\ k \in \{1, ..., K\}$. The coalitions are not randomly sampled, but rather it starts with all possible coalitions with 1 or $P-1$ predictors, then coalitions of 2 or $P-2$ predictors (and so on). For $K \geq 2^P$, the computation is exact.

- Compute the ML model prediction for the $K$ sampled coalitions, which requires replacing the "missing" predictors with values from a user-provided background dataset (typically a K-means representation of the training dataset) and taking the average model prediction of those examples.

- Compute the weight of a sampled coalition with the following Shapley kernel:

$$W = \frac{P-1}{\binom{P}{|S'|}|S'|(P-|S'|)}, \tag{6.9}$$

  where $|S'|$ is the number of present predictors.

- With the $K$ sampled coalitions weighted by their respective Shapley kernel and the averaged ML predictions as target values, fit a weighted linear regression model $g$

- The coefficients of the resulting linear model $g$ are the Shapley values $\phi$ for each predictor

This algorithm is based on a pre-existing method known as the local interpretable model-agnostic explanations (LIME; Ribeiro et al. 2016). KernelSHAP is an approximate method since the choice of $K$ is much less than that of the $2^P$ possible predictor coalitions and variance is introduced when accounting for the missing predictors. As with all permutation methods, replacing and/or permuting values can produce unphysical relationships leading to prediction instability. For this dissertation, a $K$-means representation ($K = 300$) of the training dataset is used for the background dataset, but the most appropriate choice of background dataset remains an active area of research. The default choice of $K = 2*P+2048$ in the shap python library (Lundberg and Lee 2017) was used.

Though KernelSHAP approximates Shapley values for any model, Lundberg et al. (2018) developed a fast, exact method for tree-based methods (known as TreeSHAP). Instead of simulating missing predictors by random sampling from a background dataset, the TreeSHAP method makes use of the decision tree structure by simply ignoring decision paths that rely on the missing predictors. A fuller description of the method is provided in Lundberg et al. (2018) and Molnar (2019a). The TreeSHAP method is used for the random forests and gradient-boosted trees trained in this dissertation.

# Chapter 7: Results

## 7.1 WoFS Low- and Mid-level Rotation Probabilistic Guidance

### 7.1.1 Performance Diagrams

The performance of the probabilistic low- and mid-level UH forecasts for different matching neighborhoods and forecast lead times are shown in Figure 7.1 and Figure 7.2, respectively. The location of perfect performance, indicated by a CSI of 1, is in the upper right corner, but for a probabilistic forecast with non-zero spread a perfect CSI is not possible (Hitchens et al. 2013). Additionally, the maximum CSI should correspond with POD comparable to SR (i.e., bias $\approx 1$) to discourage forecast "hedging" (e.g., overforecasting to correctly predict observations).

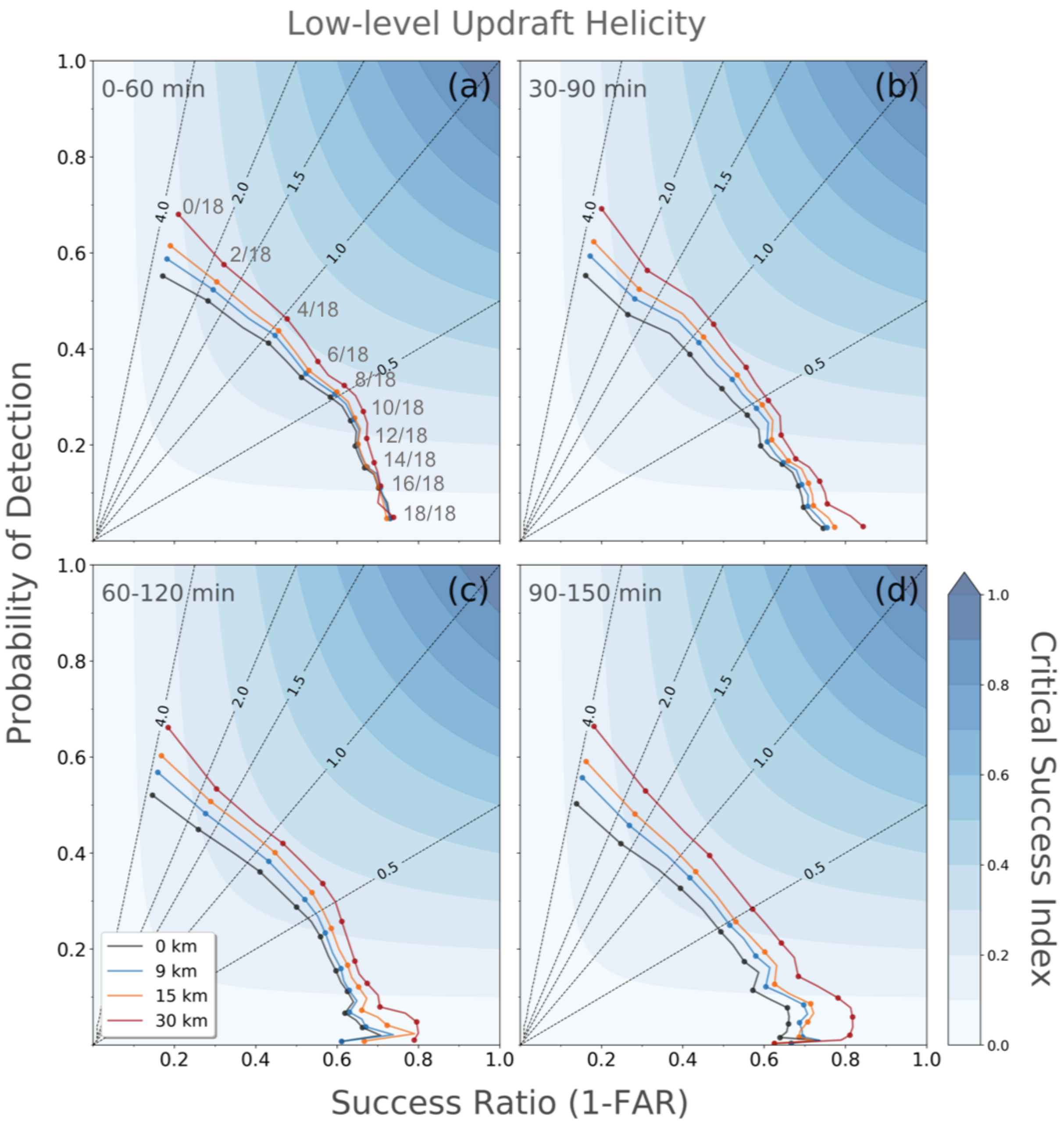


Figure 7.1: Performance diagrams for WoFS low-level (0 - 2 km AGL) mesocyclone probability swath objects using 0, 9, 15, and 30 km matching neighborhoods (gray, blue, orange, and red, respectively) and valid at a) 0-60 min, b) 30-90 min, c) 60-120 min, d) 90-150 min. The dots represent the different probability thresholds (plotted every 11.1% [2/18]).

The maximum CSI for low-level UH probability swaths tends to correspond with a probability threshold of 22.2% (4/18), independent of the lead time or matching neighborhood.

The maximum CSI value ranges from 0.26 - 0.31 (based on the matching neighborhood) in the 0-60 min period (Figure 7.1a) and drops to 0.21 - 0.27 in the 90-150 min period (Figure 7.1d). Focusing on the probability threshold = 22.2% (4/18), the POD and SR for low-level UH in the 0-60 min period at the 30 km matching neighborhood is 0.46 and 0.47 leading to a bias close to 1 ( 0.97; Figure 7.1a). These POD and SR values correspond to correct predictions of ≈50% of the observed low-level rotation tracks (with a similar success rate) out to 60 minutes of lead time. Even with a 0 km matching neighborhood (indicating overlapping forecast and observed objects), the WoFS low-level probabilistic guidance correctly predicted 40% of observed low-level rotation tracks. Looking at the different lead times for the 22.2% (4/18) probability threshold, the POD drops to 0.39 (30 km matching neighborhood) for the 90-150 min lead time (Figure 7.1d). However, the SR remains relatively unchanged as the lead time increases. One explanation for the consistent SR values with increasing lead time may be that convection initiation at later lead times is poorly forecasted, resulting in an increasing number of misses without a corresponding increase in false alarms. The trend in POD with lead time results in a steady drop in bias to 0.85 (30 km matching neighborhood) at the 90-150 min lead time (Figure 7.1d).

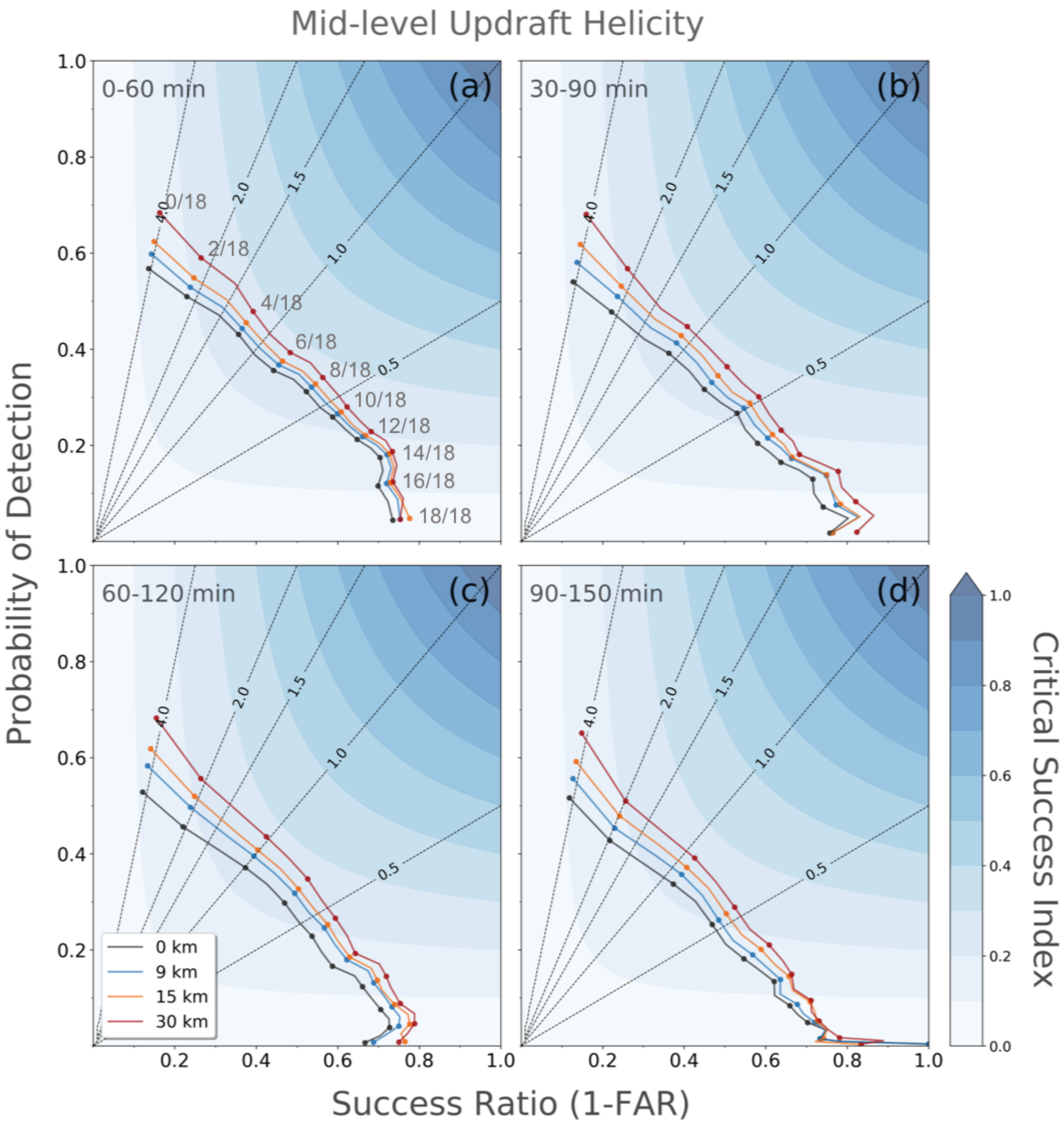


Figure 7.2: As in Figure 7.1, but for mid-level (2-5 km AGL) updraft helicity probability swath objects.

In general, as the probability threshold increases beyond 11.1% (2/18), there is a shift towards bias below 1, which is largely attributable to storm-scale predictability limits. Storm decay at later lead times in some ensemble members coupled with greater ensemble spread in mesocyclone location (increasing the likelihood of non-overlapping UH tracks in members) will cause forecast probabilities associated with an individual thunderstorm to decay with lead time (Cintineo and Stensrud 2013; Flora et al. 2018). Therefore, the maximum probability for all probability forecast objects will decrease with increasing lead time. Thus, the number of probability forecast objects at lower (higher) probability thresholds will grow (drop) with increasing lead time, effectively lowering the bias at higher probability thresholds. This increasing number of probability objects at lower probability thresholds also explains why the contingency table metrics for probability thresholds $\leq$ 11.1% (2/18) appear insensitive to forecast lead time.

Overall, the contingency table metrics and trends with increasing lead time for probabilistic forecasts of low-level UH are similar to those for mid-level UH (Figure 7.2). The probability threshold corresponding with the maximum CSI in the mid-level UH objects varies between 22.2% (4/18) and 33.3% (6/18), dependent on forecast lead time. Using the probability threshold = 33.3% (6/18), the SR is greater than the POD in the 0-60 min period, unlike the low-level UH. At later lead times, however, the maximum CSI of the mid-level UH forecasts generally have a bias of 1 (Figure 7.2c,d). The CSI for the mid-level UH forecasts tend be slightly less than corresponding thresholds in the low-level UH forecasts (cf. Figure 7.2 and Figure 7.1). This is in contrast to the results of S18 that found mid-level UH forecasts had slightly higher CSI than low-level UH in the deterministic verification. A possible explanation is that the current dissertation includes more summer-time events, where the WoFS may be overpredicting mid-level rotation. There was also a similar drop in POD in the mid-level UH forecasts as compared to the low-level UH, but nearly constant SR at the later lead times leading to the bias dropping below 1. Ultimately, the differences between UH in the two layers are very small and may not be substantial.

Lastly, some additional characteristics of low- and mid-level UH probability swath object accuracy in the performance diagrams are noted. First, separation between the performance curves at different matching neighborhoods decreases as the probability threshold increases.

This is unsurprising as increasing the probability threshold progressively reduces the number of "yes" forecasts, resulting in lower number of possible hits and a low POD regardless of the matching neighborhood. Second, separation between the performance curves at different matching neighborhoods does not change markedly with forecast lead time. As will be shown in Section 7.1.3, the centroid displacement between forecast and observed objects grows markedly with lead time. Therefore, the lack of lead time sensitivity to neighborhood in the contingency table scores is likely attributable to the minimum spatial displacement in the total interest score used for object matching ( i.e., objects may overlap but have a larger centroid displacement at longer lead times).

### 7.1.2 Attribute Diagrams

In this dissertation, the probability thresholds used for identifying probability swath objects and calculating contingency table metrics are the discrete ensemble probabilities ([1/18, 2/18,...,18/18]). Figures 7.3 and 7.4 show the reliability of the low- and mid-level UH probabilistic forecasts for the different matching neighborhoods and forecast lead times, respectively. Traditionally, for optimal reliability, the curves should lie along the diagonal from left to right with curves falling to bottom right (upper left) having an over- (under-) forecasting bias. Using the method of Bröcker and Smith (2007), we can compute consistency bars for the observed frequencies in each probability bin. Thus, we can assess how "reliable" the reliability estimates are. Additionally, the inset histograms are the number of probability objects in each probability bin (in increments of 11.1% [1/9]) for the 0 km matching neighborhood.

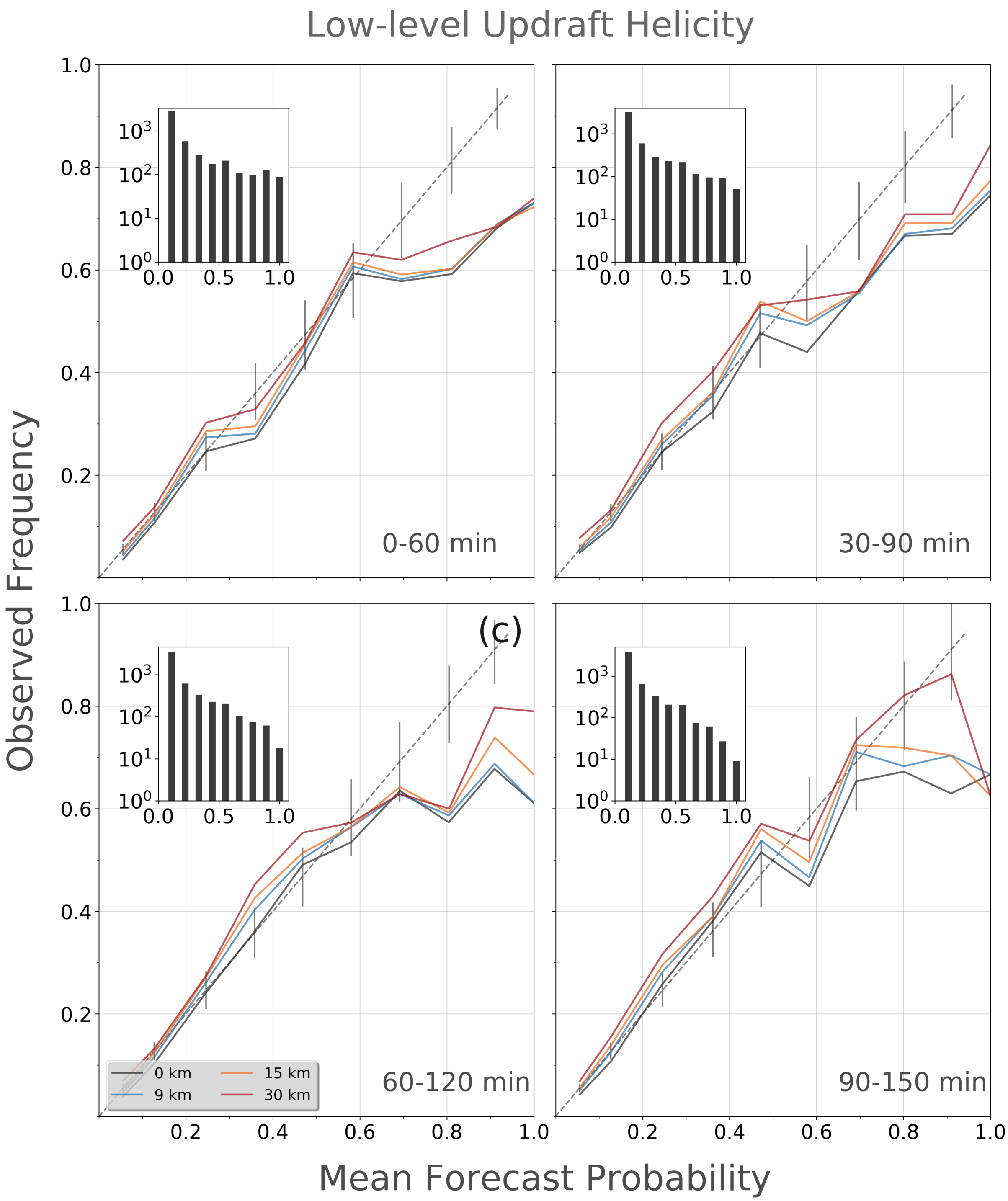


Figure 7.3: Reliability diagrams for WoFS low-level mesocyclone probability swath objects using 0, 9, 15, and 30 km matching neighborhoods (gray, blue, orange, and red, respectively) and valid at a) 0-60 min, b) 30-90 min, c) 60-120 min, d) 90-150 min. The bin increment of forecast probabilities is 11.1% (1/9). The inset (gray bar graph) is the forecast histogram for the 0 km matching neighborhood. The dashed line represents perfect reliability. The vertical line along the diagonal was the error bars for the observed frequency in each bin based on the method in Bröcker and Smith (2007).

Low-level UH forecast probability objects <60% (Figure 7.3a) have a near perfect reliability in the 0-60 min period with increasing reliability at greater matching neighborhoods, but an overprediction of mesocyclone likelihood is present for probability values greater than 60%. Overprediction of forecast probabilities greater than 60% in the 0-60 min time period are attributable to underdispersion in WoFS forecasts (Figure 4.1). In the inset histograms for both mid- and low-level UH, the forecast sharpness decays with increasing lead times as the number of probability objects at probabilities greater than 77.7% (7/9) greatly drops off. As explained above, the decay in probabilities with increasing lead time is attributable to the storm-scale predictability.

Sensitivity of the reliability for mid- and low-level UH probabilistic forecasts was generally lead-time and bin dependent. Increasing the matching neighborhood does increase the number of observed objects in a given bin, but does not necessarily improve the reliability. The greatest sensitivity to the matching neighborhood was evident for probabilities greater than >60%, especially as lead time increases. However, the probability swath values for low-level UH matched to observations using a 30 km matching neighborhood in the 60-120 and 90-150 minute periods generally deviates from the observed frequency by less than 10% (Figure 7.3c,d).

Mid-level UH forecast probabilities <30% are also reliable in the 0-60 min period, but the forecast probabilities >40% have a larger overprediction bias than low-level UH (Figure 7.4a). For example, in the 0-60 min period, probability swath objects near the 60% bin for mid-level UH only overlap with observed rotation 40% of the time. Since a similar bias does not exist for low-level UH, it is unclear why the mid-level UH has an overprediction bias. However, at later lead times, mid-level UH forecast probabilities >70% are generally more reliable than the low-level UH forecast probabilities (cf. Figure 7.3c,d and Figure 7.4c,d).

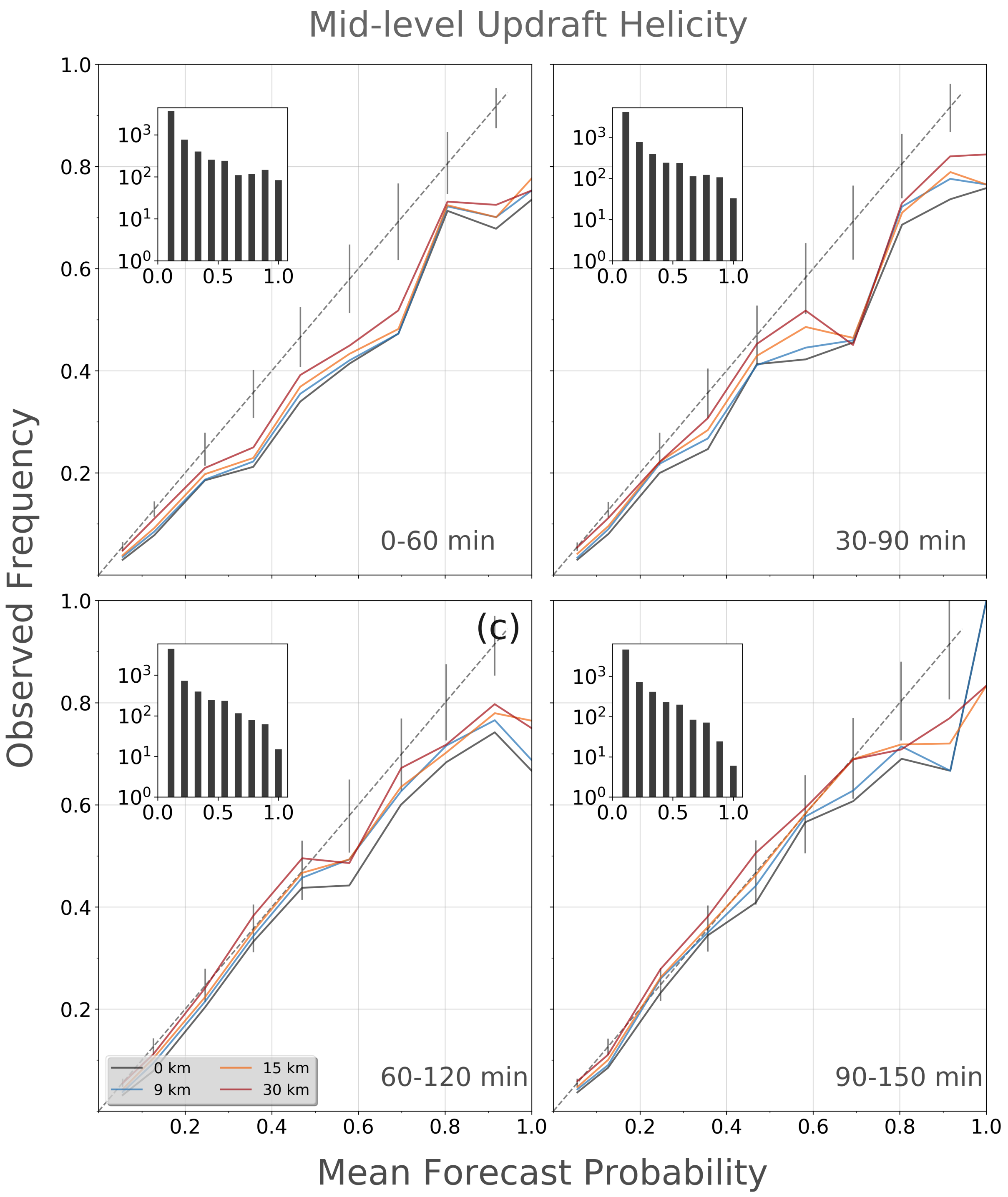


Figure 7.4: As in Figure 7.3, but for mid-level updraft helicity probability swath objects.

### 7.1.3 Centroid Displacement

Finally, centroid displacement between matched objects is examined to identify potential storm motion biases, which have been noted in subjective evaluations of WoFS probabilistic guidance (Yussouf et al. 2013b; Wheatley et al. 2015; Yussouf et al. 2015) as well as in objectively-evaluated deterministic products (Skinner et al. 2016). Figures 7.5 and 7.6 show the centroid displacement between the matched observed and forecast objects with kernel density estimate (KDE) contours overlaid for low- and mid-level UH, respectively. The KDE technique implemented here applies a Gaussian kernel with a smoothing bandwidth determined from a general optimization algorithm to each point within the parameter space (Scott 1992). Kernels for each point are summed to provide a measure of the density of points and quantify biases in the displacement between the forecast and observed objects. As discussed in section 4.1, since the enhanced watershed algorithm uses minimum area as a stopping criterion, probability swath objects in some cases will be shrunk, potentially changing their centroid and boundary displacement from observed objects. However, the impact of the enhanced watershed algorithm is primarily on the highest KDE contour when compared with probability objects identified using a single threshold method (not shown). The highest concentration of centroid displacements for both mid- and low-level UH (Figure 7.5 and 7.6) are within 30 km, consistent with S18. Deviations larger than the matching neighborhoods tested in this dissertation are a by-product of forecast probability objects in MCSs being much larger than observed rotation tracks. Often, the large probability objects associated with MCSs can have overlapping observed objects, but the centroids are displaced up to 60-90 km.

Centroid displacement for both low- and mid-level UH, based on the 99.9th percentile contour (innermost), has an inconsistent bias with forecast lead time with a slight eastward displacement (≈5 km) in the 0-60 min forecast period (Figure 7.5a and Figure 7.6a, respectively) shifting to minimal bias in the 60-120 min forecast period (Figure 7.5d and

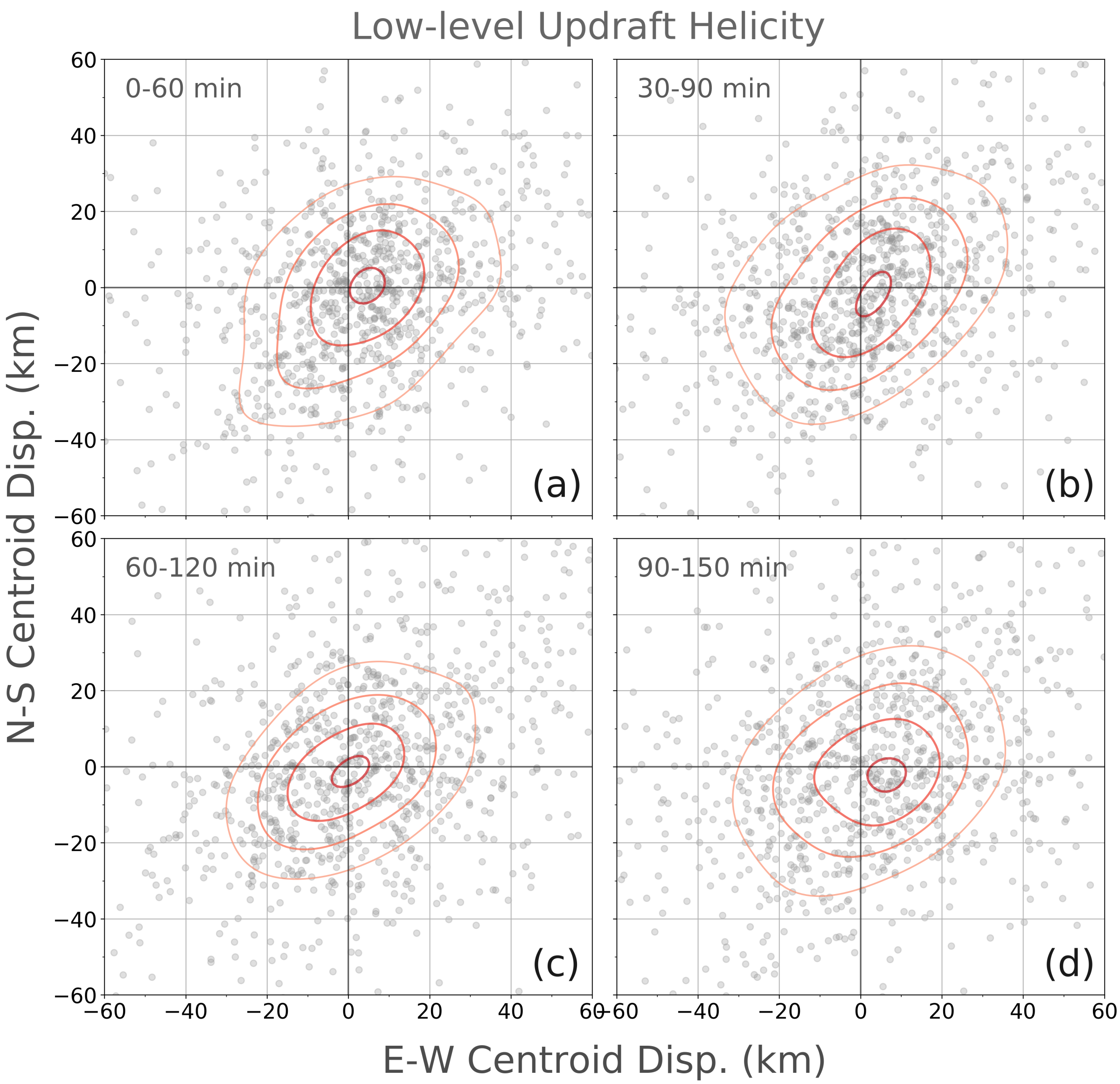


Figure 7.5: Scatterplots of the east-west and north-south centroid displacements (km) of matched objects for hour-long low-level updraft helicity probability objects valid at a) 0-60 min, b) 30-90 min, c) 60-120 min, d) 90-150 min. KDE contours of the 95, 97.5, 99, and 99.9 percentile values of each distribution are overlain to illustrate the evolution of centroid displacement with lead time.

Figure 7.6d, respectively). In the 90-150 min forecast period, there remains minimal bias in the mid-level UH forecast (Figure 7.6d), but the eastward bias returns for the low-level UH forecasts (Figure 7.5d). I suspect the bias is an artifact of different track lengths between the UH and azimuthal wind shear tracks and in addition to the object identification and matching methods. Differences between UH and azimuthal shear track lengths can be related to variation in storm motion, but also to variation in storm intensity or duration, which would also result in centroid displacement between matched object pairs. Thus, attributing centroid displacement biases solely to differences in storm motion is difficult since biases in predicted intensity or longevity could produce similar centroid displacements. At all forecast lead times, the 95 and 97.5th percentile contours (two outermost) are similar between the low- and mid-level UH and roughly centered on the origin (Figure 7.5 and Figure 7.6). The area of the 95th percentile contours are similar for low- and mid-level UH except in the 90-150 min forecast period where low-level UH is bit broader compared to the mid-level UH indicating a larger variance in centroid displacement between matched objects (cf. Figure 7.5d and Figure 7.6d). In general, the outermost KDE contour (95th percentile) expands with increasing lead time, especially for low-level UH. As noted in Section 7.1.1, the centroid displacement between forecast and observed objects grows markedly, but there was a lack of lead time sensitivity to matching neighborhood in the contingency table scores. Therefore, the minimum displacement between the forecast objects and observed azimuthal shear tracks is likely dampening the effects of the larger centroid displacements for the contingency table metrics (i.e., forecast and observed objects overlap, but have larger centroid displacement). Ultimately, the orientation of the contours are along the expected climatological storm track and there are two possible explanations:

- Given the dampening effect of the minimum displacement, the centroid displacements could represent differences in track length (and relative centroid position).

- The centroid displacements can represent a biased forecast storm motion.

Additionally, artifacts in MRMS rotation tracks are more common in the 0-2 km layer than 2-5 km (owing to more ground clutter), so the bias may be influenced by limitations of the verification dataset as well as differences in the forecasts.

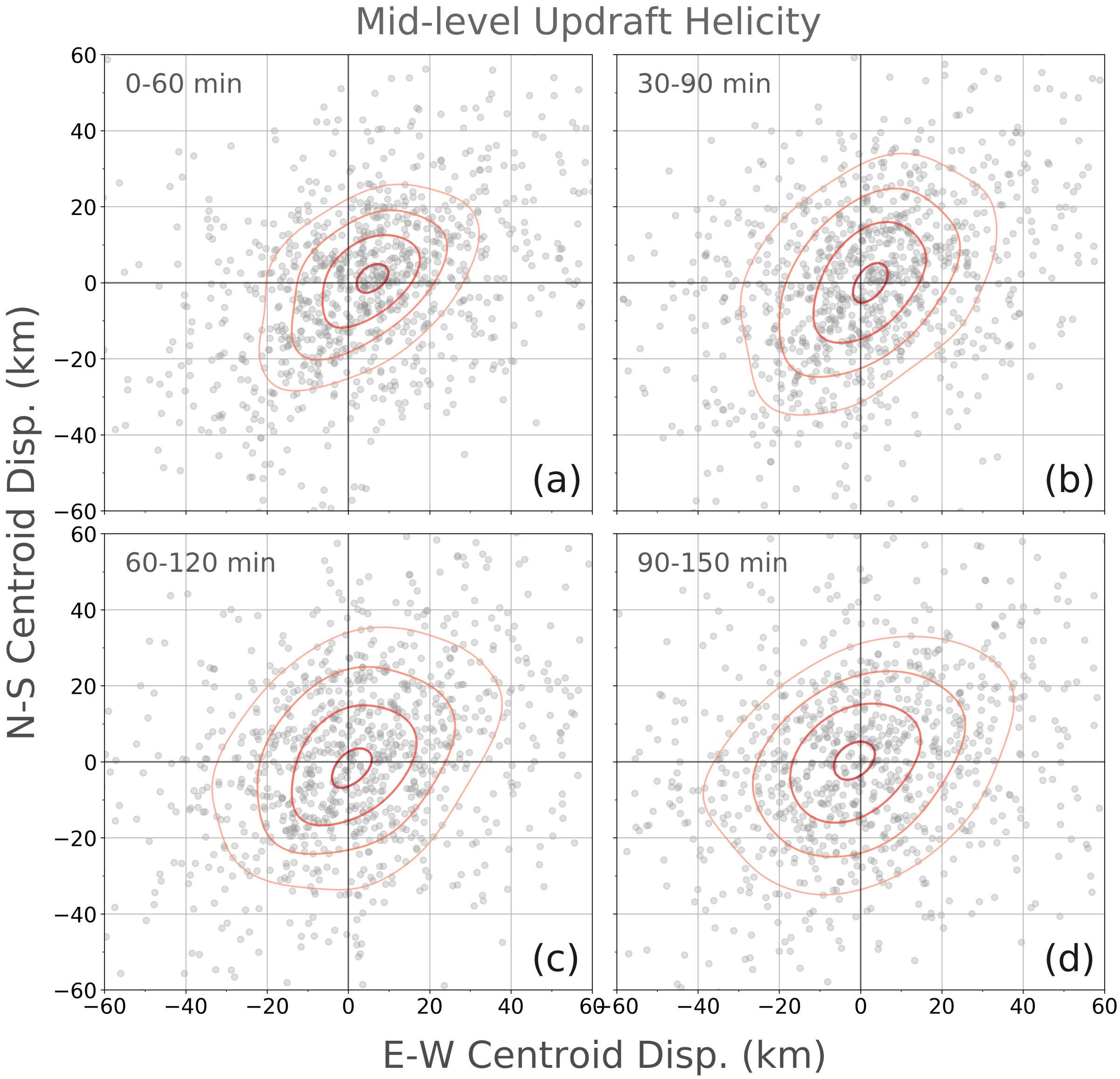


Figure 7.6: As in Figure 7.5, but for mid-level updraft helicity probability swath objects.

## 7.2 Predicting Severe Weather Hazards with ML

For the following verification results, the four components of the contingency table are redefined as

1. "hits": forecast yes for a given hazard and the ensemble storm track is matched to the corresponding LSR

2. "misses": forecast no for a given hazard, but the ensemble storm track is matched to the corresponding LSR

3. "false alarms": forecast yes for a given hazard, but the ensemble storm track is not matched to the corresponding LSR

4. "correct negatives": forecast no for a given hazard and the ensemble storm track is not matched to the corresponding LSR

### 7.2.1 Sensitivity to Class Imbalance

The full dataset (combined FIRST HOUR and SECOND HOUR) used in this dissertation is heavily imbalanced towards non-events; 1.2%, 2.5%, and 4% of ensemble storm track objects are matched to a tornado, severe hail, or severe wind report, respectively. ML algorithms often struggle to learn patterns and relationships from imbalanced datasets (Batista et al. 2004; Sun et al. 2009). One method to counteract the class imbalance is to randomly undersample the majority class (i.e., non-events) to produce a balance of events and non-events. For all three ML algorithms, randomly undersampling the majority class modestly improved tornado prediction as compared to training on the original dataset (see section Appendix D). However, for severe wind and hail, the difference in performance for all three ML algorithms training on resample data versus the original training dataset was negligible (see section Appendix D). I propose two reasons for this result. First, a large number of ensemble storm tracks are small (e.g., only composed of a single ensemble members updraft track) and are rarely matched to storm reports making them easily distinguishable as non-events. Thus, the class separation (the signal-to-noise ratio) is likely sufficient to counterbalance the class imbalance. Second, tornadoes have a lower signal-to-noise ratio than the severe wind and hail. Tornadoes are much rarer than the other two hazards and our understanding of the processes and environmental characteristics separating tornadic and non-tornadic environments remains an active area of research (e.g., Anderson-Frey et al. 2017; Coffer et al. 2017, 2019; Coniglio and Parker 2020; Flournoy et al. 2020). Therefore, eliminating a large portion of non-events (which can be associated with missing reports) from the training dataset may improve the signal-to-noise ratio more for tornadoes than the other two hazards.

### 7.2.2 Example Forecasts

Figure 7.7 shows characteristic examples of good and poor forecasts from the random forest model; these represent the other models as well (not shown). These examples include high confidence (probabilities closest to 1) forecasts matched and not matched to an event and low confidence (probabilities closest to 0) forecasts matched to an event. The skill of the ML forecasts is largely driven by the ability of the WoFS to accurately analyze ongoing convection through data assimilation. The classification, however, as we will see, is sensitive to slight changes in object location/separation. There may be minimal subjective differences between a confident match and confident false alarm (high confidence forecast not matched to the event), which is a limitation of the current method. For example, for high confidence (higher probabilities) forecasts matched to an event, the convection is fairly organized, and the WoFS matches well with the observed reflectivity (Figure 7.7a,d,g). Unfortunately, high confidence forecasts not matched to an event can exhibit similar behavior (Figure 7.7b,e,h). In Figure 7.7a and Figure 7.7b, storms in the Texas Panhandle have similar tornado probabilities despite only one of them producing tornado LSRs. It is possible that in this case the useful information for tornado forecasting in the WoFS was confined to larger spatial scales preventing discrimination of tornadic and non-tornadic storms occurring in proximity to one another. Complicating the interpretation, some of these apparent forecast busts may in fact be associated with an unreported event. For example, Potvin et al. (2019) found that over 50% of tornadoes within the central US went unreported from 1975 to 2016. For severe wind (Figure 7.7h), the timing of the higher confidence forecast was early as severe wind reports were eventually observed on the border of southern Ohio and northwest Kentucky (though

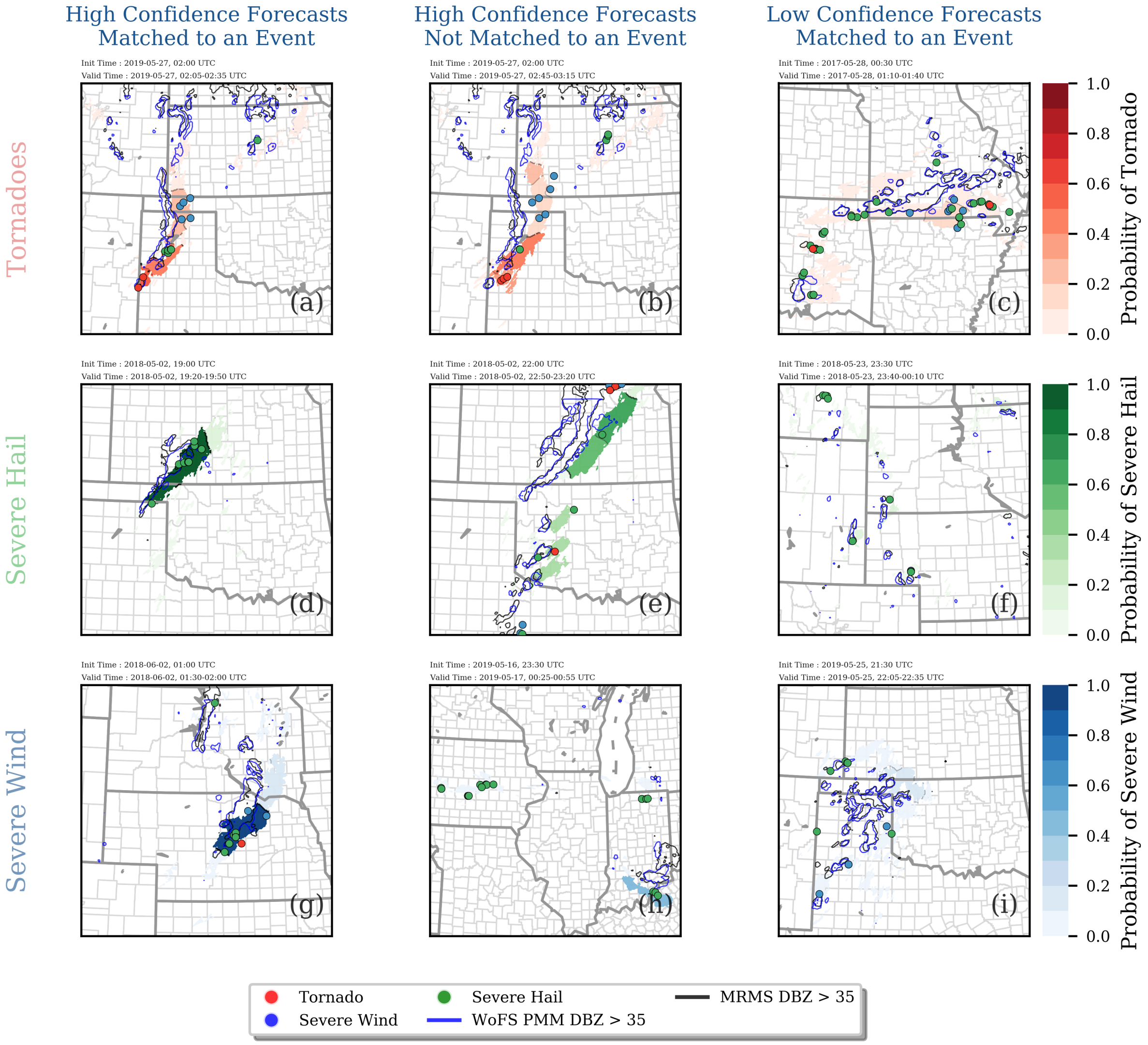

High Confidence Forecasts Matched to an Event
High Confidence Forecasts Not Matched to an Event
Low Confidence Forecasts Matched to an Event
Tornadoes
Severe Hail
Severe Wind
Init Time : 2019-05-27, 02:00 UTC
Valid Time : 2019-05-27, 02:05-02:35 UTC
(a)
Init Time : 2019-05-27, 02:00 UTC
Valid Time : 2019-05-27, 02:45-03:15 UTC
(b)
Init Time : 2017-05-28, 00:30 UTC
Valid Time : 2017-05-28, 01:10-01:40 UTC
(c)
Probability of Tornado
0.0
0.2
0.4
0.6
0.8
1.0
Init Time : 2018-05-02, 19:00 UTC
Valid Time : 2018-05-02, 19:20-19:50 UTC
(d)
Init Time : 2018-05-02, 22:00 UTC
Valid Time : 2018-05-02, 22:50-23:20 UTC
(e)
Init Time : 2018-05-23, 23:30 UTC
Valid Time : 2018-05-23, 23:40-00:10 UTC
(f)
Probability of Severe Hail
0.0
0.2
0.4
0.6
0.8
1.0
Init Time : 2018-06-02, 01:00 UTC
Valid Time : 2018-06-02, 01:30-02:00 UTC
(g)
Init Time : 2019-05-16, 23:30 UTC
Valid Time : 2019-05-17, 00:25-00:55 UTC
(h)
Init Time : 2019-05-25, 21:30 UTC
Valid Time : 2019-05-25, 22:05-22:35 UTC
(i)
Probability of Severe Wind
0.0
0.2
0.4
0.6
0.8
1.0
● Tornado
● Severe Wind
● Severe Hail
WoFS PMM DBZ > 35
MRMS DBZ > 35

Figure 7.7: Examples forecast from the random forest model predicting tornadoes (first row), severe hail (middle row), and severe wind (bottom row). These forecasts are representative instances of (first column) a high confidence forecast matched to an event (middle column) a high confidence forecast not matched to an event and (last column) a low confidence forecast matched to an event. For context, the 35-dBZ contour of the WoFS probability matched mean (blue) and Multi-Radar Multi-System (MRMS; black) composite reflectivity at forecast initialization time, respectively, are overlaid in each panel. The forecast initialization and valid forecast period are provided in the upper left hand corner of each panel. Tornado, severe hail, and severe wind reports are shown as red, green, and blues circles, respectively.

the observed storms were outside the WoFS domain).

For low confidence forecasts of severe hail and severe wind matched to an event, the convection is discrete and poorly organized (Figure 7.7f ) or disorganized and complex (Figure 7.7i). For the first case, discrete, poorly organized convection suggests a weakly forced environment that has lower predictability and in which it is more difficult to produce an accurate ensemble analysis. For the second case the WoFS reflectivity generally agrees with the observed reflectivity, but the severe wind reports are associated with the weaker, isolated convection, which can have limited predictability as well (similar for tornadoes; Figure 7.7c).

LSRs sometimes occur just outside of the boundaries of the ensemble storm tracks; see, for example, the severe hail report associated with the northernmost storm in Oklahoma in Figure 7.7e. On the other hand, the ensemble storm track areas are larger than a typical warning polygon and represent the WoFSs full range of storm location, and so our matching criterion is already relatively lenient. Given the impact of misses arising from small spatial errors in forecast storm tracks and spurious false alarms arising from missing reports, however, I argue that the following verification results likely underestimate the true skill of the ML models.

### 7.2.3 ROC Diagrams

The ROC curve results are shown in Figure 7.8. All three ML models produced, on average, an AUC greater than 0.9 for all three severe weather hazards for both lead time sets. While the ML model AUC scores were substantially better than those for the UH baseline, the latter were near or above 0.9, suggesting that the WoFS UH guidance is already a fairly good discriminator for the three severe weather hazards. While the AUC is high, its important to consider that this score is invariant to class imbalance and weighs event and non-event examples equally. Thus, the AUC provides an overly optimistic assessment of discrimination in applications where less importance is placed on correctly predicting non-events. For severe

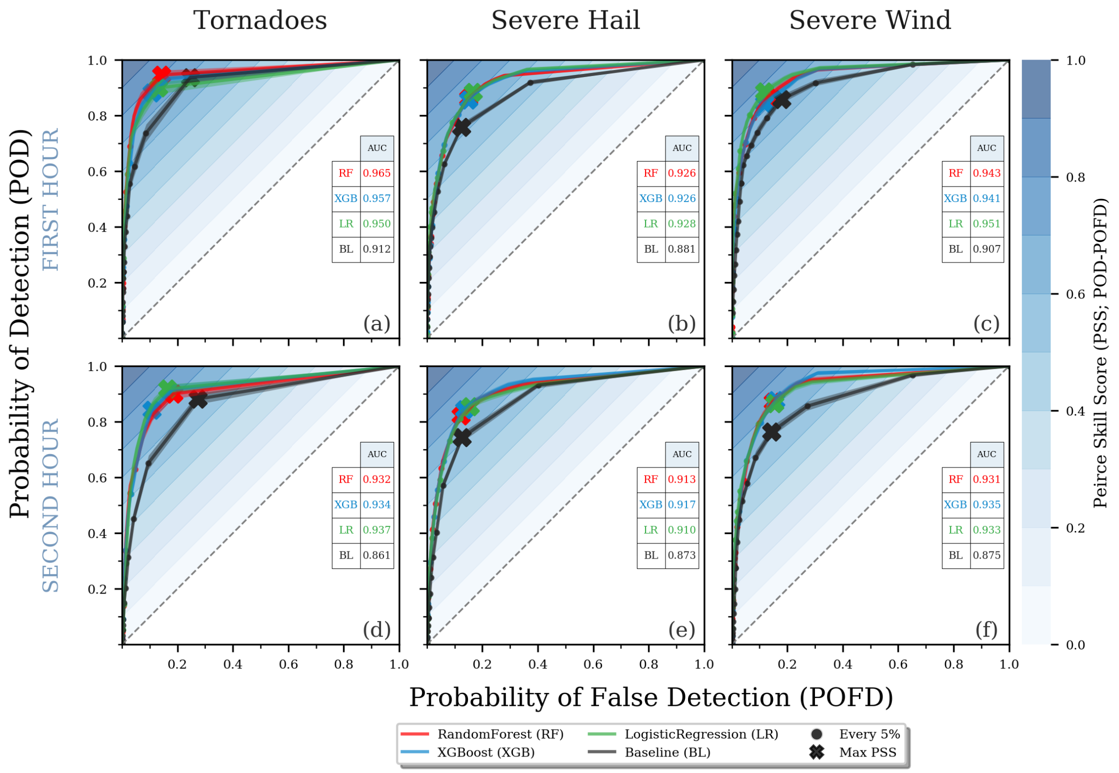


Figure 7.8: ROC curves for the random forests (RF;red), gradient-boosted classifier trees [XGBoost(XGB); blue], logistic regression (LR;green), and UH baseline (BL; black) predicting whether an ensemble storm track will contain a tornado (first column), severe hail (second column), or severe wind (third column) report. Results are combined over 30-min predictions starting within the lead times in the first hour (i.e., 0-30, 5-35, ..., 60-90 min; shown in panels a, b, c) and in the second hour (i.e., 65-95, 70-100, ..., 120-150 min; shown in panels d,e,f), respectively. Each line (shaded area) is the mean (95% confidence interval), determined by bootstrapping the testing examples (N=1000). Curves were calculated every 0.5% with dots plotted every 5%. The diagonal dashed line indicates a random classifier (no-skill). The mean AUC for each model is provided in the table in the upper right hand side of each panel. The filled contours are the Pierce skill score (PSS; also known as the true skill score) which is defined as POD-POFD. The maximum PSS is denoted on each curve with an X.

weather prediction, correct negatives are conditionally important because it is only desirable to accurately predict non-events in environments that favor severe weather (to reduce false alarms). However, a large number of ensemble storm tracks are easily distinguishable as non-events (as mentioned in section 7.2.1), which further suggests that caution be exercised when interpreting the high AUC values in this dissertation. This effect also explains why

AUC increases for severe weather hazards with lower climatological event frequencies; for rarer events, the aforementioned ensemble storm tracks become even easier to identify as non-events.

### 7.2.4 Performance Diagrams

The performance diagrams are shown in Figure 7.9. For the FIRST HOUR dataset (e.g., examples with a lead time of 0-30, 5-35, ..., 60-90 min; Figure 7.9a,b,c), the three ML models

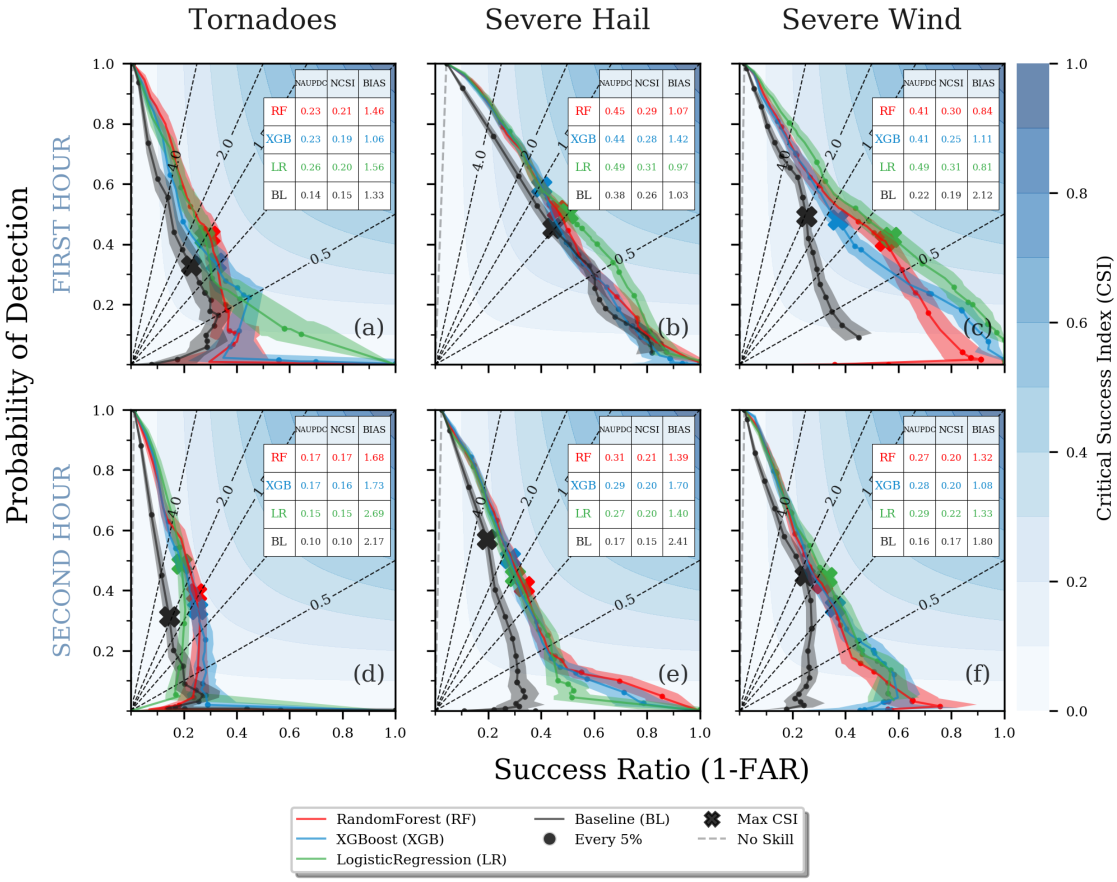


Figure 7.9: Same as in Figure 7.8, but for the performance diagram. The filled contours indicate the critical success index (CSI) while the dashed diagonal lines are the frequency bias. The dashed grey line indicates a no-skill classifier defined by equation 2.2. The mean NAUPDC, NCSI, and frequency bias (BIAS) for each model are provided in the table in the upper right hand side of each panel. The maximum CSI is denoted on each curve with an X

produced higher NAUPDC and maximum NCSI for severe hail and wind (Figure 7.9b,c) than for tornadoes (Figure 7.9a). This is unsurprising as the severe wind and hail events are more frequent than tornadoes, giving the ML more opportunities to learn from those examples. In addition, the processes governing hail growth and generation of strong near-surface winds are better resolved on a 3-km grid than the processes governing tornadogenesis, which is strongly influenced by small-scale processes in at least some cases Coffer et al. (2017); Flournoy et al. (2020). For tornadoes and severe hail, the NAUPDC and maximum NCSI of the three ML models were fairly indistinguishable from one another (Figure 7.9a,b), but for severe wind (Figure 7.9c), the random forest and logistic regression models produced substantially higher maximum NCSI than XGBoost. Other than for the severe wind random forest and logistic regression model, the frequency bias associated with maximum NCSI is greater than 1 (Figure 7.9a,b), which matches expectations for rare events (Baldwin and Kain 2006).

All three ML models substantially outperformed the UH baseline, but the magnitude of improvement varied with severe weather hazard. For tornadoes and especially severe wind, the ML predictions substantially improved upon the baseline. The superiority of the ML model severe wind forecasts is not surprising, as mid-level UH is less correlated with severe wind events (which are often produced by non-rotating storms) than with severe hail and tornado potential. The baseline predictions performed the best on severe hail, which is expected as mid-level UH is a proxy for supercells, which are the most prolific producer of severe hail (Duda and Gallus 2010) and especially significant severe hail Smith et al. (2012). This result aligns with Gagne et al. ( 2017) who found that UH predictions of severe hail competed with the ML-based predictions.

The performance curves were degraded for the SECOND HOUR dataset (e.g., examples with a lead time of 65-95, 70-100, ..., 120-150 min; Figure 7.9d,e,f). The POD remained relatively unchanged for tornadoes, but the FAR increased, which decreased the NAUPDC and maximum NCSI. The increase in FAR also led to the maximum CSI occurring with an increased over-forecasting frequency bias (especially for logistic regression). The predictability of storm-scale features relevant to tornado prediction (e.g., mid- and low-level mesocyclones) is greatly diminished at later lead times (Flora et al. 2018) and therefore this degradation in skill is not surprising. For severe hail and wind, the changes in POD and FAR relative to FIRST HOUR compensated each other such that the maximum-CSI frequency bias remained slightly above one. The major exception is the XGBoost severe hail model, which suffered from over-forecasting bias in the FIRST HOUR dataset but in the SECOND HOUR dataset has a maximum-CSI frequency bias near 1 (1.08). The difference in performance between the UH baseline predictions and the three ML models are more pronounced in SECOND HOUR than FIRST HOUR, suggesting that ML-based calibration of ensemble forecasts is more useful at longer lead times. This result suggests that the ML models are learning enough useful information from the ensemble statistics at these later lead times to partly compensate the inevitable reduction in CAM forecast skill because of intrinsically limited storm-scale predictability.

For all three severe weather hazards, the logistic regression model has a substantially higher SR (lower FAR) at higher probability thresholds (lower right-hand portion of the diagram) than the other ML models, which explains the slightly higher mean NAUPDC values. To explain why logistic regression can produce fewer false alarms for higher confidence forecasts, Figure 7.10 illustrates how predictions from a random forest and logistic regression model compare for a simple noisy 2D dataset. A classic problem in ML is the trade-off between the bias and variance of a model (Kuhn and Johnson 2013). With a high-variance model, we risk over-fitting to noisy or unrepresentative training data. In contrast, a high-bias model is typically simpler and tends to underfit the training data, failing to capture important regularities. Tree-based methods partition the predictor space and produce predictions based on the local event frequency of the training dataset. If there is sufficient noise in the classification (e.g., ensemble storm tracks mislabeled as non-events because of missing storm reports), then the local event frequency could be unrepresentative of the true local event frequency. Though the tree-based method can produce skillful high confidence

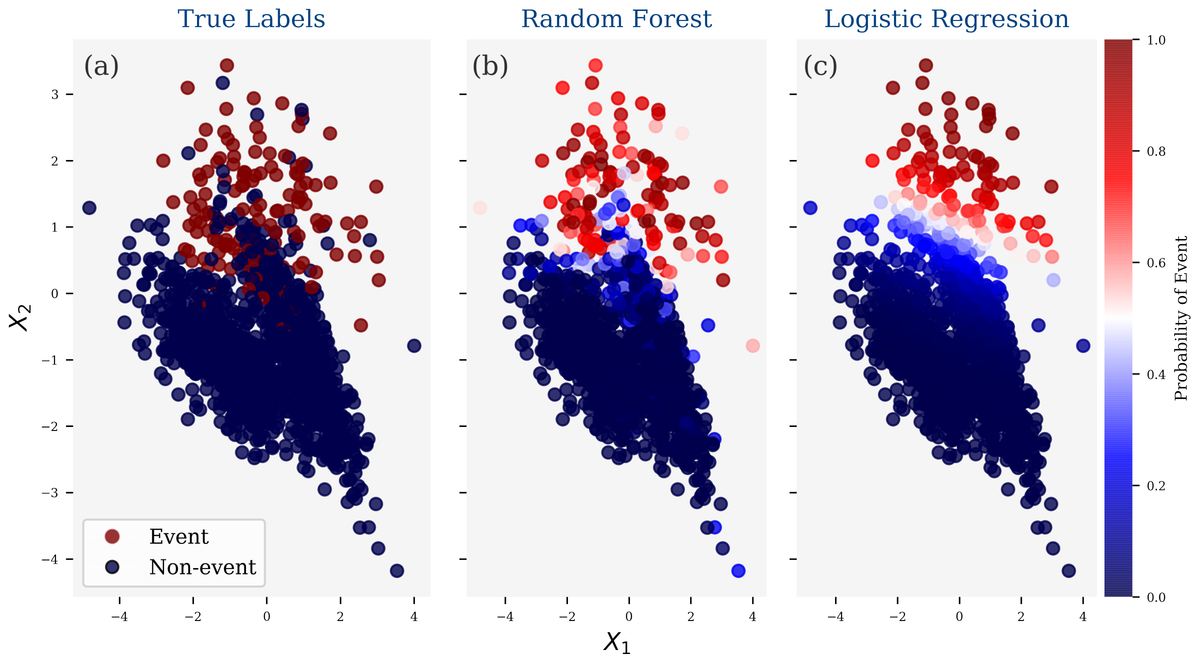


Figure 7.10: Illustration of predictions for a simple noisy 2D dataset in (shown in a) from a random forest (shown in b; tree-based models in general) and logistic regression model (shown in c).

forecasts with noisier datasets (as seen in Figure 7.10b; Hoekstra et al. 2011), they are high-variance models (more sensitive to random variations in the data) and can struggle near decision boundaries or in poorly sampled regions of the predictor space. For example, near point $(X_1; X_2) = (-1, 1)$, the random forest probabilities do not reflect the uncertainty of the true labels and for points $X_2 > 2$, the predictions have high confidence, but instances of unrepresentative uncertainty (e.g., the probability of point $(X_1; X_2) = (2, 2.5)$ is 50%, but should be 100%). Logistic regression is a lower-variance, higher-bias model compared to tree-based methods (since it is a linear model which may not sufficiently generalize a dataset) and so its predictions are not very sensitive to noisy labeling and rather, as we can see in Figure 7.10, increase (or decrease) perpendicular to the linear decision boundary. Therefore, I propose that the logistic regression models in this dissertation are producing fewer false alarms than tree-based models at higher probability thresholds since the tree-based methods are strongly impacted by the noisy labeling and are over-fitting the training dataset. However, the logistic regression models are not markedly better than the tree-based methods, so the tradeoff between bias and variance is still a relevant issue. It is likely that if the ensemble storm tracks were labeled better (improving the signal-to-noise ratio) then the tree-based methods would outperform logistic regression, since a linear decision boundary does not sufficiently generalize to the data.

### 7.2.5 Attribute Diagrams

The attribute diagram results are shown in Figure 7.11. For both lead time ranges, the severe hail and wind prediction were the most reliable (Figure 7.11b,c,e,f). The larger numbers of severe hail and wind events than tornado events in the training dataset likely contribute to increased reliability by improving the local event frequencies for the tree-based methods and the coefficients of the linear model in logistic regression. All three models produced reliable severe wind probabilities up to 40-50% with a small underforecasting bias for higher probabilities; no model produced forecast probabilities greater than 80% (Figure 7.11c). Severe hail probabilities for all three models were reliable up to 40% with a small over-forecasting bias for probabilities greater than 60% with probabilities up to 90% being produced. The under-forecasting bias was substantially higher for the logistic regression, which corresponds with the lower FAR at higher probabilities previously noted in the performance diagram (Figure 7.10). Though the logistic regression model is less reliable than the tree-based models for severe wind and hail, its resolution is higher, which explains why its BSS is higher. The logistic regression model also produced the least reliable tornado predictions, exhibiting an under-forecasting bias, and only produced forecast probabilities up to 40%. The tree-based models produced higher probabilities, but the uncertainty in the conditional event frequencies is too large to assess the forecast reliability at these higher probabilities. The smaller forecast probabilities for tornadoes is not surprising for at least two reasons. First,

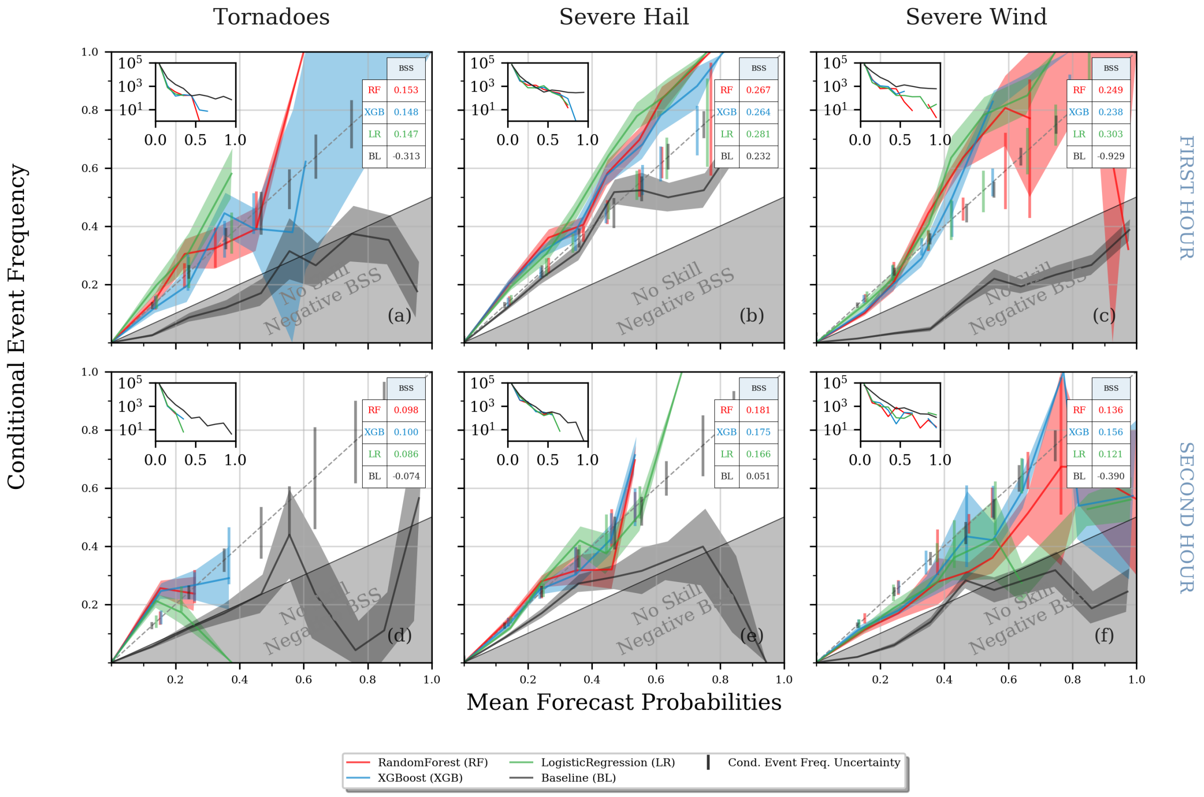


Figure 7.11: Same as in Figure 7.8, but for attribute diagrams. The bin increment of forecast probabilities is 10%. The inset figure is the forecast histogram for each model. The dashed line represents perfect reliability while the grey region separates positive and negative Brier skill score (positive Brier skill score above the grey area). The vertical lines along the diagonal are the error bars for the observed frequency for each model in each bin based on the method in Bröcker and Smith (2007). To limit figure crowding, error bars associated with an uncertainty of > 50% for a given conditional observed frequency were omitted. The mean BSS for each model is provided in the table in the upper right hand side of each panel.

missing tornado reports (Potvin et al. 2019) coupled with the rarity of tornado events limits the ability of the ML models to learn subtle patterns in the data. Second, storm-scale predictability limits (Flora et al. 2018) prevents greater confidence in tornado likelihood, especially at later lead times.

For all severe weather hazards, reliability and resolution were degraded for the SECOND HOUR dataset. The tornado probabilities are arguably reliable and the maximum probability is between 30-40%, which are fairly confident forecasts of such a rare event. For severe hail, the forecast probabilities remained relatively reliable, but the maximum forecast probability was substantially reduced, which lowered the BSS. The severe wind forecast probabilities for all three models became overconfident at later lead times (cf. Figure 7.11c and Figure 7.11f).

For tornadoes and severe wind, the UH baseline was unreliable and unskillful at all lead times (underperformed climatology; Figure 7.11a,c,d,f). Reliability is possibly improved at a higher UH threshold, but then the ranking-based metrics would have suffered. This result highlights that the simple threshold method is likely over-fitting the training dataset and is suboptimal for capturing forecast uncertainty, which is similar to the result found in Sobash et al. (2020). The UH baseline was fairly reliable for severe hail, but the ML models were still substantially more reliable (Figure 7.11b).

### 7.2.6 Performance Metrics for Individual Forecast Lead Times

Figure 7.12 and Figure 7.13 shows the performance of the three ML models for the different severe weather hazards as a function of forecast lead time (for the FIRST HOUR and SECOND HOUR dataset, respectively).

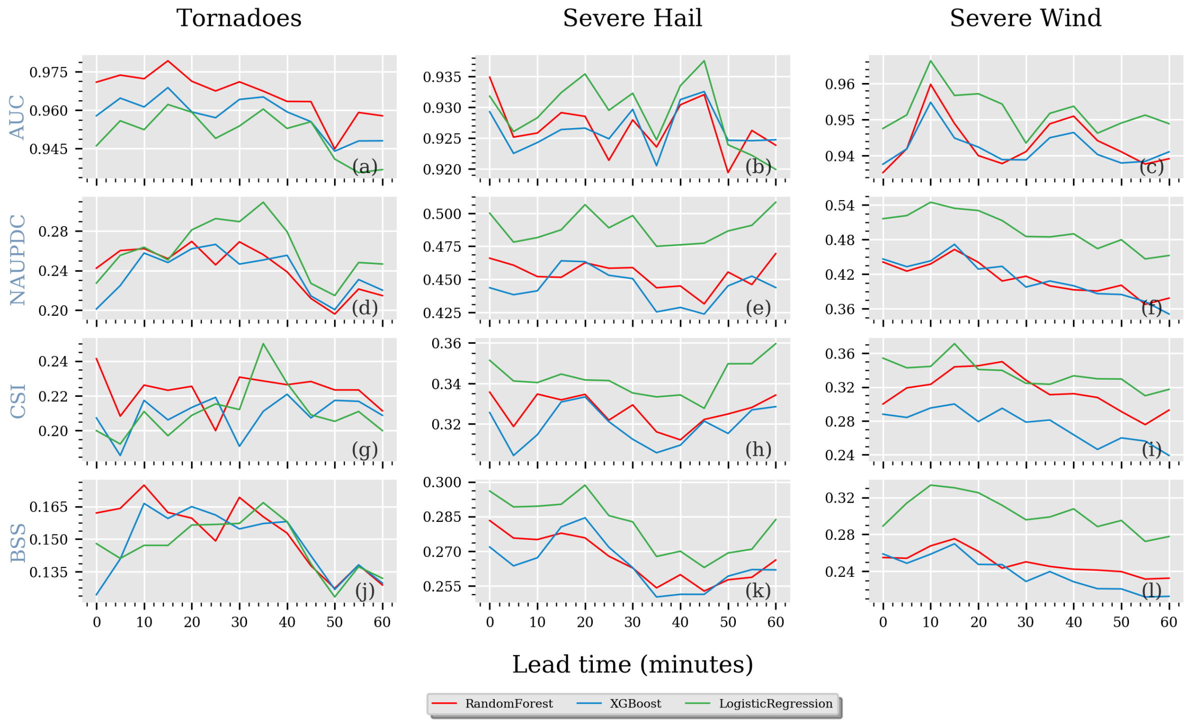


Figure 7.12: Performance of the random forest (red), XGBoost (blue), and logistic regression (green) models for predicting whether a WoFS forecast storm will produce a tornado (first column), severe hail (middle column) and/or severe wind (last column) report, respectively for lead times up to 60 minutes. The performance metrics include area under the ROC curve (AUC; first row), normalized area under the performance diagram curve (NAUPDC;second row), critical success ratio (CSI; third row), and Brier skill score (BSS; fourth row). Metrics are defined in section 2.3

Although there is some variance, all four metrics remain fairly consistent even at later lead times. In most cases, the performance does steadily degrade at later lead times. The main exception is severe hail based on NAUPDC, CSI, and BSS (Figure 7.12e,h,k) where the scores increase between 40-60 min. Given that the changes for all metrics are not substantial, it is unclear whether these trends are truly noteworthy. The skill of the three models were similar for tornadoes (Figure 7.12a,d,g,j), with slightly more separation in skill for severe hail (Figure 7.12b,e,h,k) and severe wind (Figure 7.12c,f,i,l). The results in Figure 7.13 are similar to Figure 7.12, but all the scores are lower, which is consistent with the results in Figure 7.8, Figure 7.9, and Figure 7.11. Overall, these results suggest that model performance as shown in the previous sections is representative for all lead times contained within a given dataset (either FIRST HOUR or SECOND HOUR, respectively).

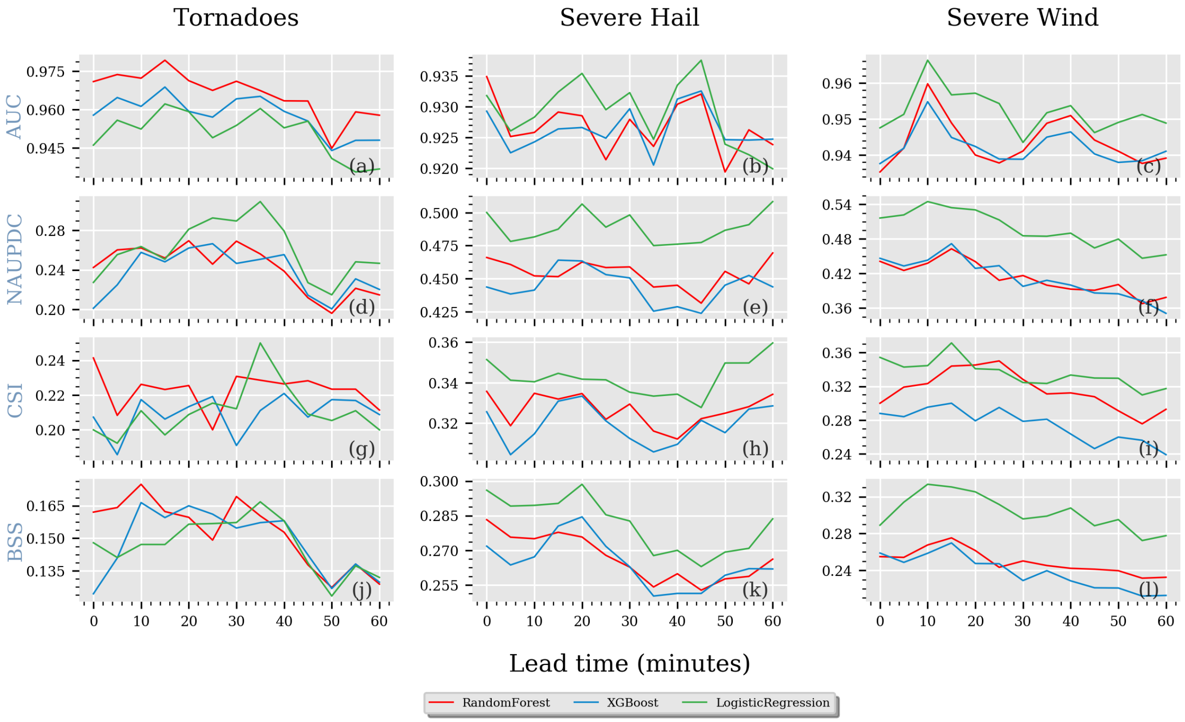


Figure 7.13: Same as in Figure 7.12, but for the SECOND HOUR dataset (forecast lead times between 60-120 min.

## 7.3 ML Interpretability

### 7.3.1 Permutation Importance and Expected Contributions

The multiple-pass permutation importance results for the FIRST HOUR dataset and corresponding ALE curves are shown in Figure 7.14 and Figures 7.15- 7.17.

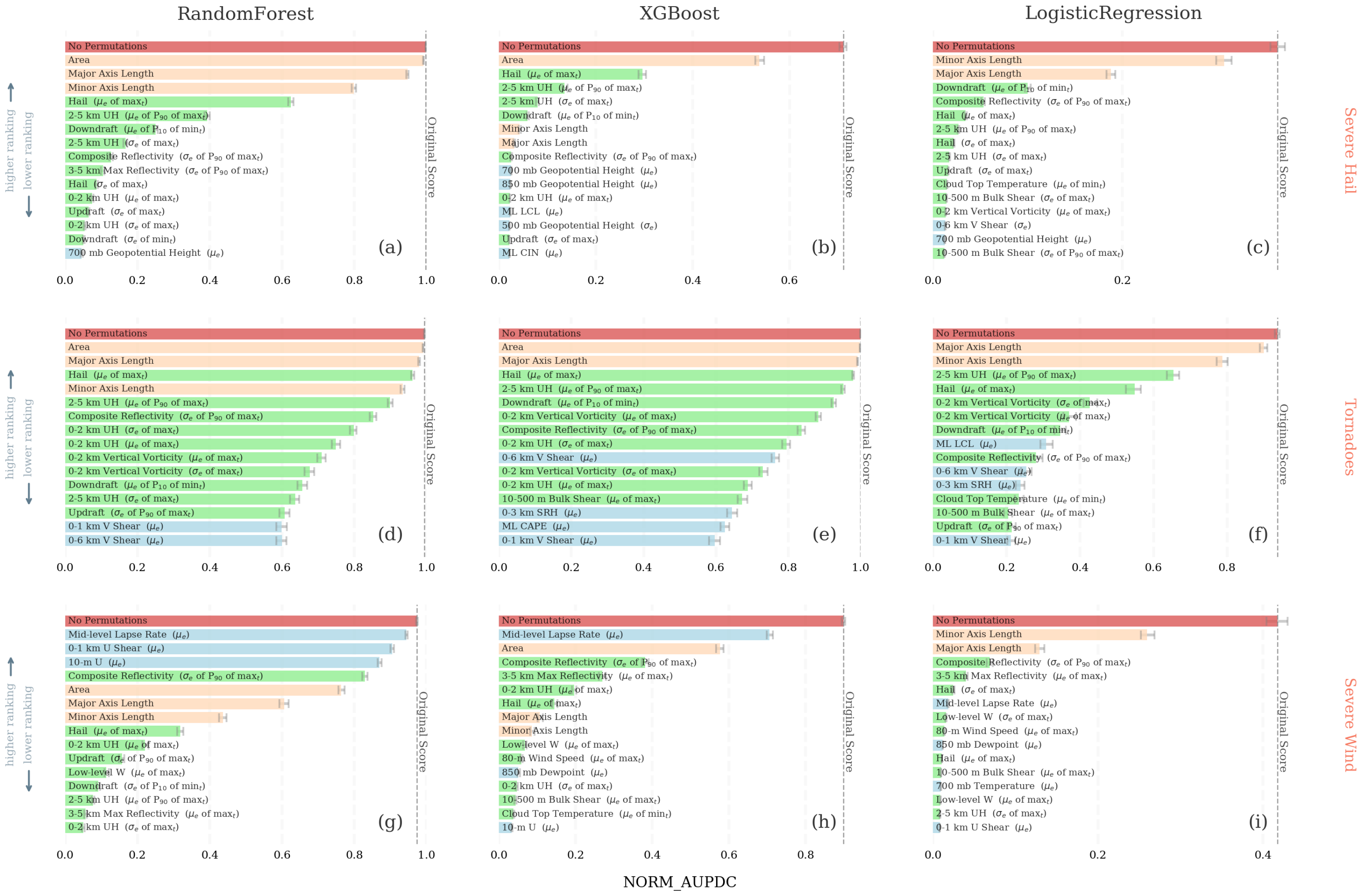

RandomForest
XGBoost
LogisticRegression
Severe Hail
Tornadoes
Severe Wind
higher ranking
lower ranking
Original Score
(a)
No Permutations
Area
Major Axis Length
Minor Axis Length
Hail ($\mu_e$ of max$_t$)
2-5 km UH ($\mu_e$ of P$_{90}$ of max$_t$)
Downdraft ($\mu_e$ of P$_{10}$ of min$_t$)
2-5 km UH ($\sigma_e$ of max$_t$)
Composite Reflectivity ($\sigma_e$ of P$_{90}$ of max$_t$)
3-5 km Max Reflectivity ($\sigma_e$ of P$_{90}$ of max$_t$)
Hail ($\sigma_e$ of max$_t$)
0-2 km UH ($\mu_e$ of max$_t$)
Updraft ($\sigma_e$ of max$_t$)
0-2 km UH ($\sigma_e$ of max$_t$)
Downdraft ($\sigma_e$ of min$_t$)
700 mb Geopotential Height ($\mu_e$)
(b)
No Permutations
Area
Hail ($\mu_e$ of max$_t$)
2-5 km UH ($\mu_e$ of P$_{90}$ of max$_t$)
2-5 km UH ($\sigma_e$ of max$_t$)
Downdraft ($\mu_e$ of P$_{10}$ of min$_t$)
Minor Axis Length
Major Axis Length
Composite Reflectivity ($\sigma_e$ of P$_{90}$ of max$_t$)
700 mb Geopotential Height ($\mu_e$)
850 mb Geopotential Height ($\mu_e$)
0-2 km UH ($\mu_e$ of max$_t$)
ML LCL ($\mu_e$)
500 mb Geopotential Height ($\sigma_e$)
Updraft ($\sigma_e$ of max$_t$)
ML CIN ($\mu_e$)
(c)
No Permutations
Minor Axis Length
Major Axis Length
Downdraft ($\mu_e$ of P$_{10}$ of min$_t$)
Composite Reflectivity ($\sigma_e$ of P$_{90}$ of max$_t$)
Hail ($\mu_e$ of max$_t$)
2-5 km UH ($\mu_e$ of P$_{90}$ of max$_t$)
Hail ($\sigma_e$ of max$_t$)
2-5 km UH ($\sigma_e$ of max$_t$)
Updraft ($\sigma_e$ of max$_t$)
Cloud Top Temperature ($\mu_e$ of min$_t$)
10-500 m Bulk Shear ($\sigma_e$ of max$_t$)
0-2 km Vertical Vorticity ($\mu_e$ of max$_t$)
0-6 km V Shear ($\sigma_e$)
700 mb Geopotential Height ($\mu_e$)
10-500 m Bulk Shear ($\sigma_e$ of P$_{90}$ of max$_t$)
0.0
0.2
0.4
0.6
0.8
1.0
(d)
No Permutations
Area
Major Axis Length
Hail ($\mu_e$ of max$_t$)
Minor Axis Length
2-5 km UH ($\mu_e$ of P$_{90}$ of max$_t$)
Composite Reflectivity ($\sigma_e$ of P$_{90}$ of max$_t$)
0-2 km UH ($\sigma_e$ of max$_t$)
0-2 km UH ($\mu_e$ of max$_t$)
0-2 km Vertical Vorticity ($\mu_e$ of max$_t$)
0-2 km Vertical Vorticity ($\sigma_e$ of max$_t$)
Downdraft ($\mu_e$ of P$_{10}$ of min$_t$)
2-5 km UH ($\sigma_e$ of max$_t$)
Updraft ($\sigma_e$ of P$_{90}$ of max$_t$)
0-1 km V Shear ($\mu_e$)
0-6 km V Shear ($\mu_e$)
(e)
No Permutations
Area
Major Axis Length
Hail ($\mu_e$ of max$_t$)
2-5 km UH ($\mu_e$ of P$_{90}$ of max$_t$)
Downdraft ($\mu_e$ of P$_{10}$ of min$_t$)
0-2 km Vertical Vorticity ($\mu_e$ of max$_t$)
Composite Reflectivity ($\sigma_e$ of P$_{90}$ of max$_t$)
0-2 km UH ($\sigma_e$ of max$_t$)
0-6 km V Shear ($\mu_e$)
0-2 km Vertical Vorticity ($\sigma_e$ of max$_t$)
0-2 km UH ($\mu_e$ of max$_t$)
10-500 m Bulk Shear ($\mu_e$ of max$_t$)
0-3 km SRH ($\mu_e$)
ML CAPE ($\mu_e$)
0-1 km V Shear ($\mu_e$)
(f)
No Permutations
Major Axis Length
Minor Axis Length
2-5 km UH ($\mu_e$ of P$_{90}$ of max$_t$)
Hail ($\mu_e$ of max$_t$)
0-2 km Vertical Vorticity ($\sigma_e$ of max$_t$)
0-2 km Vertical Vorticity ($\mu_e$ of max$_t$)
Downdraft ($\mu_e$ of P$_{10}$ of min$_t$)
ML LCL ($\mu_e$)
Composite Reflectivity ($\sigma_e$ of P$_{90}$ of max$_t$)
0-6 km V Shear ($\mu_e$)
0-3 km SRH ($\mu_e$)
Cloud Top Temperature ($\mu_e$ of min$_t$)
10-500 m Bulk Shear ($\mu_e$ of max$_t$)
Updraft ($\sigma_e$ of P$_{90}$ of max$_t$)
0-1 km V Shear ($\mu_e$)
(g)
No Permutations
Mid-level Lapse Rate ($\mu_e$)
0-1 km U Shear ($\mu_e$)
10-m U ($\mu_e$)
Composite Reflectivity ($\sigma_e$ of P$_{90}$ of max$_t$)
Area
Major Axis Length
Minor Axis Length
Hail ($\mu_e$ of max$_t$)
0-2 km UH ($\mu_e$ of max$_t$)
Updraft ($\sigma_e$ of P$_{90}$ of max$_t$)
Low-level W ($\mu_e$ of max$_t$)
Downdraft ($\sigma_e$ of P$_{10}$ of min$_t$)
2-5 km UH ($\mu_e$ of P$_{90}$ of max$_t$)
3-5 km Max Reflectivity ($\mu_e$ of max$_t$)
0-2 km UH ($\sigma_e$ of max$_t$)
(h)
No Permutations
Mid-level Lapse Rate ($\mu_e$)
Area
Composite Reflectivity ($\sigma_e$ of P$_{90}$ of max$_t$)
3-5 km Max Reflectivity ($\mu_e$ of max$_t$)
0-2 km UH ($\mu_e$ of max$_t$)
Hail ($\mu_e$ of max$_t$)
Major Axis Length
Minor Axis Length
Low-level W ($\mu_e$ of max$_t$)
80-m Wind Speed ($\mu_e$ of max$_t$)
850 mb Dewpoint ($\mu_e$)
0-2 km UH ($\sigma_e$ of max$_t$)
10-500 m Bulk Shear ($\mu_e$ of max$_t$)
Cloud Top Temperature ($\mu_e$ of min$_t$)
10-m U ($\mu_e$)
(i)
No Permutations
Minor Axis Length
Major Axis Length
Composite Reflectivity ($\sigma_e$ of P$_{90}$ of max$_t$)
3-5 km Max Reflectivity ($\mu_e$ of max$_t$)
Hail ($\sigma_e$ of max$_t$)
Mid-level Lapse Rate ($\mu_e$)
Low-level W ($\sigma_e$ of max$_t$)
80-m Wind Speed ($\mu_e$ of max$_t$)
850 mb Dewpoint ($\mu_e$)
Hail ($\mu_e$ of max$_t$)
10-500 m Bulk Shear ($\mu_e$ of max$_t$)
700 mb Temperature ($\mu_e$)
Low-level W ($\mu_e$ of max$_t$)
2-5 km UH ($\sigma_e$ of max$_t$)
0-1 km U Shear ($\mu_e$)
NORM_AUPDC

Figure 7.14: The 15 most important predictors, according to the multiple-pass permutation method, for random forest (first row), logistic regression (second row) and gradient-boosted classifier trees (XGBoost; last row) for predicting tornadoes (first column), severe hail (second column), and severe wind (last column). Predictor importance was measured using the normalized area under the performance diagram (NORM_AUPDC; defined in Section 7.2.4). Values are averaged over 100 bootstrapping replicates, and error bars show the 95% confidence interval. Object properties are orange, environmental parameters are blue, and intra-storm state predictors are green. The original score before any permutations is shown as the top red bar and as a vertical dashed line. ($\mu_e$) refers to spatial-average ensemble mean of the environmental variables, ($\mu_e$ of $\max_t$) is spatial-average ensemble mean of the time-composite intra-storm variables, ($\mu_e$ of $P_{90}$ of $\max_t$) is the ensemble-average of the spatial 90th percentile values extracted from ensemble members within the ensemble storm tracks, and ($\sigma_e$ of $\max_t$) is spatial-average ensemble standard deviation of the time-composite intra-storm variables. SRH is the storm-relative helicity, and Hail refers to maximum hail diameter from WRF-HAILCAST.

To limit the analysis, only the top 15 predictors were computed for each model. The top predictors are fairly similar for the random forest, XGBoost, and logistic regression, but order varies. This is unsurprising as the rankings within a model are not unambiguous (Marzban et al. 1999) and because of the "Rashomon" effect (Breiman 2001a; Fisher et al. 2018) different models can fit the data equally well, but focus on different multivariate relationships in the data. The most important predictors for all three models, however, were based on the storm morphology (e.g., area, minor axis length, and/or major axis length). Increasing the ensemble track size increased the probability of an event for all three ML Models (cf. Figure 7.15d, Figure 7.16b, and Figure 7.17a,b). This is not unexpected as ensemble tracks size can range from a single updraft track from a single ensemble member to a composite of MCSs or supercells from several (if not all) ensemble members and a larger composite area is more likely to capture an event than a single updraft track. With a few exceptions, the importance of the storm morphology predictors was limited as permuting their data did not substantially decrease model performance. This is not surprising, as information about small ensemble storm tracks is redundant in other predictors. For example, if only one ensemble member predicts an updraft track in a particular location, the ensemble mean and spread for all intra-storm predictors will be near zero. To support this claim, the ML

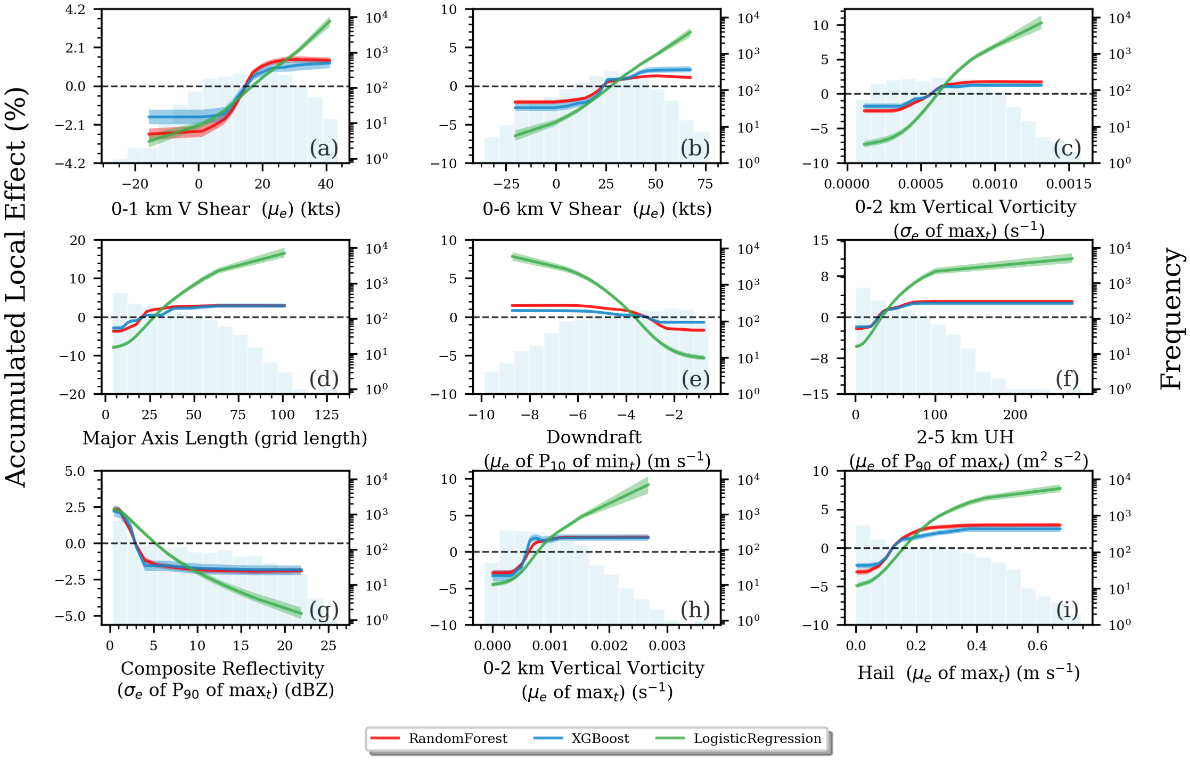


Figure 7.15: Accumulated local effect (ALE) curves for top predictors of a random forest (RandomForest; red), gradient-boosted classifier tree (XGBoost; blue), and logistic regression (LogisticRegression; green) trained to predict tornado likelihood. The ALE is the expected contribution of a predictor when it takes a particular value (where contributions are additive). Marginal distribution of the predictors in the training shown in light blue. The ALE values were computed through bootstrap iteration (N=100) with mean and 95% confidence interval contours shown.

models were retrained with the storm morphology predictors missing and it was found that the performance did not substantially decrease, if at all (see Appendix C).

The degradation of performance as more predictors are permuted varies based on model and severe weather hazard. For example, the NAUPDC degradation for the logistic regression and XGBoost models are much greater than that of the random forest models, especially for severe hail and severe wind (cf. last two columns of Figure 7.14 with the first column). In Figure 7.14b, Figure 7.14c, and Figure 7.14i, the NAUPDC asymptotes close to zero after the top 7-8 predictors are removed, suggesting that these few predictors make the

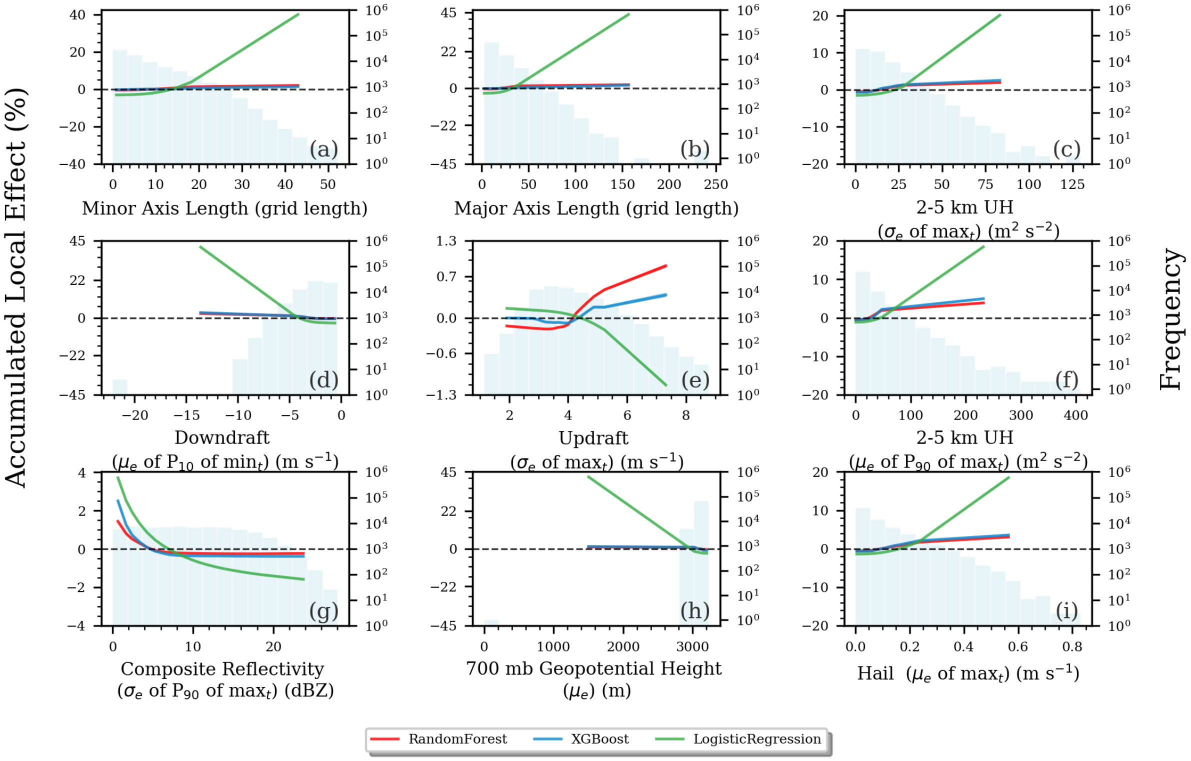


Figure 7.16: Same as Figure 7.15, but for severe hail.

biggest contribution to the overall model performance. The shallower decay in NAUPDC for the tornado-based models (Figure 7.14d,e,f) is likely related to the weaker interaction strength (see Table 7.1). If an ML model has learned strong multivariate relationships, then permuting any of those predictors can substantially impact the model performance. However, if the interaction strength is weaker, then there is more reliance on first-order effects, which requires permuting more predictors to reduce model performance. It is possible that a low signal-to-noise ratio coupled with misclassification prohibited the ML models from capturing strong multivariate relationships for tornado prediction. A more definitive answer, however, is beyond the scope of this dissertation, but warrants future exploration.

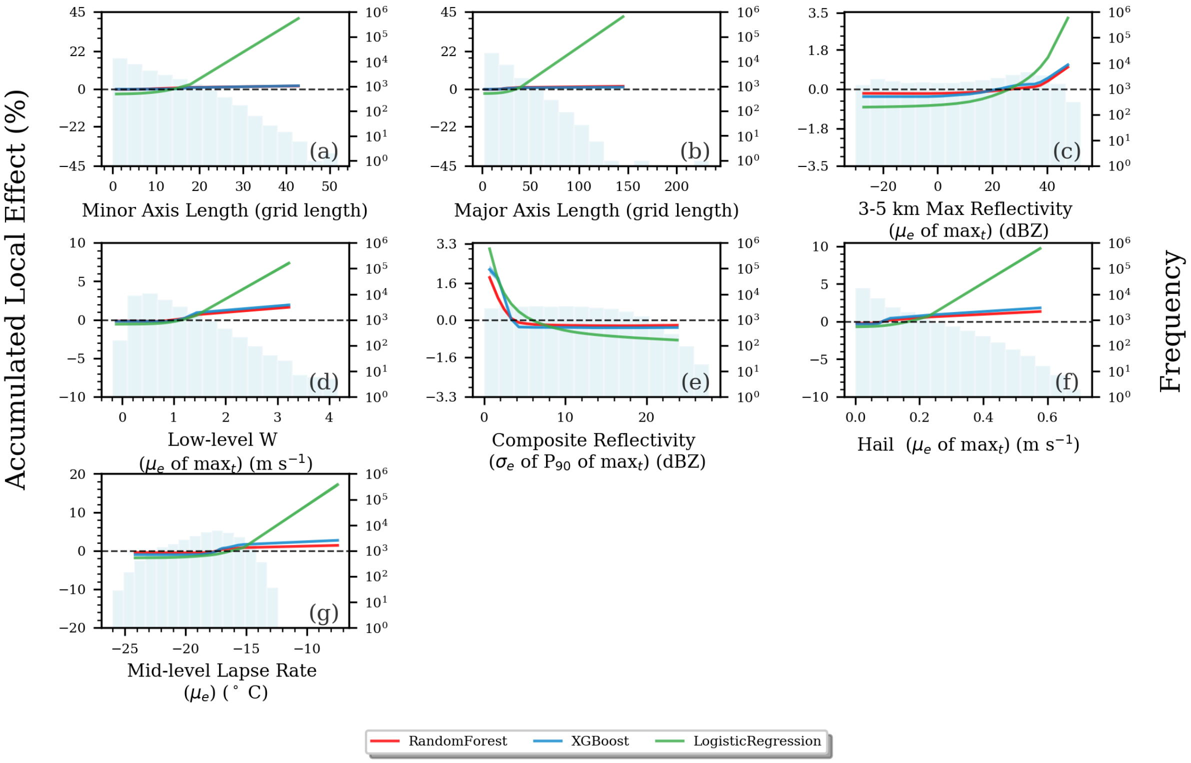


Figure 7.17: Same as Figure 7.15, but for severe wind.

| | Random Forest | XGBoost | Logistic Regression |
|---|---|---|---|
| Severe Hail | 0.62 | 0.73 | 0.68 |
| Tornadoes | 0.18 | 0.23 | 0.14 |
| Severe Wind | 0.83 | 0.68 | 0.88 |

Table 7.1: Interaction strength (see equation 6.6) for the random forest (first column), XGBoost (middle column), and logistic regression (last column) for predicting severe hail (first row), tornadoes (middle row), and severe wind (last row) in the FIRST HOUR dataset.

The majority of the top predictors are intra-storm variables (Figure 7.14). The greater importance of the intra-storm predictors is not surprising, as the lead times used in this dissertation are short enough such that the storm-scale predictability has not fully degraded the useful storm-scale information in the WoFS forecasts (e.g., Flora et al. 2018). The environmental predictors may be redundant information as a fully developed intra-storm state is a product of its environment. As will be shown in the following section, there is evidence that environmental parameters are useful when storms are poorly spun-up in the WoFS domain. However, the contributions of environmental parameters are overwhelmed by the negative contributions of the poorly developed storm state. A final consideration is that the WoFS domain is purposely centered on the most favorable conditions for severe weather (as determined by the Storm Prediction Center), so distinguishing between event and non-event may require hyperfine distinctions. Learning such distinctions may be difficult given model and LSR reporting errors in the training dataset.

In terms of the specific severe weather hazards, tornado prediction relied on hail size, reflectivity, low-level vertical vorticity, updraft helicity, downdraft and the vertical wind structure (e.g., 0-6 km and 0-1 km V-component of wind shear and 0-3 km SRH). Though a 3-km grid may not properly resolve features such as the low-level mesocyclone (Potvin and Flora 2015), these results suggest that signatures from the mid-level mesocyclone and low-level rotation (or in the low-level wind profile) are useful predictors for severe weather likelihood. Based on Figure 7.15c, all three models found that an increase in ensemble spread of vertical vorticity increased tornado likelihood while increasing ensemble spread for composite reflectivity decreased tornado likelihood (Figure 7.15g). Increasing tornado likelihood with increasing ensemble spread for vertical vorticity (and other intra-storm variables like updraft helicity) is not unexpected as an ensemble of weak storms will have zero spread for intra-storm variables. As the storms increase in intensity/strength, the ensemble spread will also increase. One exception is the composite reflectivity, which can have non-zero spread for weak, non-severe storms. The interpretation of ensemble spread of composite reflectivity, however, is more nuanced. It is likely that younger (or poorly spun up) storms within an ensemble track will be associated with a higher ensemble spread, which should decrease the tornado likelihood. The predictability regime then dictates whether the ensemble spread will drop [e.g., the storms are inheriting predictability from large scale forcing which improves the confidence in the forecast (Flora et al. 2018)] or remain higher (e.g., a weakly forced environment) as the storms develop. In the former case, one would expect a positive contribution to tornado likelihood, while in the latter case a negative contribution.

Figure 7.18 shows the ALE curves for select environmental predictors of tornado likelihood to determine if the ML models have learned known physical relationships.

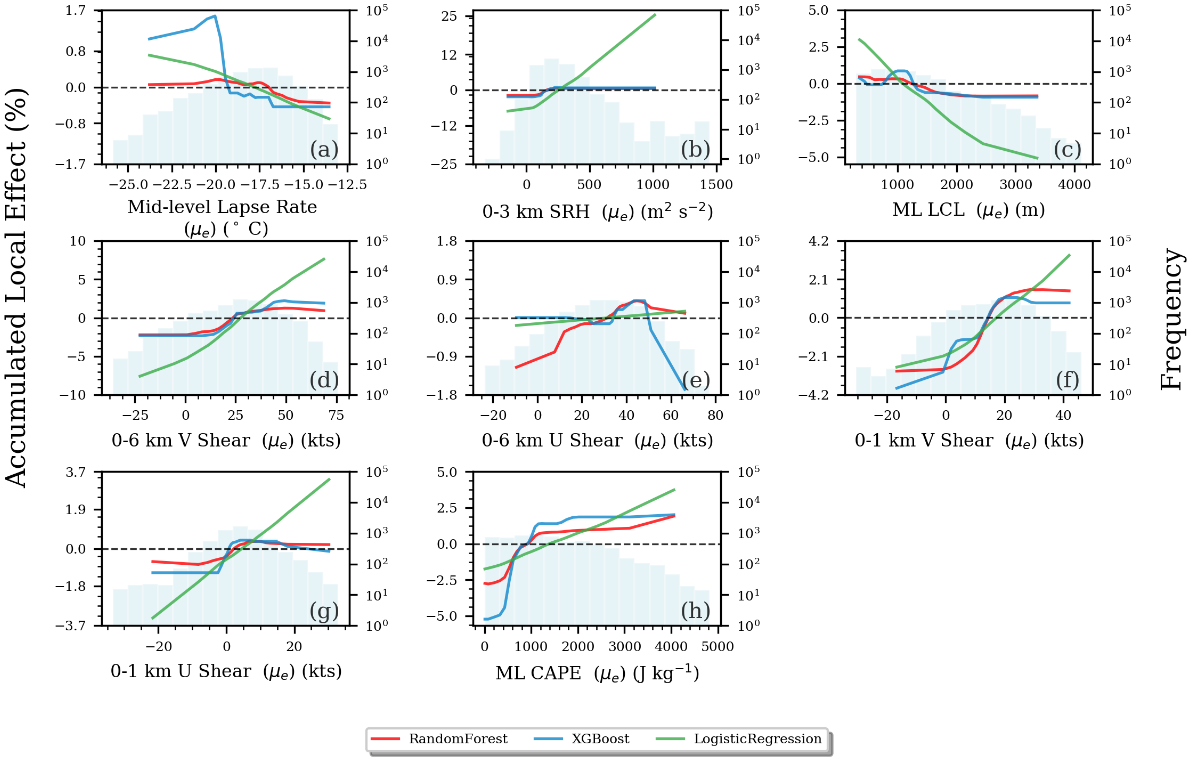


Figure 7.18: Same as in Figure 7.15, but for select environmental predictors

All three models correctly learned that increasing atmospheric instability (based on mid-level lapse rate and mixed-layer CAPE; Figure 7.18a,h) and low- and deep-layer wind shear (0-3 km SRH, 0-1 km and 0-6 km V-component of wind shear; Figure 7.18b,d,f) while lowering LCL heights increases tornado likelihood. The decision boundaries for mixed-layer CAPE, 0-3 km SRH, and LCL height are approximately 750 J kg$^{-1}$, 100-150 m$^2$ s$^{-2}$, and 1000-1250 m respectively, which are consistent with the thresholds used in the significant tornado parameter (Thompson et al. 2003). All three models have a stronger response to the v-component of the wind shear versus the u-component of the wind shear. This is not surprising as the typical synoptic-scale set-up for tornado potential in the Great Plains is a strong southerly flow from the Gulf Coast region while strong northerly wind shear is more typical for MCS development, which are not large producers of tornadoes.

The first-order effects for the top predictors are physically sound, but are the ML models correctly modeling predictor interactions as well? Figure 7.19 shows the purely 2D ALE (1D ALE effects have been removed) for the interaction between ensemble mean mixed-layer CAPE and ensemble mean 0-6 km v-component of wind shear for all three ML models. Overall, the interaction effects are quite weak with the average contributions less than 0.5%. However, there are $\frac{P!}{2!(P-1)!}$ (= 6328) possible second-order interactions between predictors and it is unknown what an appropriate magnitude of second-order effects ought to be for ML models with many predictors. The IAS for tornado prediction is comparatively low (Table 7.1), so we can assume that interactions in Figure 7.19 are likely modest. It is important to keep in mind that these are the expected contributions and therefore the effect is expected to be higher (or lower) in certain situations.

Though the interaction strength between CAPE and deep-layer shear is weak for tornado prediction, the overall patterns are consistent with our physical understanding and the training dataset. To interpret Figure 7.19, we must be mindful of the decomposition of the model's prediction into terms of increasing dimensionality (see equation 6.5) and that first-order effects have been removed. For example, Figure 7.19 indicates that lower values of CAPE and deep-layer wind shear interact together to increase the tornado likelihood. However, for all three models, the expected contribution based on the first-order ALE is negative for low values of both predictors (Figure 7.18f,d). In these situations where the environment is unfavorable to tornadogenesis, the first order effect for most predictors is negative and the sum total is a negative value. Since the final probability cannot be negative, the second-order effects are positive in response. As for when CAPE and deep-layer wind shear are both high, the overall second-order effect is positive for tree-based models, but negative for logistic regression. Here, the tree-based models are based on the local event frequency (as discussed for Figure 7.10), which is well below 100% for well-sampled regions (note the histograms in Figure 7.19a,b). Therefore, the second-order effect can contribute positively to the tornado likelihood, albeit rather unsubstantially. However, the first-order effects for logistic regression for high ensemble mean CAPE and deep-layer wind shear are much stronger, and therefore the second-order effect is negative.

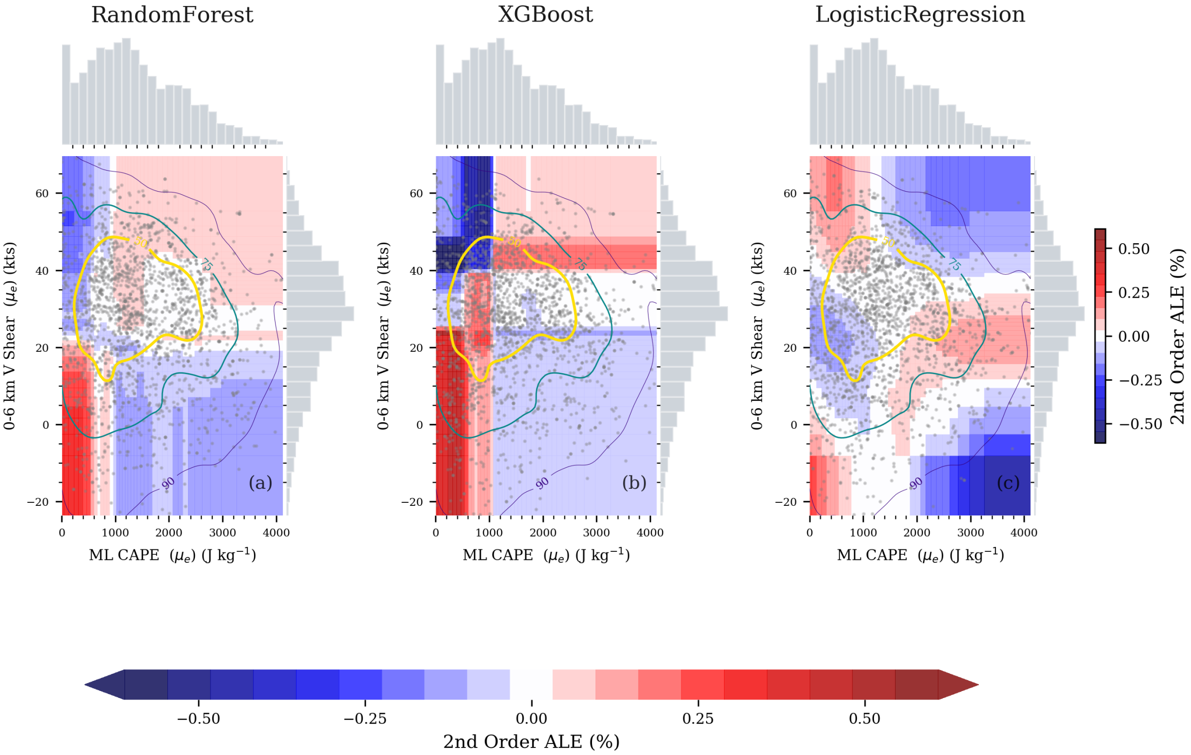


Figure 7.19: 2D Accumulated local effect (ALE) contours for interaction effects between ensemble mean mixed-layer CAPE and ensemble mean 0-6 km v-component of wind shear. A random 2000 points from the training dataset are shown as scatter points. Kernal density estimates for the whole training dataset are overlaid. contours are labelled by percentile of the data they capture. For example, 95 indicates that 95% of the examples fall with that contour. The marginal distribution for the mixed-layer CAPE and 0-6 km V-component of wind shear shown on the top and right axes, respectively.

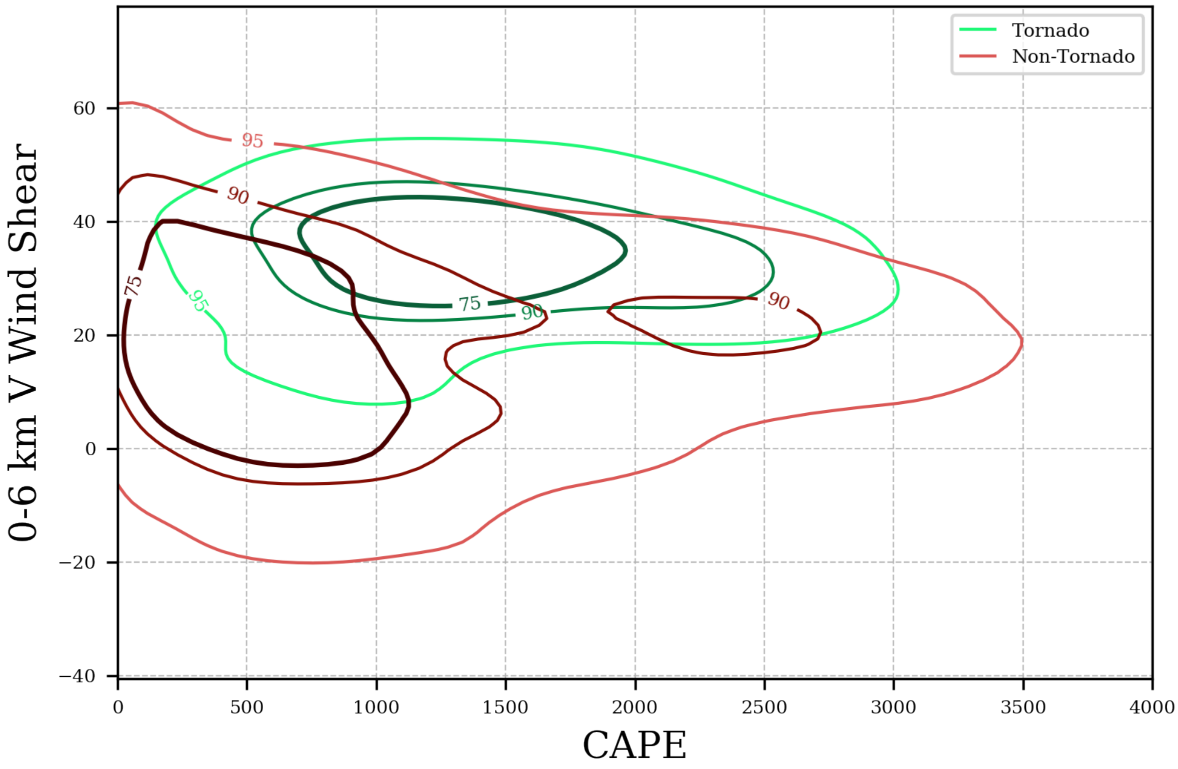


Figure 7.20: Kernel density estimate of tornado (red) and non-tornado (green) examples in the FIRST HOUR training dataset in ensemble mean CAPE - 0-6 km V-component of wind shear space. Contours are labelled by percentile of the data they capture. For example, 95 indicates that 95% of the examples fall with that contour.

The top predictors for severe hail were maximum predicted hail size, vertical velocity strength (e.g., column-minimum downdraft, column-maximum updraft), updraft helicity, and reflectivity (Figure 7.14a,b,c), while top predictors for severe wind were reflectivity, 80-m wind speed, hail size, mid-level lapse rate, and the low-level updraft (Figure 7.14g,h,i). For severe hail, the ALE patterns are similar to those for tornadoes, with increasing predicted hail size (Figure 7.16i) and increasing mid-level mesocyclone strength (Figure 7.16c,f), while lowering the ensemble spread in composite reflectivity (Figure 7.16g) increases the severe hail likelihood. The clear outlier is that logistic regression has negative slope ALE for updraft (Figure 7.16e). As discussed in Section 6.1, strong correlations can cause instability in the logistic regression coefficients. In this case, the peak updraft speed is highly correlated with the hail size (which logistic regression learned the correct sign for; Figure 7.16i). However, this sign error did not affect model performance as there are compensating effects (as we have seen for second-order interactions), but it muddles the interpretation.

Though the logistic regression ALE curve was comparable to the random forest and XGBoost for tornado prediction, it is noticeably larger for severe hail and wind prediction (cf. Figure 7.16, Figure 7.17, and Figure 7.15). This is related to the overconfidence logistic regression can have compared to the tree-based methods, which was shown in Figure 7.10 in Section  7.2.4. The interaction between CAPE and deep-layer shear is absent for the random forest severe hail prediction (Figure 7.21a), which may be surprising given that the IAS was higher for hail prediction (see Table 7.1). However, based on Figure 7.22, there is more overlap of severe hail and non-severe hail examples in CAPE-deep-layer shear space than for tornadoes, which is likely a strong contributor. It is possible that given the higher IAS for hail prediction that other predictors were more strongly linked, possibly reducing the second order effect in Figure 7.21a,b. Fully summarizing the degree of feature interactions is beyond the scope of this dissertation, but should be explored in future work.

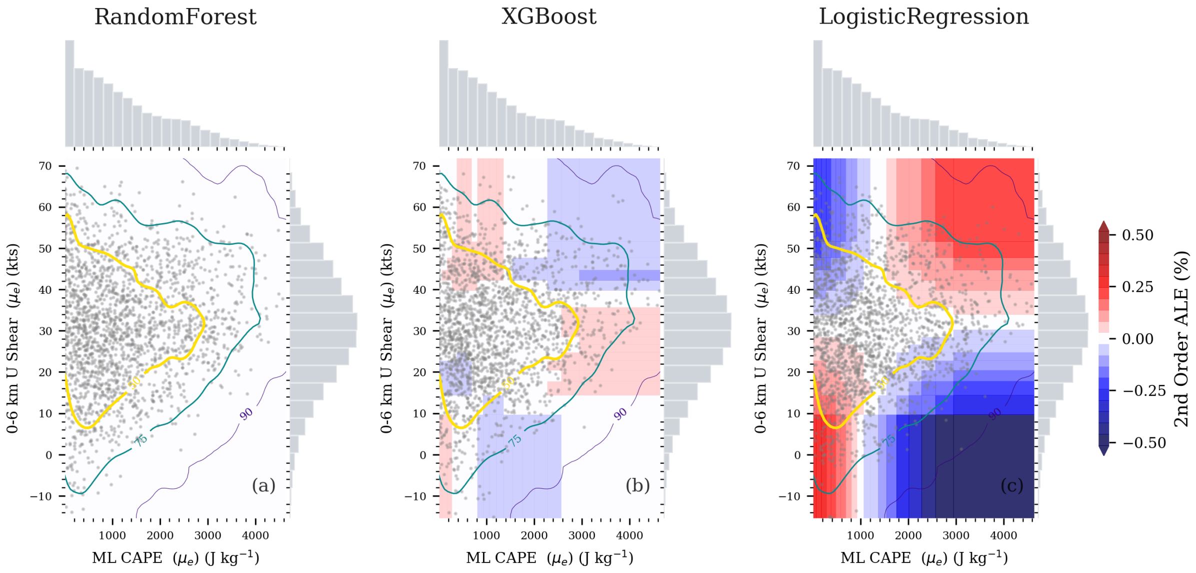


Figure 7.21: Same as in Figure 7.19, but for severe hail

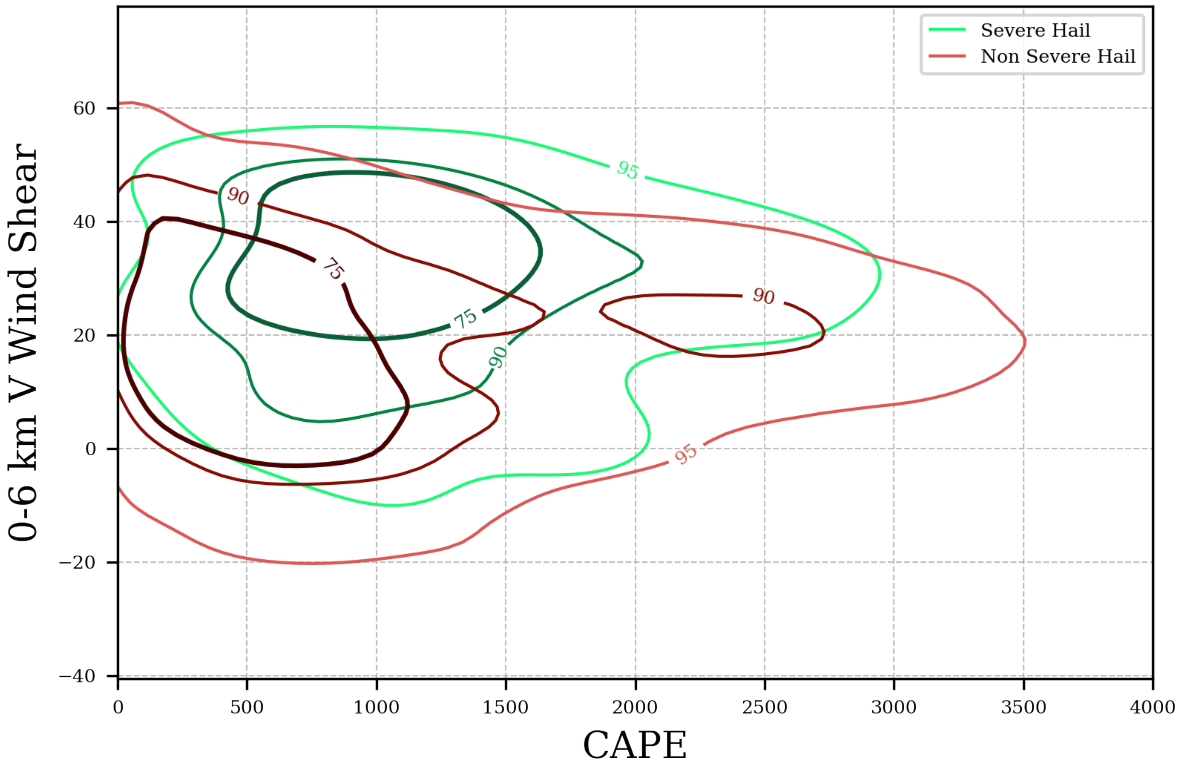


Figure 7.22: Same as in Figure 7.20, but for severe hail

For severe wind, in addition to similar patterns as for severe hail and tornadoes, the ML models found that decreasing mid-level instability (Figure 7.17g) and increasing the strength of the low-level updraft increased the severe wind likelihood (Figure 7.17d).

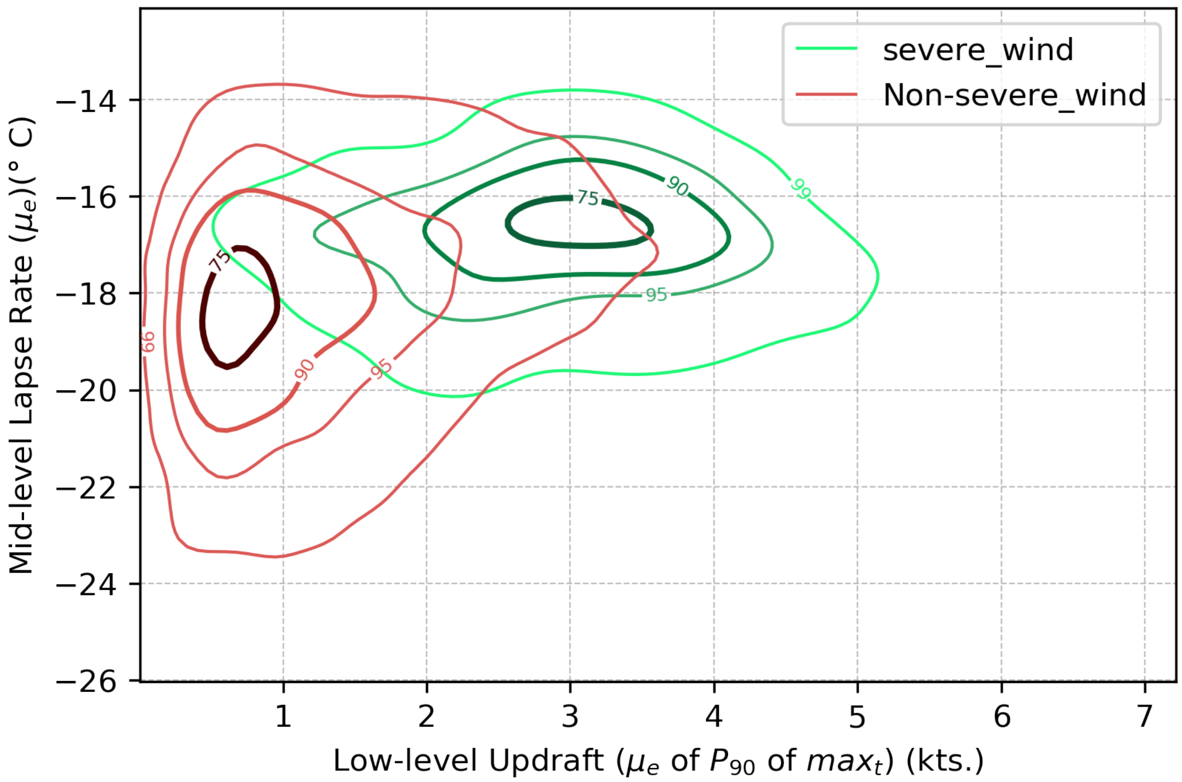


Figure 7.23: Same as in Figure 7.20, but for severe wind and plot is valid for low-level updraft versus mid-level lapse rate.

As we can see in Figure 7.23, there is a substantial separation between severe and non-severe wind examples based on the strength of the low-level updraft (computed in the lowest 1 km AGL), which we suspect is associated with the gust fronts from MCSs. Supercells can produce strong low-level updrafts (through the low-level mesocyclone) such that the storm inflow can produce damaging straight-line winds, but it is fairly uncommon. As for the mid-level lapse rates (the difference between 500 and 700 mb temperature), there is slight skew with greater values ( $<$ -20° C) being associated with non-severe wind. However, for a majority of examples in the training dataset (i.e., -16 to -20° C) the relationship between mid-level lapse rates and severe wind likelihood is negligible, which is consistent with Kuchera and Parker (2006). A recent study by Taszarek et al. (2020) also found that for severe wind, mid-level lapse rates were a poor discriminator (Figure 7.24). Though it

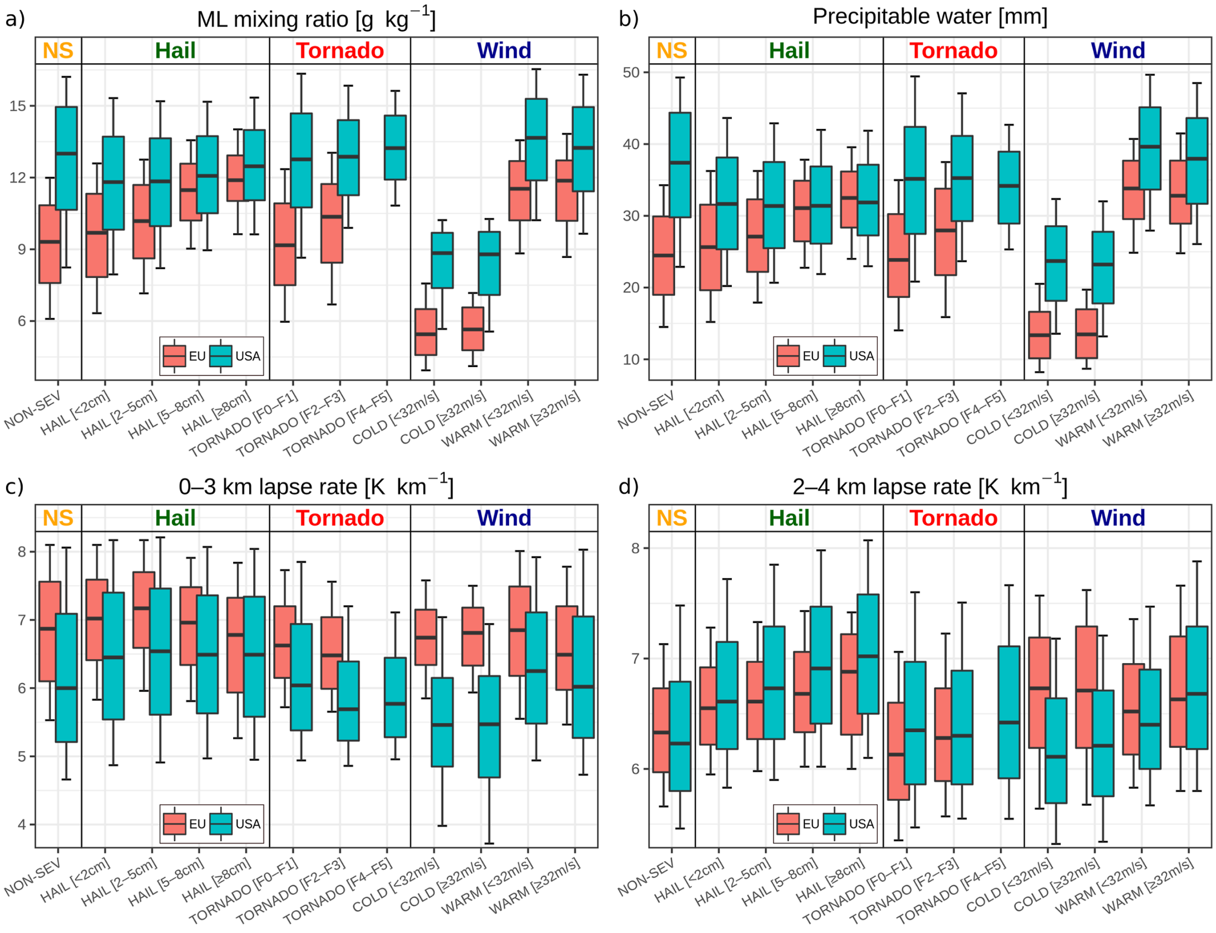


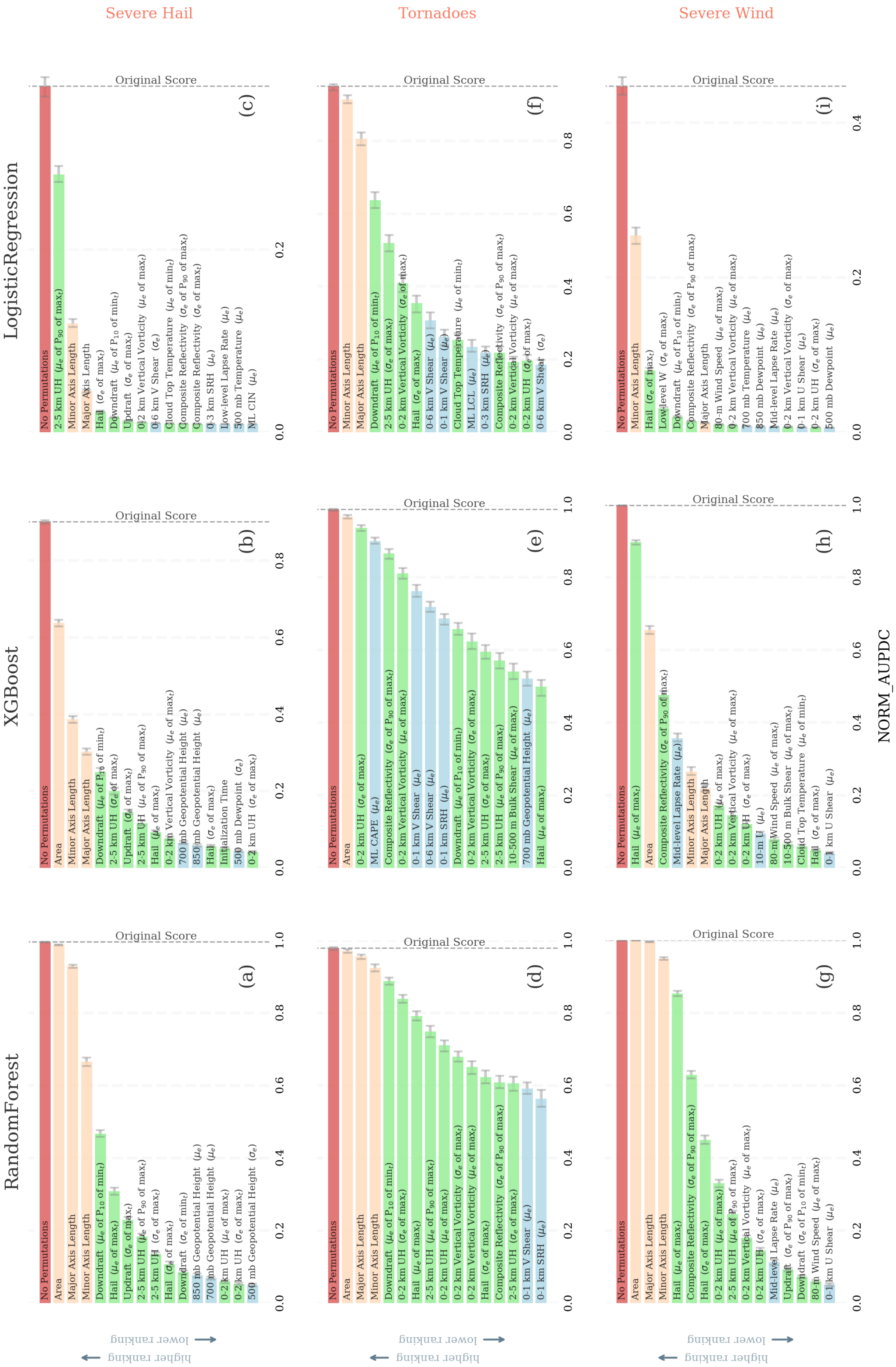
Figure 7.24: From Taszarek et al. (2020), their Figure 3. Box-and-whisker plots of (a) ML CAPE, (b) 03-km ML CAPE, (c) ML LCL, (d) ML LFC, (e) ML CIN, and (f) convective cloud depth (ML EL and ML LFC difference). The median is represented as a horizontal line inside the box, the edges of the box represent the 25th and 75th percentiles, and whiskers represent the 10th and 90th percentiles. Categories are defined as in Tables 1 and 2. Convective variables are derived from ERA5 proximity grid points.

is hypothesized that steeper mid-level lapse rates should be important for damaging winds (Wakimoto 1985), these results suggest that they are poorly correlated with severe wind reports. This discrepancy is possibly related to the noisy severe wind LSRs (Trapp et al. 2006), but unfortunately, no compelling evidence was discovered to explain this issue and warrants further exploration.

The results for the SECOND HOUR dataset can be summarized as follows:

- Though the rankings shift, the top predictors in the FIRST HOUR and SECOND HOUR datasets (Figure 7.26) were fairly consistent. Environmental predictors became more important while the storm morphology predictors became less important which is expected given storm-scale predictability limits. One noticeable difference is that intra-storm predictors became more important in the SECOND HOUR dataset for severe wind models.

- The ALE curves (Figure 7.26- 7.29) have the same orientation as for the FIRST HOUR dataset, but the effects were smaller in magnitude as expected because of storm-scale predictability limit precluding having a greater confidence at later lead times. The decision thresholds, especially for environmental predictors (Figure 7.29), remained fairly unchanged in the SECOND HOUR dataset.



122

Figure 7.25: Same as in Figure 7.14, but the SECOND HOUR dataset

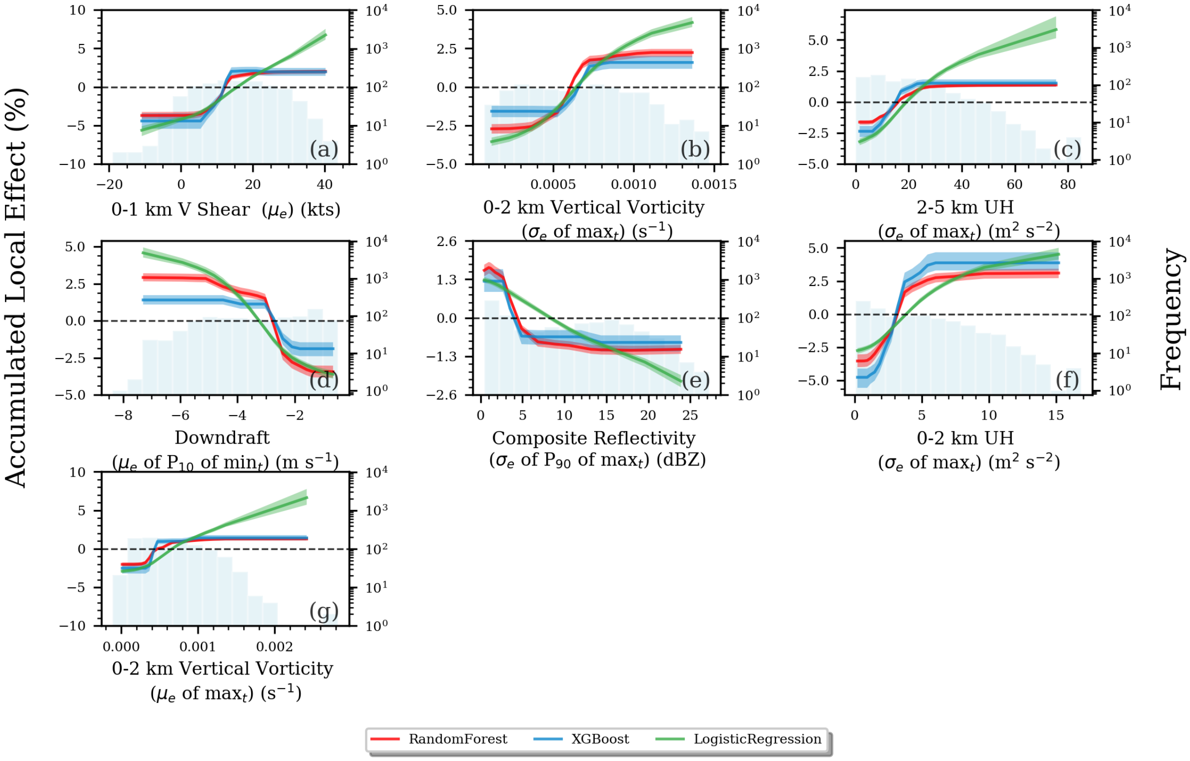


Figure 7.26: Same as Figure 7.15, but SECOND HOUR tornado prediction.

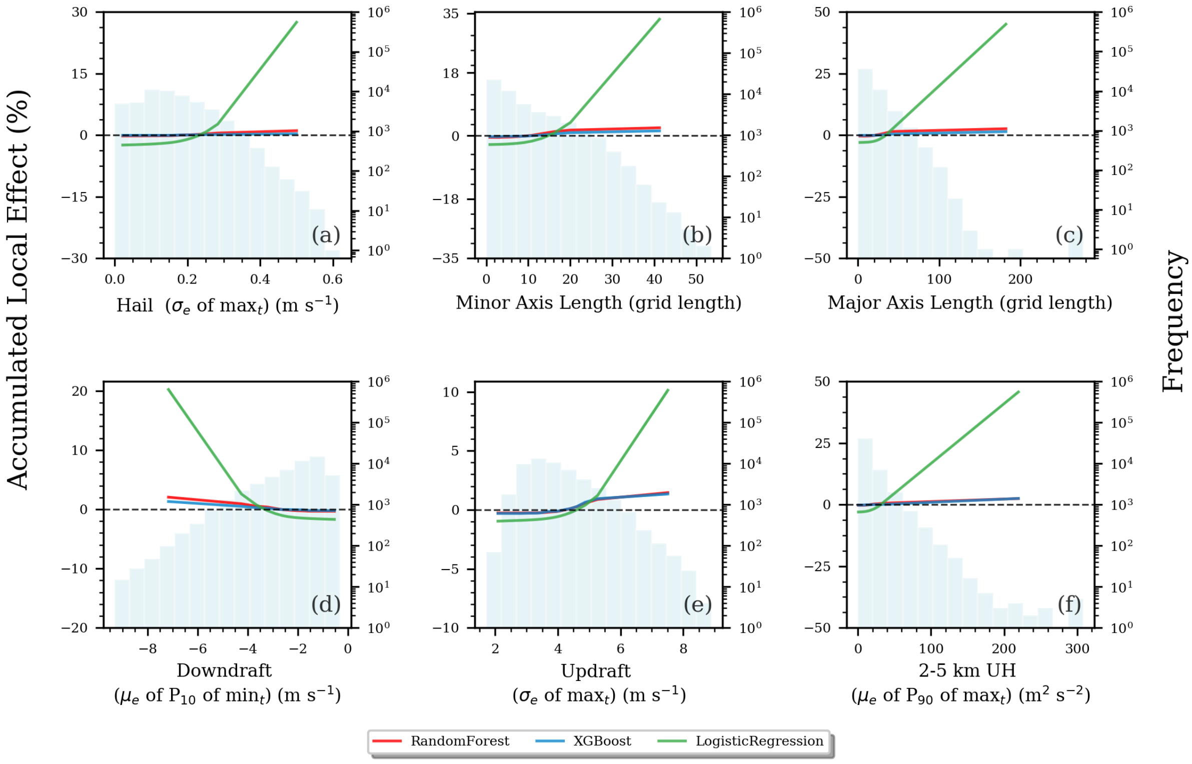


Figure 7.27: Same as Figure 7.15, but SECOND HOUR severe hail prediction.

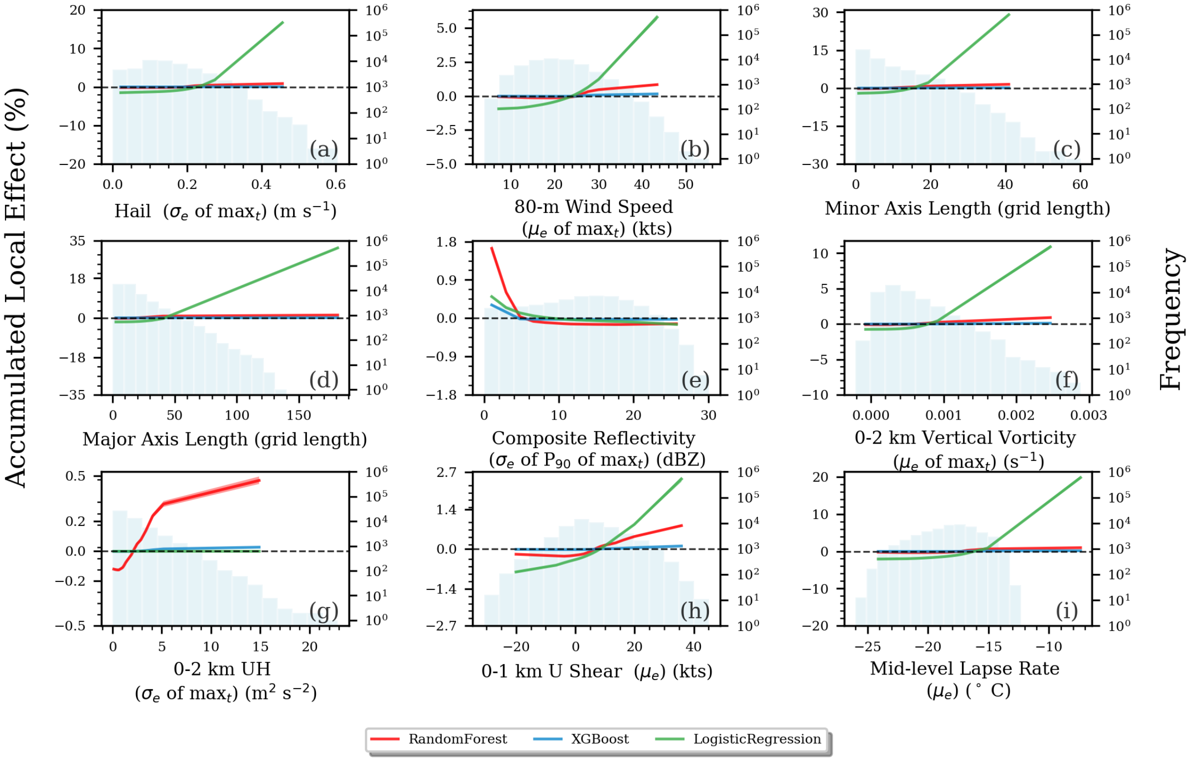


Figure 7.28: Same as Figure 7.15, but SECOND HOUR severe wind prediction.

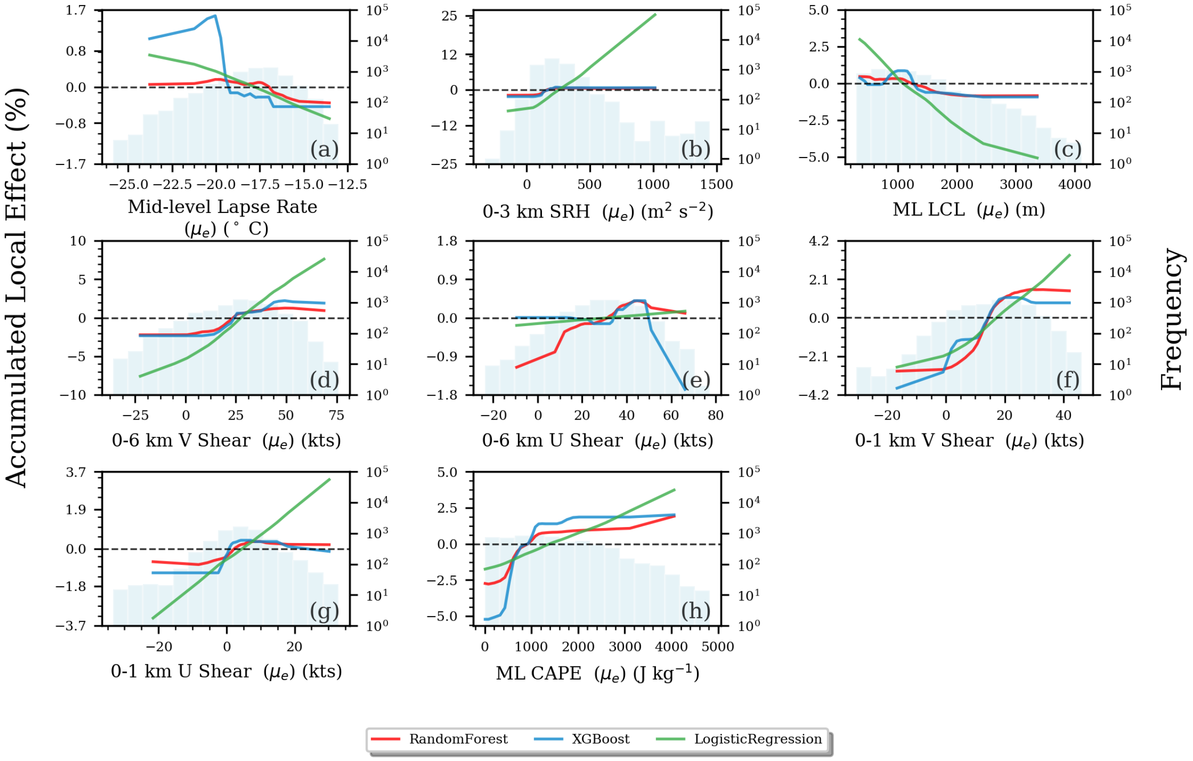


Figure 7.29: Same as in Figure 7.15, but for select environmental predictors from the models trained on the SECOND HOUR dataset.

### 7.3.2 Predictor Contributions

Global approaches to ML model interpretability, such as predictor importance and expected predictor contributions, were shown in the previous section. In this section, I adopt a more local approach to analyze predictor contributions based on model performance. For example, how does a predictor's contribution vary from high confidence forecasts matched to an event versus not matched to an event (e.g., a false alarm)?

Figure 7.30 shows the average predictor contributions based on forecast performance for severe hail prediction for all three ML models in the FIRST HOUR dataset.

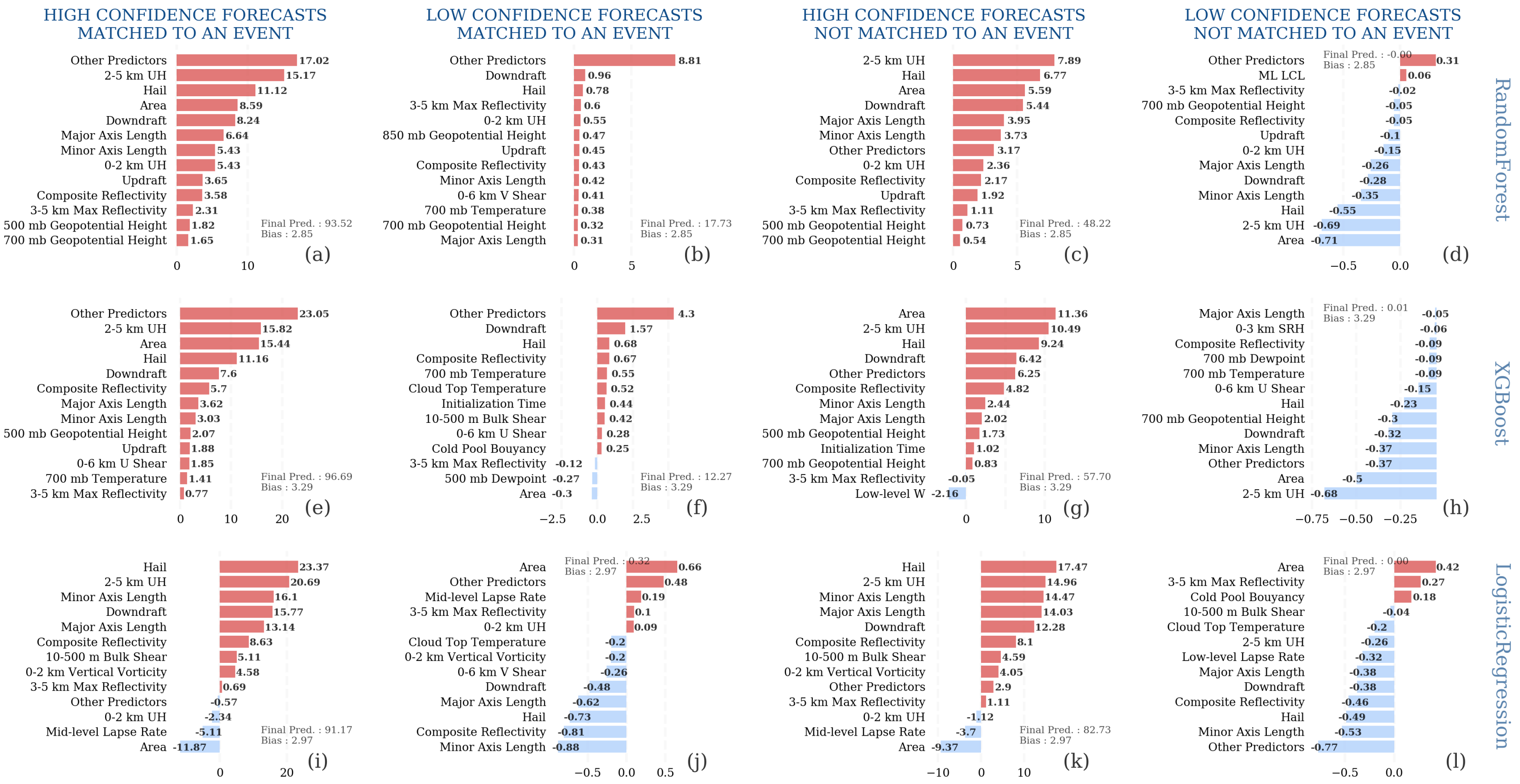

HIGH CONFIDENCE FORECASTS
MATCHED TO AN EVENT
LOW CONFIDENCE FORECASTS
MATCHED TO AN EVENT
HIGH CONFIDENCE FORECASTS
NOT MATCHED TO AN EVENT
LOW CONFIDENCE FORECASTS
NOT MATCHED TO AN EVENT
RandomForest
XGBoost
LogisticRegression
(a)
Other Predictors 17.02
2-5 km UH 15.17
Hail 11.12
Area 8.59
Downdraft 8.24
Major Axis Length 6.64
Minor Axis Length 5.43
0-2 km UH 5.43
Updraft 3.65
Composite Reflectivity 3.58
3-5 km Max Reflectivity 2.31
500 mb Geopotential Height 1.82
700 mb Geopotential Height 1.65
Final Pred. : 93.52
Bias : 2.85
0
10
(b)
Other Predictors 8.81
Downdraft 0.96
Hail 0.78
3-5 km Max Reflectivity 0.6
0-2 km UH 0.55
850 mb Geopotential Height 0.47
Updraft 0.45
Composite Reflectivity 0.43
Minor Axis Length 0.42
0-6 km V Shear 0.41
700 mb Temperature 0.38
700 mb Geopotential Height 0.32
Major Axis Length 0.31
Final Pred. : 17.73
Bias : 2.85
0
5
(c)
2-5 km UH 7.89
Hail 6.77
Area 5.59
Downdraft 5.44
Major Axis Length 3.95
Minor Axis Length 3.73
Other Predictors 3.17
0-2 km UH 2.36
Composite Reflectivity 2.17
Updraft 1.92
3-5 km Max Reflectivity 1.11
500 mb Geopotential Height 0.73
700 mb Geopotential Height 0.54
Final Pred. : 48.22
Bias : 2.85
0
5
(d)
Other Predictors 0.31
ML LCL 0.06
3-5 km Max Reflectivity -0.02
700 mb Geopotential Height -0.05
Composite Reflectivity -0.05
Updraft -0.1
0-2 km UH -0.15
Major Axis Length -0.26
Downdraft -0.28
Minor Axis Length -0.35
Hail -0.55
2-5 km UH -0.69
Area -0.71
Final Pred. : -0.00
Bias : 2.85
−0.5
0.0
(e)
Other Predictors 23.05
2-5 km UH 15.82
Area 15.44
Hail 11.16
Downdraft 7.6
Composite Reflectivity 5.7
Major Axis Length 3.62
Minor Axis Length 3.03
500 mb Geopotential Height 2.07
Updraft 1.88
0-6 km U Shear 1.85
700 mb Temperature 1.41
3-5 km Max Reflectivity 0.77
Final Pred. : 96.69
Bias : 3.29
0
10
20
(f)
Other Predictors 4.3
Downdraft 1.57
Hail 0.68
Composite Reflectivity 0.67
700 mb Temperature 0.55
Cloud Top Temperature 0.52
Initialization Time 0.44
10-500 m Bulk Shear 0.42
0-6 km U Shear 0.28
Cold Pool Bouyancy 0.25
3-5 km Max Reflectivity -0.12
500 mb Dewpoint -0.27
Area -0.3
Final Pred. : 12.27
Bias : 3.29
−2.5
0.0
2.5
(g)
Area 11.36
2-5 km UH 10.49
Hail 9.24
Downdraft 6.42
Other Predictors 6.25
Composite Reflectivity 4.82
Minor Axis Length 2.44
Major Axis Length 2.02
500 mb Geopotential Height 1.73
Initialization Time 1.02
700 mb Geopotential Height 0.83
3-5 km Max Reflectivity -0.05
Low-level W -2.16
Final Pred. : 57.70
Bias : 3.29
0
10
(h)
Major Axis Length -0.05
0-3 km SRH -0.06
Composite Reflectivity -0.09
700 mb Dewpoint -0.09
700 mb Temperature -0.09
0-6 km U Shear -0.15
Hail -0.23
700 mb Geopotential Height -0.3
Downdraft -0.32
Minor Axis Length -0.37
Other Predictors -0.37
Area -0.5
2-5 km UH -0.68
Final Pred. : 0.01
Bias : 3.29
−0.75
−0.50
−0.25
(i)
Hail 23.37
2-5 km UH 20.69
Minor Axis Length 16.1
Downdraft 15.77
Major Axis Length 13.14
Composite Reflectivity 8.63
10-500 m Bulk Shear 5.11
0-2 km Vertical Vorticity 4.58
3-5 km Max Reflectivity 0.69
Other Predictors -0.57
0-2 km UH -2.34
Mid-level Lapse Rate -5.11
Area -11.87
Final Pred. : 91.17
Bias : 2.97
0
20
(j)
Area 0.66
Other Predictors 0.48
Mid-level Lapse Rate 0.19
3-5 km Max Reflectivity 0.1
0-2 km UH 0.09
Cloud Top Temperature -0.2
0-2 km Vertical Vorticity -0.2
0-6 km V Shear -0.26
Downdraft -0.48
Major Axis Length -0.62
Hail -0.73
Composite Reflectivity -0.81
Minor Axis Length -0.88
Final Pred. : 0.32
Bias : 2.97
−0.5
0.0
0.5
(k)
Hail 17.47
2-5 km UH 14.96
Minor Axis Length 14.47
Major Axis Length 14.03
Downdraft 12.28
Composite Reflectivity 8.1
10-500 m Bulk Shear 4.59
0-2 km Vertical Vorticity 4.05
Other Predictors 2.9
3-5 km Max Reflectivity 1.11
0-2 km UH -1.12
Mid-level Lapse Rate -3.7
Area -9.37
Final Pred. : 82.73
Bias : 2.97
−10
0
10
(l)
Area 0.42
3-5 km Max Reflectivity 0.27
Cold Pool Bouyancy 0.18
10-500 m Bulk Shear -0.04
Cloud Top Temperature -0.2
2-5 km UH -0.26
Low-level Lapse Rate -0.32
Major Axis Length -0.38
Downdraft -0.38
Composite Reflectivity -0.46
Hail -0.49
Minor Axis Length -0.53
Other Predictors -0.77
Final Pred. : 0.00
Bias : 2.97
−0.5
0.0

Figure 7.30: Top predictor contributions averaged over different sets of 250 training examples, respectively, of (first column) high confidence forecasts matched to an event (second column) low confidence forecasts matched to an event (third column) high confidence forecast not matched to an event, and (last column) low confidence forecasts not matched to an event. The results are valid for (top row) random forest, (middle row) XGBoost, and (last row) logistic regression models predicting severe hail likelihood. The average base rate prediction (known as the bias) and average final prediction are provided in each panel. **Other Predictors** indicates the sum total of remaining predictor contributions not explicitly stated.

To ease interpretation, I sum together contributions for a base variable. For example, the contributions from all predictors containing 2-5 km UH are added together. For confident forecasts matched to an event (Figure 7.30a,e,i), the top contributors include updraft helicity, hail, storm morphology, and downdraft for all three models, which roughly agrees with the permutation importance (Figure 7.14a,b,c). For the random forest and XGBoost, the largest contributor was the sum total of minor contributions from less important predictors. These predictors were not insignificant, as removing them resulted in a substantial decrease in model performance (see Appendix C). The random forest and XGBoost, however, are likely over-fitting some of these additional predictors as the logistic regression, which has similar model performance, largely relies on the WoFS-predicted hail and 2-5 km UH with minimal contribution from the “Other Predictors”.

For confident forecasts not matched to a severe hail report (e.g., false alarms; Figure 7.30c,g,k), the overall contributions are similar to the confident forecasts matched to a severe hail report (Figure 7.30a,e,i), which is a similar result found for the other two hazards (Figure 7.31c,g,k and Figure 7.32c,g,k). There a couple reasons why predictor contributions can be similar for hits and false alarms. First, for a well-calibrated system, some portion of examples with high forecast probabilities ought to be associated with non-events. These examples can represent situations where multiple elements of the environment are favorable for severe weather (high CAPE, strong deep-layer shear, abundant low-level moisture), but where one or more key factors can introduce uncertainty (a modest capping inversion) and ultimately result in an non-event. Second, as discussed in section 7.2.2, WoFS may not discriminate between tornadic and non-tornadic storms occurring in proximity to one another, which can also be the case for other hazards. Lastly, some of these apparent forecast busts may in fact be associated with an unreported event.

As a sanity check, all three models can correctly identify non-events (Figure 7.30d,h,l; Figure 7.31d,h,l, and Figure 7.32d,h,l ). Being able to reliably determine which storms will be non-severe is a useful result for forecasters. However, more work is required to determine if the ML models can discriminate well in environments that are conditional severe and if the discrimination ability between severe and non-severe storms is more related to the WoFS forecasts than the ML models themselves.

As for low confident forecasts matched to an event (e.g., missed event; Figure 7.31b,f,j,), the results were model- and hazard-dependent. For example, though the environmental predictors were found to be less important than the storm predictors (see Figure 7.14), the random forest and XGBoost models do have positive contributions from the environmental predictors for low confidence forecasts matched to a tornado (e.g., missed events; Figure 7.31b,f,j). These examples likely represent when the WoFS struggles to analyze ongoing convection. In these cases, some environmental predictors make positive contributions to the tornado likelihood, but given the poorly forecasted storm properties, the probabilities ultimately remain low. When presented with this information, human forecasters could account for situations where the WoFS fails to analyze ongoing convection and mentally increase the tornado likelihood. For severe hail and tornado, the top contributor for the random forest and XGBoost models for these "missed events" is also the sum total of small contributions from less important predictors resulting in higher forecast probabilities as opposed to the logistic regression model, which is interpreted as a sign of over-fitting. Additional predictors from the WoFS can extract useful forecast information, but can also lead to over-fitting the training dataset.

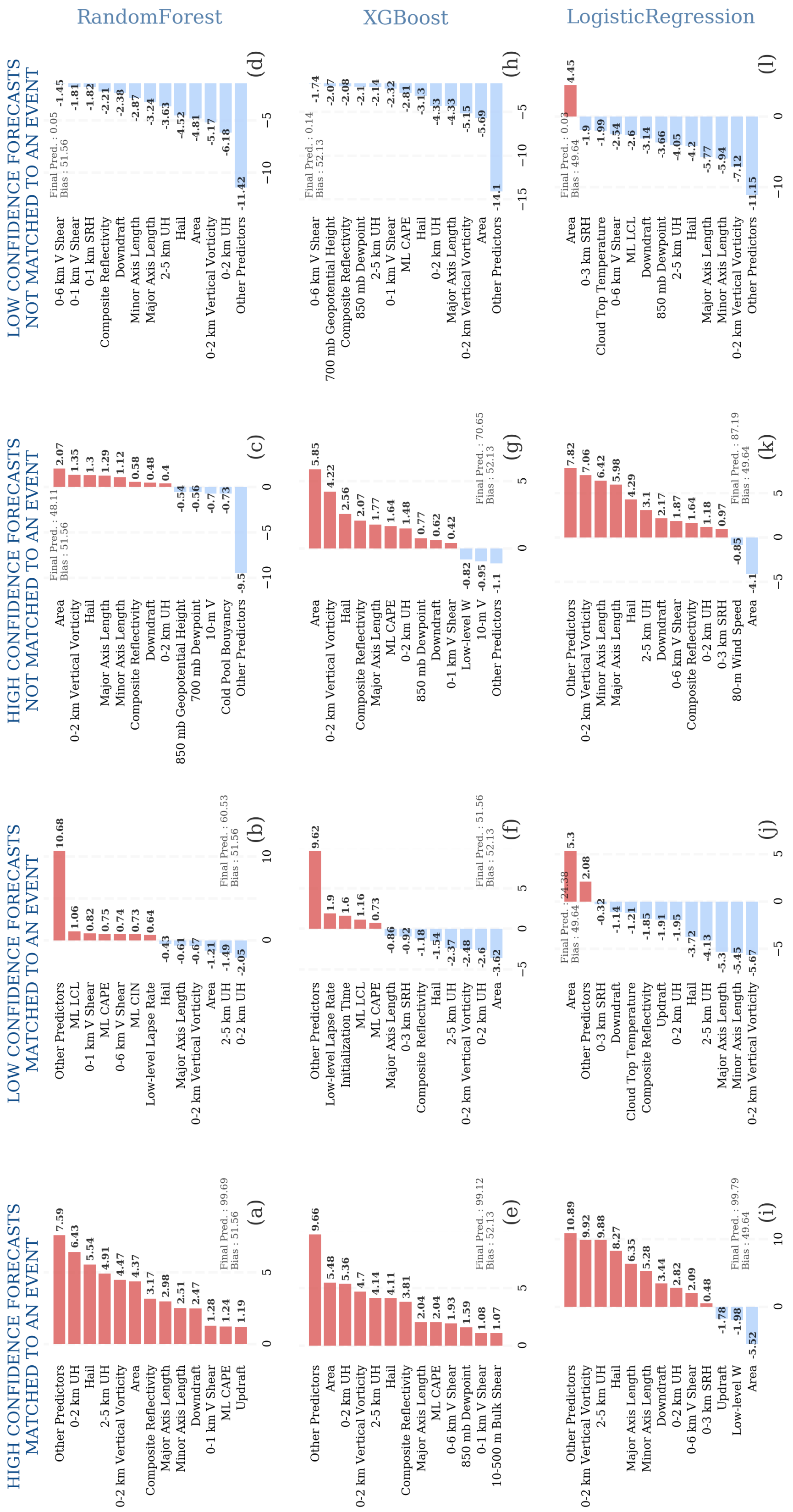


Figure 7.31: Same as Figure 7.30, but for FIRST HOUR tornado prediction.

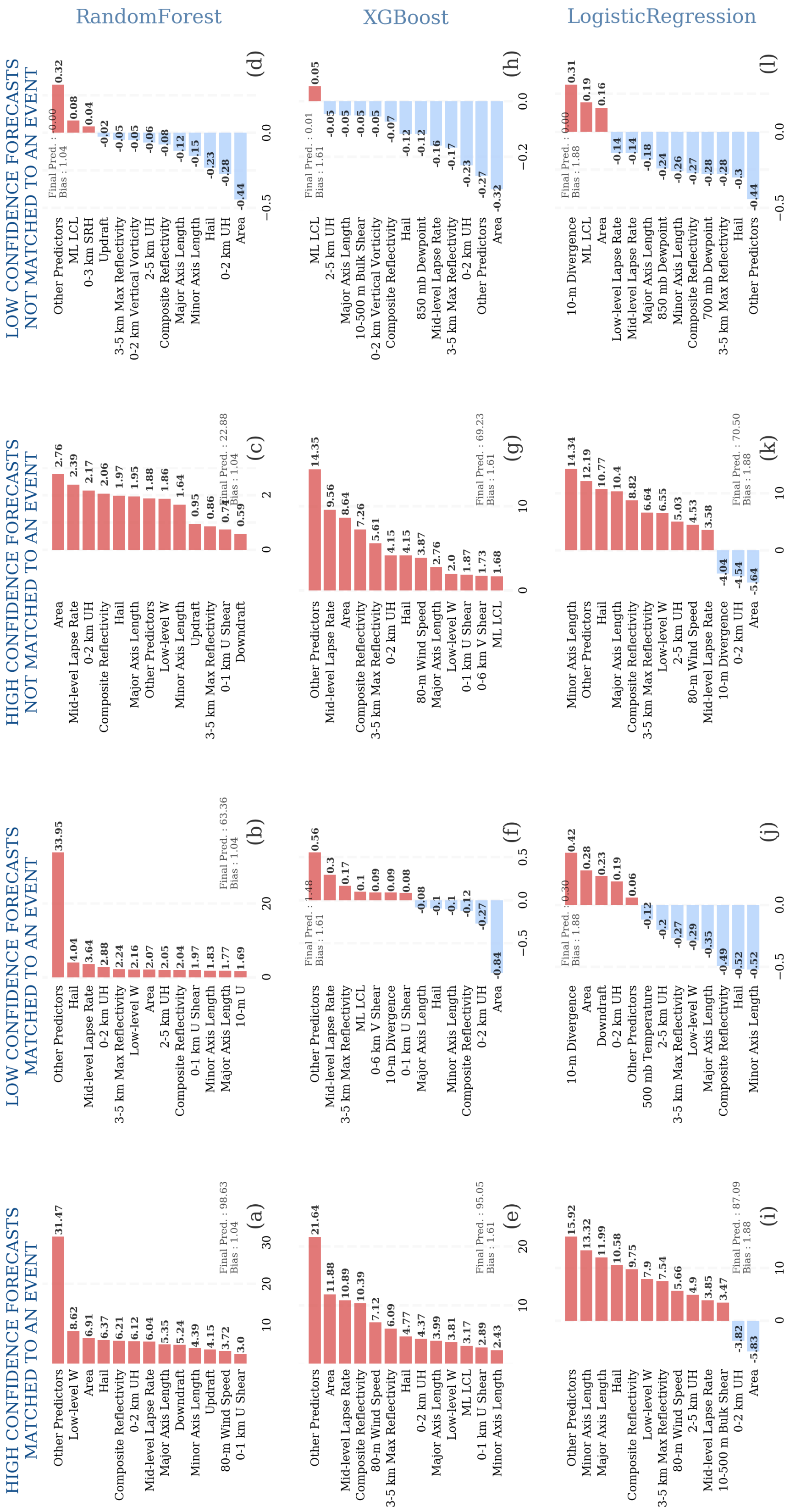


Figure 7.32: Same as Figure 7.30, but for FIRST HOUR severe wind prediction.

In addition to comparing the average predictor contribution based on model performance, I can stratify the examples by environmental parameters. Figure 7.33 shows the average predictor contribution (based on model performance) for training examples in environments with a significant tornado parameter (STP) greater than 1. The top contributors (and the magnitude of their contributions) are fairly similar between Figure 7.33a,e,i and Figure 7.31a,e,i. This is not surprising as best "hits" and/or worst "false alarms" are likely to be in highly favourable tornadic environments. However, the contributions from the environmental predictors for "misses" (Figure 7.33b,f,j) become larger, which aligns with the first-order ALE from Figure 7.18. This further supports the claim that forecasters may be able to mentally account for situations where the WoFS fails to analyze ongoing convection and increase the tornado likelihood if provided with this information.

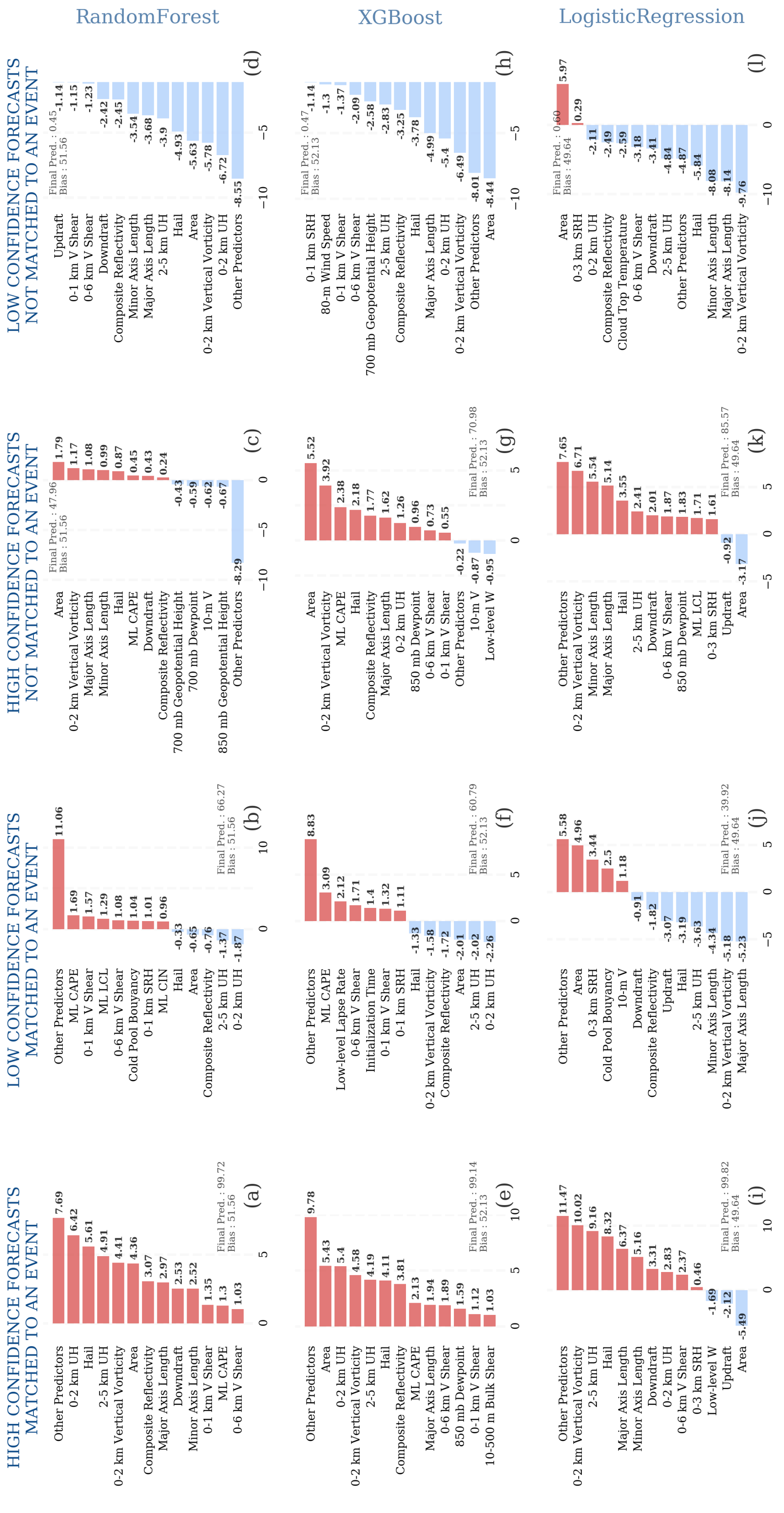


Figure 7.33: Same as Figure 7.30, but for examples in the FIRST HOUR tornado prediction dataset with an significant tornado parameter > 1 (Thompson et al. 2003).

# Chapter 8: Conclusions, Limitations, and Future Work

A fundamental goal of the WoF project is to provide probabilistic guidance of severe weather hazards associated with individual thunderstorms. This dissertation developed a novel probability object verification framework and extends upon Skinner et al. (2018) by verifying the accuracy and reliability of WoFS hour-long probabilistic mesocyclone track forecasts. As grid-based verification showed, the WoFS probabilistic mesocyclone guidance on the native 3-km grid greatly over-predicts the likelihood of a mesocyclone impacting a particular point. This over-prediction bias indicates considerable underdispersion in the WoFS. It is possible to improve the grid-based reliability by upscaling the forecasts and observations, but doing so obscures probabilities associated with individual storms.

Despite the over-prediction bias, WoFS probabilistic guidance on the native 3-km grid has been found to be useful in operational settings (Wilson et al. 2019). For example, Choate et al. (2019) found that paintball plots, which show the separate rotation tracks for all ensemble members on a single figure, were, by far, the most commonly used products in the SFE. These differences between grid-based verification metrics and forecaster usage have motivated the development of a novel, complementary verification method for evaluating short-term, storm-scale probabilistic guidance. The verification method used in this dissertation uses an object-based framework where probability swaths associated with individual storms are treated as forecast objects and prescribed a single, representative probability. This approach tolerates spatial differences between forecasts and observations by defining a user-specified matching distance. Importantly, unlike in the grid-based framework, the forecast probabilities are not smoothed or upscaled, which preserves forecast likelihood of mesocyclones occurring within individual thunderstorms. Lastly, this verification method was designed with the human forecast decision model for WoFS probabilistic guidance in mind and is intended to match the expected forecaster usage of probability swaths (e.g., Wilson et al. 2019). The primary findings from applying the object-based verification technique to WoFS probabilistic mesocyclone guidance forecasts for 63 cases during 2017 and

2018 are as follows:

- The highest skill, in terms of CSI, of the WoFS mesocyclone probabilistic guidance was approximately associated with a probability threshold of 22.2% (4/18).

- The highest skill in the 0-60 minute forecast period for low-level UH probabilistic forecasts had a POD, SR, and CSI of 0.47, 0.46, and 0.31, respectively. In the 90-150 minute forecast period, the POD and CSI dropped to 0.39 and 0.27 while SR remained relatively unchanged.

- WoFS probabilistic low-level mesocyclone guidance is reliable for forecast probabilities <60% at all forecast lead times using a 0-km matching neighborhood size, but an overprediction of mesocyclone likelihood is present at probability values >60%.

- Mid-level and low-level probabilistic mesocyclone forecasts had similar contingency table metrics, reliability, and centroid displacement of matched pairs.

- The highest concentrations of centroid displacements (as indicated by KDE contours greater than the 99.9th percentile) in matched objects remained under 30 km (which is the approximate size of the NWS warning polygon) up to lead times of 90-150 min.

Though WoFS guidance could skillfully and reliably predict observed mesocyclones, the guidance was not calibrated for the separate severe weather hazards. An emerging approach to solving this problem are ML models, which can easily incorporate many predictors, are well-suited for complex, noisy datasets, and have been shown to produce calibrated, skillful probabilistic guidance for a variety of meteorological phenomena.

In this dissertation, gradient-boosted classification trees, random forests, and logistic regression models were trained on WoFS forecasts from the 2017-2019 HWT-SFEs to predict which 30-min forecast storm tracks in the WoFS domain will produce a tornado, severe hail, and/or severe wind report up to lead times of 150 min. The ensemble storm track identification method was used to extract ensemble statistics of intra-storm and environmental parameters. The ensemble storm tracks were labeled based on local storm reports, which, while error prone, are the best available severe weather database for individual hazards. The ML predictions were compared against the probability of mid-level UH exceeding a threshold that was tuned for each severe weather hazard. The primary conclusions of that work are as follows:

- The ML models produced substantially higher maximum Normalized Critical Success Index (NCSIs) and normalized area under the performance diagram than the UH baselines, especially at later lead times. This result is especially encouraging since observation-based severe weather prediction methods rapidly degrade beyond nowcasting lead times.

- The ML models produced markedly more reliable predictions than the UH baselines, which were unreliable and produced negative BSS scores.

- The ML models discriminated well (AUCs $> 0.9$) for all three severe weather hazards up to a lead time of 150 min.

- For a given severe weather hazard, the contingency table metrics for the three ML algorithms were fairly similar. The severe hail predictions had the highest NCSI while tornado predictions had the lowest NCSI, especially at later lead times.

- Severe hail and wind predictions were more reliable than tornado predictions at all lead times. All three models produced fairly reliable hail and wind probabilities up to 50% while hail (wind) forecasts were under-confident (overconfident) for higher probabilities. At later lead times, severe hail forecast probabilities were reliable up to 60% while severe wind forecast probabilities became more overconfident.

Besides evaluating the ML performance, this dissertation explored a suite of state-of-the-art ML interpretability methods. Using these methods we can gain a global perspective of the relationships learned by ML models and even explain individual predictions, which should reduce the concern that ML models are “black boxes.” Being able to explain individual predictions should help build human forecasters’ trust in ML predictions and maximize the use of automated guidance. The primary conclusions of the interpretability work are:

- The models learned physically sound relationships for the respective severe weather hazards. In addition, the models learned appropriate responses to the ensemble statistics.

- The top predictors were fairly consistent for the different models respective to the severe weather hazard. The ML models trained to predict severe wind and hail relied on a few predictors for overall model performance, while tornado prediction relied on much more.

- Intra-storm predictors were generally found to be more important than environmental predictors. The greater importance of the intra-storm predictors is not surprising, as the lead times used in this dissertation are short enough such that the useful storm-scale information in the WoFS forecasts has not been limited by storm-scale predictability limits.

- Though intra-storm predictors were overall more important to the ML models, the environmental predictors were found to make modest positive contributions for low confidence forecasts not matched to events (missed events). If presented with real-time visualizations of the predictor contributions, forecasters may mentally account for this and correct for situations when the WoFS poorly analyzes or forecasts ongoing convection.

The object-based framework I developed herein can be adapted to evaluate the performance and reliability of any severe weather hazards (or other phenomena such as tropical cyclones or heavy rainfall events) and changes in performance across different WoFS system configurations. In future work, it will be important to distinguish between the skill and reliability of probabilistic rotation forecasts in MCSs versus supercells. I expect mesocyclone forecasts will be more skillful for discrete supercells in a favorable environment than for rotation associated with MCSs (e.g., S18). The processes related to supercell mesocyclones are sufficiently resolved on a 3-km grid (Potvin and Flora 2015) while the intricate processes associated with rotation in MCSs may not be. It is also important to explore the impact of timing errors on the performance and reliability of WoFS mesocyclone guidance. In future work, 15- or 30-min probability swath objects could be used to explore the impact of timing errors.

Other techniques beyond simple object-based verification should be explored in future work. No single verification method adequately describes the unique attributes of forecast performance, and it is crucial to develop complementary verification measures. For example, in a WoF framework, Skinner et al. (2016) explored multiple verification techniques of deterministic forecasts of low-level mesocyclones. Although the object-based methods were favored in that dissertation, more work exploring different spatial verification methods is warranted. There are also promising new techniques such as ensemble structure-amplitude-location (eSAL; Radanovics et al. 2018) or verification that leverages information theory (Lawson et al. 2018b) which could be suited for short term, storm-scale probabilistic guidance.

There are limitations of the current method, which will need to be improved upon in future iterations. First, I am using imperfect observation data, coupled with an imperfect object identification method. Though extensive efforts were made to tune the object identification algorithms used in this dissertation, the number of objects identified is sensitive to the scale of the phenomena to be identified. Observed rotation tracks and probability swaths, especially when considering different storm modes, can span a wide spectrum of spatial scales. Thus, it is difficult to find universal parameter settings for any object identification algorithm that covers all relevant scales in this problem. This limitation, however, might be mitigated by improving observations of mesocyclones and accurately categorizing storm mode in simulated and observed reflectivity. It will also be possible to mitigate limitations in the object identification method at higher resolution where discriminating between intense and weak rotation is improved.

While these ML-calibration results are promising, there are some limitations to this dissertation that should be considered. First, since I am operating in an event-based framework, I am not correcting for instances when the WoFS fails to accurately analyze ongoing convection or exhibits biases in storm location. In future studies, I plan to adopt a hybrid gridpoint-based/event-based framework that, for near missed storms, produces a complementary forecast that is largely based on environmental parameters. Second, the labeling of ensemble storm tracks was based on whether they contain a local storm report. I showed that because of small spatial errors in forecast storm tracks, reports may fall just outside the boundary of an ensemble storm track. Given these near-misses, and the spurious false alarms arising from missing storm reports, the verification results likely underestimate the potential ML skill. Third, I did not evaluate the ML models for different geographic regions (e.g., Gagne et al. 2014; Herman and Schumacher 2018; Sobash et al. 2020), diurnal times, or initialization time. The data in this dissertation were largely sampled from the Great Plains (Figure 3.1) so it is important to assess the ML model performance in other regions. In future work, I plan to expand upon the verification of the ML predictions to highlight any potential failure modes.

There are additional potential extensions of this work. First, though the ML predictions outperformed a competitive baseline, they were not compared against any preexisting method for predicting severe weather hazards (e.g., ProbSevere; Cintineo et al. 2014, 2018) nor were they compared against a more hazard-specific baseline like WRF-HAILCAST (Adams-Selin and Ziegler 2016; Adams-Selin et al. 2019) for severe hail or model low-level wind gusts for severe wind. To further assess the potential operational value of our prediction algorithms, and to increase forecaster trust in the algorithms, it will be necessary to evaluate the ML models against existing methods. Second, the labels used in this dissertation are based on error-prone local storm reports. It will be crucial as a community to address these deficiencies in severe weather reporting. An alternative to storm reports would be to use radar-observed azimuthal shear (Smith and Elmore 2004; Miller et al. 2013; Smith et al. 2016; Mahalik et al. 2019) as a proxy for severe weather, but this approach has its own limitations. Third, the different ML algorithms were similarly skillful, but tended to over- and under-predict in different situations. The best forecast may therefore be a weighted average of the different ML predictions, just as ensembles outperform deterministic forecasts in numerical weather prediction. Ensemble approaches can also provide estimates of forecast uncertainty, which can improve the trustworthiness of ML methods. Future work should therefore explore the use of ML model ensembles for severe weather prediction. Lastly, I did not evaluate the ability of the ML models to differentiate between severe weather hazards. In future work, it is worth exploring multi-class approaches (i.e., will a forecast storm produce hail or a tornado or both?).

Though this dissertation explored multiple interpretation methods for traditional ML algorithms, there were some limitations worth discussing. First, this dissertation primarily focused on the global approach to ML model interpretation. The global approach is a necessary first-step in evaluating an ML model, but future research should stratify the training dataset as mentioned above for verification (by time of day, environmental conditions, etc) and further develop local approaches to explore the learned relationships in particular regimes. For example, this dissertation briefly explored predictor contributions to tornado-based ML models in environments where the significant tornado parameter was greater than 1. Understanding how and why a model performs well or poorly in particular situations is valuable information for forecasters. Second, though predictor interactions were briefly assessed, a more comprehension study on predictor interactions is a necessary next step. There are methods that were not discussed such as the $H$-statistic (Friedman 2002; Molnar 2019a) which can characterize predictor interactions, but they are often computationally expensive to compute, especially for ML models with many predictors. Lastly, these interpretation methods allow us to peek in the black box, but we cannot ignore that these are complex models with high dimensionality and therefore may not lend themselves to being easily understood or conceptualized.

A goal of this dissertation was not only to assess current WoFS probabilistic guidance but also provide a framework to objectively assess the impacts of potential post-processing techniques (e.g., machine learning calibration). Applications of artificial intelligence methods are becoming more common in the meteorological community with methods spanning from traditional machine learning algorithms to sophisticated deep learning methods (McGovern et al. 2017). Post-processing techniques using machine learning can potentially improve the skill and reliability of WoFS probability swath objects by correcting model biases. Developing verification techniques suited to short-term, storm-scale probabilistic guidance is a necessary first step to evaluating machine learning and other promising post-processing methods.

# References

Adams-Selin, R. D., A. J. Clark, C. J. Melick, S. R. Dembek, I. L. Jirak, and C. L. Ziegler, 2019: Evolution of WRF-HAILCAST during the 201416 NOAA/Hazardous Weather Testbed Spring Forecasting Experiments. *Wea. Forecasting*, **34 (1)**, 61–79, doi:10.1175/waf-D-18-0024.1.

Adams-Selin, R. D. and C. L. Ziegler, 2016: Forecasting hail using a one-dimensional hail growth model within WRF. *Mon. Wea. Rev.*, **144 (12)**, 4919–4939, doi:10.1175/mwr-d-16-0027.1.

Adrianto, I., T. B. Trafalis, and V. Lakshmanan, 2009: Support vector machines for spatiotemporal tornado prediction. *International Journal of General Systems*, **38 (7)**, 759–776, doi:10.1080/03081070601068629.

Ahijevych, D., E. Gilleland, B. G. Brown, and E. E. Ebert, 2009: Application of Spatial Verification Methods to Idealized and NWP-Gridded Precipitation Forecasts. *Wea. Forecasting*, **24 (6)**, 1485–1497, doi:10.1175/2009waf2222298.1.

Alaka, M. A., W. D. Bonner, J. P. Charba, R. L. Crisci, R. C. Elvander, and R. M. Reap, 1973: Objective techniques for forecasting thunderstorms and severe weather. Tech. rep., Techniques Development Laboratory, 97 pp., Silver Spring, Md.

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

Alexiuk, M., N. Pizzi, P. C. Li, and W. Pedrycz, 2000: Classification of Volumetric Storm Cell Patterns. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, **4 (3)**, 206–211, doi:10.20965/jaciii.2000.p0206.

Alexiuk, M., N. Pizzi, and W. Pedrycz, 1999: Classification of volumetric storm cell patterns. *Engineering Solutions for the Next Millennium. 1999 IEEE Canadian Conference on Electrical and Computer Engineering (Cat. No.99TH8411)*, **2 (3)**, 1081–1085 vol.2, doi:10.1109/ccece.1999.808201.

Allen, J. T., M. K. Tippett, and A. H. Sobel, 2015: An empirical model relating u.s. monthly hail occurrence to large-scale meteorological environment. *Journal of Advances in Modeling Earth Systems*, **7 (1)**, 226–243, doi:10.1002/2014MS000397, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2014MS000397, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2014MS000397.

Anderson, J. L., 2001: An ensemble adjustment kalman filter for data assimilation. *Mon. Wea. Rev.*, **129 (12)**, 2884–2903, doi:10.1175/1520-0493(2001)129⟨2884:AEAKFF⟩2.0.CO;2.

Anderson-Frey, A. K., Y. P. Richardson, A. R. Dean, R. L. Thompson, and B. T. Smith, 2017: Self-Organizing Maps for the Investigation of Tornadic Near-Storm Environments. *Wea. Forecasting*, doi:10.1175/waf-d-17-0034.1.

Apley, D. W. and J. Zhu, 2016: Visualizing the Effects of Predictor Variables in Black Box Supervised Learning Models. *arXiv*, 1612.08468.

Baldwin, M. E. and J. S. Kain, 2006: Sensitivity of Several Performance Measures to Displacement Error, Bias, and Event Frequency. *Wea. Forecasting*, **21 (4)**, 636–648, doi:10.1175/WF933.1, https://journals.ametsoc.org/waf/article-pdf/21/4/636/4638227/waf933_1.pdf.

Barthold, F. E., T. E. Workoff, B. A. Cosgrove, J. J. Gourley, D. R. Novak, and K. M. Mahoney, 2015: Improving flash flood forecasts: The hmt-wpc flash flood and intense rainfall experiment. *Bulletin of the American Meteorological Society*, **96 (11)**, 1859–1866, doi:10.1175/BAMS-D-14-00201.1.

Batista, G. E. A. P. A., R. C. Prati, and M. C. Monard, 2004: A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, **6 (1)**, 20, doi:10.1145/1007730.1007735.

Bergstra, J., Y. D., and D. D. Cox, 2013: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *Proc. of the 30th International Conference on Machine Learning*, ICML, Vol. 28, 115–123.

Billet, J., M. DeLisi, B. G. Smith, and C. Gates, 1997: Use of Regression Techniques to Predict Hail Size and the Probability of Large Hail. *Wea. Forecasting*, **12 (1)**, 154–164, doi:10.1175/1520-0434(1997)012⟨0154:uorttp⟩2.0.co;2.

Boyd, K., V. S. Costa, J. Davis, and D. Page, 2012: Unachievable Region in Precision-Recall Space and Its Effect on Empirical Evaluation. *arXiv*, 1206.4667.

Breiman, L., 2001a: Random forests. *Machine Learning*, **45**, 5–32, doi:10.1023/A:1010933404324.

Breiman, L., 2001b: Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, **16 (3)**, 199–231, doi:10.1214/ss/1009213726.

Brimelow, J. C., G. W. Reuter, and E. R. Poolman, 2002: Modeling Maximum Hail Size in Alberta Thunderstorms. *Wea. Forecasting*, **17 (5)**, 1048–1062, doi:10.1175/1520-0434(2002)017⟨1048:MMHSIA⟩2.0.CO;2, https://journals.ametsoc.org/WF/article-pdf/17/5/1048/4632236/1520-0434(2002)017_1048_mmhsia_2_0_co_2.pdf.

Bröcker, J. and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22 (3)**, 651–661, doi:10.1175/waf993.1.

Brooks, H. E. and J. Correia, James, 2018: Long-Term Performance Metrics for National Weather Service Tornado Warnings. *Wea. Forecasting*, **33 (6)**, 1501–1511, doi:10.1175/waf-D-18-0120.1, URL https://doi.org/10.1175/waf-D-18-0120.1, https://journals.ametsoc.org/waf/article-pdf/33/6/1501/4666347/waf-d-18-0120_1.pdf.

Brooks, H. E., C. A. Doswell, and M. P. Kay, 2003: Climatological estimates of local daily tornado probability for the united states. *Wea. Forecasting,*, **18 (4)**, 626–640, doi:10.1175/1520-0434(2003)018⟨0626:CEOLDT⟩2.0.CO;2.

Brooks, H. E., I. Doswell, Charles A., and J. Cooper, 1994: On the Environments of Tornadic and Nontornadic Mesocyclones. *Wea. Forecasting*, **9 (4)**, 606–618, doi:10.1175/1520-0434(1994)009⟨0606:OTEOTA⟩2.0.CO;2, https://journals.ametsoc.org/WF/article-pdf/9/4/606/4651318/1520-0434(1994)009_0606_oteota_2_0_co_2.pdf.

Brotzge, J. and W. Donner, 2013: The tornado warning process: Areview of current research, challenges, and opportunities. *BAMS.*, **94**, 1715–1733.

Brown, M. and C. J. Nowotarski, 2019: The Influence of Lifting Condensation Level on Low-Level Outflow and Rotation in Simulated Supercell Thunderstorms. *Journal of the Atmospheric Sciences*, **76 (5)**, 1349–1372, doi:10.1175/JAS-D-18-0216.1, https://journals.ametsoc.org/jas/article-pdf/76/5/1349/4829453/jas-d-18-0216_1.pdf.

Bryan, G. H., J. C. Wyngaard, and J. M. Fritsch, 2003: Resolution Requirements for the Simulation of Deep Moist Convection. *Mon. Wea. Rev.*, **131 (10)**, 2394–2416, doi:10.1175/1520-0493(2003)131⟨2394:RRFTSO⟩2.0.CO;2, https://journals.ametsoc.org/mwr/article-pdf/131/10/2394/4206189/1520-0493(2003)131_2394_rrftso_2_0_co_2.pdf.

Burke, A., N. Snook, D. J. Gagne, S. McCorkle, and A. McGovern, 2019: Calibration of Machine Learning-Based Probabilistic Hail Predictions for Operational Forecasting. *Wea. Forecasting*, doi:10.1175/waf-d-19-0105.1.

Cai, H. and R. E. Dumais, 2015: Object-based evaluation of a numerical weather prediction models performance through forecast storm characteristic analysis. *Wea. Forecasting*, **30 (6)**, 1451–1468, doi:10.1175/waf-D-15-0008.1.

Center, D. T., 2017: Ensemble kalman filter (enkf) user's guide for version 1.2. 86, URL http://www.dtcenter.org/EnKF/users/docs/index.php.

Charba, J. P., 1979: Two to Six Hour Severe Local Storm Probabilities: An Operational Forecasting System. *Mon. Wea. Rev.*, **107 (3)**, 268–282, doi:10.1175/1520-0493(1979)107⟨0268:ttshsl⟩2.0.co;2.

Chen, T. and C. Guestrin, 2016: XGBoost: A Scalable Tree Boosting System. *arXiv*, doi:10.1145/2939672.2939785, 1603.02754.

Choate, J. J., P. S. Skinner, K. A. Wilson, E. Grimes, B. T. Gallo, P. L. Heinselman, and A. J. Clark, 2019: Examining the use of the NSSL experimental warn-on-forecast system for ensembles for the prediction of severe storms through short-term forecast outlooks during the 2018 spring forecasting experiment. *99th AMS Annual Meeting*, Phoenix, AZ, Amer. Meteor. Soc., 5B.6.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, L. Cronce, and J. Brunner, 2020: Noaa probsevere v2.0 probhail, probwind, and probtor. *Wea. Forecasting*, **0 (0)**, null, doi:10.1175/waf-D-19-0242.1.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, and D. T. Lindsey, 2014: An empirical model for assessing the severe weather potential of developing convection. *Wea. Forecasting*, **29 (3)**, 639–653, doi:10.1175/waf-D-13-00113.1, URL https://doi.org/10.1175/waf-D-13-00113.1.

Cintineo, J. L., T. M. Smith, V. Lakshmanan, H. E. Brooks, and K. L. Ortega, 2012: An objective high-resolution hail climatology of the contiguous united states. *Wea. Forecasting*, **27 (5)**, 1235–1248, doi:10.1175/waf-D-11-00151.1, URL https://doi.org/10.1175/waf-D-11-00151.1, https://doi.org/10.1175/waf-D-11-00151.1.

Cintineo, J. L., et al., 2018: The NOAA/CIMSS ProbSevere Model incorporation of total lightning and validation. *Wea. Forecasting*, **33 (1)**, 331–345, doi:10.1175/waf-d-17-0099.1.

Cintineo, R. M. and D. J. Stensrud, 2013: On the predictability of supercell thunderstorm evolution. *JAS*, **70**, 1993–2011, doi:10.1175/jas-d-12-0166.1.

Clark, A. J., R. G. Bullock, T. L. Jensen, M. Xue, and F. Kong, 2014: Application of object-based time-domain diagnostics for tracking precipitation systems in convection-allowing models. *Wea. Forecasting*, **29 (3)**, 517–542, doi:10.1175/waf-D-13-00098.1.

Clark, A. J., J. Gao, P. T. Marsh, T. Smith, J. S. Kain, J. Correia, James, M. Xue, and F. Kong, 2013: Tornado Pathlength Forecasts from 2010 to 2011 Using Ensemble Updraft Helicity. *Wea. Forecasting*, **28 (2)**, 387–407, doi:10.1175/waf-D-12-00038.1, URL https://doi.org/10.1175/waf-D-12-00038.1, https://journals.ametsoc.org/waf/article-pdf/28/2/387/4653179/waf-d-12-00038_1.pdf.

Clark, A. J., J. S. Kain, P. T. Marsh, J. Correia, James, M. Xue, and F. Kong, 2012: Forecasting Tornado Pathlengths Using a Three-Dimensional Object Identification Algorithm Applied to Convection-Allowing Forecasts. *Wea. Forecasting*, **27 (5)**, 1090–1113, doi:10.1175/waf-D-11-00147.1, URL https://doi.org/10.1175/waf-D-11-00147.1, https://journals.ametsoc.org/waf/article-pdf/27/5/1090/4651995/waf-d-11-00147_1.pdf.

Clark, A. J., C. J. Schaffer, J. Gallus, William A., and K. Johnson-OMara, 2009: Climatology of Storm Reports Relative to Upper-Level Jet Streaks. *Wea. Forecasting*, **24 (4)**, 1032–1051, doi:10.1175/2009waf2222216.1, URL https://doi.org/10.1175/2009waf2222216.1, https://journals.ametsoc.org/waf/article-pdf/24/4/1032/4645360/2009waf2222216_1.pdf.

Coffer, B. E. and M. D. Parker, 2016: Simulated Supercells in Nontornadic and Tornadic VORTEX2 Environments. *Mon. Wea. Rev.*, **145 (1)**, 149–180, doi:10.1175/MWR-D-16-0226.1, URL https://doi.org/10.1175/MWR-D-16-0226.1, https://journals.ametsoc.org/mwr/article-pdf/145/1/149/4361091/mwr-d-16-0226_1.pdf.

Coffer, B. E., M. D. Parker, J. M. L. Dahl, L. J. Wicker, and A. J. Clark, 2017: Volatility of Tornadogenesis: An Ensemble of Simulated Nontornadic and Tornadic Supercells in VORTEX2 Environments. *Mon. Wea. Rev.*, **145 (11)**, 4605–4625, doi:10.1175/MWR-D-17-0152.1, URL https://doi.org/10.1175/MWR-D-17-0152.1, https://journals.ametsoc.org/mwr/article-pdf/145/11/4605/4361600/mwr-d-17-0152_1.pdf.

Coffer, B. E., M. D. Parker, R. L. Thompson, B. T. Smith, and R. E. Jewell, 2019: Using Near-Ground Storm Relative Helicity in Supercell Tornado Forecasting. *Wea. Forecasting*, **34 (5)**, 1417–1435, doi:10.1175/waf-D-19-0115.1, URL https://doi.org/10.1175/waf-D-19-0115.1, https://journals.ametsoc.org/waf/article-pdf/34/5/1417/4883068/waf-d-19-0115_1.pdf.

Coleman, T. A., K. R. Knupp, J. Spann, J. B. Elliott, and B. E.Peters, 2011: The history (and future) of tornado warningdissemination in the united states. *BAMS*, **92**, 567–582.

Coniglio, M. C., J. Y. Hwang, and D. J. Stensrud, 2010: Environmental Factors in the Upscale Growth and Longevity of MCSs Derived from Rapid Update Cycle Analyses. *Mon. Wea. Rev.*, **138 (9)**, 3514–3539, doi:10.1175/2010MWR3233.1, URL https://doi.org/10.1175/2010MWR3233.1, https://journals.ametsoc.org/mwr/article-pdf/138/9/3514/4259253/2010mwr3233_1.pdf.

Coniglio, M. C. and M. D. Parker, 2020: Insights into supercells and their environments from three decades of targeted radiosonde observations. *Mon. Wea. Rev.*, 1–68, doi:10.1175/MWR-D-20-0105.1, https://journals.ametsoc.org/mwr/article-pdf/doi/10.1175/MWR-D-20-0105.1/4998226/mwrd200105.pdf.

Corfidi, S. F., 2003: Cold Pools and MCS Propagation: Forecasting the Motion of Downwind-Developing MCSs. *Wea. Forecasting*, **18 (6)**, 997–1017, doi:10.1175/1520-0434(2003)018⟨0997:CPAMPF⟩2.0.CO;2, https://journals.ametsoc.org/WF/article-pdf/18/6/997/4634824/1520-0434(2003)018_0997_cpampf_2_0_co_2.pdf.

Cortes, C. and V. Vapnik, 1995: Support-vector networks. *Machine learning*, **20 (3)**, 273–297.

Czernecki, B., M. Taszarek, M. Marosz, M. Prolniczak, L. Kolendowicz, A. Wyszogrodzki, and J. Szturc, 2019: Application of machine learning to large hail prediction - The importance of radar reflectivity, lightning occurrence and convective parameters derived from ERA5. *Atmospheric Research*, **227**, 249–262, doi:10.1016/j.atmosres.2019.05.010.

Davies-Jones, R., 1984: Streamwise Vorticity: The Origin of Updraft Rotation in Supercell Storms. *Journal of the Atmospheric Sciences*, **41 (20)**, 2991–3006, doi:10.1175/1520-0469(1984)041⟨2991:svtoou⟩2.0.co;2.

Davies-Jones, R. and H. Brooks, 1993: *Mesocyclogenesis from a Theoretical Perspective*, 105–114. American Geophysical Union (AGU), doi:10.1029/GM079p0105, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/GM079p0105, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/GM079p0105.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. Part I: Methodology and application to mesoscale rain areas. *Mon. Wea. Rev.*, **134 (7)**, 1772–1784, doi:10.1175/mwr3145.1.

Davis, C. A., B. G. Brown, R. Bullock, and J. Halley-Gotway, 2009: The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. *Wea. Forecasting*, **24 (5)**, 1252–1267, doi:10.1175/2009waf2222241.1.

Davis, J. and M. Goadrich, 2006: The relationship between precision-recall and roc curves. *Proceedings of the 23rd International Conference on Machine Learning*, 233240, doi:10.1145/1143844.1143874, URL https://doi.org/10.1145/1143844.1143874.

Dawson, D. T., L. J. Wicker, E. R. Mansell, and R. L. Tanamachi, 2012: Impact of the environmental low-level wind profile on ensemble forecasts of the 4 may 2007 greensburg, kansas, tornadic storm and associated mesocyclones. *Mon. Wea. Rev.*, **140 (2)**, 696–716, doi:10.1175/mwr-d-11-00008.1.

Dawson, L. C., G. S. Romine, R. J. Trapp, and M. E. Baldwin, 2017: Verifying Supercellular Rotation in a Convection-Permitting Ensemble Forecasting System with Radar-Derived Rotation Track Data. *Wea. Forecasting*, **32 (2)**, 781–795, doi:10.1175/waf-d-16-0121.1.

Dennis, E. J. and M. R. Kumjian, 2017: The impact of vertical wind shear on hail growth in simulated supercells. *Journal of the Atmospheric Sciences*, **74 (3)**, 641–663, doi:10.1175/jas-d-16-0066.1.

Doshi-Velez, F. and B. Kim, 2017: Towards A Rigorous Science of Interpretable Machine Learning. *arXiv*, 1702.08608.

Doswell, C. A., H. E. Brooks, and M. P. Kay, 2005: Climatological estimates of daily local nontornadic severe thunderstorm probability for the united states. *Wea. Forecasting,*, **20 (4)**, 577–595, doi:10.1175/WF866.1.

Doswell, C. A. and J. S. Evans, 2003: Proximity sounding analysis for derechos and supercells: an assessment of similarities and differences. *Atmospheric Research*, **67-68**, 117 – 133, doi:https://doi.org/10.1016/S0169-8095(03)00047-4, URL http://www.sciencedirect.com/science/article/pii/S0169809503000474, european Conference on Severe Storms 2002.

Dowell, D. and Coauthors, 2016: Development of a high-resolution rapid refresh ensemble (HRRRE) for severe weather forecasting. Preprints. *28th Conf. on Severe Local Storms*, Portland, OR, Amer. Meteor. Soc.,, 8B.2.

Duda, J. D. and W. A. Gallus, 2010: Spring and Summer Midwestern Severe Weather Reports in Supercells Compared to Other Morphologies. *Wea. Forecasting*, **25 (1)**, 190–206, doi:10.1175/2009waf2222338.1.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications*, **15 (1)**, 51–64, doi:10.1002/met.25, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.25.

Ebert, E. E., M. Turk, S. J. Kusselson, J. Yang, M. Seybold, P. R. Keehn, and R. J. Kuligowski, 2011: Ensemble Tropical Rainfall Potential (eTRaP) Forecasts. *Wea. Forecasting*, **26 (2)**, 213–224, doi:10.1175/2010waf2222443.1.

Edwards, R. and R. L. Thompson, 1998: Nationwide Comparisons of Hail Size with WSR-88D Vertically Integrated Liquid Water and Derived Thermodynamic Sounding Data. *Wea. Forecasting*, **13**, 277–285.

Evans, J. S. and I. Doswell, Charles A., 2001: Examination of Derecho Environments Using Proximity Soundings. *Wea. Forecasting*, **16 (3)**, 329–342, doi:10.1175/1520-0434(2001)016⟨0329:EODEUP⟩2.0.CO;2, https://journals.ametsoc.org/WF/article-pdf/16/3/329/4631193/1520-0434(2001)016_0329_eodeup_2_0_co_2.pdf.

Fisher, A., C. Rudin, and F. Dominici, 2018: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *arXiv*, 1801.01489.

Flora, M. and S. Handler, 2020: Model interpretability in python (mintpy). GitHub, https://github.com/monte-flora/mintpy.

Flora, M., C. K. Potvin, P. S. Skinner, S. Handler, and A. McGovern, 2020: Using machine learning to calibrate storm-scale probabilistic guidance of severe weather hazards in the warn-on-forecast system. 2012.00679.

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

Flora, M. L., P. S. Skinner, C. K. Potvin, A. E. Reinhart, T. A. Jones, N. Yussouf, and K. H. Knopfmeier, 2019: Object-based verification of short-term, storm-scale probabilistic mesocyclone guidance from an experimental warn-on-forecast system. *Wea. Forecasting*, **34 (6)**, 1721–1739, doi:10.1175/WF-D-19-0094.1.

Flournoy, M. D., M. C. Coniglio, E. N. Rasmussen, J. C. Furtado, and B. E. Coffer, 2020: Modes of Storm-Scale Variability and Tornado Potential in VORTEX2 Near- and Far-Field Tornadic Environments. *Mon. Wea. Rev.*, **148 (10)**, 4185–4207, doi:10.1175/MWR-D-20-0147.1, https://journals.ametsoc.org/mwr/article-pdf/148/10/4185/5002680/mwrd200147.pdf.

Friedman, J., 2002: Stochastic gradient boosting. *Comput. Stat. Data Anal.*, **38**, 367–378, doi:https://doi.org/10.1016/S0167-9473(01)00065-2.

Friedman, J. H., 2001: GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE. *The Annals of Statistics*, **29 (5)**, 1189–1232, doi:10.1214/aos/1013203451.

Gagne, D. J., A. McGovern, J. B. Basara, and R. A. Brown, 2012: Tornadic supercell environments analyzed using surface and reanalysis data: A spatiotemporal relational data-mining approach. *Journal of Applied Meteorology and Climatology*, **51 (12)**, 2203–2217, doi:10.1175/JAMC-D-11-060.1, URL https://doi.org/10.1175/JAMC-D-11-060.1, https://doi.org/10.1175/JAMC-D-11-060.1.

Gagne, D. J., A. McGovern, N. Snook, R. Sobash, J. Labriola, J. K. Williams, S. E. Haupt, and M. Xue, 2016: Hagelslag: Scalable object-based severe weather analysis and forecasting. *Proceedings of the Sixth Symposium on Advances in Modeling and Analysis Using Python*, New Orleans, LA, Amer. Meteor. Soc.,, 447.

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

Gagne, I., David John, 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. *Wea. Forecasting*, **32 (5)**, 1819–1840, doi:10.1175/waf-D-17-0010.1, URL https://doi.org/10.1175/waf-D-17-0010.1, https://journals.ametsoc.org/waf/article-pdf/32/5/1819/4661035/waf-d-17-0010_1.pdf.

Gagne, J. D., A. McGovern, J. Brotzge, and M. Xue, 2013: Severe Hail Prediction within a Spatiotemporal Relational Data Mining Framework. *IEEE 13th International Conference on Data Mining Workshops*, 994–1001, doi:10.1109/icdmw.2013.121.

Gagne II, D. J., S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable Deep Learning for Spatial Analysis of Severe Hailstorms. *Mon. Wea. Rev.*, **147 (8)**, 2827–2845, doi:10.1175/MWR-D-18-0316.1.

Gallo, B. T., A. J. Clark, and S. R. Dembek, 2016: Forecasting tornadoes using convection-permitting ensembles. *Wea. Forecasting*, **31 (1)**, 273–295, doi:10.1175/waf-D-15-0134.1.

Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2018: Blended Probabilistic Tornado Forecasts: Combining Climatological Frequencies with NSSL-WRF Ensemble Forecasts. *Wea. Forecasting*, **33 (2)**, 443–460, doi:10.1175/waf-d-17-0132.1.

Gallo, B. T., A. J. Clark, B. T. Smith, R. L. Thompson, I. Jirak, and S. R. Dembek, 2019: Incorporating UH Occurrence Time to Ensemble-Derived Tornado Probabilities. *Wea. Forecasting*, **0 (0)**, null, doi:10.1175/waf-D-18-0108.1.

Gallo, B. T., et al., 2017: Breaking new ground in severe weather prediction: The 2015 NOAA/Hazardous Weather Testbed Spring Forecasting Experiment. *Wea. Forecasting*, **32 (4)**, 1541–1568, doi:10.1175/waf-D-16-0178.1.

Gallus, W. A., 2010: Application of object-based verification techniques to ensemble precipitation forecasts. *Wea. Forecasting*, **25 (1)**, 144–158, doi:10.1175/2009waf2222274.1.

Gatzen, C., 2011: A 10-year climatology of cold-season narrow cold-frontal rainbands in germany. *Atmospheric Research - ATMOS RES*, **100**, 366–370, doi:10.1016/j.atmosres.2010.09.018.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of Spatial Forecast Verification Methods. *Wea. Forecasting*, **24 (5)**, 1416–1430, doi:10.1175/2009waf2222269.1.

Gilleland, E., D. A. Ahijevych, B. G. Brown, and E. E. Ebert, 2010: Verifying forecasts spatially. *Bulletin of the American Meteorological Society*, **91 (10)**, 1365–1376, doi:10.1175/2010BAMS2819.1.

Gilmore, M. S. and L. J. Wicker, 1998: The Influence of Midtropospheric Dryness on Supercell Morphology and Evolution. *Mon. Wea. Rev.*, **126 (4)**, 943–958, doi:10.1175/1520-0493(1998)126⟨0943:TIOMDO⟩2.0.CO;2, https://journals.ametsoc.org/mwr/article-pdf/126/4/943/4180405/1520-0493(1998)126_0943_tiomdo_2_0_co_2.pdf.

Glahn, H. R. and D. A. Lowry, 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting. *Journal of Applied Meteorology*, **11 (8)**, 1203–1211, doi:10.1175/1520-0450(1972)011⟨1203:tuomos⟩2.0.co;2.

Grant, L. D. and S. C. van den Heever, 2014: Microphysical and Dynamical Characteristics of Low-Precipitation and Classic Supercells. *Journal of the Atmospheric Sciences*, **71 (7)**, 2604–2624, doi:10.1175/JAS-D-13-0261.1, URL https://doi.org/10.1175/JAS-D-13-0261.1.

Greenwell, B. M., B. C. Boehmke, and A. J. McCarthy, 2018: A Simple and Effective Model-Based Variable Importance Measure. *arXiv*, 1805.04755.

Groenemeijer, P. and A. v. Delden, 2007: Sounding-derived parameters associated with large hail and tornadoes in the Netherlands. *Atmospheric Research*, **83 (2-4)**, 473–487, doi:10.1016/j.atmosres.2005.08.006.

Hastie, T., R. Tibshirani, and J. Friedman, 2001: *The Elements of Statistical Learning*. Springer Series in Statistics, Springer New York Inc., New York, NY, USA.

Hepper, R. M., I. L.Jirak, and J. M.Milne, 2016: Assessing the skill of convection-allowing ensemble forecasts of severe mcs winds from the sseo Preprints. *28th Conf. on Severe Local Storms,*, Portland, OR, Amer. Meteor. Soc.,, 16B.2, URL https://ams.confex.com/ams/28SLS/webprogram/Paper300134.html.

Herman, G. R. and R. S. Schumacher, 2018: Money Doesnt Grow on Trees, But Forecasts Do: Forecasting Extreme Precipitation with Random Forests. *Mon. Wea. Rev.*, doi:10.1175/mwr-d-17-0250.1.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting Severe Weather with Random Forests. *Mon. Wea. Rev.*, **148 (5)**, 2135–2161, doi:10.1175/MWR-D-19-0344.1, https://journals.ametsoc.org/mwr/article-pdf/148/5/2135/4928197/mwrd190344.pdf.

Hitchens, N. M., H. E. Brooks, and M. P. Kay, 2013: Objective limits on forecasting skill of rare events. *Wea. Forecasting*, **28 (2)**, 525–534, doi:10.1175/waf-D-12-00113.1.

Hoekstra, S., K. Klockow, R. Riley, J. Brotzge, H. Brooks, and S. Erickson, 2011: A preliminary look at the social perspective of warn-on-forecast: Preferred tornado warning lead time and the general publics perceptions of weather risks. *Weather, Climate, and Society*, **3 (2)**, 128–140, doi:10.1175/2011WCAS1076.1, URL https://doi.org/10.1175/2011WCAS1076.1.

Hoffman, R. R., D. S.LaDue, H. M.Mogil, P. J.Roebber, and J. G.Trafton, 2017: *Minding the Weather: How Expert Forecasters Think*, chap. Forecaster-Computer Interdependence, 470 pp. The MIT Press.

Homeyer, C. R., T. N. Sandml, C. K. Potvin, and A. M. Murphy, 2020: Distinguishing Characteristics of Tornadic and Nontornadic Supercell Storms from Composite Mean Analyses of Radar Observations. *Mon. Wea. Rev.*, 1–64, doi:10.1175/MWR-D-20-0136.1, https://journals.ametsoc.org/mwr/article-pdf/doi/10.1175/MWR-D-20-0136.1/5008102/mwrd200136.pdf.

Houze Jr., R. A., 2004: Mesoscale convective systems. *Reviews of Geophysics*, **42 (4)**, doi:10.1029/2004RG000150, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2004RG000150.

Hsu, W.-r. 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, URL https://EconPapers.repec.org/RePEc:eee:intfor:v:2:y:1986:i:3:p:285-293.

Huntrieser, H., H. H. Schiesser, W. Schmid, and A. Waldvogel, 1997: Comparison of Traditional and Newly Developed Thunderstorm Indices for Switzerland. *Wea. Forecasting*, **12 (1)**, 108–125, doi:10.1175/1520-0434(1997)012⟨0108:cotand⟩2.0.co;2.

Jain, A., 1989: *Fundamentals of Digital Image Processing.* Prentice Hall, 569 pp.

Jergensen, G. E., A. McGovern, R. Lagerquist, and T. Smith, 2020: Classifying Convective Storms Using Machine Learning. *Wea. Forecasting*, **35 (2)**, 537–559, doi:10.1175/waf-d-19-0170.1.

Jewell, R. and J. Brimelow, 2009: Evaluation of Alberta Hail Growth Model Using Severe Hail Proximity Soundings from the United States. *Wea. Forecasting*, **24 (6)**, 1592–1609, doi:10.1175/2009waf2222230.1.

Jirak, I. L., C. J.Melick, and S. J.Weiss, 2014: Combining probabilistic ensemble information from the environment with simulated storm attributes to generate calibrated probabilities of severe weather hazards. Preprints. *27th Conf. on Severe Local Storms*, Madison, WI,, Amer. Meteor. Soc.,, 2.5, URL https://ams.confex.com/ams/27SLS/webprogram/Paper254649.html.

Johnson, A. and K. Sugden, 2014: Evaluation of Sounding-Derived Thermodynamic and Wind-Related Parameters Associated with Large Hail Events. *Electronic Journal of Severe Storms Meteorology*, **9 (5)**, 1–42.

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

Jones, T., P. Skinner, N. Yussouf, K. Knopfmeier, A. Reinhart, and D. Dowell, 2019: Forecasting high-impact weather in landfalling tropical cyclones using a warn-on-forecast system. *BAMS*, **100 (8)**, 1405–1417, doi:10.1175/BAMS-D-18-0203.1, https://doi.org/10.1175/BAMS-D-18-0203.1.

Jones, T. A., K. Knopfmeier, D. Wheatley, G. Creager, P. Minnis, and R. Palikonda, 2016: Storm-scale data assimilation and ensemble forecasting with the NSSL experimental Warn-on-Forecast system Part II: Combined radar and satellite data experiments. *Wea. Forecasting*, **31 (1)**, 297–327, doi:10.1175/waf-d-15-0107.1.

Jones, T. A., P. Skinner, K. Knopfmeier, E. Mansell, P. Minnis, R. Palikonda, and W. S. Jr., 2018: Comparison of cloud microphysics schemes in a Warn-on-Forecast system using synthetic satellite objects Comparison of cloud microphysics schemes in a Warn-on-Forecast system using synthetic satellite objects. *Wea. Forecasting*, doi:10.1175/waf-d-18-0112.1.

Karstens, C. D., et al., 2018: Development of a HumanMachine Mix for Forecasting Severe Convective Events. *waf*, **33 (3)**, 715–737, doi:10.1175/waf-D-17-0188.1, URL https://doi.org/10.1175/waf-D-17-0188.1, https://doi.org/10.1175/waf-D-17-0188.1.

Kitzmiller, D. H., W. E. McGovern, and R. F. Saffle, 1995: The WSR-88D Severe Weather Potential Algorithm. *Wea. Forecasting*, **10 (1)**, 141–159, doi:10.1175/1520-0434(1995)010⟨0141:twswpa⟩2.0.co;2.

Klein, W. H. and H. R. Glahn, 1974: Forecasting local weather by means of model output statistics. *Bull. Amer. Meteor. Soc.*, **55**, 1217– 1227.

Kuchera, E. L. and M. D. Parker, 2006: Severe Convective Wind Environments. *Wea. Forecasting*, **21 (4)**, 595–612, doi:10.1175/waf931.1.

Kuhn, M. and K. Johnson, 2013: *Applied predictive modeling*. Springer, New York, NY, doi:10.1007/978-1-4614-6849-3, URL http://www.amazon.com/Applied-Predictive-Modeling-Max-Kuhn/dp/1461468485/.

Kumjian, M. R., Z. J. Lebo, and A. M. Ward, 2019: Storms Producing Large Accumulations of Small Hail Storms Producing Large Accumulations of Small Hail. *Journal of Applied Meteorology and Climatology*, **58 (2)**, 341–364, doi:10.1175/jamc-d-18-0073.1.

Kumjian, M. R. and K. Lombardo, 2020: A Hail Growth Trajectory Model for Exploring the Environmental Controls on Hail Size: Model Physics and Idealized Tests. *Journal of the Atmospheric Sciences*, **77 (8)**, 2765–2791, doi:10.1175/jas-d-20-0016.1.

Kunz, M., E. Fluck, S. Baumstark, J. Wandel, S. Ritz, S. Schemm, and M. Schmidberger, 2017: Hail frequency in central europe estimated from 2d radar data and the relation to atmospheric characteristics. *9th European Conference on Severe Storms (ECSS)*, Pula, Croatia. Wessling, Germany: European Severe Storms Laboratory (via Copernicus).

Labriola, J., N. Snook, Y. Jung, B. Putnam, and M. Xue, 2017: Ensemble hail prediction for the storms of 10 may 2010 in south-central oklahoma using single- and double-moment microphysical schemes. *Mon. Wea. Rev.*, **145 (12)**, 4911–4936, doi:10.1175/MWR-D-17-0039.1.

Labriola, J., N. Snook, Y. Jung, and M. Xue, 2019: Explicit Ensemble Prediction of Hail in 19 May 2013 Oklahoma City Thunderstorms and Analysis of Hail Growth Processes with Several Multimoment Microphysics Schemes. *Mon. Wea. Rev.*, **147 (4)**, 1193–1213, doi:10.1175/MWR-D-18-0266.1.

Labriola, J., N. Snook, Y. Jung, and M. Xue, 2020: Evaluating Ensemble Kalman Filter Analyses of Severe Hailstorms on 8 May 2017 in Colorado: Effects of State Variable Updating and Multimoment Microphysics Schemes on State Variable Cross Covariances. *Mon. Wea. Rev.*, **148 (6)**, 2365–2389, doi:10.1175/MWR-D-19-0300.1, https://journals.ametsoc.org/mwr/article-pdf/148/6/2365/4944250/mwrd190300.pdf.

Lagerquist, R., A. McGovern, C. R. Homeyer, I. Gagne, David John, and T. Smith, 2020: Deep Learning on Three-dimensional Multiscale Data for Next-hour Tornado Prediction. *Mon. Wea. Rev.*, **148 (7)**, 2837–2861, doi:10.1175/MWR-D-19-0372.1.

Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine Learning for Real-Time Prediction of Damaging Straight-Line Convective Wind. *Wea. Forecasting*, doi:10.1175/waf-d-17-0038.1.

Lakshmanan, V., I. Adrianto, T. Smith, and G. Stumpf, 2005: A Spatiotemporal Approach to Tornado Prediction. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, **3**, 1642–1647, doi:10.1109/ijcnn.2005.1556125.

Lakshmanan, V., K. Hondl, and R. Rabin, 2009: An efficient, general-purpose technique for identifying storm cells in geospatial images. *JAOT*, **26 (3)**, 523–537, doi:10.1175/2008JTECHA1153.1.

Lakshmanan, V., C. Karstens, J. Krause, K. Elmore, A. Ryzhkov, and S. Berkseth, 2015: Which Polarimetric Variables Are Important for Weather/No-Weather Discrimination? *Journal of Atmospheric and Oceanic Technology*, **32 (6)**, 1209–1223, doi: 10.1175/jtech-d-13-00205.1.

Lawson, J. R., J. S. Kain, N. Yussouf, D. C. Dowell, D. M. Wheatley, K. H. Knopfmeier, and T. A. Jones, 2018a: Advancing from Convection-Allowing NWP to Warn-on-Forecast: Evidence of Progress. *Wea. Forecasting*, **33 (2)**, 599–607, doi:10.1175/waf-D-17-0145.1.

Lawson, J. R., C. Potvin, and M. Flora, 2018b: Information, Predictability, and Verification at the Thunderstorm Scale. *29th Conference on Severe Local Storms*, Stowe, VT, Amer. Meteor. Soc., 124.

LeCun, Y., B. E. Boser, J. S. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, 1990: Handwritten digit recognition with a back-propagation network. advances in neural information processing systems. *Advances in Neural Information Processing Systems*, 396404, URL http://papers.nips.cc/.

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. *Wea. Forecasting*, doi:10.1175/WF-D-19-0258.1, https://journals.ametsoc.org/waf/article-pdf/doi/10.1175/waf-D-19-0258.1/4951271/wafd190258.pdf.

Lopez, L., E. Garca-Ortega, and J. L. Snchez, 2007: A short-term forecast model for hail. *Atmospheric Research*, **83 (2-4)**, 176–184, doi:10.1016/j.atmosres.2005.10.014.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21 (3)**, 289–307, doi:10.1111/j.2153-3490.1969.tb00444.x.

Lundberg, S. M., G. G. Erion, and S.-I. Lee, 2018: Consistent Individualized Feature Attribution for Tree Ensembles. *arXiv*, 1802.03888.

Lundberg, S. M. and S.-I. Lee, 2017: A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 4765–4774, URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Mahalik, M. C., B. R. Smith, K. L. Elmore, D. M. Kingfield, K. L. Ortega, and T. M. Smith, 2019: Estimates of gradients in radar moments using a linear least-squares derivative technique. *Wea. Forecasting*, **0 (0)**, null, doi:10.1175/waf-D-18-0095.1.

Manning, C. and H. Schtze, 1999: *Foundations of Statistical Natural Language Processing.* MIT Press., Cambridge, MA.

Mansell, E. R., C. L. Ziegler, and E. C. Bruning, 2010: Simulated Electrification of a Small Thunderstorm with Two-Moment Bulk Microphysics. *Journal of the Atmospheric Sciences*, **67 (1)**, 171–194, doi:10.1175/2009JAS2965.1, URL https://doi.

org/10.1175/2009JAS2965.1, https://journals.ametsoc.org/jas/article-pdf/67/1/171/3514078/2009jas2965_1.pdf.

Manzato, A., 2013: Hail in Northeast Italy: A Neural Network Ensemble Forecast Using Sounding-Derived Indices. *Wea. Forecasting*, **28 (1)**, 3–28, doi:10.1175/waf-d-12-00034.1.

Markowski, P., C. Hannon, J. Frame, E. Lancaster, A. Pietrycha, R. Edwards, and R. L. Thompson, 2003: Characteristics of Vertical Wind Profiles near Supercells Obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18 (6)**, 1262–1272, doi:10.1175/1520-0434(2003)018⟨1262:COVWPN⟩2.0.CO;2, https://journals.ametsoc.org/WF/article-pdf/18/6/1262/4634685/1520-0434(2003)018_1262_covwpn_2_0_co_2.pdf.

Markowski, P. M. and Y. P. Richardson, 2013: The Influence of Environmental Low-Level Shear and Cold Pools on Tornadogenesis: Insights from Idealized Simulations. *Journal of the Atmospheric Sciences*, **71 (1)**, 243–275, doi:10.1175/JAS-D-13-0159.1, URL https://doi.org/10.1175/JAS-D-13-0159.1, https://journals.ametsoc.org/jas/article-pdf/71/1/243/3643392/jas-d-13-0159_1.pdf.

Markowski, P. M., J. M. Straka, and E. N. Rasmussen, 2002: Direct Surface Thermodynamic Observations within the Rear-Flank Downdrafts of Nontornadic and Tornadic Supercells. *Mon. Wea. Rev.*, **130 (7)**, 1692–1721, doi:10.1175/1520-0493(2002)130⟨1692:dstowt⟩2.0.co;2.

Marzban, C., E. D. W. Mitchell, and G. J. Stumpf, 1999: The Notion of Best Predictors: An Application to Tornado Prediction. *Wea. Forecasting*, **14 (6)**, 1007–1016, doi:10.1175/1520-0434(1999)014⟨1007:TNOBPA⟩2.0.CO;2, https://journals.ametsoc.org/WF/article-pdf/14/6/1007/4629347/1520-0434(1999)014_1007_tnobpa_2_0_co_2.pdf.

Marzban, C. and G. J. Stumpf, 1996: A Neural Network for Tornado Prediction Based on Doppler Radar-Derived Attributes. *Journal of Applied Meteorology*, **35 (5)**, 617–626, doi:10.1175/1520-0450(1996)035⟨0617:annftp⟩2.0.co;2.

Marzban, C. and G. J. Stumpf, 1998: A Neural Network for Damaging Wind Prediction. *Wea. Forecasting*, **13 (1)**, 151–163, doi:10.1175/1520-0434(1998)013⟨0151:annfdw⟩2.0.co.

Marzban, C. and A. Witt, 2001: A Bayesian Neural Network for Severe-Hail Size Prediction. *Wea. Forecasting*, **16 (5)**, 600–610, doi:10.1175/1520-0434(2001)016⟨0600:abnnfs⟩2.0.co;2.

Mashiko, W., 2016a: A Numerical Study of the 6 May 2012 Tsukuba City Supercell Tornado. Part I: Vorticity Sources of Low-Level and Midlevel Mesocyclones. *Mon. Wea. Rev.*, **144 (3)**, 1069–1092, doi:10.1175/MWR-D-15-0123.1, URL https://doi.org/10.1175/MWR-D-15-0123.1, https://journals.ametsoc.org/mwr/article-pdf/144/3/1069/4331606/mwr-d-15-0123_1.pdf.

Mashiko, W., 2016b: A Numerical Study of the 6 May 2012 Tsukuba City Supercell Tornado. Part II: Mechanisms of Tornadogenesis. *Mon. Wea. Rev.*, **144 (9)**, 3077–3098, doi:10.1175/MWR-D-15-0122.1, URL https://doi.org/10.1175/MWR-D-15-0122.1, https://journals.ametsoc.org/mwr/article-pdf/144/9/3077/4351878/mwr-d-15-0122_1.pdf.

McGovern, A., K. L. Elmore, J. I. David, S. 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*, doi:10.1175/bams-d-16-0123.1.

McGovern, A., C. D. Karstens, T. Smith, and R. Lagerquist, 2019a: Quasi-operational testing of real-time storm-longevity prediction via machine learning. *Wea. Forecasting*, **34 (5)**, 1437–1451, doi:10.1175/waf-D-18-0141.1.

McGovern, A., R. Lagerquist, D. J. G. II, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019b: Making the black box more transparent: Understanding the physical implications of machine learning Making the black box more transparent: Understanding the physical implications of machine learning. *BAMS*, doi:10.1175/bams-d-18-0195.1.

McGovern, A., N. Troutman, R. A. Brown, J. K. Williams, and J. Abernethy, 2013: Enhanced spatiotemporal relational probability trees and forests. *Data Mining and Knowledge Discovery*, **26 (2)**, 398–433, doi:10.1007/s10618-012-0261-2.

Metz, C. E., 1978: Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, **8 (4)**, 283–298, doi:10.1016/s0001-2998(78)80014-2.

Milbrandt, J. A. and H. Morrison, 2013: Prediction of Graupel Density in a Bulk Microphysics Scheme. *Journal of the Atmospheric Sciences*, **70 (2)**, 410–429, doi:10.1175/JAS-D-12-0204.1, URL https://doi.org/10.1175/JAS-D-12-0204.1, https://journals.ametsoc.org/jas/article-pdf/70/2/410/3815558/jas-d-12-0204_1.pdf.

Milbrandt, J. A. and M. K. Yau, 2006: A Multimoment Bulk Microphysics Parameterization. Part III: Control Simulation of a Hailstorm. *Journal of the Atmospheric Sciences*, **63 (12)**, 3114–3136, doi:10.1175/JAS3816.1, https://journals.ametsoc.org/jas/article-pdf/63/12/3114/3486367/jas3816_1.pdf.

Miller, M. L., V. Lakshmanan, and T. M. Smith, 2013: An automated method for depicting mesocyclone paths and intensities. *Wea. Forecasting*, **28 (3)**, 570–585, doi:10.1175/waf-D-12-00065.1.

Molnar, C., 2019a: *Interpretable Machine Learning*. Bookdown, https://christophm.github.io/interpretable-ml-book/.

Molnar, C., 2019b: *Limitations of Interpretable Machine Learning Methods*. Bookdown, https://compstat-lmu.github.io/iml_methods_limitations/.

Molnar, C., G. Casalicchio, and B. Bischl, 2019: Quantifying Model Complexity via Functional Decomposition for Better Post-Hoc Interpretability. *arXiv*, 1904.03867.

Molnar, C., et al., 2020: Pitfalls to Avoid when Interpreting Machine Learning Models. *arXiv*, 2007.04131.

Morrison, H. and J. A. Milbrandt, 2015: Parameterization of Cloud Microphysics Based on the Prediction of Bulk Ice Particle Properties. Part I: Scheme Description and Idealized Tests. *Journal of the Atmospheric Sciences*, **72 (1)**, 287–311, doi:10.1175/JAS-D-14-0065.1, URL https://doi.org/10.1175/JAS-D-14-0065.1, https://journals.ametsoc.org/jas/article-pdf/72/1/287/3838519/jas-d-14-0065_1.pdf.

Murphy, A. H., 1991: Probabilities, Odds, and Forecasts of Rare Events. *Wea. Forecasting*, **6 (2)**, 302–307, doi:10.1175/1520-0434(1991)006⟨0302:poafor⟩2.0.co;2.

Murphy, A. H., 1993: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Wea. Forecasting*, **8 (2)**, 281–293, doi:10.1175/1520-0434(1993)008⟨0281:WIAGFA⟩2.0.CO;2, https://journals.ametsoc.org/WF/article-pdf/8/2/281/4650824/1520-0434(1993)008_0281_wiagfa_2_0_co_2.pdf.

NCEI, N., 2020: U.s. billion-dollar weather and climate disasters. URL https://www.ncdc.noaa.gov/billions/, doi:DOI:10.25921/stkw-7w73, URL https://www.ncdc.noaa.gov/billions/.

Neuhäuser, M., 2011: *Wilcoxon–Mann–Whitney Test*, 1656–1658. Springer Berlin Heidelberg, Berlin, Heidelberg, doi:10.1007/978-3-642-04898-2_615, URL https://doi.org/10.1007/978-3-642-04898-2_615.

Niculescu-Mizil, A. and R. Caruana, 2005: Predicting good probabilities with supervised learning. *Proceedings of the 22Nd International Conference on Machine Learning*, ACM, New York, NY, USA, 625–632, ICML '05, doi:10.1145/1102351.1102430.

Pedregosa, F., et al., 2011: Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Platt, J. C., 1999: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *ADVANCES IN LARGE MARGIN CLASSIFIERS*, MIT Press, 61–74.

Potvin, C. K., C. Broyles, P. S. Skinner, H. E. Brooks, and E. Rasmussen, 2019: A Bayesian Hierarchical Modeling Framework for Correcting Reporting Bias in the U.S. Tornado Database. *Wea. Forecasting*, **34 (1)**, 15–30, doi:10.1175/waf-D-18-0137.1.

Potvin, C. K. and M. L. Flora, 2015: Sensitivity of idealized supercell simulations to horizontal grid spacing: Implications for Warn-on-Forecast. *Mon. Wea. Rev.*, **143 (8)**, 2998–3024, doi:10.1175/mwr-d-14-00416.1.

Potvin, C. K. and L. J. Wicker, 2013: Assessing Ensemble Forecasts of Low-Level Supercell Rotation within an OSSE Framework. *Wea. Forecasting*, **28**, 940–960, doi: 10.1175/waf-D-12-00122.1.

Pucik, T., P. Groenemeijer, D. Rva, and M. Kol, 2015: Proximity Soundings of Severe and Nonsevere Thunderstorms in Central Europe. *Mon. Wea. Rev.*, **143 (12)**, 4805–4821, doi:10.1175/mwr-d-15-0104.1.

Radanovics, S., J.-P. Vidal, and E. Sauquet, 2018: Spatial verification of ensemble precipitation: An ensemble version of sal. *Wea. Forecasting*, **33 (4)**, 1001–1020, doi: 10.1175/waf-D-17-0162.1.

Rasmussen, E. N. and D. O. Blanchard, 1998: A Baseline Climatology of Sounding-Derived Supercell andTornado Forecast Parameters. *Wea. Forecasting*, **13 (4)**, 1148–1164, doi:10.1175/1520-0434(1998)013⟨1148:ABCOSD⟩2.0.CO;2, https://journals.ametsoc.org/WF/article-pdf/13/4/1148/4627972/1520-0434(1998)013_1148_abcosd_2_0_co_2.pdf.

Reap, R., 1974: Thunderstorm and severe weather probabilities based on model output statistics. preprints. Amer. Meteor. Soc., Fifth Conference on Forecasting and Analysis, St. Louis, MO, 266–269.

Reap, R. M. and D. S. Foster, 1979: Automated 1236 Hour Probability Forecasts of Thunderstorms and Severe Local Storms. *Journal of Applied Meteorology*, **18 (10)**, 1304–1315, doi:10.1175/1520-0450(1979)018⟨1304:ahpfot⟩2.0.co;2.

Ribeiro, M. T., S. Singh, and C. Guestrin, 2016: Model-Agnostic Interpretability of Machine Learning. *arXiv*, 1606.05386.

Roberts, B., M. Xue, and I. Dawson, Daniel T., 2020: The Effect of Surface Drag Strength on Mesocyclone Intensification and Tornadogenesis in Idealized Supercell Simulations. *Journal of the Atmospheric Sciences*, **77 (5)**, 1699–1721, doi:10.1175/JAS-D-19-0109.1, https://journals.ametsoc.org/jas/article-pdf/77/5/1699/4927478/jasd190109.pdf.

Roberts, B., M. Xue, A. D. Schenkman, and D. T. Daniel, 2016: The Role of Surface Drag in Tornadogenesis within an Idealized Supercell Simulation. *Journal of the Atmospheric Sciences*, **73 (9)**, 3371–3395, doi:10.1175/jas-d-15-0332.1.

Roebber, P., D. Schultz, B.A.Colle, and D.J.Stensrud, 2004: Toward improved prediction: High-resolution and ensemble modeling systems in operations. *Wea. Forecasting,*, **19**, 936–949.

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24 (2)**, 601–608, doi:10.1175/2008waf2222159.1, URL https://doi.org/10.1175/2008waf2222159.1.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams, 1985: Learning Internal Representations by Error Propagation. *The Problem, The Generalized Delta Rule, Simulation Results, Some Further Generalizations, Conclusion*, doi:10.21236/ada164453.

Schwartz, C. S. and R. A. Sobash, 2017: Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: A review and recommendations. *Mon. Wea. Rev.*, **145 (9)**, 3397–3418, doi:10.1175/mwr-d-16-0400.1.

Scott, D. W., 1992: *Multivariate Density Estimation: Theory,Practice, and Visualization.*, 360 pp. John Wiley and Sons.

Shapley, L. S., 1953: A value for n-person games. *Contributions to the Theory of Games*, **2.28**, 307–317.

Skamarock, W. and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note,NCAR/TN-4751STR, NCAR/MMM. 113 pp., doi:10.5065/D68S4MVH.

Skinner, P. S., C. C. Weiss, L. J. Wicker, C. K. Potvin, and D. C. Dowell, 2015: Forcing Mechanisms for an Internal Rear-Flank Downdraft Momentum Surge in the 18 May 2010 Dumas, Texas, Supercell. *Mon. Wea. Rev.*, **143 (11)**, 4305–4330, doi:10.1175/mwr-d-15-0164.1.

Skinner, P. S., L. J. Wicker, D. M. Wheatley, and K. H. Knopfmeier, 2016: Application of two spatial verification methods to ensemble forecasts of low-level rotation. *Wea. Forecasting*, **31 (3)**, 713–735, doi:10.1175/waf-D-15-0129.1.

Skinner, P. S., et al., 2018: Object-based verification of a prototype warn-on-forecast system. *Wea. Forecasting*, **33 (5)**, 1225–1250, doi:10.1175/waf-D-18-0020.1.

Smith, B. T., T. E. Castellanos, A. C. Winters, C. M. Mead, A. R. Dean, and R. L. Thompson, 2013: Measured Severe Convective Wind Climatology and Associated Convective Modes of Thunderstorms in the Contiguous United States, 200309. *Wea. Forecasting*, **28 (1)**, 229–236, doi:10.1175/waf-d-12-00096.1.

Smith, B. T., R. L. Thompson, J. S. Grams, C. Broyles, and H. E. Brooks, 2012: Convective Modes for Significant Severe Thunderstorms in the Contiguous United States. Part I: Storm Classification and Climatology. *Wea. Forecasting*, **27 (5)**, 1114–1135, doi:10.1175/waf-d-11-00115.1.

Smith, T. M. and K. L. Elmore, 2004: The use of radial velocity derivatives to diagnose rotation and divergence. *11th Conf. onAviation, Range, and Aerospace*, Hyannis, MA, Amer. Meteor. Soc.,, P5.6, URL https://ams.confex.com/ams/pdfpapers/81827.pdf.

Smith, T. M., et al., 2016: Multi-radar multi-sensor (mrms) severe weather and aviation products: Initial operating capabilities. *BAMS*, **97 (9)**, 1617–1630, doi:10.1175/BAMS-D-14-00173.1.

Snook, N., Y. Jung, J. Brotzge, B. Putnam, and M. Xue, 2016: Prediction and ensemble forecast verification of hail in the supercell storms of 20 may 2013. *Wea. Forecasting*, **31 (3)**, 811–825, doi:10.1175/waf-D-15-0152.1.

Snook, N., M. Xue, and Y. Jung, 2012: Ensemble probabilistic forecasts of a tornadic mesoscale convective system from ensemble kalman filter analyses using wsr-88d and casa radar data. *Mon. Wea. Rev.*, **140 (7)**, 2126–2146, doi:10.1175/MWR-D-11-00117.1.

Sobash, R. A. and J. S. Kain, 2017: Seasonal Variations in Severe Weather Forecast Skill in an Experimental Convection-Allowing Model. *Wea. Forecasting*, **32 (5)**, 1885–1902, doi:10.1175/waf-d-17-0043.1.

Sobash, R. A., J. S. Kain, D. R. Bright, A. R. Dean, M. C. Coniglio, and S. J. Weiss, 2011: Probabilistic Forecast Guidance for Severe Thunderstorms Based on the Identification of Extreme Phenomena in Convection-Allowing Model Forecasts. *Wea. Forecasting*, **26 (5)**, 714–728, doi:10.1175/waf-D-10-05046.1, URL https://doi.org/10.1175/waf-D-10-05046.1, https://journals.ametsoc.org/waf/article-pdf/26/5/714/4648500/waf-d-10-05046_1.pdf.

Sobash, R. A., G. S. Romine, and C. S. Schwartz, 2020: A comparison of neural-network and surrogate-severe probabilistic convective hazard guidance derived from a convection-allowing model. *Wea. Forecasting*, 1–54, doi:10.1175/waf-d-20-0036.1.

Sobash, R. A., G. S. Romine, C. S. Schwartz, D. J. Gagne, and M. L. Weisman, 2016a: Explicit forecasts of low-level rotation from convection-allowing models for next-day tornado prediction. *Wea. Forecasting*, **31 (5)**, 1591–1614, doi:10.1175/waf-D-16-0073.1.

Sobash, R. A., C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016b: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Wea. Forecasting*, **31 (1)**, 255–271, doi:10.1175/waf-D-15-0138.1.

Sobash, R. A., C. S. Schwartz, G. S. Romine, and M. L. Weisman, 2019: Next-Day Prediction of Tornadoes Using Convection-Allowing Models with 1-km Horizontal Grid Spacing. *Wea. Forecasting*, **34 (4)**, 1117–1135, doi:10.1175/waf-D-19-0044.1, URL https://doi.org/10.1175/waf-D-19-0044.1, https://journals.ametsoc.org/waf/article-pdf/34/4/1117/4883083/waf-d-19-0044_1.pdf.

SPC, 2020: URL https://www.spc.noaa.gov/new/SVRclimo/climo.php?parm=anySvr, URL https://www.spc.noaa.gov/new/SVRclimo/climo.php?parm=anySvr.

Steinkruger, D., P. Markowski, and G. Young, 2020: An Artificially Intelligent System for the Automated Issuance of Tornado Warnings in Simulated Convective Storms. *Wea. Forecasting*, **35 (5)**, 1939–1965, doi:10.1175/waf-D-19-0249.1, https://journals.ametsoc.org/waf/article-pdf/35/5/1939/4995927/wafd190249.pdf.

Stensrud, D. J., et al., 2009: Convective-scale warn-on-forecast system. *BAMS*, **90 (10)**, 1487–1500, doi:10.1175/2009BAMS2795.1, URL https://doi.org/10.1175/2009BAMS2795.1.

Stensrud, D. J., et al., 2013: Progress and challenges with Warn-on-Forecast. *AR*, **123**, 2–16, doi:10.1016/j.atmosres.2012.04.004.

Stratman, D. R. and K. A. Brewster, 2017: Sensitivities of 1-km forecasts of 24 may 2011 tornadic supercells to microphysics parameterizations. *Mon. Wea. Rev.*, **145 (7)**, 2697–2721, doi:10.1175/MWR-D-16-0282.1.

Stumpf, G. J., A. Witt, E. D. Mitchell, P. L. Spencer, J. T. Johnson, M. D. Eilts, K. W. Thomas, and D. W. Burgess, 1998: The National Severe Storms Laboratory Mesocyclone Detection Algorithm for the WSR-88D*. *Wea. Forecasting*, **13 (2)**, 304–326, doi:10.1175/1520-0434(1998)013⟨0304:TNSSLM⟩2.0.CO;2, https://journals.ametsoc.org/WF/article-pdf/13/2/304/4626931/1520-0434(1998)013_0304_tnsslm_2_0_co_2.pdf.

Sun, Y., A. K. C. Wong, and M. S. KAMEL, 2009: CLASSIFICATION OF IMBALANCED DATA: A REVIEW. *International Journal of Pattern Recognition and Artificial Intelligence*, **23 (04)**, 687–719, doi:10.1142/s0218001409007326.

Taszarek, M., J. T. Allen, T. Pik, K. A. Hoogewind, and H. E. Brooks, 2020: Severe Convective Storms across Europe and the United States. Part II: ERA5 Environments Associated with Lightning, Large Hail, Severe Wind, and Tornadoes. *Journal of Climate*, **33 (23)**, 10 263–10 286, doi:10.1175/JCLI-D-20-0346.1, https://journals.ametsoc.org/jcli/article-pdf/33/23/10263/5014217/jclid200346.pdf.

Thompson, R. L., R. Edwards, J. A. Hart, K. L. Elmore, and P. Markowski, 2003: Close Proximity Soundings within Supercell Environments Obtained from the Rapid Update Cycle. *Wea. Forecasting*, **18 (6)**, 1243–1261, doi:10.1175/1520-0434(2003)018⟨1243:cpswse⟩2.0.co;2.

Trafalis, T., M. Richman, A. White, and B. Santosa, 2013: Data mining techniques for improved WSR-88D rainfall estimation. *Computers & Industrial Engineering*, **43 (4)**, 775–786, doi:10.1016/s0360-8352(02)00139-0.

Trafalis, T., B. Santosa, and Richman, 2005: Learning networks for tornado forecasting: a Bayesian perspective. *WIT Transactions on Information and Communication Technologies*, **35**.

Trafalis, T. B., I. Adrianto, and M. B. Richman, 2007: Active Learning with Support Vector Machines for Tornado Prediction. *International Conference on Computational Science*, **4487**, 1130–1137, doi:10.1007/978-3-540-72584-8_148.

Trafalis, T. B., H. Ince, and M. B. Richman, 2003: Computational Science  ICCS 2003, International Conference, Melbourne, Australia and St. Petersburg, Russia, June 24, 2003 Proceedings, Part IV. *ICCS*, 289–298, doi:10.1007/3-540-44864-0_30.

Trapp, R. J., G. J. Stumpf, and K. L. Manross, 2005: A reassessment of the percentage of tornadic mesocyclones. *Wea. Forecasting,*, **20 (4)**, 680–687, doi:10.1175/WF864.1.

Trapp, R. J., D. M. Wheatley, N. T. Atkins, R. W. Przybylinski, and R. Wolf, 2006: Buyer beware: Some words of caution on the use of severe wind reports in postevent assessment and research. *Wea. Forecasting,*, **21 (3)**, 408–415, doi:10.1175/WF925.1.

Tuovinen, J.-P., J. Rauhala, and D. M. Schultz, 2015: Significant-Hail-Producing Storms in Finland: Convective-Storm Environment and Mode. *Wea. Forecasting*, **30 (4)**, 1064–1076, doi:10.1175/waf-d-14-00159.1.

van der Laan, M. J., 2006: Statistical inference for variable importance. *The International Journal of Biostatistics*, **2 (1)**, doi:https://doi.org/10.2202/1557-4679.1008.

Verbout, S. M., H. E. Brooks, L. M. Leslie, and D. M. Schultz, 2006: Evolution of the u.s. tornado database: 1954-2003. *Wea. Forecasting,*, **21 (1)**, 86–93, doi:10.1175/WF910.1.

Wakimoto, R., 2001: *Convectively Driven High Wind Events*, 255–298. doi:10.1007/978-1-935704-06-5_7.

Wakimoto, R. M., 1985: Forecasting Dry Microburst Activity over the High Plains. *Mon. Wea. Rev.*, **113 (7)**, 1131–1143, doi:10.1175/1520-0493(1985)113⟨1131:FDMAOT⟩2.0.CO;2, https://journals.ametsoc.org/mwr/article-pdf/113/7/1131/4168770/1520-0493(1985)113_1131_fdmaot_2_0_co_2.pdf.

Weisman, M. L., 1993: The Genesis of Severe, Long-Lived Bow Echoes. *Journal of the Atmospheric Sciences*, **50 (4)**, 645–670, doi:10.1175/1520-0469(1993)050⟨0645:TGOSLL⟩2.0.CO;2, https://journals.ametsoc.org/jas/article-pdf/50/4/645/3426741/1520-0469(1993)050_0645_tgosll_2_0_co_2.pdf.

Wheatley, D. M., K. H. Knopfmeier, T. A. Jones, and G. J. Creager, 2015: Storm-scale data assimilation and ensemble forecasting with the NSSL experimental Warn-on-Forecast system. Part I: Radar data experiments. *Wea. Forecasting*, **30 (6)**, 1795–1817, doi:10.1175/waf-d-15-0043.1.

Wilson, K. A., et al., 2019: Exploring applications of storm-scale probabilistic warn-on-forecast guidance in weather forecasting. *Lecture Notes in Computer Science*, **11575**, 577–572, doi:10.1007/978-3-030-21565-1_39.

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, doi:10.1016/j.atmosres.2020.105093.

Yokota, S., H. Niino, H. Seko, M. Kunii, and H. Yamauchi, 2018: Important Factors for Tornadogenesis as Revealed by High-Resolution Ensemble Forecasts of the Tsukuba Supercell Tornado of 6 May 2012 in Japan. *Mon. Wea. Rev.*, **146**, doi:10.1175/MWR-D-17-0254.1.

Young, H. P., 1985: Monotonic solutions of cooperative games. *International Journal of Game Theory*, **14.2**, 65–72.

Yussouf, N., D. C. Dowell, L. J. Wicker, K. H. Knopfmeier, and D. M. Wheatley, 2015: Storm-scale data assimilation and ensemble forecasts for the 27 April 2011 severe weather outbreak in Alabama. *Mon. Wea. Rev.*, **143**, 3044–3066, doi:10.1175/MWR-D-14-00268.1.

Yussouf, N., J. Gao, D. J. Stensrud, and G. Ge, 2013a: The impact of mesoscale environmental uncertainty on the prediction of a tornadic supercell storm using ensemble data assimilation approach. *Advances in Meteorology*, **2013**, 1–15, doi:10.1155/2013/731647.

Yussouf, N., J. S. Kain, and A. J. Clark, 2016: Short-term probabilistic forecasts of the 31 May 2013 Oklahoma tornado and flash flood event using a continuous-update-cycle storm-scale ensemble system. *Wea. Forecasting*, **31 (3)**, 957–983, doi:10.1175/waf-d-15-0160.1.

Yussouf, N., E. R. Mansell, L. J. Wicker, D. M. Wheatley, and D. J. Stensrud, 2013b: The ensemble kalman filter analyses and forecasts of the 8 may 2003 oklahoma city tornadic supercell storm using single- and double-moment microphysics schemes. *Mon. Wea. Rev.*, 130318145938008, doi:10.1175/mwr-d-12-00237.1.

Yussouf, N., K. A. Wilson, S. M. Martinaitis, H. Vergara, P. L. Heinselman, and J. J. Gourley, 2020: The coupling of nssl warn-on-forecast and flash systems for probabilistic flash flood prediction. *Journal of Hydrometeorology*, **21 (1)**, 123–141, doi:10.1175/JHM-D-19-0131.1.

# APPENDICES

# Appendix A: Deriving the Maximum Critical Success Index for a No-Skill System

This theorem constitutes an original work as I am unaware of any prior attempt to prove the maximum critical success index for a no-skill system is equal to the climatological event frequency.

**Theorem 1.** *For a forecast system that predicts 0 for all $y = 1$ (i.e., a no-skill system) where $y$ is the set of binary outcome variables, the corresponding maximum CSI is equal to the climatological event frequency, $c$.*

*Proof.* From Roebber (2009), the CSI can be defined as a function of success ratio ($s$) and probability of detection ($p$):

$$CSI = \frac{1}{s^{-1} + p^{-1} - 1} \tag{A.1}$$

Substitute the minimum success ratio for a no-skill system (equation 2.2), into equation 2.6

$$CSI = \frac{1}{\frac{1-c+cp}{cp} + \frac{1}{p} - 1}. \tag{A.2}$$

Multiply the numerator and denominator by $cp$,

$$CSI = \frac{\pi p}{1 - c + cp + c - cp} \tag{A.3}$$

Cancel the terms in the denominator:

$$CSI = cp. \tag{A.4}$$

Based on equation 2.9, the maximum CSI of a no-skill system occurs for $p = 1$ and is equal to climatological event frequency ($c$). □

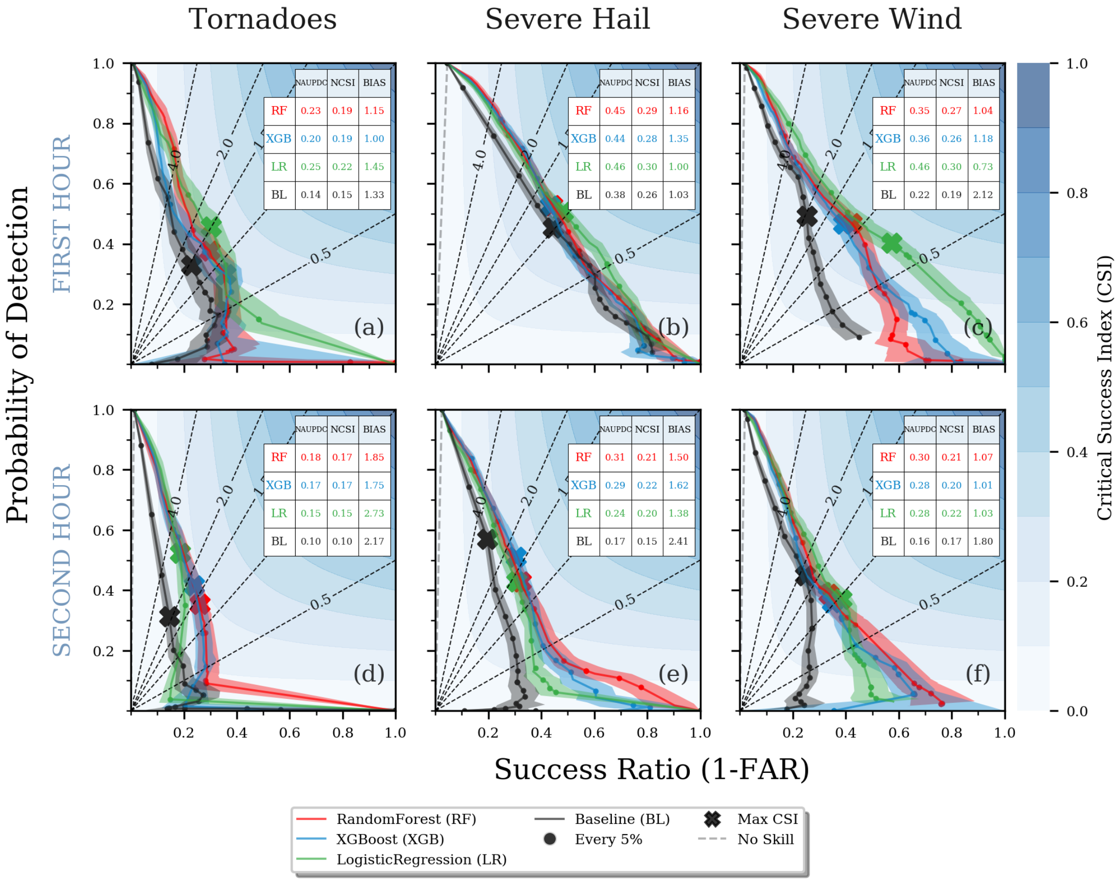
## Appendix B: WoFS forecast dates used for ML

| | | | | |
|---|---|---|---|---|
| 01 May 2017 | 02 May 2017 | 03 May 2017 | 04 May 2017 | 08 May 2017 |
| 09 May 2017 | 11 May 2017 | 15 May 2017 | 16 May 2017 | 17 May 2017 |
| 18 May 2017 | 19 May 2017 | 22 May 2017 | 23 May 2017 | 24 May 2017 |
| 25 May 2017 | 26 May 2017 | 27 May 2017 | 30 May 2017 | 01 June 2017 |
| 02 June 2017 | 29 April 2018 | 01 May 2018 | 02 May 2018 | 03 May 2018 |
| 04 May 2018 | 07 May 2018 | 09 May 2018 | 10 May 2018 | 11 May 2018 |
| 12 May 2018 | 14 May 2018 | 15 May 2018 | 16 May 2018 | 19 May 2018 |
| 21 May 2018 | 23 May 2018 | 24 May 2018 | 25 May 2018 | 27 May 2018 |
| 28 May 2018 | 29 May 2018 | 30 May 2018 | 31 May 2018 | 01 June 2018 |
| 19 June 2018 | 20 June 2018 | 21 June 2018 | 22 June 2018 | 23 June 2018 |
| 24 June 2018 | 25 June 2018 | 27 June 2018 | 28 June 2018 | 29 June 2018 |
| 30 June 2018 | 30 April 2019 | 01 May 2019 | 02 May 2019 | 03 May 2019 |
| 06 May 2019 | 07 May 2019 | 08 May 2019 | 09 May 2019 | 10 May 2019 |
| 13 May 2019 | 14 May 2019 | 15 May 2019 | 16 May 2019 | 17 May 2019 |
| 18 May 2019 | 20 May 2019 | 21 May 2019 | 22 May 2019 | 23 May 2019 |
| 24 May 2019 | 25 May 2019 | 26 May 2019 | 28 May 2019 | 29 May 2019 |
| 30 May 2019 | | | | |

Table B.1: Complete list of dates used from the WoFS used.

# Appendix C: Performance after Dropping Predictors

Figure C.1 and Figure C.2 shows the performance and attribute diagram, respectively, once correlated predictors were removed (see section 6.1). When comparing with Figure 7.9 and Figure C.2, respectively, we can see that any drop in skill is not substantial. However, retraining the ML models on their respective top 15 predictors (see section 7.3.1) caused a substantial drop in skill (even below the baseline predictions; Figure C.3). The drop in skill is most substantial for tornadoes (Figure C.3a,d), and less so for severe wind (Figure C.3b,e) and severe hail (Figure C.3c,f). As was shown in section 7.3.1, the tornado prediction rely on more than the top 15 predictors while severe wind and severe hail strongly rely on less than 10 predictors.



(a)

|  | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.23 | 0.19 | 1.15 |
| XGB | 0.20 | 0.19 | 1.00 |
| LR | 0.25 | 0.22 | 1.45 |
| BL | 0.14 | 0.15 | 1.33 |

(b)

|  | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.45 | 0.29 | 1.16 |
| XGB | 0.44 | 0.28 | 1.35 |
| LR | 0.46 | 0.30 | 1.00 |
| BL | 0.38 | 0.26 | 1.03 |

(c)

|  | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.35 | 0.27 | 1.04 |
| XGB | 0.36 | 0.26 | 1.18 |
| LR | 0.46 | 0.30 | 0.73 |
| BL | 0.22 | 0.19 | 2.12 |

(d)

|  | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.18 | 0.17 | 1.85 |
| XGB | 0.17 | 0.17 | 1.75 |
| LR | 0.15 | 0.15 | 2.73 |
| BL | 0.10 | 0.10 | 2.17 |

(e)

|  | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.31 | 0.21 | 1.50 |
| XGB | 0.29 | 0.22 | 1.62 |
| LR | 0.24 | 0.20 | 1.38 |
| BL | 0.17 | 0.15 | 2.41 |

(f)

|  | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.30 | 0.21 | 1.07 |
| XGB | 0.28 | 0.20 | 1.01 |
| LR | 0.28 | 0.22 | 1.03 |
| BL | 0.16 | 0.17 | 1.80 |

Figure C.1: Same as in Figure 7.9, but when highly correlated predictors were removed.

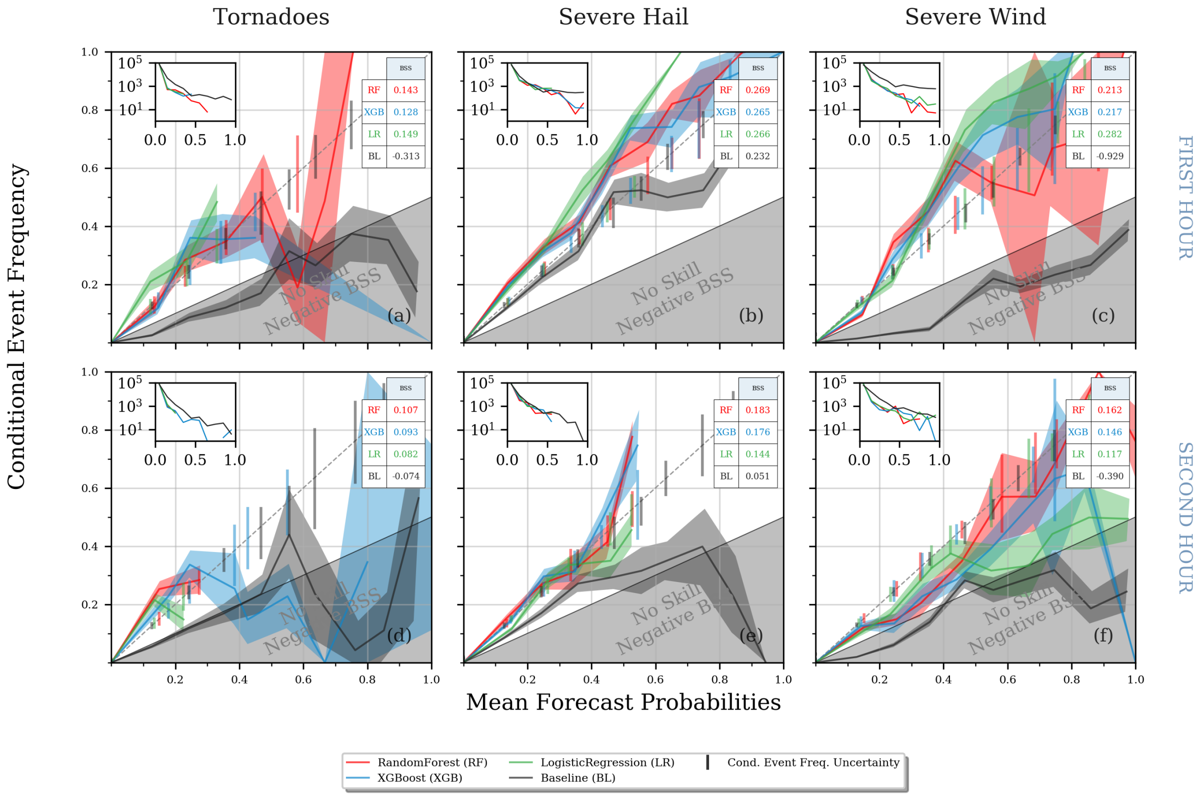
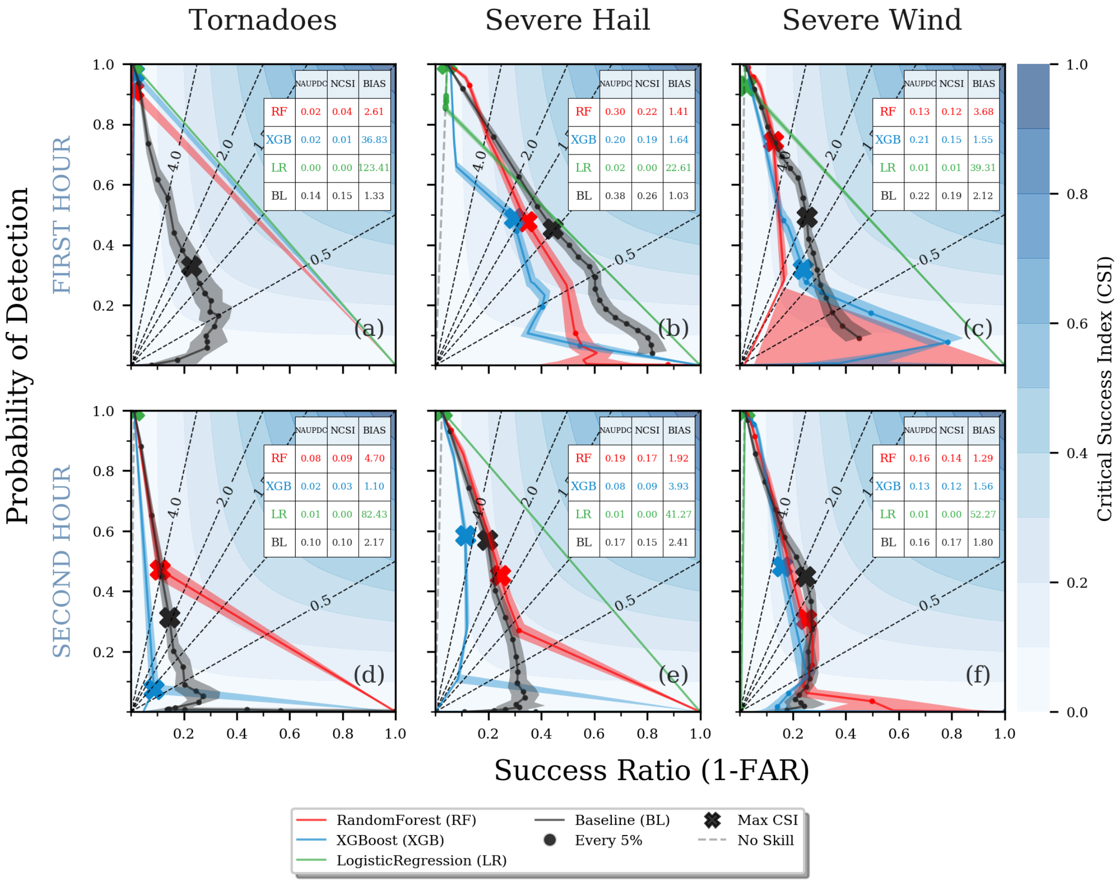


Figure C.2: Same as in Figure 7.11, but when highly correlated predictors were removed.

Figure C.4 and Figure C.5 show the performance and attribute diagrams after the storm morphological predictors were dropped (see Table 5.1). Any drop in skill as compared to Figure 7.9 and Figure C.2, respectively, is unsubstantial.



| | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.02 | 0.04 | 2.61 |
| XGB | 0.02 | 0.01 | 36.83 |
| LR | 0.00 | 0.00 | 123.41 |
| BL | 0.14 | 0.15 | 1.33 |

| | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.30 | 0.22 | 1.41 |
| XGB | 0.20 | 0.19 | 1.64 |
| LR | 0.02 | 0.00 | 22.61 |
| BL | 0.38 | 0.26 | 1.03 |

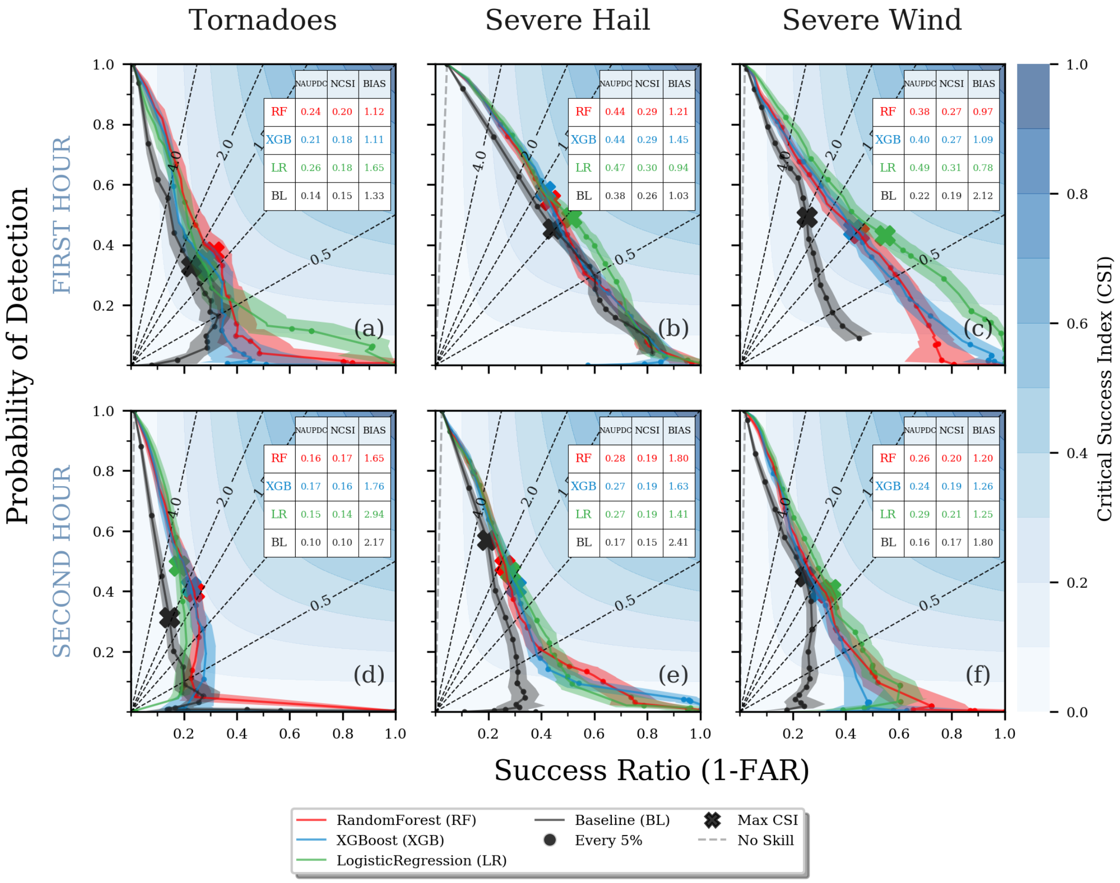
| | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.13 | 0.12 | 3.68 |
| XGB | 0.21 | 0.15 | 1.55 |
| LR | 0.01 | 0.01 | 39.31 |
| BL | 0.22 | 0.19 | 2.12 |

| | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.08 | 0.09 | 4.70 |
| XGB | 0.02 | 0.03 | 1.10 |
| LR | 0.01 | 0.00 | 82.43 |
| BL | 0.10 | 0.10 | 2.17 |

| | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.19 | 0.17 | 1.92 |
| XGB | 0.08 | 0.09 | 3.93 |
| LR | 0.01 | 0.00 | 41.27 |
| BL | 0.17 | 0.15 | 2.41 |

| | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.16 | 0.14 | 1.29 |
| XGB | 0.13 | 0.12 | 1.56 |
| LR | 0.01 | 0.00 | 52.27 |
| BL | 0.16 | 0.17 | 1.80 |

Figure C.3: Same as in Figure 7.9, but when only the top 15 predictors as determined by the multi-pass permutation importance were retained.



(a)

| | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.24 | 0.20 | 1.12 |
| XGB | 0.21 | 0.18 | 1.11 |
| LR | 0.26 | 0.18 | 1.65 |
| BL | 0.14 | 0.15 | 1.33 |

(b)

| | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.44 | 0.29 | 1.21 |
| XGB | 0.44 | 0.29 | 1.45 |
| LR | 0.47 | 0.30 | 0.94 |
| BL | 0.38 | 0.26 | 1.03 |

(c)

| | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.38 | 0.27 | 0.97 |
| XGB | 0.40 | 0.27 | 1.09 |
| LR | 0.49 | 0.31 | 0.78 |
| BL | 0.22 | 0.19 | 2.12 |

(d)

| | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.16 | 0.17 | 1.65 |
| XGB | 0.17 | 0.16 | 1.76 |
| LR | 0.15 | 0.14 | 2.94 |
| BL | 0.10 | 0.10 | 2.17 |

(e)

| | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.28 | 0.19 | 1.80 |
| XGB | 0.27 | 0.19 | 1.63 |
| LR | 0.27 | 0.19 | 1.41 |
| BL | 0.17 | 0.15 | 2.41 |

(f)

| | NAUPDC | NCSI | BIAS |
|---|---|---|---|
| RF | 0.26 | 0.20 | 1.20 |
| XGB | 0.24 | 0.19 | 1.26 |
| LR | 0.29 | 0.21 | 1.25 |
| BL | 0.16 | 0.17 | 1.80 |

Figure C.4: Same as in Figure 7.9, but when highly correlated predictors were removed.

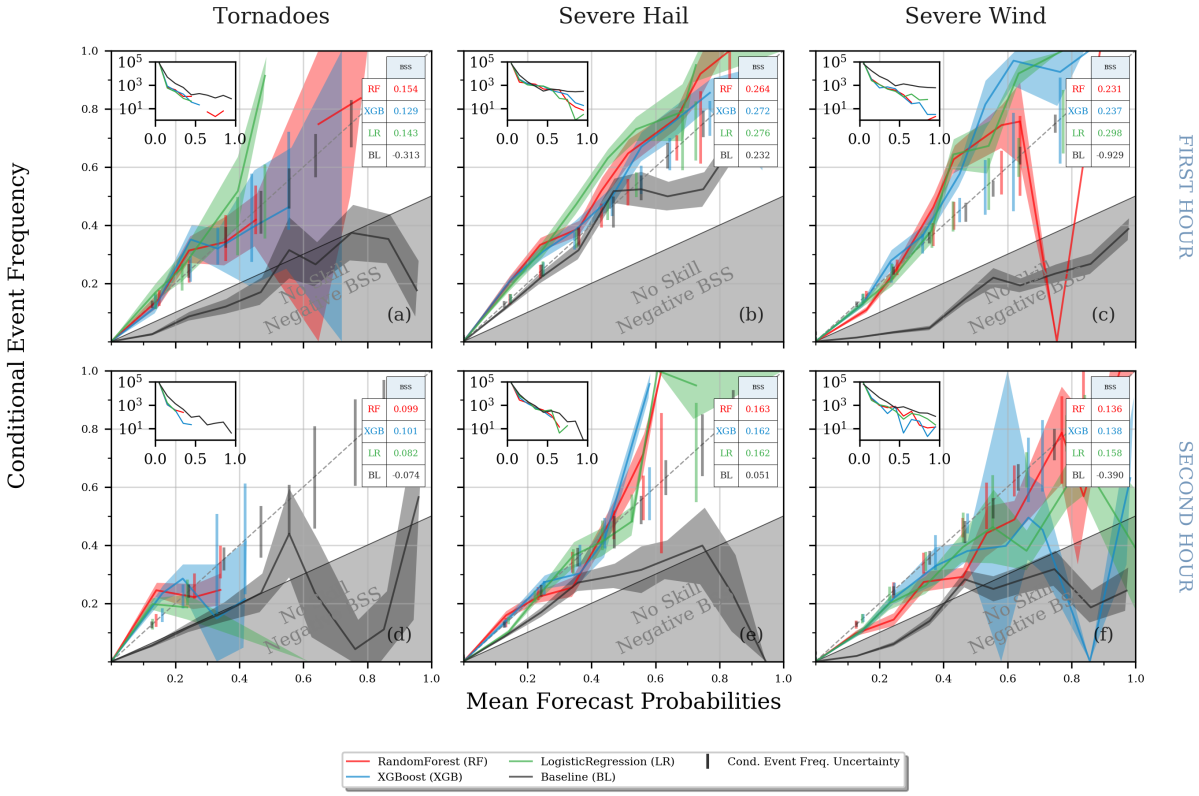


Figure C.5: Same as in Figure 7.11, but when highly correlated predictors were removed.

# Appendix D: Performance based on Resampling Training Dataset

In section 7.2.1, I discussed sensitivity to the class imbalance on the testing performance of the ML models. Figure D.1 and Figure D.2 shows the performance and attribute diagrams for the severe hail and severe wind fit on a training dataset where the minority class had been randomly undersampled and tornado prediction where the models were fit on the original training dataset. Training on resampled data made a negligible impact for the severe wind and severe hail prediction (cf. Figure 7.9b,c,e,f and Figure D.1b,c,e,f) while training on the resampled data for tornado prediction did make a small improvement (cf. Figure 7.9a,d and Figure D.1a,d).

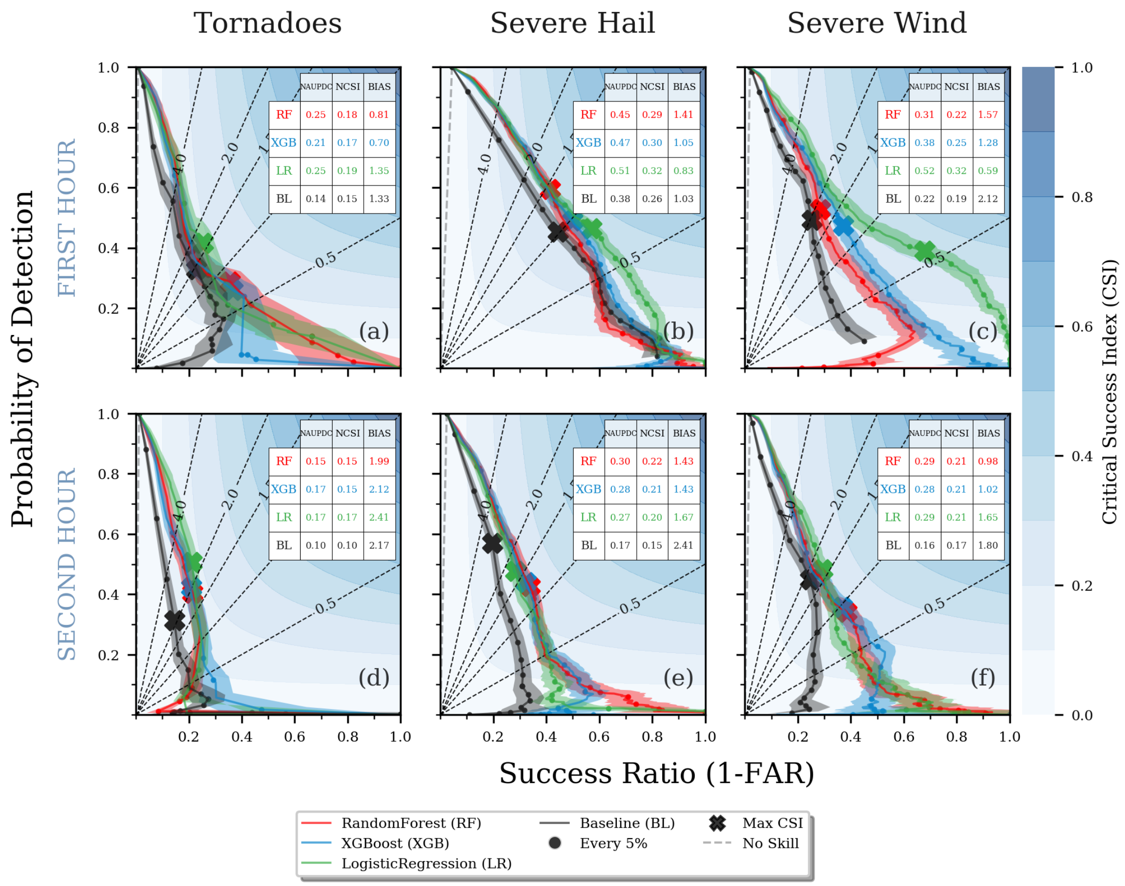


Figure D.1: Same as in Figure 7.9, but for tornado prediction where the ML models were fit on the original, unaltered training dataset and the severe wind and severe hail were fit on a training dataset where the minority class was randomly subsampled.

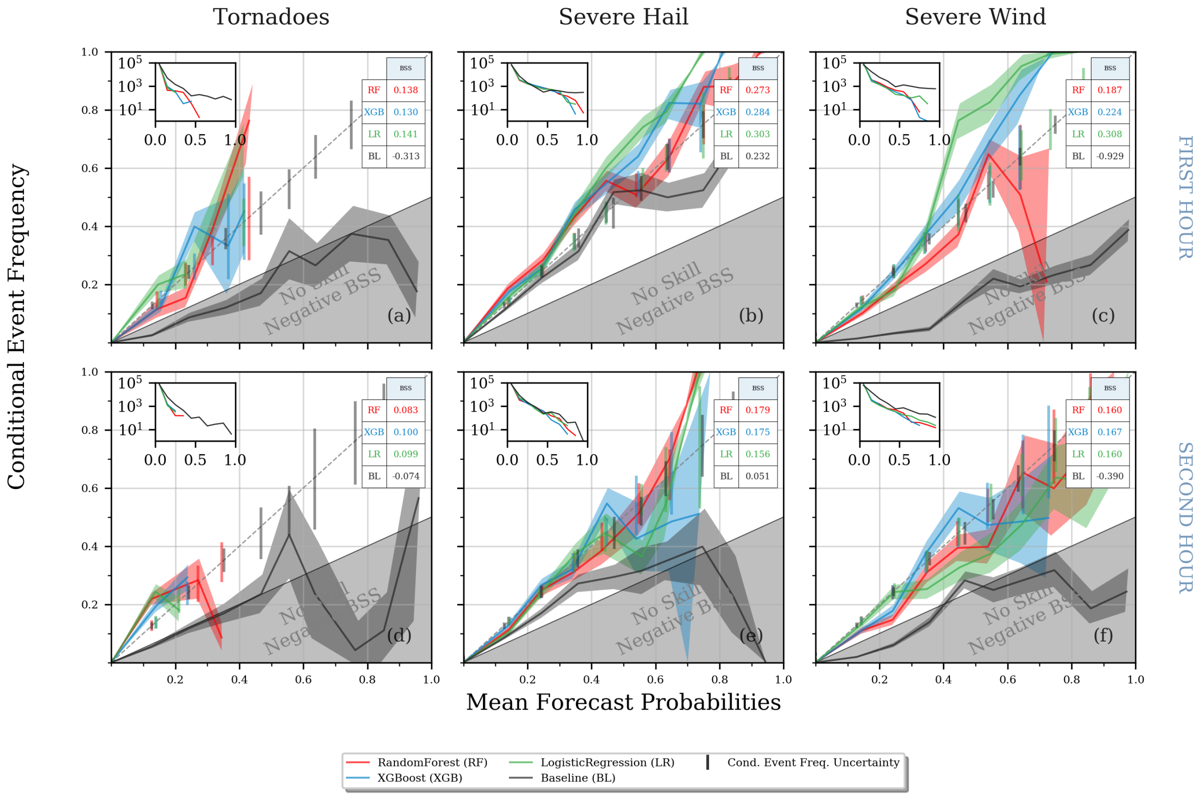


Figure D.2: Same as in Figure 7.11, but for tornado prediction where the ML models were fit on the original, unaltered training dataset and the severe wind and severe hail were fit on a training dataset where the minority class was randomly subsampled.

# Appendix E: Additional Figures

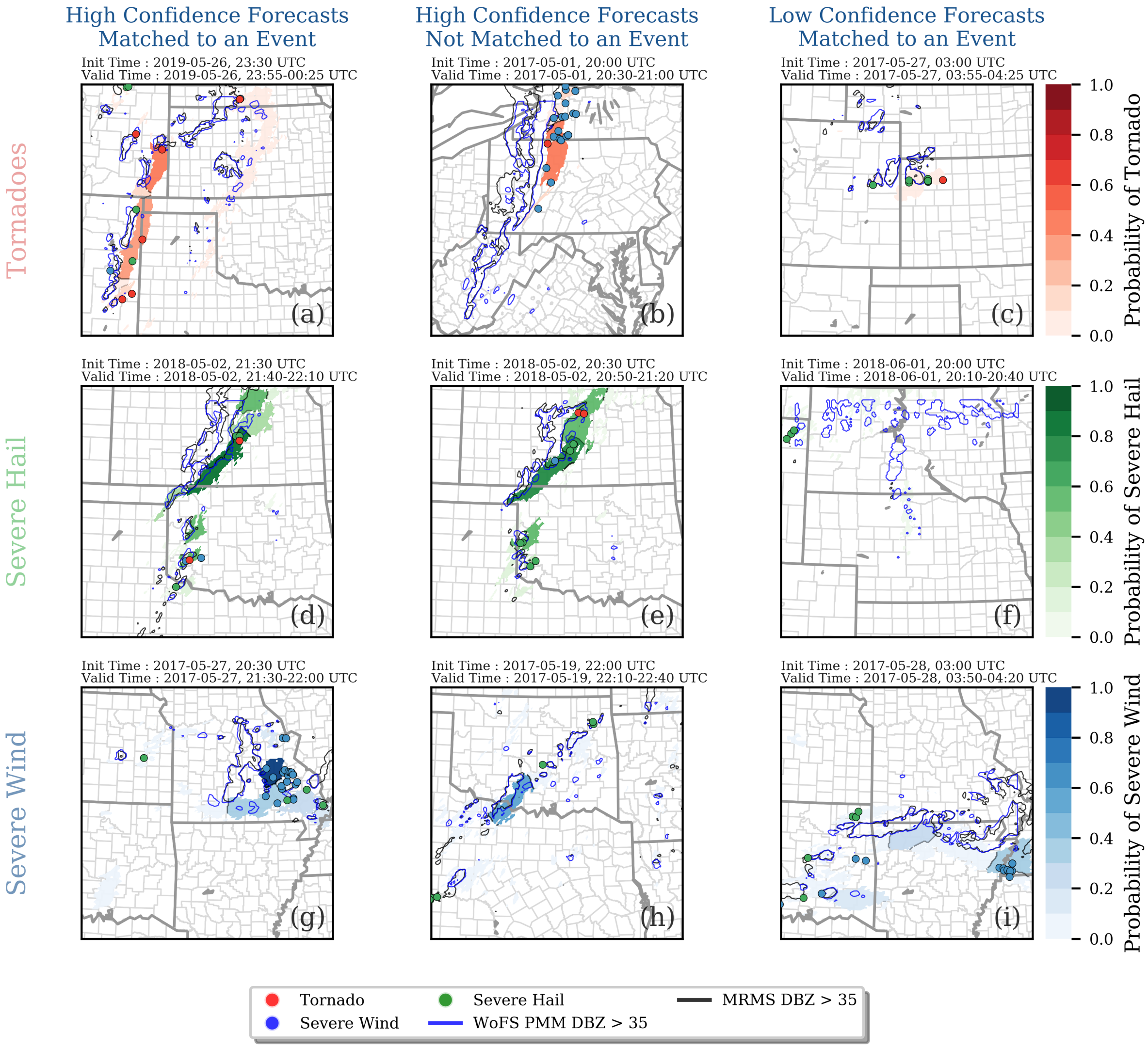

High Confidence Forecasts
Matched to an Event
High Confidence Forecasts
Not Matched to an Event
Low Confidence Forecasts
Matched to an Event
Tornadoes
Severe Hail
Severe Wind
Init Time : 2019-05-26, 23:30 UTC
Valid Time : 2019-05-26, 23:55-00:25 UTC
(a)
Init Time : 2017-05-01, 20:00 UTC
Valid Time : 2017-05-01, 20:30-21:00 UTC
(b)
Init Time : 2017-05-27, 03:00 UTC
Valid Time : 2017-05-27, 03:55-04:25 UTC
(c)
Probability of Tornado
1.0
0.8
0.6
0.4
0.2
0.0
Init Time : 2018-05-02, 21:30 UTC
Valid Time : 2018-05-02, 21:40-22:10 UTC
(d)
Init Time : 2018-05-02, 20:30 UTC
Valid Time : 2018-05-02, 20:50-21:20 UTC
(e)
Init Time : 2018-06-01, 20:00 UTC
Valid Time : 2018-06-01, 20:10-20:40 UTC
(f)
Probability of Severe Hail
1.0
0.8
0.6
0.4
0.2
0.0
Init Time : 2017-05-27, 20:30 UTC
Valid Time : 2017-05-27, 21:30-22:00 UTC
(g)
Init Time : 2017-05-19, 22:00 UTC
Valid Time : 2017-05-19, 22:10-22:40 UTC
(h)
Init Time : 2017-05-28, 03:00 UTC
Valid Time : 2017-05-28, 03:50-04:20 UTC
(i)
Probability of Severe Wind
1.0
0.8
0.6
0.4
0.2
0.0
● Tornado
● Severe Wind
● Severe Hail
WoFS PMM DBZ > 35
MRMS DBZ > 35

Figure E.1: Same as Figure 7.7, but for the XGBoost model.

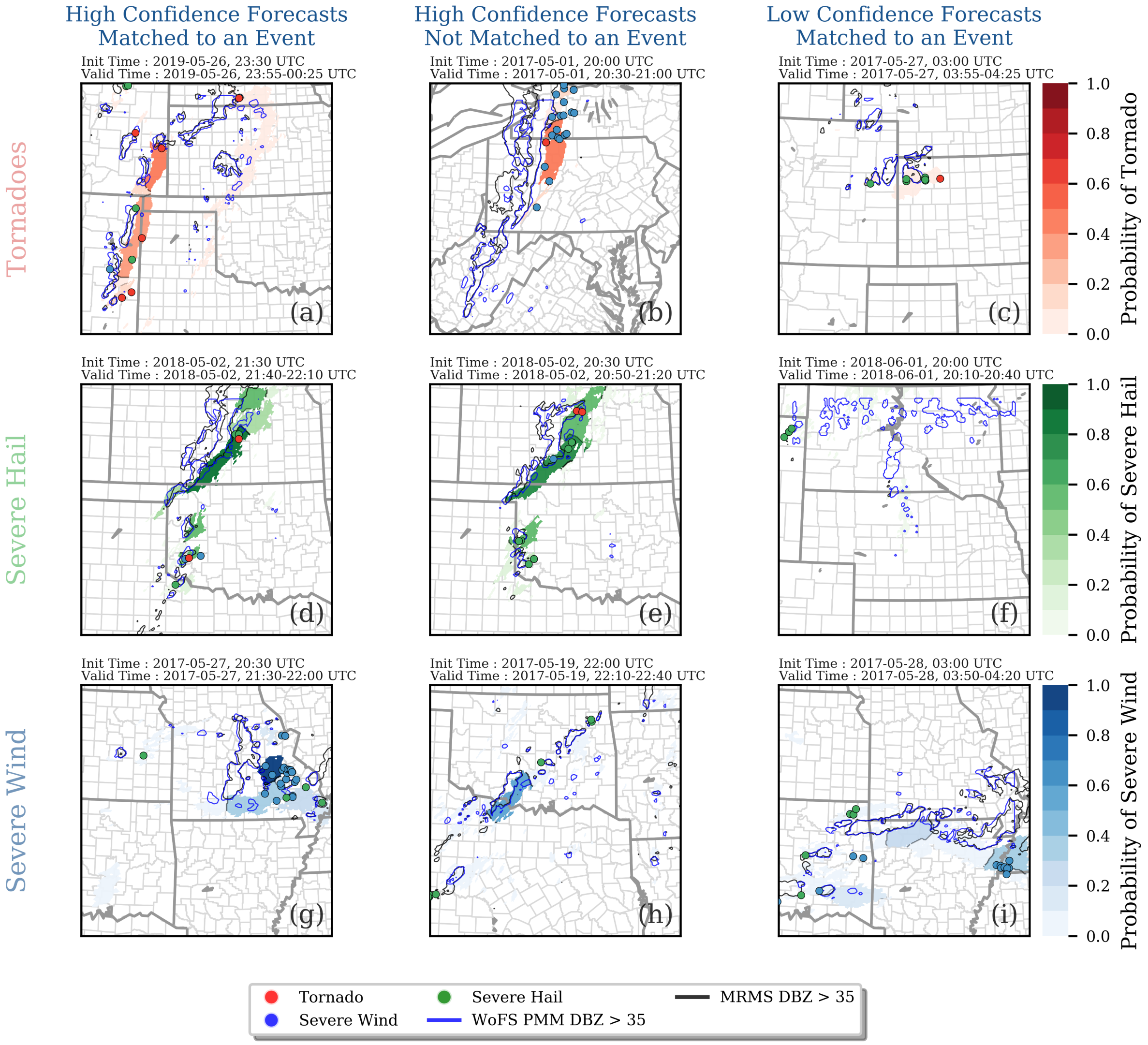

High Confidence Forecasts Matched to an Event
High Confidence Forecasts Not Matched to an Event
Low Confidence Forecasts Matched to an Event
Tornadoes
Severe Hail
Severe Wind
Init Time : 2019-05-26, 23:30 UTC
Valid Time : 2019-05-26, 23:55-00:25 UTC
(a)
Init Time : 2017-05-01, 20:00 UTC
Valid Time : 2017-05-01, 20:30-21:00 UTC
(b)
Init Time : 2017-05-27, 03:00 UTC
Valid Time : 2017-05-27, 03:55-04:25 UTC
(c)
Probability of Tornado
Init Time : 2018-05-02, 21:30 UTC
Valid Time : 2018-05-02, 21:40-22:10 UTC
(d)
Init Time : 2018-05-02, 20:30 UTC
Valid Time : 2018-05-02, 20:50-21:20 UTC
(e)
Init Time : 2018-06-01, 20:00 UTC
Valid Time : 2018-06-01, 20:10-20:40 UTC
(f)
Probability of Severe Hail
Init Time : 2017-05-27, 20:30 UTC
Valid Time : 2017-05-27, 21:30-22:00 UTC
(g)
Init Time : 2017-05-19, 22:00 UTC
Valid Time : 2017-05-19, 22:10-22:40 UTC
(h)
Init Time : 2017-05-28, 03:00 UTC
Valid Time : 2017-05-28, 03:50-04:20 UTC
(i)
Probability of Severe Wind
0.0
0.2
0.4
0.6
0.8
1.0
Tornado
Severe Wind
Severe Hail
WoFS PMM DBZ > 35
MRMS DBZ > 35

Figure E.2: Same as Figure 7.7, but for the Logistic Regression model.