UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

CREATING GRID-BASED MACHINE LEARNING SEVERE WEATHER GUIDANCE FOR WATCH-TO-WARNING LEAD TIMES IN THE WARN-ON-FORECAST SYSTEM

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

SAMUEL VARGA
Norman, Oklahoma
2024

CREATING GRID-BASED MACHINE LEARNING SEVERE WEATHER GUIDANCE FOR WATCH-TO-WARNING LEAD TIMES IN THE WARN-ON-FORECAST SYSTEM

A THESIS APPROVED FOR THE
SCHOOL OF METEOROLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Corey Potvin, Chair

Dr. Montgomery Flora

Dr. Aaron Hill

Dr. Cameron Homeyer

Acknowledgments

To begin, I thank my research advisors: Dr. Corey Potvin and Dr. Montgomery Flora. Throughout my two years at the University of Oklahoma, they have continued to challenge me to grow professionally and academically. Their continuous support and guidance have been instrumental to my success and growth over the last two years. I also thank the members of the WoFS team for welcoming me and for all of the opportunities they have provided. The ability to consult with them has been paramount to both my growth and the success of this project. Quite literally, this project would not be possible without their hard work. Finally, I thank my family and friends for their support. They may not have understood what I was saying, yet they listened all the same.

This material is based upon work supported by the Joint Technology Transfer Initiative Program within the NOAA/OAR Weather Program Office under Award No. NA22OAR4590171.

Table of Contents

List Of Tables

List Of Figures

Abstract

The Warn-on-Forecast System (WoFS) is a rapidly updating convection-allowing ensemble focused on providing numerical guidance at Watch-to-Warning lead times (0-6 hours). Previous studies (e.g., Flora et al. 2021; Clark and Loken 2022) have incorporated machine learning (ML) to take advantage of the unique benefits of the WoFS and produce skillful guidance for severe weather hazards at lead times of 0-3 hours. This study evaluates the use of multiple ML architectures to produce 2-6 hour severe weather guidance using data from the WoFS. This represents the first use of machine learning to produce WoFS-based guidance at these lead times and the first use of deep learning to produce severe weather guidance using WoFS data.

Predictors are created using WoFS forecasts from the 2018-2023 Hazardous Weather Testbed Spring Forecasting Experiments. Data from forecast hours 2 through 6 are processed into predictors of multiple scales, incorporating both storm and environmental fields. We utilize three ML architectures: logistic regression, histogram-based gradient boosting trees, and U-nets. These models are trained to predict severe wind, severe hail, tornadoes, or any-severe hazard during the 2-6 hour window. Target data comes from the NOAA Storm Events database. The four-hour ML guidance is compared to rigorous baselines consisting of optimized Neighborhood Maximum Ensemble Probabilities for each hazard.

All ML methods evaluated outperform the NMEP baselines with tree-based methods achieving the highest performance of the traditional architectures. The largest improvement occurs for severe wind, followed by severe hail, tornado, and any-severe. Feature ablation shows that skill primarily comes from the intrastorm predictors and that the inclusion of multi-scale features exhibits little effect on skill. Despite the inclusion of additional features, the U-nets are unable to surpass the skill of the tree-based architectures. Similar to prior studies, this work shows the benefits of using the WoFS and ML to produce skillful guidance during the Watch-to-Warning period.

Chapter 1

Introduction

Research related to severe weather is often focused on lead times on the order of a few minutes to a half hour (e.g., Brooks 2004; Simmons and Sutter 2006; Brotzge et al. 2013) as this is the typical lead time provided for tornadoes (Brooks and Correia Jr 2018). Guidance issued during these lead times, such as tornado warnings, has been shown to reduce the loss of life to severe weather hazards (Simmons and Sutter 2008). A substantial amount of research is also devoted to next-day severe weather forecasting (e.g., Hill et al. 2023; Loken et al. 2022). However, the lead times between these extremes are not as well understood or robustly researched. An even smaller section of the literature is devoted to a subdivision of these middle lead times; the watch-to-warning window. Watch-to-warning lead times are herein defined as hours 0-6 preceding a severe weather event (Heinselman et al. 2024b). While these lead times are important, there are only a few types of information, such as observations or the output of CAMs, available to forecasters during this period. As such, increasing the quality and availability of guidance during the watch-to-warning period is paramount to supporting decision-makers.

The Warn-on-Forecast System (WoFS) is a forecast system being developed by the National Severe Storms Laboratory with a focus on watch-to-warning lead times (Stensrud et al. 2009, 2013). The WoFS is a regional system that produces six-hour forecasts with an emphasis on rapid assimilation of various observation types (Heinselman et al.

2024b). While still a developing system, the numerical guidance from the WoFS has been shown to be a useful tool for forecasters (Gallo et al. 2022).

To supplement the numerical guidance provided by WoFS, multiple WoFS-based post-processed machine learning (ML) products have previously been developed (e.g., Flora et al. 2021; Clark and Loken 2022; Heinselman et al. 2024a). These products leverage WoFS forecasts and ML to produce skillful guidance for severe weather hazards, such as wind, hail, and tornadoes. These types of products have been shown to provide more skillful guidance than non-ML baselines derived from WoFS forecasts (Flora et al. 2021). Additionally, evaluation of these products has shown that they improve the skill of forecasts (Clark et al. 2023; Flora et al. 2024).

However, the current WoFS ML products generally focus on lead times of 0-3 hours. As such, there is no currently available machine-learning-based supplement to the numerical WoFS guidance at lead times greater than 3 hours. Additionally, there has been no prior work with the WoFS to determine what skill (if any) ML guidance would provide over non-ML products at these lead times.

The primary goal of this thesis is to assess the skill of ML-based severe weather guidance produced from WoFS forecasts at lead times of 2-6 hours. This product addresses the current gap in ML guidance produced for the WoFS. The end goal of this study is to determine what advantage (if any) the ML guidance provides over guidance produced by non-ML baselines.

The secondary objectives of this study are numerous. First, we aim to discern which ML architecture produces the most skillful guidance for this task. Second, we aim to determine which fields provide the highest skill benefit for the guidance. Additionally, we evaluate multiple spatial scales of predictors to determine what benefit is gained from including multi-scale features. Finally, we evaluate the performance of deep learning models with respect to traditional machine learning architectures.

Chapter 2

Literature Review

2.1 Watch-to-Warning Guidance

As no accepted definition exists within the broader literature, watch-to-warning lead times are herein defined as the six hours preceding a severe weather event to maintain consistency with Heinselman et al. (2024b). The mean lead time for a tornado warned in advance is about 13 minutes, although the lead time of a warning can vary based on convective mode and numerous other factors (Brooks and Correia Jr 2018; Brotzge et al. 2013). As such, a substantial amount of the watch-to-warning period favors watches rather than warnings. However, many of the tools available to decision-makers during this period favor short-term guidance and nowcasting (e.g., Zinner et al. 2008; Cintineo et al. 2014; James et al. 2018). Given that the mean lead time for tornadoes warned in advance has been between 15-20 minutes for much of the last 40 years (Brooks and Correia Jr 2018), a substantial paradigm shift may be necessary to best utilize longer-term guidance during this window.

This is a key concern within the Warn-on-Forecast program due to its focus on the watch-to-warning window. Gallo et al. (2024) focuses on the use of WoFS by forecasters during the 2021 Hazardous Weather Testbed Spring Forecasting Experiment (HWT SFE). During this experiment, a multi-group study was conducted where one set of forecasters had access to WoFS data and the other set did not. Both sets produced a series of forecasts at lead times of 2-3 hours for severe wind, severe hail, and tornadoes.

Access to WoFS data improved the skill of these forecasts. This supports the findings of Gallo et al. (2022) which showed that forecasters could use WoFS guidance to produce skillful forecasts at a variety of lead times. Flora et al. (2024) details an experiment performed during the 2022 SFE. Two groups were given access to WoFS products. However, only one of these groups was given access to the WoFS ML products. The forecasts produced by both groups were then evaluated (Clark et al. 2023; Flora et al. 2024). The findings showed that access to the WoFS ML products resulted in more skillful severe weather guidance. Thus, both WoFS and its ML products are beneficial to forecasts during the watch-to-warning period. Discussion of these products is left to the next section of the literature review.

While access to WoFS data has been shown to increase the quality of forecasts during the watch-to-warning window, it is unclear how to best use information during these lead times. Simmons and Sutter (2008) showed that tornado warnings reduced fatalities when issued within 15 minutes of a tornado. However, lead times above 15 minutes were correlated with increased fatalities from tornadoes. The authors state that the increase in fatalities beyond 15 minutes was primarily driven by a small, yet dangerous, subset of strong tornadoes. In Hoekstra et al. (2011), a survey showed that members of the public would prefer lead times on the order of a half hour. However, the respondents also indicated that providing longer lead times may increase unsafe behaviors, such as securing belongings or fleeing, rather than taking shelter. Additionally, they reported that an increase in lead time would lead to a decrease in perceived threat.

An alternative is to target products providing advance notice of severe weather towards decision-makers, rather than the general public. Obermeier et al. (2023) provided probabilistic hazard information to broadcast meteorologists to assess how they would use the data. Their results suggest that the information would be valuable for

decision-making, although it may not be explicitly communicated to the public due to the risk of confusion. Additionally, more information was desired between the issuance of watches and warnings. While the primary focus of Obermeier et al. (2023) was on probabilistic information, products designed to provide extended lead times may receive similar feedback. Thus, while targeting decision-makers is possible, this approach would require more interaction with prospective users to receive their feedback on the product.

2.2 Machine Learning for Severe Weather

Machine learning (ML) is a multi-function tool that fills a multitude of roles within meteorology. It is commonly used as a post-processing tool for the output of numerical weather prediction (NWP) models. Recent advancements have led to the development of fully AI-driven NWP systems (e.g., GraphCast; Lam et al. 2022). However, AI NWP systems capable of competing with traditional NWP methods at convection-allowing scales do not yet exist. As such, within this review, we focus on the use of machine learning for post-processing NWP output to produce severe weather guidance.

Hill et al. (2020) investigated the use of machine learning for producing severe weather guidance using data from the Global Ensemble Forecast System Reforecast ensemble (GEFS/R). Guidance was produced by random forests for lead times of days 1-3 (i.e., 12-84 hours before an event) on a 55 km grid. Severe wind, severe hail, and tornadoes are predicted separately for day 1, while days 2 and 3 consist of predictions for any of these hazards (dubbed "any-severe" for the remainder of this work). The day 1 skill was highest for severe wind, followed by severe hail, and tornadoes. However, the day 1 guidance produced by random forests was marginally less skillful than SPC outlooks with some regional variation. This was remedied by producing a weighted

blend of the random forest guidance and the SPC outlooks. Comparatively, the day 2 guidance and day 3 guidance produced by random forests were more skillful than the corresponding SPC outlooks. As an extension of this work, Hill et al. (2023) showed that random forests could provide skillful severe weather guidance out to days 4 and 5. Hill and Schumacher (2021) utilized a similar framework to create predictions of day 1 excessive rainfall. However, the guidance produced by random forests was found to be generally less skillful than the Excessive Rainfall Outlooks issued by the Weather Prediction Center.

Loken et al. (2020) also focused on the use of random forests for producing day 1 severe weather guidance. Guidance was created on an 80 km grid using data from the Storm-Scale Ensemble of Opportunity (Jirak et al. 2012, 2016). Random forests were trained to produce any-severe guidance, as well as guidance for severe wind, severe hail, and tornadoes. The random forests generally outperformed both an updraft helicity baseline and SPC forecasts. However, the tornado guidance from random forests did not outperform the SPC tornado predictions. Sobash et al. (2020) performed a related study; any-severe guidance was once again produced on an 80 km grid for four-hour windows. However, rather than random forests, a feed-forward neural network was the architecture of choice. The guidance produced by this neural network outperformed an updraft helicity surrogate severe forecast. However, the difference between the neural network and baseline was reduced when evaluating on smaller target scales. The neural networks also produced more skillful guidance than the updraft helicity baseline when the convective regime was not supercellular. Sha et al. (2024) is an extension of this work using data from the High-Resolution Rapid Refresh. In this study, a conditional generative adversarial network was used to expand the deterministic HRRR forecast into a synthetic ensemble. This was paired with a convolutional neural network (CNN) to produce skillful four-hour guidance for any-severe hazards.

While Gagne II et al. (2019) falls within the scope of using machine learning to post-process NWP output and predict severe weather, the study took a slightly different approach. Day 1 forecast data was gathered from the NCAR convection-allowing ensemble using 96x96 km patches on a 3 km grid. A logistic regression model and CNN were trained to predict the occurrence of severe hail. However, rather than using storm reports, the severe hail was diagnosed using the maximum diameter of hail produced by the Thompson microphysics scheme. Thus, the predictions were for simulated hail rather than real. The CNN was the highest-performing method and was able to leverage spatial information to learn the storm modes. As such, Gagne II et al. (2019) suggests that deep learning may be able to improve on traditional machine learning methods when it comes to severe weather guidance. Finally, we consider the NOAA ProbSevere products (Cintineo et al. 2014, 2018, 2020). ProbSevere uses data from the Multi-Radar/Multi-Sensor (MRMS) suite, lightning data, satellites, and NWP output to produce 0-1 hour severe weather guidance. ProbSevere makes use of naïve Bayesian classifiers. While the initial version of ProbSevere only produced guidance for any-severe hazard (Cintineo et al. 2014), the second iteration produces guidance for severe wind, severe hail, tornadoes, and any-severe hazard. As these products are shown to be skillful, this suggests that MRMS data may be a skillful predictor to incorporate for other prediction tasks within the watch-to-warning window. Additionally, Cintineo et al. (2022) showed that a CNN could make use of satellite data to provide skillful, short-term guidance for lightning within the next hour. This reinforces that deep learning can be used as a tool to create skillful predictions of short-term hazardous weather. For a more comprehensive review of the recent usage of traditional ML and deep learning for producing guidance for severe convective hazards, the readers are encouraged to consult McGovern et al. (2023).

Given the studies discussed previously, both traditional ML and deep learning have been shown to be useful for predicting severe weather at a variety of timescales. We now consider the two most relevant works, as they are focused on the use of ML within the WoFS framework. Flora et al. (2021) represents the first application of ML to the Warn-on-Forecast System for producing severe weather guidance. Rather than a grid-based approach, the study employed an object-based approach where predictions are applied to ensemble storm tracks rather than for each grid point. The storm tracks are derived from the WoFS ensemble forecasts of peak vertical velocity over a 30-minute window. Logistic regression, random forests, and gradient-boosted trees were then used to produce the probability that each storm track would overlap a storm report. This guidance was produced for severe wind, severe hail, and tornadoes and was valid over 30-minute windows out to a maximum lead time of 2.5 hours. The results of this study showed that the ML guidance was capable of outperforming ensemble-probability baselines for each hazard and all considered metrics. The improvements over the baseline were largest for severe wind, followed by tornadoes and severe hail. Additionally, when comparing within an individual hazard, the variation in skill between the different ML architectures was fairly small (e.g., random forests and gradient-boosted trees had similar levels of skill). The main limitation of the object-oriented method was the restrictions it placed on the dataset. As predictions could only be evaluated for the objects, the verification was restricted to the areas where the WoFS forecasts had storms.

Rather than the object-oriented approach of Flora et al. (2021), Clark and Loken (2022) focused on a grid-based approach to create products for the entire 0-3 hour window. This guidance was produced by random forests trained to predict the occurrence of any-severe, rather than individual hazards. The random forests' guidance was capable of outperforming an updraft helicity-based baseline. Additionally, this study

showed that the most important features for the skill of the ML model were intrastorm predictors rather than environmental. Furthermore, the inclusion of smoothed updraft helicity fields from the WoFS ensemble members increased the random forest's skill. Recent ML work with the WoFS has included a focus on probabilistic hazard information (PHI). The resulting product, WoFS-PHI, uses WoFS forecasts and data from ProbSevere to produce PHI for wind, hail, tornadoes, and lightning at lead times of 0.5-3 hours (Heinselman et al. 2024a; Calhoun et al. 2024).

It is notable that Clark and Loken (2022) only evaluated the performance of random forests trained to produce guidance for any-severe hazards. As such, the work presented in this thesis can be considered an extension of Clark and Loken (2022) to longer lead times, additional hazards, and more ML architectures. This includes the use of deep learning, which was not evaluated in either Clark and Loken (2022) or Flora et al. (2021).

Chapter 3

Data & Methods

The primary objective of this thesis is to produce severe weather guidance at lead times of two-to-six hours using data from the Warn-on-Forecast system. This is established as a binary classification task, where the guidance provides the probability of a severe weather hazard occurring within 36 km of a given point during the two-to-six hour guidance window. The following sections describe the data and methods used to achieve this goal.

3.1 Warn-on-Forecast System

The Warn-on-Forecast System (WoFS) was initially proposed in 2009 as a flagship model that could eventually lead a shift away from the warn-on-detection paradigm (Stensrud et al. 2009). The novel warn-on-forecast framework envisioned a scenario where model guidance was skillful enough that warnings could be issued largely based on model forecasts. These warnings would extend lead times far beyond warnings issued when severe storms became detectable with observations. As expected of such a monumental undertaking, the development of the WoFS had to overcome numerous challenges (Stensrud et al. 2013). While the Warn-on-Forecast paradigm has not yet been fully realized, the current WoFS has proven capable of providing skillful, rapidly updating guidance during the watch-to-warning period and is routinely used by the

Weather Prediction Center, Storm Prediction Center, and various weather forecast offices (Heinselman et al. 2024b; Wilson et al. 2024; Wilson 2023).

The current iteration of the WoFS is a convection-allowing ensemble with a 3 km grid spacing. The ensemble domain covers a 900 x 900 km region and can be relocated between every case. Additionally, the WoFS supports multiple domains per case when requested. The WoFS is only run for specific cases, with the first and last forecasts typically initializing at 1700 and 0300 UTC respectively. The exception to this is the month of May; the WoFS is generally run daily during this period to support the Hazardous Weather Testbed Spring Forecasting Experiments (Clark et al. 2023).

The ensemble component of the WoFS consists of 36 members. These members utilize various physics schemes as outlined in Skinner et al. (2018). Satellite, radar, and mesonet data assimilation occur every 15 minutes using a variation of the grid-point statistical interpretation ensemble Kalman filter (Heinselman et al. 2024b). This is augmented by hourly assimilation of conventional observation types. While the ensemble consists of 36 members, only 18 are utilized for forecasts. New forecasts are initialized every half hour while the WoFS is running. Forecasts launched at the top (bottom) of the hour extend six (three) hours with output available every five minutes (Heinselman et al. 2024b). As we are interested in 2-6-hour predictions, we only use the six-hour forecasts that are initialized hourly.

3.2 Dataset

3.2.1 Warn-on-Forecast Datasets

The data used for this project consist primarily of ensemble forecasts from the Warn-on-Forecast System. Due to the focus on severe convective weather, we only use forecasts from May. These forecasts were produced as part of the annual Hazardous Weather

Testbed Spring Forecasting Experiments from 2018 to 2023 and are split into two datasets. As this project began in 2022, the first dataset only consists of forecasts through 2021. When more recent data became available, it was incorporated into the second dataset.

The first dataset, referred to as the initial dataset, consists of 644 forecasts across 83 cases ranging from 2018 - 2021. Much of the preliminary work was conducted on this dataset, with these results being reviewed in Sections 4.1-4.3. The secondary dataset includes 1154 forecasts from 111 cases ranging from 2019-2023. This represents the most recent forecast data available. Additionally, the data from 2018 was dropped from this dataset as the WoFS had a different configuration at that time. The results of this dataset are reviewed in Sections 4.4 and 4.5. In both datasets, the initialization times of the WoFS range from 1700-0300 for each case with forecasts extending six hours after initialization.

Figure 3.1 shows heatmaps of the WoFS domains in these datasets. Both datasets have a strong bias towards the south-central plains and limited data over the eastern seaboard. Additionally, both datasets have extremely few cases covering the West Coast and Southwestern regions of CONUS. However, this is expected given the lack of severe weather in these regions during the warm season (Farney and Dixon 2015).

Both datasets were split into training and testing sets, with 70% of the cases being used for the training process and the remaining being used for the evaluation of the models' performance. Cases were split into these sets based on the date of the forecast, with all data from each case being isolated to either the training set or the testing set. Combined with the near-daily relocation of the WoFS domain, this is a reasonable safeguard against cross-contamination between the training and testing data. In total, this results in approximately 4.2 million (2.1 million) training (testing) samples for

the initial dataset, and 8 million (3.5 million) training (testing) samples for the final dataset.

3.2.2 Feature Engineering

Multiple fields are extracted from the WoFS forecasts and processed to create predictors (Table 3.1). While many of the fields rely implicitly on temperature (e.g., CAPE), it is notable that temperature is not explicitly included as a predictor unlike in Clark and Loken (2022). Following the precedent set by Flora et al. (2021) and Clark and Loken (2022), the fields in Table 3.1 are categorized as either intrastorm fields (Fields 1-8) or environmental (Fields 9-21).

While both types of field undergo a similar process of refinement into predictors, the filters used to process a field vary based on the field's category. Intrastorm fields are processed using spatial and temporal maximum filters, as their distributions tend to be highly skewed. Conversely, environmental fields are processed using spatial and temporal mean filters as their distributions tend towards Gaussian.

The raw forecast fields are transformed into predictors via a multi-step process. First, the 3 km forecasts from each ensemble member are coarsened to a 9 km grid using a maximum (uniform) filter for the intrastorm (environmental fields). A k-dimensional (KD) tree is used to remap the 3 km grid to a 9 km grid using a nearest neighbor approach; the KD tree is preferred over other methods due to its efficiency. Next, the time average (maximum) value is calculated for each environmental (intrastorm) field at every grid point. This is performed separately for each ensemble member and uses the forecast data over a window starting two hours after forecast initialization and ending six hours after initialization. At each step, any missing values are replaced with the mean value of the field for that ensemble member.

Intrastorm Fields:	
1) 2-5 km Updraft Helicity	2) 0-2 km Updraft Helicity
3) Composite Reflectivity	4) 80 m Wind Speed
5) 0-2 km Average Vertical Vorticity	6) HAILCAST
7) Column-maximum Updraft	8) Okubo-Weiss Number
Ensemble Statistics: mean, interquartile range,	2nd highest, 2nd lowest members.
Environmental Fields:	
9) Mid-level Lapse Rate	10) Low-level Lapse Rate
11) 0-3 km Storm Relative Helicity	12-13) 0-1 km U and V shear
14-15) 0-6 km U and V Shear	16-17) 3-6 km U and V Shear
18) Significant Tornado Parameter	19) Mixed-layer CIN
20) Supercell Composite Parameter	21) Mixed-layer Cape
Ensemble Statistics: mean, standard deviation	

Table 3.1: Fields extracted from WoFS ensemble forecasts and used to create features. Fields 1-8 are intrastorm while the remaining fields are classed as environmental. The ensemble statistics for each field type are also listed. Each statistic is calculated three times per field due to the three different smoothing radii, resulting in 174 predictor fields.

These time composites are then smoothed using a spatial maximum or spatial mean filter with a radius of 9 km. Ensemble statistics are then calculated for each field using the smoothed time composites. As seen in Table 3.1, the ensemble mean, interquartile range, second highest ensemble member's value, and second lowest ensemble member's value are calculated for each intrastorm field. The second highest and second lowest member's values roughly correspond to the 90th and 10th percentiles of the ensemble. For environmental fields, only the ensemble mean and standard deviation are calculated. Similarly to the choice of filter, the choice of ensemble statistics is primarily motivated by the different distributions of intrastorm and environmental fields. Hence, the environmental statistics describe Gaussian distributions while the

intrastorm statistics better capture skewed distributions. Fewer statistics are calculated for the environmental fields as they are generally less skillful predictors than the intrastorm fields (Flora et al. 2021; Clark and Loken 2022; Loken et al. 2022).

This process results in 58 fields that have been smoothed using a filter with a 9 km radius. The smoothing and calculation of ensemble statistics is repeated twice more using a radius of 27 km and then 45 km. The smoothing is performed in parallel, rather than sequentially (e.g., the 9, 27, and 45 km smoothing are all performed on the time composite fields and not fields that have been previously smoothed at a smaller scale). This results in a final set of 174 predictors; 58 from each spatial scale. These ensemble statistics are then used as predictors for the machine learning discussed in later sections. The inclusion of predictors at multiple spatial scales attempts to address some of the spatial uncertainty that occurs at 2-6 hour lead times. Additionally, the inclusion of multi-scale predictors is meant to capture and separate larger meso-β scale information from smaller scales of information.

3.2.3 Target Data

While the predictors discussed previously only include data from the WoFS forecasts, target fields are produced using *Storm Data* reports from the National Centers for Environmental Information. For a given forecast, all reports located within the WoFS domain during forecast hours 2-6 are identified. To account for any temporal inaccuracies in the reports, we also include reports that occur within 15 minutes of the start or end of the 2-6 hour forecast window. Only reports associated with severe weather (i.e., wind $\geq$ 50 knots, hail $\geq$ 1 inch, or any tornado) are retained. The locations of these reports are mapped to the nearest grid point on the native 3 km WoFS domain using a K-dimensional tree. The reports are then mapped to the 9 km grid and a maximum filter of radius 36 km is applied to account for spatial uncertainty. This is

done to mitigate the effects of phase errors within the WoFS forecasts (i.e. allowing some displacement of WoFS storms from the geographic location of the report), to allow for some spatial uncertainty in the report's location, and to account for errors in storm location within the WoFS. The final target field is a boolean field with 9-km grid spacing. We adopted this approach to roughly match the SPC's probabilistic outlook of severe weather within 26 miles (40 km) of a point.

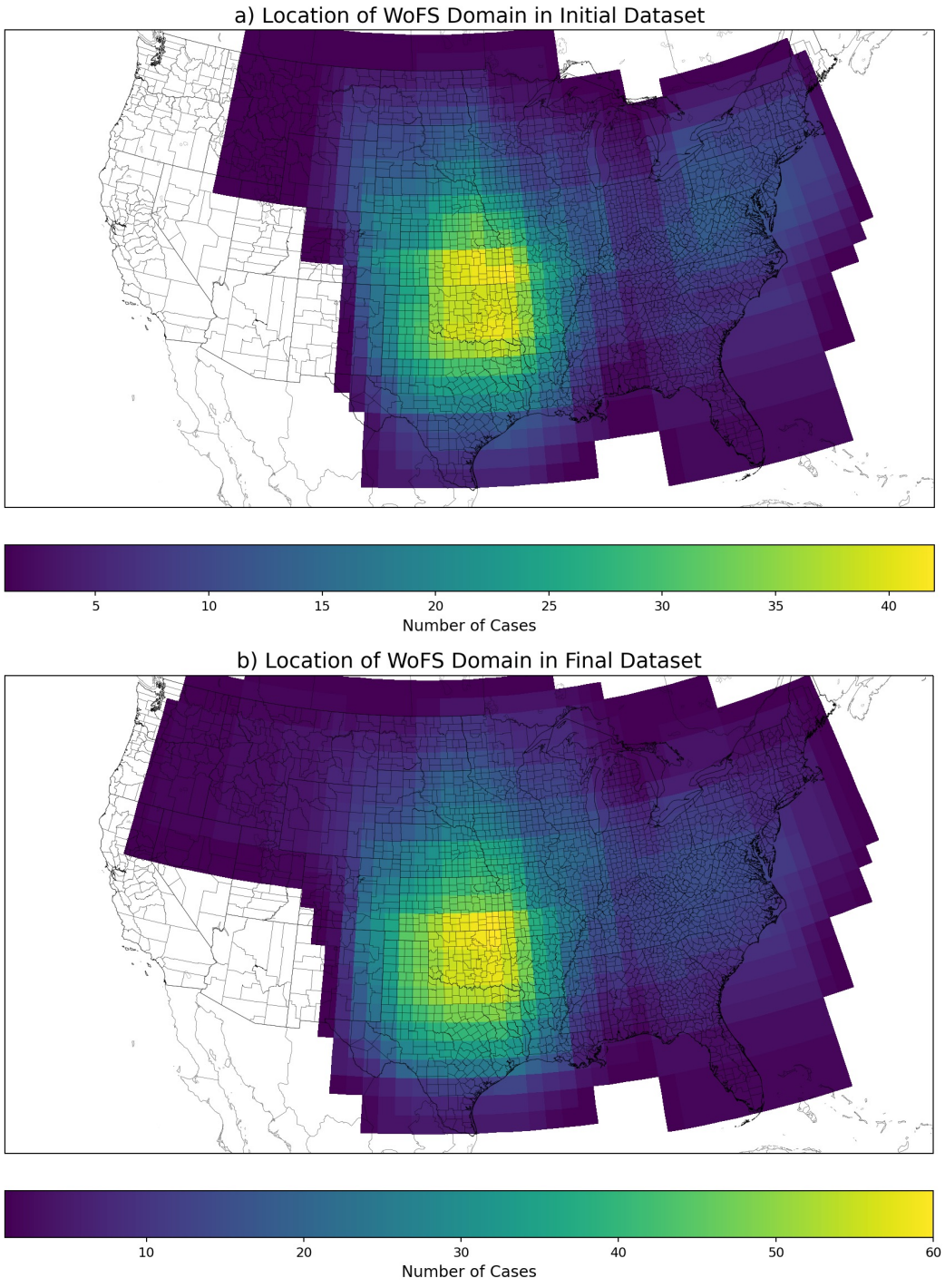

Figure 3.1: Heatmaps showing the 900 x 900 km domain locations on a one-degree grid for the a) initial dataset and b) final dataset. The initial dataset contains 83 SFE cases from 2018 - 2021, while the final dataset contains 111 SFE cases from 2019 - 2023. In both datasets, the majority of cases are located across the central plains.

3.3 Baseline and Machine Learning

3.3.1 Baseline Models

To place the performance of the ML models into perspective, we create a set of rigorous non-ML baselines to compare against. These consist of Neighborhood Maximum Ensemble Probabilities (NMEP) produced from the WoFS forecasts. The ensemble probability is calculated as:

$$EP = \frac{1}{N} \sum_{n=1}^{N} f_i \geq t$$

where N is the number of ensemble members, f_i is the value of a forecast field, and t is some threshold. Thus, the EP represents the fraction of ensemble members with a forecast value exceeding the threshold. To convert the EP to an NMEP, we apply a spatial maximum filter to each ensemble member before calculating the EP. Similar to the smoothing of the predictor fields, this is done to account for spatial uncertainty in the forecasts. Both the size of the spatial filter and the threshold value are chosen based on values that optimize the Brier Skill Score (BSS) of the BL models on the training set using 5-fold cross-validation.

Figure 3.2 shows the various thresholds and neighborhoods considered for each hazard during this process, as well as the optimal combination for each. A unique NMEP BL is produced for each hazard, with each BL only using one forecast field. As seen in Figure 3.2, both the any-severe and tornado BL use 2-5 km Updraft Helicity (UH) while the severe wind and severe hail products use 80-m wind speed and HAILCAST respectively. 2-5 km Updraft Helicity has been shown to be a skillful indicator of both tornadoes and severe weather in similar work (Sobash et al. 2016, 2020), while 80-m wind speed and HAILCAST have been shown as skillful for their respective hazards (Flora et al. 2021). While the any-severe and tornado BLs are produced from the same

field, the optimal any-severe BL uses both a lower threshold and a larger neighborhood than the tornado BL.

While raw NMEP baselines have been utilized by prior work (e.g., Clark and Loken 2022), we also use isotonic regression to improve the calibration of our baselines compared to the raw ensemble probabilities. This choice is motivated by previous studies that also used isotonic regression to calibrate the output of machine learning models (Burke et al. 2020; Lagerquist et al. 2017). Once the optimal NMEP parameters are selected, the isotonic regression fits a non-decreasing function to the NMEP output. This is done using the approach of Platt et al. (1999); the validation data from each cross-validation fold is concatenated and used to learn the calibration function.

3.3.2 Logistic Regression

The first machine learning architecture that we implement is logistic regression (LR). The LR architecture outputs a probability ($\hat{y}$) given by:

$$\hat{y} = \sigma(\mathbf{W}^T\mathbf{X})$$

where $\mathbf{W}$ is the weight vector, $\mathbf{X}$ is the predictor vector, and σ is the logistic (or sigmoid) function (Géron 2022). Thus, the output probability is given by the logistic function applied to the linear combination of each predictor's weight and value with the addition of a bias term. For large positive (negative) values of ($\mathbf{W}^T\mathbf{X}$), $\hat{y}$ will approach 1 (0). $\hat{y}$ is then the probability of a sample being the positive class (i.e. severe weather occurring within 36 km).

Each logistic regression instance is implemented using the scikit-learn Python package (Pedregosa et al. 2011). The weight vector is learned through gradient descent during the training process by minimizing the log loss of the training dataset. During

the training process, 5-fold cross-validation is used to ensure that the set of selected hyperparameters is robust. We use the set of hyperparameters that achieves the highest mean critical success index (CSI) on the validation folds. CSI is defined in Section 3.5. Isotonic regression is once again implemented to improve the calibration of the output probabilities. The implementation follows the same methods as discussed in Section 3.3.1.

3.3.3 Tree-based Machine Learning

The foundation of tree-based machine learning is the decision tree. In essence, a decision tree is a series of optimized thresholds that act to split the dataset's examples of the positive class from the remaining samples (Breiman 2017). These splits are referred to as nodes. At each node, the dataset can be split based on some criterion. This criterion is usually the optimization of a quantity, such as impurity in the node's samples or the value of a loss function (Géron 2022). By finding a variable and threshold that optimize this quantity, and then continuing this process with more nodes, a decision tree begins to separate the positive class samples from other samples. Once another criterion is met (e.g., too little data to split further or the maximum number of nodes is met), a terminal node (known as a leaf node) is established. For any given leaf node, the probability of the positive class is estimated as the ratio of positive class samples in the node to total samples in the node.

When used in isolation, single decision trees are prone to overfitting the training dataset (Géron 2022). Rather than using a single decision tree, many tree-based methods combine multiple trees to form an ensemble. While random forests have been extensively used in meteorology (e.g., Clark and Loken 2022; Loken et al. 2020; Hill et al. 2023), Flora et al. (2021) showed that gradient-boosted trees could achieve similar

levels of skill to random forests. As such, we opt to use a variant of gradient-boosted trees known as histogram-based gradient boosting trees.

The basis of the histogram-based gradient boosting trees (HGBT) architecture is the process of gradient boosting. The goal of boosting methods is to combine multiple weak learners sequentially into a strong learner (Géron 2022). Specifically, gradient boosting entails adding additional trees to account for errors made by the previous trees (Friedman 2001). Unlike random forests, the trees in a gradient-boosted ensemble are in series. Thus, every tree after the initial is sensitive to the output from trees in the sequence before it. The basic structure of the tree is unchanged. However, later trees now attempt to fix the residual errors of prior trees rather than providing a pure class probability. The output probability is then a sum of the contributions from all trees, rather than the average (Géron 2022).

Gradient boosting trees (GBTs) are generally just as skillful as random forests for a wide range of problems and are generally faster to train (Bentéjac et al. 2021). We opt to use the scikit-learn implementation of histogram-based GBTs, as this is the same library we use to implement LR. The histogram refers to how data is processed during the training of a GBT. Rather than maintaining and sorting all unique values, the data are separated into 256 bins. Instead of repeatedly sorting through every value of every feature when creating nodes, these bins are referenced. This drastically expedites the training process (Bentéjac et al. 2021).

The training process and implementation of the HGBT are identical to the LR discussed in Section 3.3.2. While tree-based methods normally do not require data scaling, we preprocess all predictors using standard scaling to maintain parity with the data used by LR. We use 5-fold cross-validation during training to find the hyper-parameter set that optimizes the critical success index (CSI; defined in Section 3.5).

Additionally, we again include posterior isotonic regression to better calibrate the output probabilities. All hyperparameter optimization is performed using the training set.

3.3.4 Convolutional Neural Networks and U-nets

Following our discussion of traditional ML, such as LR and HGBT, we now consider an ML architecture from the realm of deep learning: U-nets. To discuss U-nets, we must first establish their foundation: convolutional neural networks (CNNs). CNNs are a type of neural network used for processing images or other spatial fields through the use of convolutional layers (LeCun et al. 1989, 1995). Rather than layers composed of neurons connected to all neurons in the previous layer, as seen in a standard fully connected neural network, convolutional layers only connect neurons to a subsection of the previous layer. In a convolutional layer, the weights of neurons can be represented as a series of convolutional filters. These filters can be specialized for specific tasks, such as detecting edges (LeCun et al. 2015). The size of the filter is referred to as the kernel size. By passing these filters across the input field, the convolutional layers can detect features in various locations within the image (Géron 2022). These convolutional layers are often followed by pooling layers. Pooling layers aggregate the input from all neurons within the kernel. The most common form of aggregation is retaining the maximum value within that kernel (Géron 2022). By varying the distance between pooling kernels, CNNs progressively reduce the size of the image. Repeating the series of convolutional and pooling layers then allows CNNs to learn progressively higher-level features. While CNNs are often skillful for classification-based tasks (e.g., classifying handwritten digits or predicting severe hail; LeCun et al. 1989; Gagne II et al. 2019, respectively), they generally perform classification for an entire image. As we want to

classify every pixel within an image (i.e. semantic segmentation), we employ U-nets rather than a standard CNN.

U-nets are related to a subsection of CNNs referred to as autoencoders. Autoencoders retain the same basic ideas of convolutional layers and pooling layers from CNNs. However, autoencoders are designed to perform image-to-image translation rather than image-to-scalar. To do this, autoencoders make use of an encoder-decoder pair (Kramer 1991). The encoder is similar to the CNN discussed previously; a series of convolutional layers and pooling layers coarsen the input. To pass information through this bottleneck, the network must construct a higher-level latent representation of the input data. The decoder effectively mirrors this process by converting the latent representation back to the original fields. Rather than coarsening the image, the decoder repeatedly upscales the representation until it reaches the original image's resolution. Thus, the output of an autoencoder has identical spatial dimensions to the input.

As mentioned, U-nets are closely related to autoencoders. The distinction is nuanced, if somewhat pedantic. Both architectures can share the same shape. However, autoencoders are generally intended to reproduce the input fields, while U-nets are intended to create a new product with the same shape as the input fields. In fact, U-nets were initially designed to perform semantic segmentation of biomedical images (Ronneberger et al. 2015). Additionally, U-nets generally feature skip connections while pure autoencoders do not. These are alternative pathways that allow information to bypass the central bottleneck imposed by autoencoders. In other words, skip connections provide routes for information to pass directly from one side of the network to the other without being forced through the pooling layers.

The final configuration of our U-net is discussed in the following and a simplified representation of the architecture is displayed in Figure 3.3. While some of the parameters (e.g., network size, loss, and activation functions) are defined by our use case, many of the hyperparameters (e.g., kernel size, filter number) are the result of hyperparameter optimization as discussed in Section 3.4. Our U-net is five layers deep and uses binary cross-entropy as the loss function. We start with a normalization layer to standardize all data using standard scaling. This is followed by an input layer for a 96x96 image. While it is common practice to use powers of 2, we opt for a patch size of 96x96 to incorporate the largest amount of data. Restricting this to a power of 2, the images would be 64x64 or smaller. Given the 36 km target radius, we elected to break convention and use a larger patch size. Following the input layer, each step of the network is composed of 2 convolutional layers followed by a max pooling (upsampling) layer on the encoder (decoder) side. The pooling layers use 2x2 kernels and a stride of 2 to reduce the patch dimensions (area) by a half (quarter). As such, the encoder layers have shapes of 96x96, 48x48, 24x24, 12x12, 6x6, 3x3. These are then reversed on the decoding side. The convolutional layers start with 64 convolutional filters; the number of filters increases by a factor of 2 after every pooling layer up to a maximum of 2048 filters. The convolutional kernel size alternates between 2 and 3, again changing after every pooling layer. Each convolutional layer uses the rectified linear unit (ReLU) as the activation function. Additionally, spatial dropout layers with a dropout probability of 0.33 are incorporated after all convolutional layers. All of these hyperparameter choices, with the exception of ReLU as the activation function, were determined through hyperparameter optimization performed using 5-fold cross-validation on the training set. Skip connections are also utilized, but they pass directly across the U-net between the pooling and concatenation layers. This is a notably simpler approach than the skip connections used in the U-net++ and U-net 3+

architectures which connect to multiple layers (Zhou et al. 2018; Huang et al. 2020). The final convolutional layer uses one filter and a 1x1 kernel paired with a sigmoid activation function to output the probability of the positive class at each pixel. Thus, the interpretation of the U-net output is identical to those of the previous architectures. In all, the final U-net has 40,502,849 parameters to learn.

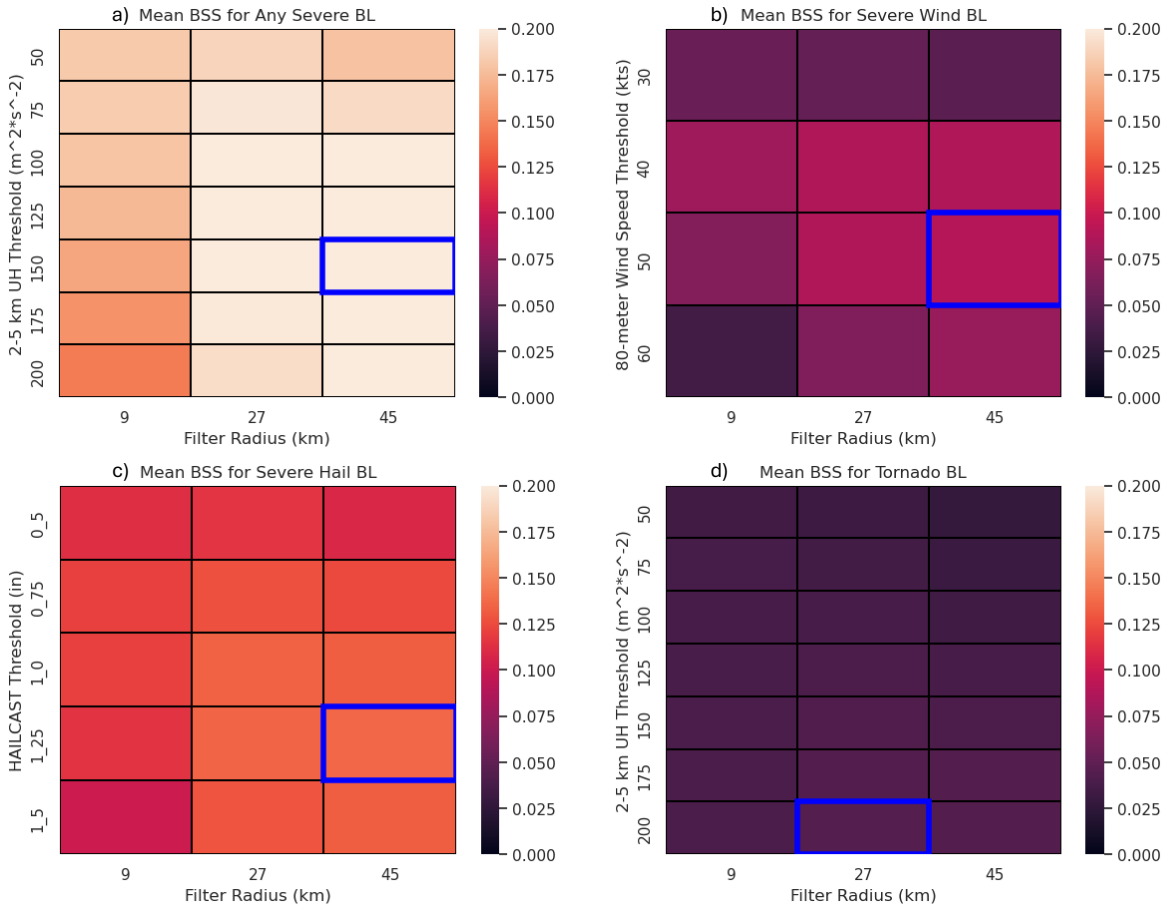

Figure 3.2: Heatmaps of the NMEP Brier Skill Score (BSS) for various neighborhoods and thresholds for a) any-severe hazards, b) severe wind, c) severe hail, and d) tornadoes. The BSS is calculated as the mean BSS across all validation folds using 5-fold cross-validation on the training set. The optimal combinations are outlined in blue. These optimal combinations are used to produce the baseline NMEPs. The shading indicates the range of BSS that the NMEPs achieve on each hazard.

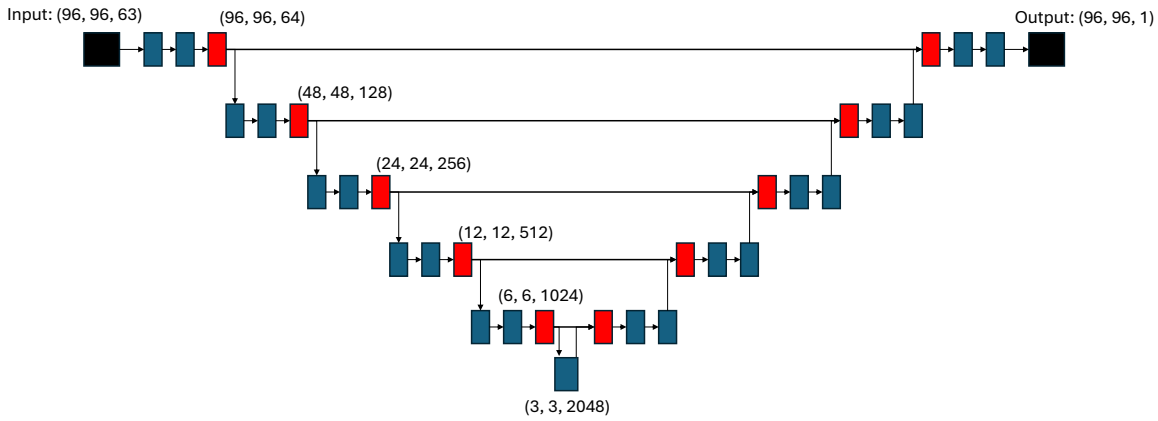

Figure 3.3: A simplified representation of the U-net architecture used within this study. Convolutional layers are represented as blue boxes while pooling and upscaling layers are represented as red boxes. The shape is given as (y,x, features) at each level of the U-net. the leftmost (rightmost) black box represents the input (output) layer. Skip connections are only connected across the U-net to the level with a similar spatial dimension, rather than the full-scale skip connections used by the U-net 3+ architecture.

3.4 Deep Learning Methods

The general implementation of the u-net method is similar to the previous ML architectures. The data is identical to the final dataset (2019-2023) discussed in Section 3.2.1. Dates are split into the training and testing sets identically to that dataset. However, there are some notable differences from the feature engineering discussed in Section 3.2.2. The data is once again coarsened to a 9 km grid as discussed previously. However, rather than retaining the full domain, we drop data within 18 km of the domain boundary. This changes the data from a 100x100 field to a 96x96 patch. Following this, we again take the time composite of each field over forecast hours 2-6. However, unlike the traditional ML predictors, we do not follow this step with smoothing. Instead, we directly calculate the ensemble statistics at each grid point without smoothing. As such, any smoothing or pooling is left for the convolutional filters. The ensemble statistics for each field are identical to those discussed previously.

All features included in Table 3.1 are retained in the deep learning dataset, as are the storm data targets. We add the WoFS forecasts of cloud top temperature as an additional intrastorm field to the predictors discussed previously. Furthermore, two additional types of data are used to supplement the predictors and targets. First, as an additional predictor field, we include composite reflectivity from the multi-radar/multi-sensor (MRMS) product suite (Smith et al. 2016). This reflectivity is valid at the initialization time of the forecast, rather than forecast hours two through six. This is meant to indicate convection that has already initiated and to reduce the impact of spurious convection within the WoFS forecasts. The MRMS composite reflectivity is regridded to the base 3 km grid of the WoFS. After this, the field is coarsened to 9 km. As there is no ensemble component, the composite reflectivity field itself is used as a predictor rather than ensemble statistics.

Furthermore, we incorporate the maximum expected size of hail (MESH) as an additional target field. MESH is a radar-based product shown to be useful as an indicator of severe hail (Witt et al. 1998; Wendt and Jirak 2021). While MESH estimates hail size, we translate this into a binary field by thresholding the values at 30 mm. For U-nets trained to predict any-severe hazards, this serves as a supplement to the storm data reports rather than a replacement (i.e. any-severe is now the union of MESH, severe wind reports, severe hail reports, and tornado reports). We use the same smoothing process described previously to extend both the MESH and storm data targets into 36 km radii.

The training of the U-net is identical to that discussed previously. We use 5-fold cross-validation on the training set for both training and hyperparameter tuning. The any-severe U-net's final structure and parameter choice are outlined in Section 3.3.4. However, unlike the previous ML architectures, we do not use isotonic regression to calibrate the probabilities.

3.5 Verification Metrics

This section introduces the general techniques and metrics used to verify and compare the performance of the BL and ML architectures. Discussion of the metrics is heavily inspired by Brooks et al. (2024), and as such, we adopt their notation when applicable. Many of the verification metrics are derived from the 2x2 table of forecasts and observations. Adopting the same style and notation as Brooks et al. (2024), the table is given in Table 3.2.

While this table works for binary observations and forecasts, some alterations must be made for probabilistic output. By setting some forecast threshold t, we can convert

	Observations	
Forecasts	Yes ($y = 1$)	No ($y = 0$)
Yes	a	b
No	c	d

Table 3.2: The standard 2x2 contingency table, also known as a confusion matrix, for binary forecasts and outcomes. The terminology for the table elements varies widely across fields and are therefore left unnamed within this paper. Readers interested in the history of the table and its associated metrics are encouraged to consult Brooks et al. (2024).

the probabilistic forecasts into a binary outcome. Thus, for a given threshold t and forecast probability x, the table becomes Table 3.3

While the threshold t can be fixed arbitrarily, we can allow t to vary to gather information on how the skill of the models changes over their entire output range. For our case, we evaluate the table at 11 thresholds ranging from [0,1] every 0.1. At the limit of $t = 0$, all forecasts become a yes while (nearly) all forecasts become a no at $t = 1$. Using the prior definition of the 2x2 table elements, we can define verification metrics that we calculate at every threshold.

The probability of detection (POD) is the conditional probability that an event was forecast given an event occurred and is defined as:

$$POD = \frac{a}{a + c} \tag{3.1}$$

Similarly, the success ratio (SR) is the conditional probability of an event occurring given a yes forecast and is defined as:

$$SR = \frac{a}{a + b} \tag{3.2}$$

	Observations	
Forecasts	Yes ($y = 1$)	No ($y = 0$)
Yes	$x \geq t$	$x \geq t$
No	$x < t$	$x < t$

Table 3.3: A 2x2 contingency table for the verification of probabilistic forecasts. The forecasts (x) are categorized as *yes* or *no* depending on their relationship to the threshold value (t). If only one value of t is utilized, this reduces to Table 3.2. However, by using a series of values for t, multiple contingency tables are constructed detailing the forecast performance across various probabilities.

The false alarm ratio (FAR) is given by $1 - SR$. The POD and SR can be related to each other through the critical success index (CSI) which is given by:

$$CSI = \frac{a}{a + b + c} = \frac{POD * SR}{POD + SR - POD * SR} \tag{3.3}$$

Given their relationship, POD, SR, and CSI are plotted together on performance diagrams. To mitigate the effects of base rate on CSI and facilitate easier comparison between datasets, Flora et al. (2021) proposed the normalized CSI (NCSI) which is calculated as:

$$NCSI = \frac{CSI - \overline{y}}{1 - \overline{y}} \tag{3.4}$$

where $\overline{y}$ is the base rate of the dataset (e.g., the ratio of target grid points in the testing data set to total grid points in the testing data set). For any-severe, severe wind, severe hail, and tornadoes, the base rates (expressed here as a percentage) are approximately 4.9%, 2.9%, 2.3%, and 0.8% respectively. An additional metric that we use to summarize the curves on a performance diagram is the Area Under the Performance Diagram Curve (AUPDC). This is calculated via the difference in POD for bin k and $k - 1$ weighted by the SR of the kth bin.

$$AUPDC = \sum_{k=1}^{K} (POD_k - POD_{k-1}) * SR_k \tag{3.5}$$

Similarly to NCSI, we scale the AUPDC by the unobtainable area to the left of the base rate to give the normalized AUPDC or NAUPDC (Flora et al. 2021; Miller et al. 2022).

$$NAUPDC = \frac{AUPDC - \overline{y}}{1 - \overline{y}} \tag{3.6}$$

We also calculate the Probability of False Detection, POFD, which is paired with POD to create Receiver Operating Characteristic (ROC) curves:

$$POFD = \frac{b}{b + d} \tag{3.7}$$

While NCSI and NAUPDC are useful as summary metrics, the information contained therein does not explicitly provide details on the skill of predictions at various thresholds (i.e. various curves on a performance diagram can yield identical values of NCSI and NAUPDC).

While not derived from the 2x2 table, we also consider the Brier Score as a metric. The Brier Score is given by the sum of three terms:

$$BS = \frac{1}{N} \sum_{k=1}^{K} n_k (x_k - \overline{y}_k)^2 - \frac{1}{N} \sum_{k=1}^{k} n_K (\overline{y}_k - \overline{y})^2 + \overline{y}(1 - \overline{y}) \tag{3.8}$$

where the predictions (x) are sorted and broken into K bins. $\overline{y}_k$ is the base rate of the kth bin, while $\overline{y}$ is the base rate of the dataset. The first term is known as the reliability term and is a weighted average of the squared error between the kth bin's mean forecast value x_k and the bin's base rate. The second term, resolution, is a weighted average of the squared difference between the kth bin's base rate and the dataset's base rate. The

final term (uncertainty) is only a function of the dataset's base rate, and as such, is identical for the BL and ML. Unlike the previous metrics, a perfect Brier Score holds a value of zero. Thus, we generally prefer the Brier Skill Score (BSS) as it is a positively oriented skill score. The BSS is given by the following as a function of the Brier Score of a forecast system (BS) and the Brier Score of a reference forecast BS_{ref}:

$$BSS = \frac{BS - BS_{ref}}{0 - BS_{ref}} = \frac{resolution - reliability}{uncertainty} \tag{3.9}$$

As indicated, the BSS can also be written as a function of the resolution, reliability, and uncertainty terms. The reference forecast's Brier Score (BS_{ref}) is the Brier Score associated with forecasting the base rate ($\overline{y}$) of the dataset. As such, BS_{ref} is identical for the BL and ML of a specific hazard, but each hazard has a unique BS_{ref}.

Chapter 4

Results

4.1 Performance of Traditional Machine Learning Techniques

4.1.1 Objective Skill

To begin the evaluation of the products, we first consider the performance diagrams displayed in Figure 4.1. These plot the SR and POD for three different products: HGBT (red), LR (blue), and the baseline (black). Each panel corresponds to a different hazard. Within each diagram, the probability threshold for a yes forecast increases from top left to bottom right. The location of the maximum CSI is denoted by a marker on each curve. A perfect model will have a curve in the upper right-hand corner of the diagram.

The evaluation is rather straightforward; as indicated in Figure 4.1, the ML products outperform the baselines in every hazard and for all considered summary metrics. Additionally, this holds true for both ML architectures across all probability thresholds. The largest increase in performance compared to the baseline is seen for severe wind; the ML products nearly double the baseline performance for this hazard. While the HGBT outperforms the LR for any-severe and severe wind, the LR remains competitive across all four hazard types (i.e. the 2σ confidence intervals overlap). Of the

hazards, the any-severe and severe wind products generally exhibit the highest success ratios. The severe hail and tornado products generally exhibit a smaller increase in success as the probability threshold increases.

Figure 4.2 plots the receiver operating characteristic (ROC) curves for each of the hazard types. Similar to the performance diagram, the y-axis plots POD. However, the x-axis is now the POFD. The AUC, or area under the ROC curve, is a summary statistic that measures how well the products discriminate between event and non-event samples. Given the context of severe weather, it is fairly easy to achieve a high AUC by correctly ranking non-events. This is due to the high number of clear-air grid points; e.g., predicting a probability of 1 at every grid point with reflectivity values exceeding 40 dBZ achieves an AUC of 0.76 for any-severe hazards. However, these diagrams still retain useful information.

Notably, the ML products generally exhibit higher PODs and lower POFDs than the baseline. As such, the ML again provides more skillful guidance than the baseline. The difference between HGBT and LR is marginal. Similar to the performance diagram, the largest difference between HGBT and LR occurs in the any-severe product's performance metrics. The severe wind ML products show the largest improvement in discriminative ability over the baseline (Fig. 4.2b). This is followed closely by the ML's improvement over the baseline for the tornado guidance (Fig. 4.2d). The ML tornado guidance achieves the highest overall discrimination, as measured by the AUC. However, as previously mentioned, this is primarily due to the correct ranking of non-events.

The final verification diagram we utilize is Figure 4.3, known as the reliability diagram. The guidance is sorted and broken into bins with a bin width of 0.1; the number of samples in each bin is displayed as a histogram in the background. The curves then plot the mean forecast probability of each bin against the conditional event

frequency of that bin. The one-to-one line is displayed in each panel as a dashed line. A curve's distance from this line is a measure of its reliability or calibration. Any system along this line would be identified as having a perfect reliability component of the Brier Score (e.g., 0). Systems above this line are identified as having an underprediction bias, while systems below the line are overpredictive.

As seen in Figure 4.3, the ML guidance for each of the hazards is reasonably reliable with a tendency to overpredict. Again, the results are nearly identical to those discussed previously. Both ML architectures handily outperform the baselines across each hazard. While the BL probabilities are well calibrated, they are also low amplitude. The largest improvements are again seen for the severe wind and tornado products. However, unlike the previous diagrams, the reliability diagrams reveal a fairly large distinction between the HGBT and LR; the LR tends to output higher probabilities than the HGBT. However, these high probabilities are over-predictive. As such, the LR is generally a less reliable system than the HGBT. Some users may be willing to trade off the decreased reliability for the increased amplitudes of the LR. The severe hail products are an exception to the trend of the HGBT output having better reliability than the LR. As seen in Figure 4.3c, the HGBT outputs poorly calibrated probabilities at the top end. As such, the HGBT's BSS is decreased compared to the LR. A major limitation of the interpretation of these results is the small sample size in higher probability bins. As a result, the true performance of the ML guidance at higher probabilities is fairly nebulous given the high uncertainty. This is addressed in part in Section 4.4 through the use of a substantially larger dataset.

4.1.2 Case Study

As a case study, we consider the guidance created from the 2200 UTC WoFS forecast on 23 May 2019. The SPC issued a moderate risk for this day with a primarily hail and tornado-based risk (SPC 2019). The WoFS domain was located over the Texas panhandle. The guidance is valid from 00-04 UTC on 24 May 2019; Figure 4.4 shows the guidance created by the baselines for each hazard, while Figures 4.5 and 4.6 show the guidance created by the LR and HGBT respectively. The case studies selected within this work are not meant to highlight specific failure modes of the guidance. Rather, they are intended to provide visualizations of how the guidance from each architecture behaves on a typical case.

As shown in Figure 4.4a, the any-severe UH baseline correctly highlights the central region of the domain where tornadoes were reported as the primary risk area. However, there is little variation in the probabilities within that region. The any-severe baseline also highlights the northeast and southwest clusters of reports, but incorrectly places higher probabilities on the northeastern region where only a few hail reports were made. Additionally, the any-severe baseline misses the northernmost cluster of reports within the domain. The wind baseline has an extremely high FAR for this case; much of the central to northeastern domain is highlighted, while only three reports are made during this window. The baseline guidance suggests that the wind threat is farther northeast. The hail baseline misses the same cluster of reports in the northern region of the domain as the any-severe. While the hail baseline highlights the central corridor, it does not feature a region of elevated probabilities for the southern cluster of hail reports. The tornado baseline correctly captures the tornado events; however, it features similar amplitude probabilities across a large portion of the domain.

Figure 4.5 displays the LR guidance for the aforementioned case. Once again, the any-severe guidance captures the central region where tornadoes occurred. However,

the LR also extends the probabilities to capture the hail reports in the northern region of the domain. Additionally, probabilities are raised near the southwestern cluster of reports. While the probabilities are fairly expansive across the central corridor, the guidance correctly highlights the tornadoes with higher probabilities. The severe wind guidance experiences a similar issue to the baseline. A substantial false alarm is produced for the central and northeastern regions of the domain. While the LR produces a hotter false alarm in the central region, it also produces slightly fewer false alarms in the southwestern and northern regions of the domain when compared to the baseline. The hail LR clearly produces two regions of interest associated with the central and southwestern clusters of reports. Additionally, the LR expands the probabilities to include the reports in the northern region that the baseline misses. The tornado guidance features slivers of elevated probabilities along the central corridor; this is more confident than the baseline, but not as confident as the HGBT guidance shown in Figure 4.6d. The LR tornado guidance also experiences a higher FAR than the baseline and HGBT in the northeastern region of the domain.

Finally, we consider the HGBT guidance in Figure 4.6. The any-severe guidance captures both the central corridor where tornadoes occurred, as well as the northeastern and southwestern regions. The probabilities in the northeastern and southwestern regions specifically are elevated above the baseline and LR. Similar to the LR, the any-severe guidance again captures the hail reports in the north of the domain that the baseline misses. As discussed previously, the severe wind guidance experiences a large false alarm over much of the domain. While the amplitude of the false alarm is smaller than the LR, it occurs over a larger region. The hail guidance creates a region of elevated probabilities along the main corridor and captures the southernmost reports better than the LR. However, the HGBT also misses the hail reports in the central northern region of the domain. The HGBT tornado guidance is the most skillful for

this case; it features higher probabilities collocated with the tornado reports than the LR and baseline. Additionally, it has a smaller false alarm in the northeastern region of the domain. However, it produces a secondary erroneous maximum in the southwest that neither the LR nor the baseline produces.

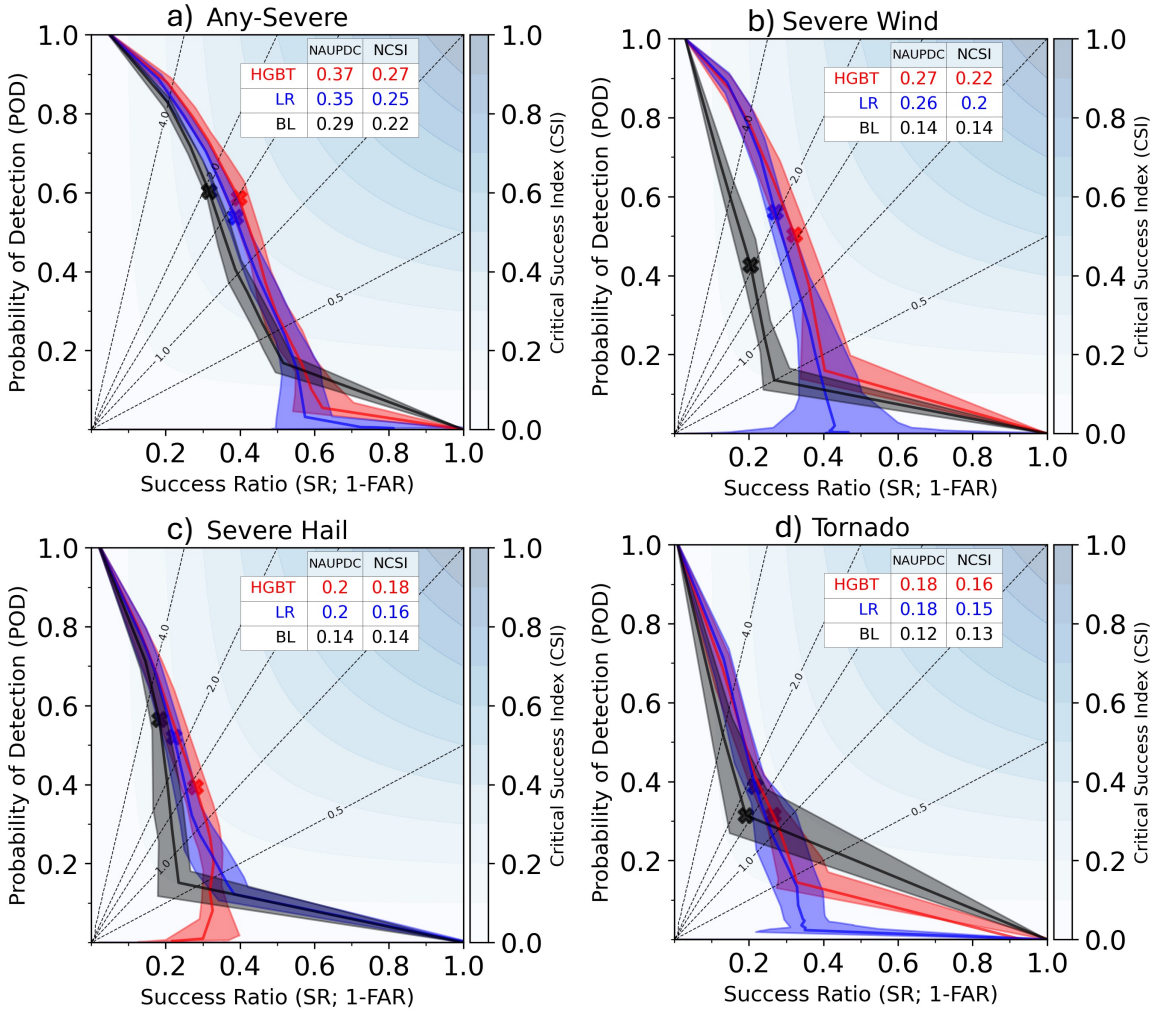

Figure 4.1: Performance diagrams evaluating HGBT (red), LR (blue), and baselines (black) on the initial testing set for a) any-severe, b) severe wind, c) severe hail, and d) tornado. The threshold for a yes forecast increases from the top left to the bottom right. Shading indicates the 2σ confidence interval. For all four hazards, HGBT is generally the best-performing architecture. The largest improvement over the baseline occurs for severe wind; HGBT is a 50% increase over the baseline for severe hail and tornado.

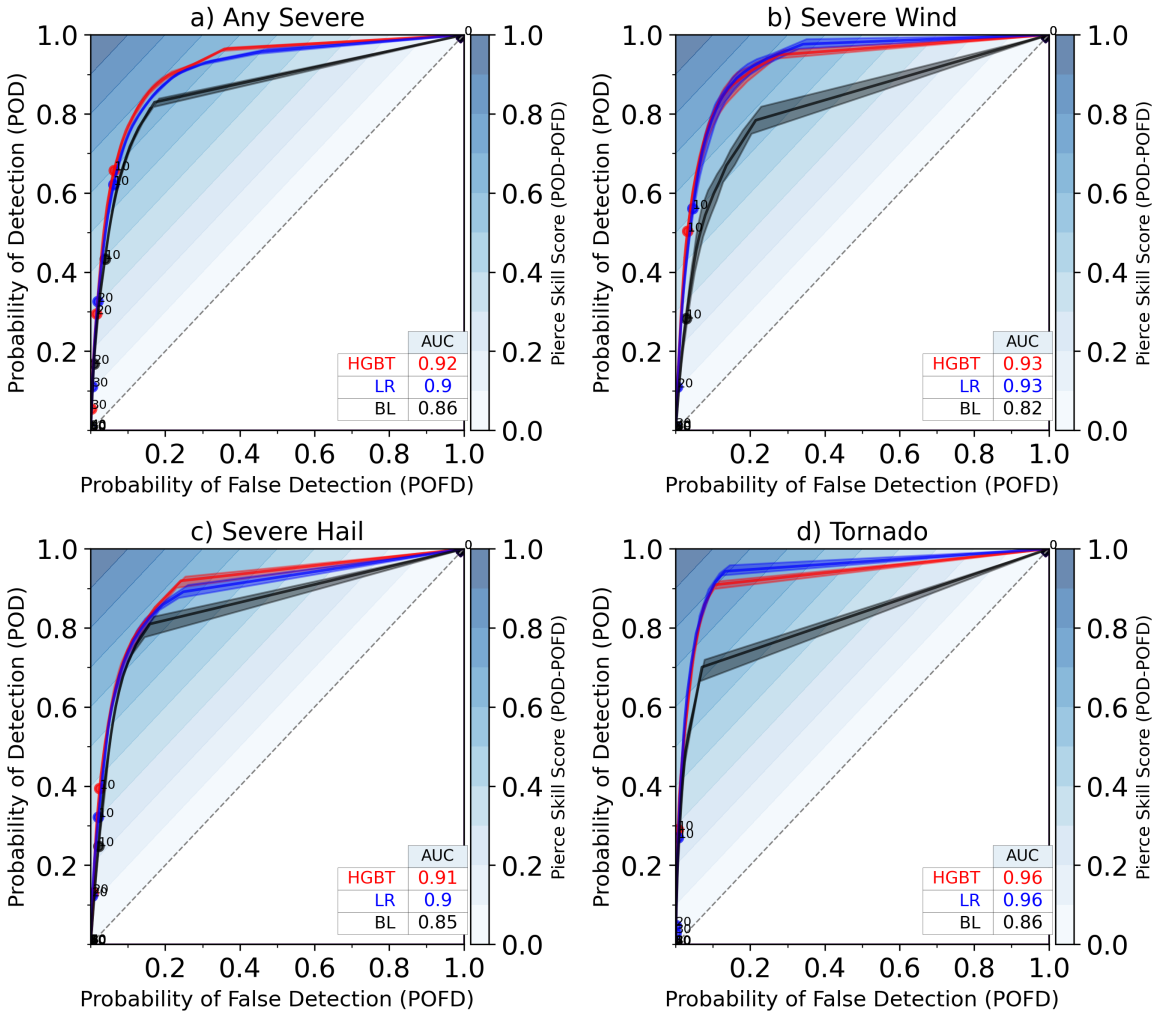

Figure 4.2: Receiver-Operating Characteristic (ROC) diagrams evaluating HGBT (red), LR (blue), and baselines (black) on the initial testing set for a) any-severe, b) severe wind, c) severe hail, and d) tornado. Shading indicates the 2σ confidence interval. The severe wind and tornado ML products experience the largest improvement in discrimination compared to their respective baselines.

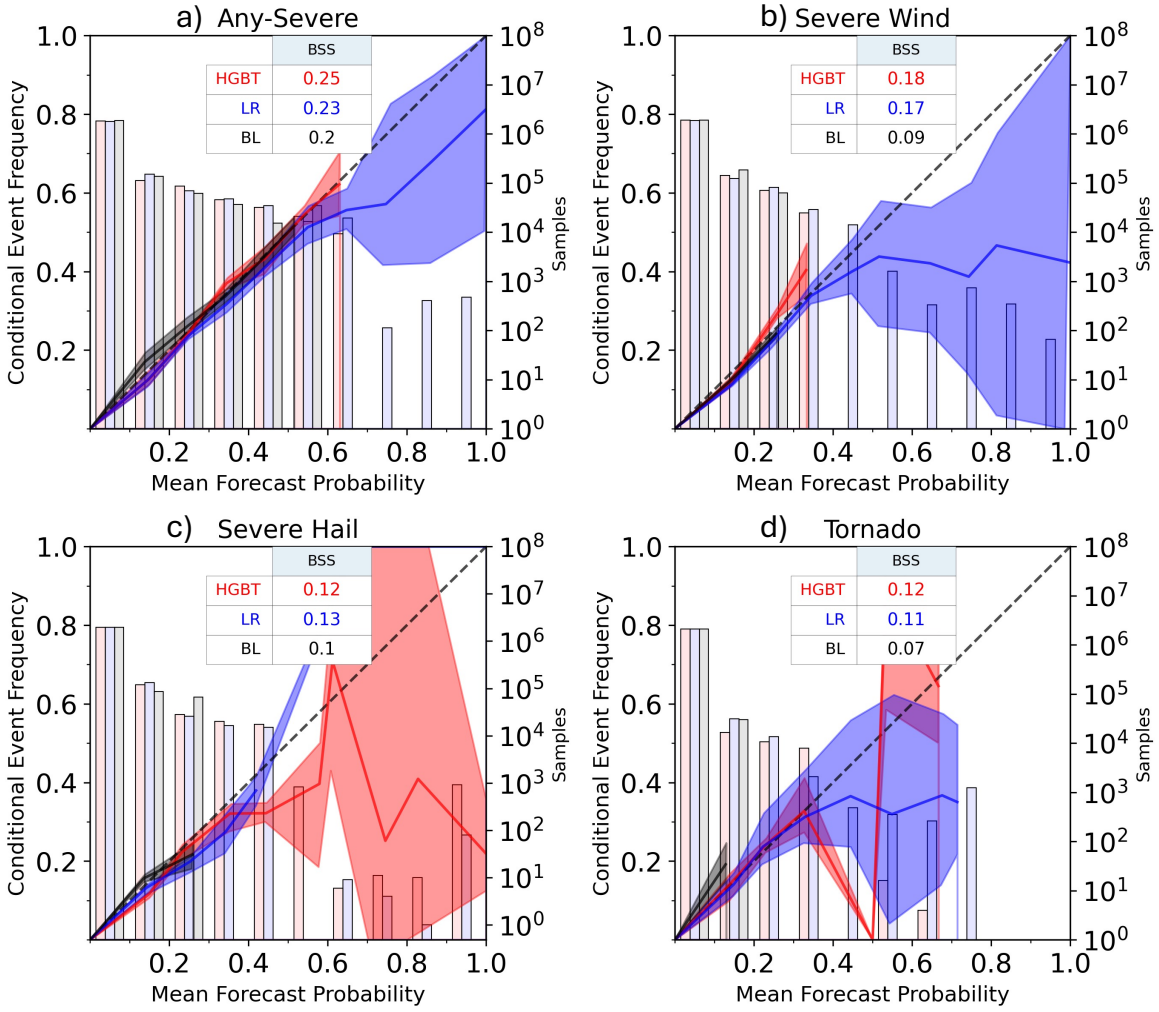

Figure 4.3: Reliability diagrams evaluating HGBT (red), LR (blue), and baselines (black) on the initial testing set for a) any-severe, b) severe wind, c) severe hail, and d) tornado. Shading indicates the 2σ confidence interval. The baselines are reliable but have limited amplitudes. LR and HGBT both tend to overpredict at higher probabilities. With the exception of severe hail, the HGBT are susceptible to overprediction than LR. As measured by the BSS, the HGBT again is generally the best-performing architecture.

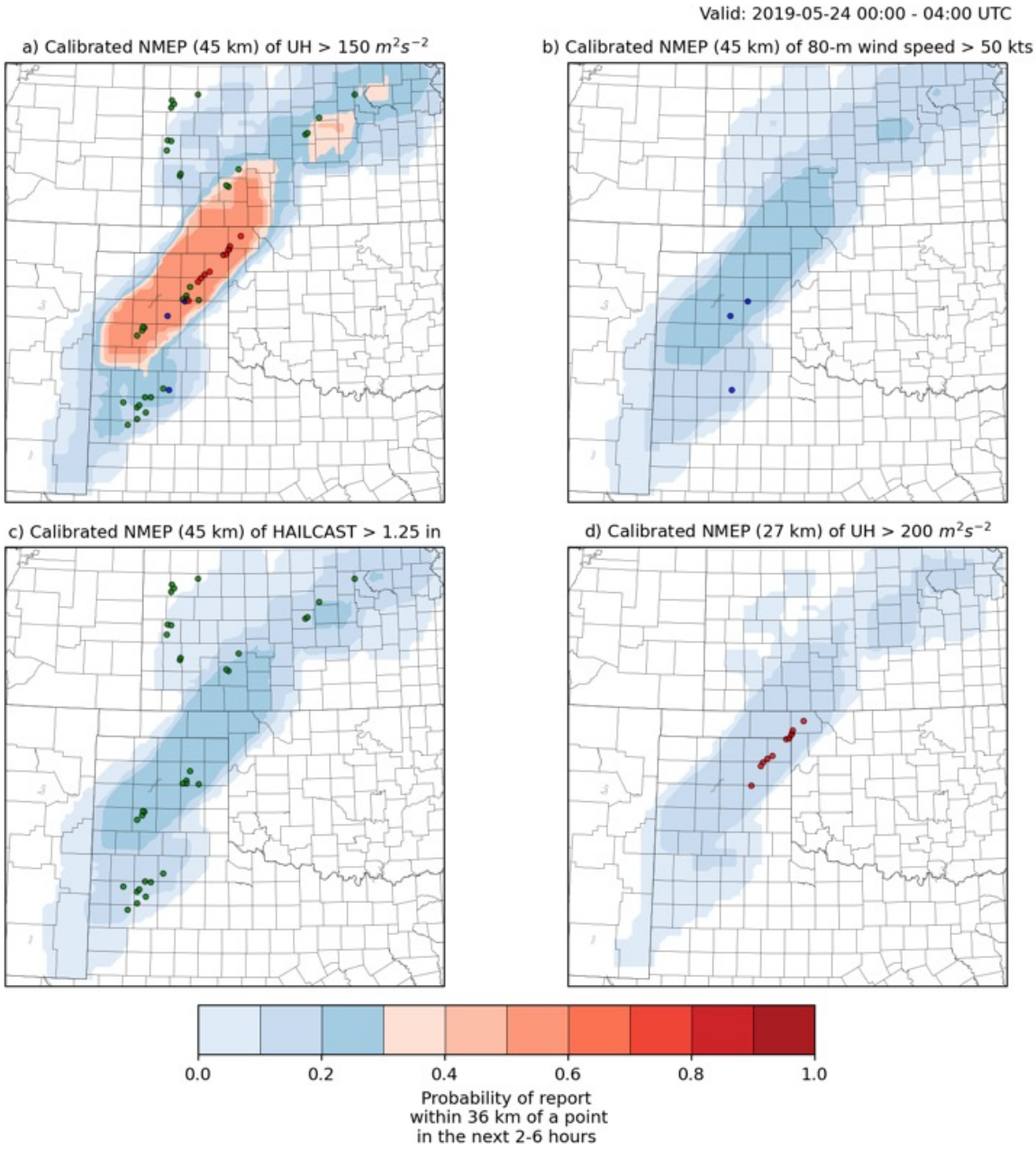

Figure 4.4: Watch-to-warning guidance produced using the WoFS forecast initialized at 2200 UTC on 23 May 2019. The panels correspond to the NMEP baselines for a) any-severe, b) severe wind, c) severe hail, and d) tornado. The guidance is valid from 00-04 UTC on 24 May 2019. Reports from the guidance window are also plotted, with the blue, green, and red points corresponding to wind, hail, and tornado reports. While the any-severe product performs well, the wind guidance exhibits a large false alarm.

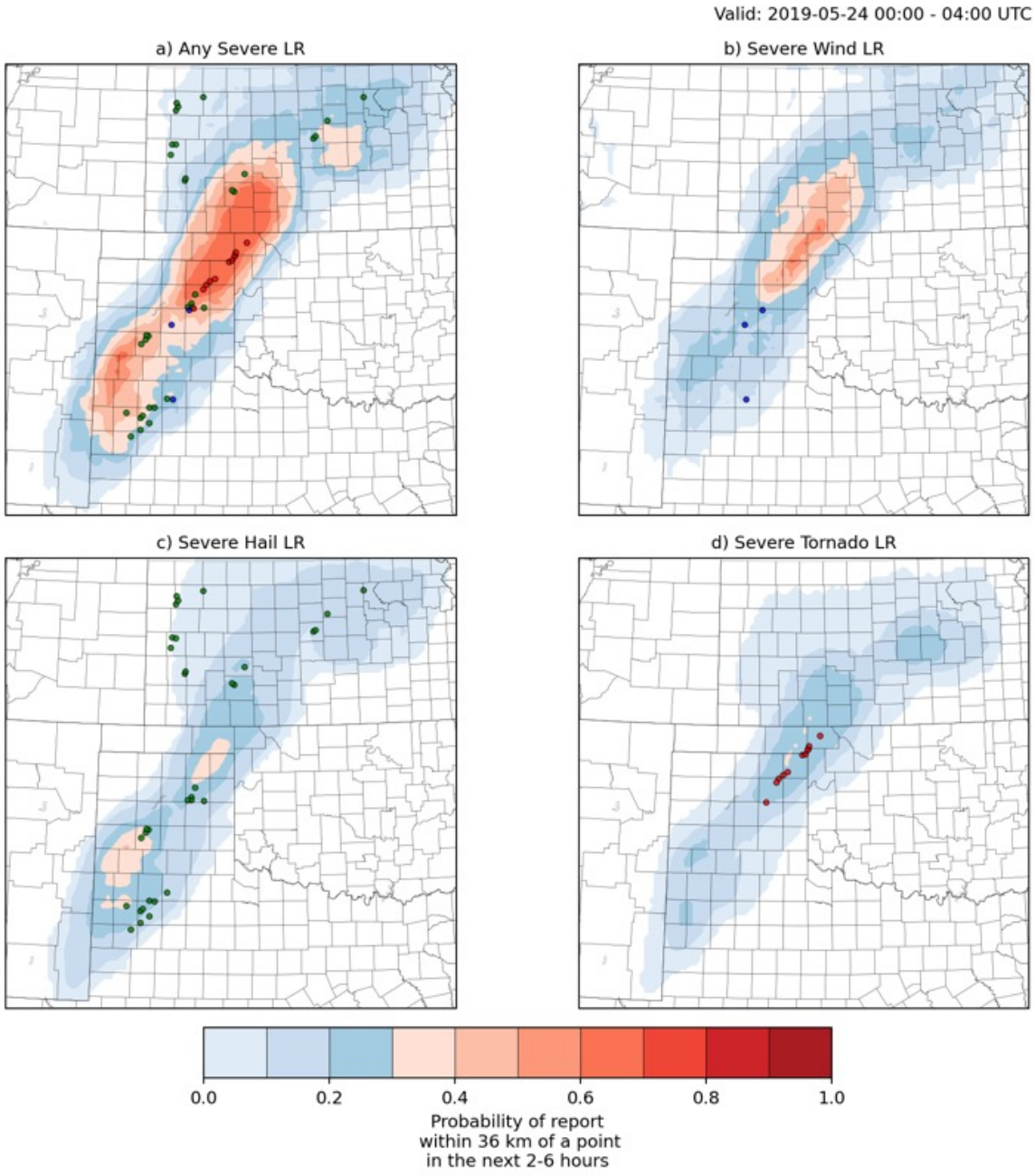

Figure 4.5: As in Fig. 4.4, but for logistic regression. While the severe hail and tornado guidance correctly highlight regions of interest, the severe wind guidance exhibits a substantial false alarm.

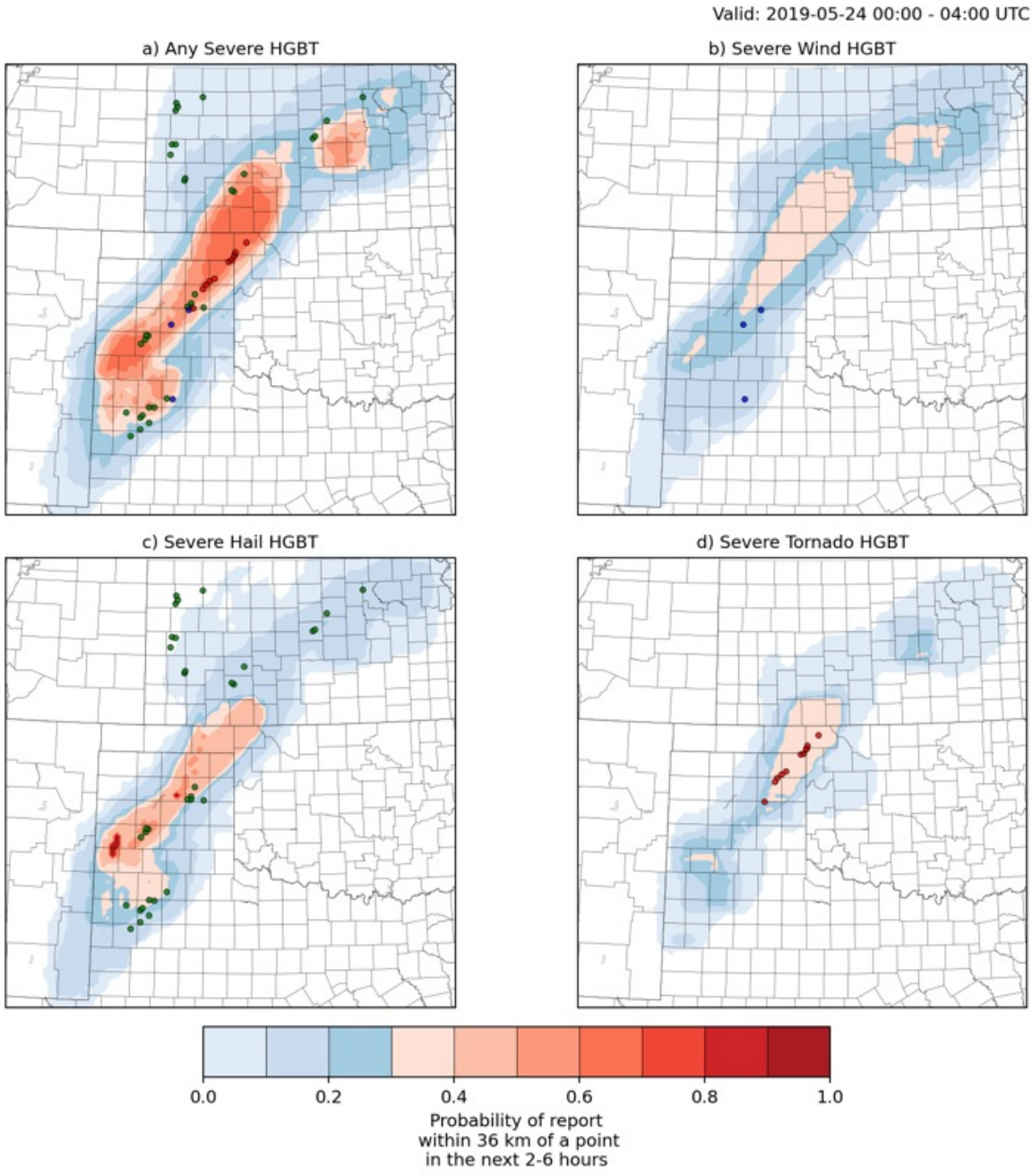

Figure 4.6: As in Fig. 4.4, but for HGBT. For this case, the tree-based any-severe and tornado products perform better than those produced by other methods.

4.2 Stratified Verification

4.2.1 Skill by Initialization Time

As an additional measure of verification, we consider the performance of the ML and BL guidance at various initialization times for each of the hazards. We again use the testing dataset, but rather than a bulk verification over the entire dataset, we calculate performance metrics as a function of initialization time. This provides insight into diurnal variations in skill within the guidance. However, as the WoFS domain can be relocated between cases (e.g., Figure 3.1), one initialization time may include data from a range of local times. As an example, initialization times of 2200 UTC could correspond to both 1900-2300 CDT and 2000-0000 EDT. Additionally, rather than consulting verification diagrams for each of the initialization times, we consider only the associated verification metrics. While this allows for a more succinct analysis, it is not exhaustive as curves can have similar metrics while being associated with substantially different characteristics.

Figure 4.7 shows the results of this process for the any-severe guidance. In general, the skill of the guidance decreases as the initialization time increases. Both the ML and BL guidance experience a large reduction in skill after 1800 UTC; however, the amplitude of this skill reduction is substantially larger for the ML. This is likely driven by a decrease in skill on severe wind, as the severe wind products mirror this trend. The cause of this trend is currently unidentified. After 1900 UTC, the ML performance generally exhibits a positive or constant trend until 0000 UTC. After 0000 UTC, the HGBT and BL guidance again experience a reduction in skill. However, the LR NAUPDC and NCSI remain fairly stable after 1900 UTC. Despite the loss of skill over time, the ML guidance outperforms the BL at each initialization time. The HGBT typically outperforms the LR, although the LR obtains an advantage at the

latest initialization times. The generally decreasing skill of the guidance with initialization time is fairly unexpected. The expected performance would increase over time as convection is initiated and becomes better assimilated within the forecast system. Although not shown within this work, this behavior is observed when conducting a similar verification on the final dataset.

The stratified verification of the severe wind products is shown in Figure 4.8. The severe wind products experience a similar trend to the any-severe guidance; a large reduction in skill occurs for initializations after 1800 UTC. In this case, the ML guidance decreases in skill by nearly 40%. This is likely accentuated by the test dataset having fewer total samples at the earliest initialization times as shown in Figure 4.8d. Once again, the ML guidance outperforms the baseline at every initialization time. The baseline NAUPDC and BSS remain fairly stable from 2000 UTC to 0100 UTC. However, the baseline's NCSI and BSS both decrease after 0100 UTC. Comparatively, the ML guidance exhibits a general increase in skill from 2300 UTC onwards. The HGBT once again outperforms the LR at most initialization times.

Figure 4.9 displays the stratified verification results for the severe hail guidance. Unlike the any-severe and severe wind products, the severe hail products exhibit a marked increase in skill for initialization after 1800 UTC. The severe hail ML guidance outperforms the baseline at most times. The LR guidance generally increases in skill for later initialization times and is typically more skillful than the baseline. The HGBT is also typically more skillful than the baseline, but the BSS experiences a sharp decrease at 0100 UTC resulting in the HGBT being less reliable than the baseline at that time. Similar to the any-severe guidance, the LR outperforms the HGBT at later initialization times. The baseline experiences a sharp dropoff in skill after 0000 UTC, resulting in a generally parabolic trend. Conversely, the ML products have a positive trend.

Finally, we consider the stratified verification of tornado guidance as shown in Figure 4.10. Due to the substantially lower base rate of tornadoes, these results generally have a higher degree of variability than the previous hazards. Similar to the previous hazards, the ML guidance outperforms the baseline for most initialization times. The baseline exhibits a notable decrease in NAUPDC and NCSI after 2000 UTC when the base rate begins to decline. Neither the LR nor the HGBT share this trend. While the LR has a higher NAUPDC and NCSI than the HGBT at most initialization times, the HGBT generally achieves a higher BSS. Both the ML and the baseline experience a drastic reduction in skill after 0000 UTC. This is expected, as the base rate is so low at these times that the performance is dictated by only a few positive target samples. However, while the baseline eventually achieves a negative BSS, the HGBT guidance retains a positive BSS.

4.2.2 Impact of Removing Cases

As additional insurance, we check to ensure that the testing set does not overly favor the ML. This is done to verify that the ML guidance is a robust system and that its improvements over the baselines are systematic and not the result of good performance on a few, isolated cases. To that end, we perform two additional evaluations of the any-severe product. The first removes the five cases where the ML achieved the highest BSS and evaluates the remaining data. The second removes the five cases where the BL achieved the lowest BSS and evaluates the remaining data.

Figure 4.11 shows the performance of the any-severe guidance on the initial testing set after removing the five best cases for the ML products. Comparing against Figures

4.1a, 4.2a, and 4.3a, all of the performance metrics are lowered. The discriminative ability of the guidance (as measured by the AUC) experiences the least change across both the ML and BL guidance. Comparatively, the NAUPDC and NCSI of all systems decrease by about 28%; the baseline experiences the smallest reduction in magnitude, while the HGBT experiences the largest. This is expected, as we specifically removed cases where the ML performed well. While the magnitude of the reduction is smaller, the BL experiences a similar percentage change in its performance metrics. Similarly, all three systems experience a similar decrease in BSS. The removal of the cases enhances the overprediction bias that was already present. Again, this was expected as cases were removed based on the BSS. The performance and reliability of all systems degrading suggests that the BL and ML perform well in similar scenarios; these scenarios are likely when the WoFS forecasts correspond well with the observations. However, even removing these cases, the ML remains a more skillful system than the baselines.

Conversely, Figure 4.12 shows the performance of the any-severe guidance when the dataset favors the BL. Here, we drop the five cases where the BL achieves the lowest BSS. As seen in Figure 4.12a, dropping these cases slightly increases the performance of all systems. The improvement for the BL is nearly 10%, while the ML guidance experiences little change. The AUC of the ML guidance deteriorates slightly when removing these cases. Two of the five removed cases had no associated reports within the WoFS domain; as such, removing these cases removed many points that were easy to correctly rank as non-events. This results in a slightly lower level of discrimination. The reliability of the guidance remains unchanged from the full testing dataset. As previously mentioned, two of the five cases had no associated reports and low probabilities. Thus, removing these cases had a marginal effect on the BSS. Once again, the ML outperforms the baselines even when the dataset favors the baseline. As such, we ensure that the ML is providing a routine and robust improvement over the baselines.

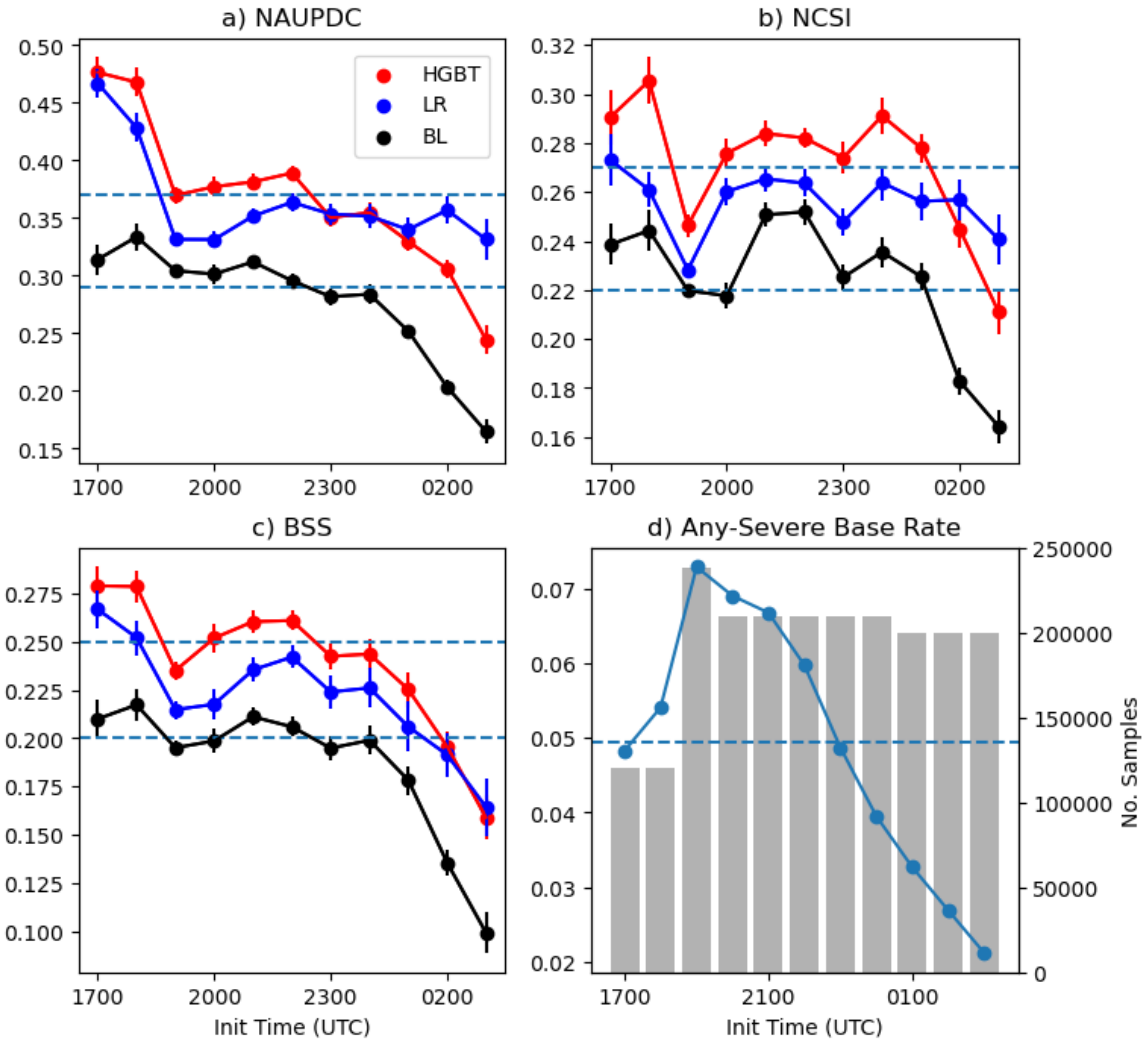

Figure 4.7: Panels a-c show the any-severe verification metrics calculated on the initial testing set stratified by initialization time for HGBT (red), LR (blue), and BL (black). The lower (upper) dashed lines indicate the lower (upper) bounds of the metric from the unstratified verification. Error bars indicate the 2σ confidence interval. Panel d) shows the base rate ($\overline{y}$; see Section 3.5) of each initialization time (blue), as well as the number of samples in the test set. A large drop in skill is observed after 1800 UTC, likely a result of severe wind.

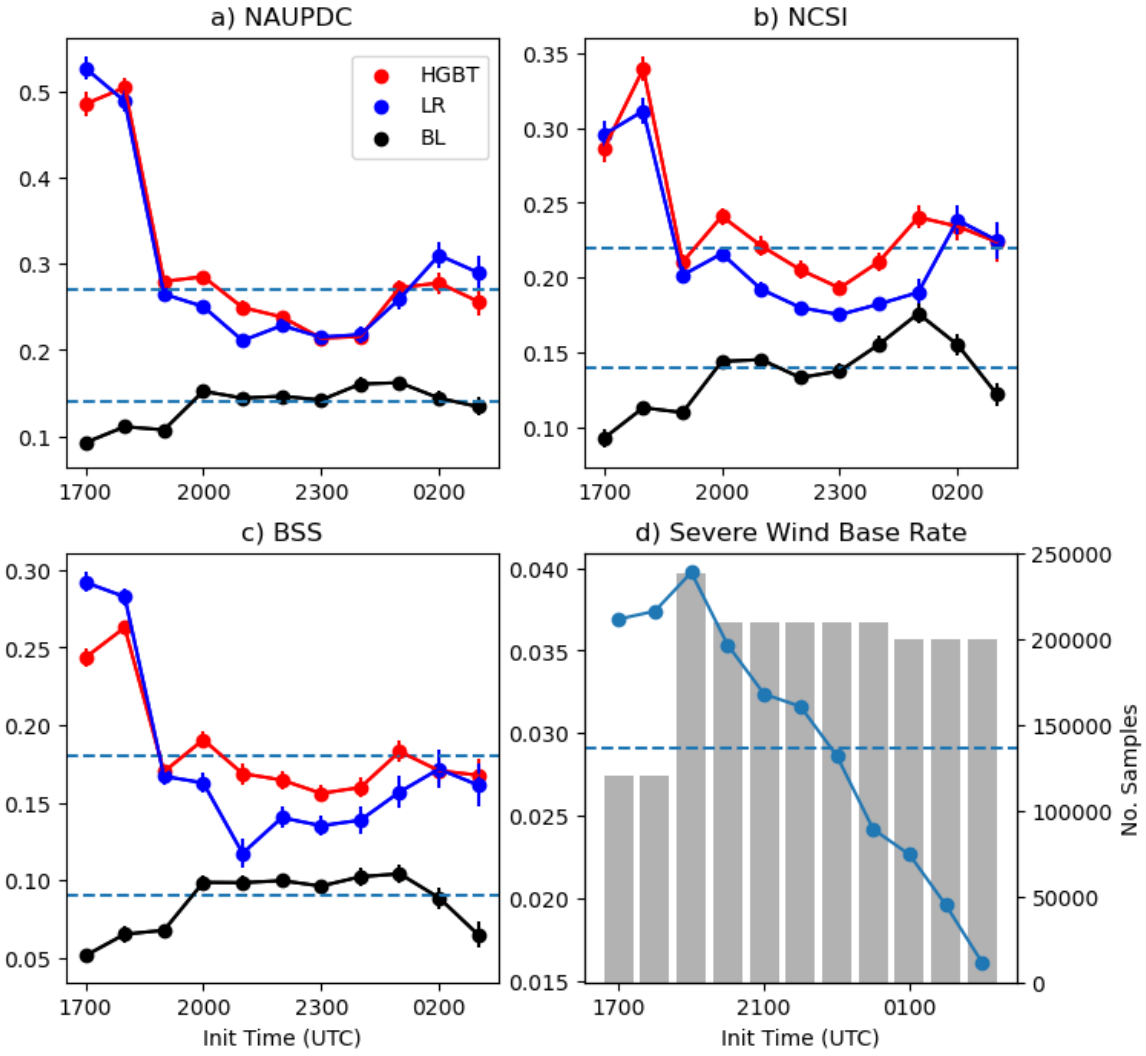

Figure 4.8: As in Figure 4.7, but for severe wind. After 1800 UTC, the ML performance remains fairly stable.

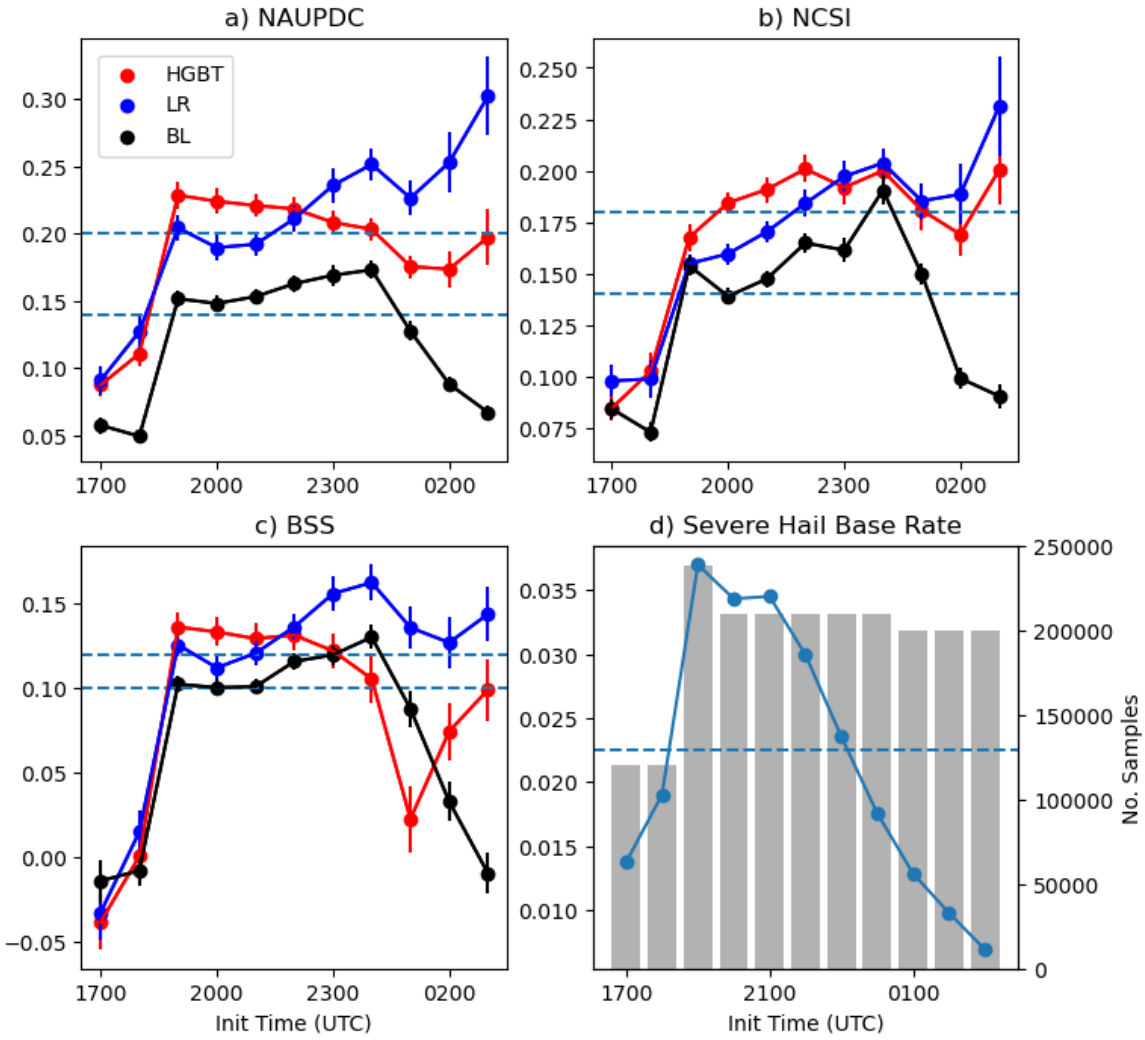

Figure 4.9: As in Figure 4.7, but for severe hail. The LR tends to outperform the HGBT at later initializations, especially given the degradation of reliability in the HGBT after 0000 UTC.

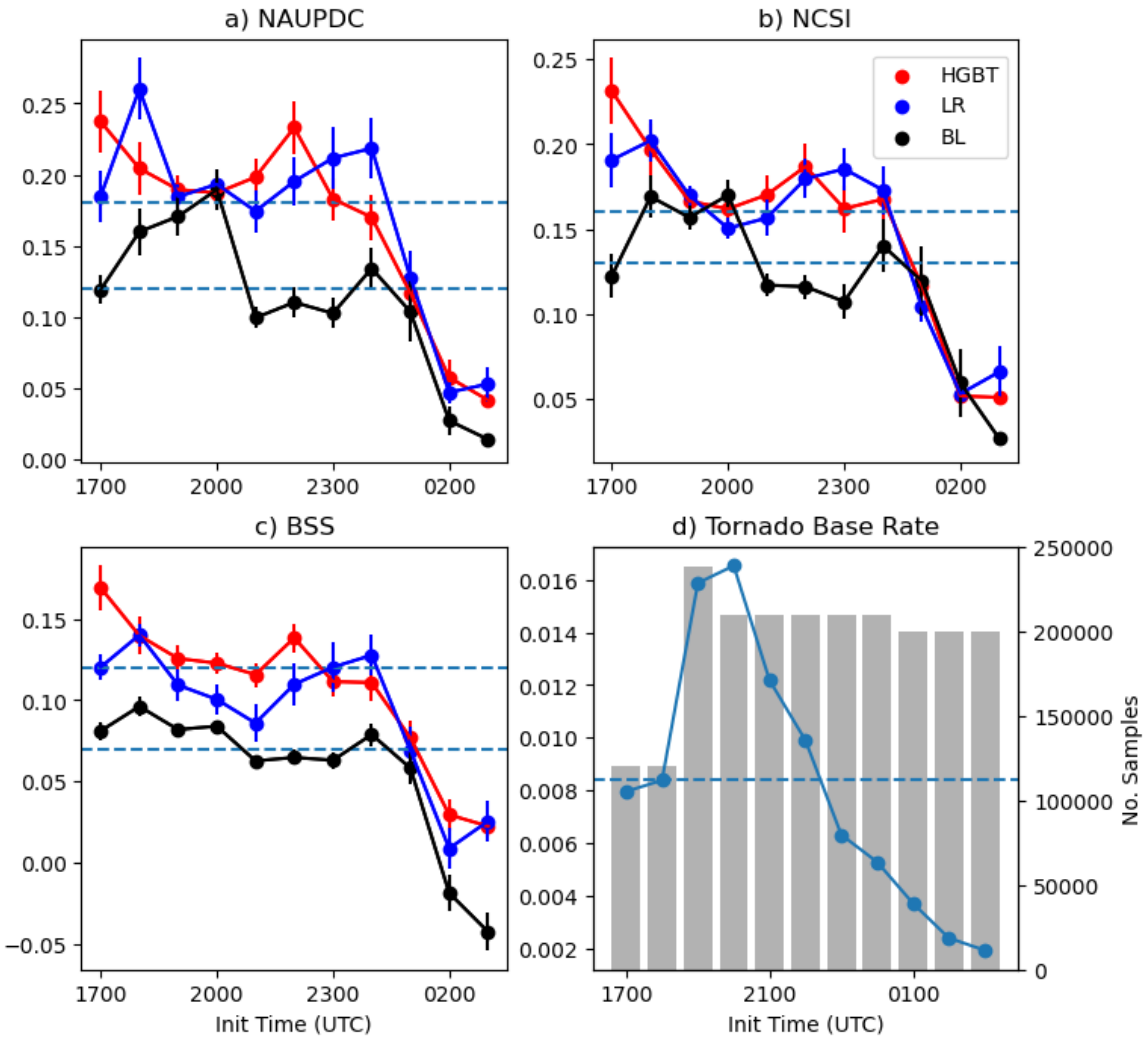

Figure 4.10: As in Figure 4.7, but for tornadoes. All models exhibit a substantial decrease in skill after 0000 UTC, likely due to the lower base rate of tornadic events.

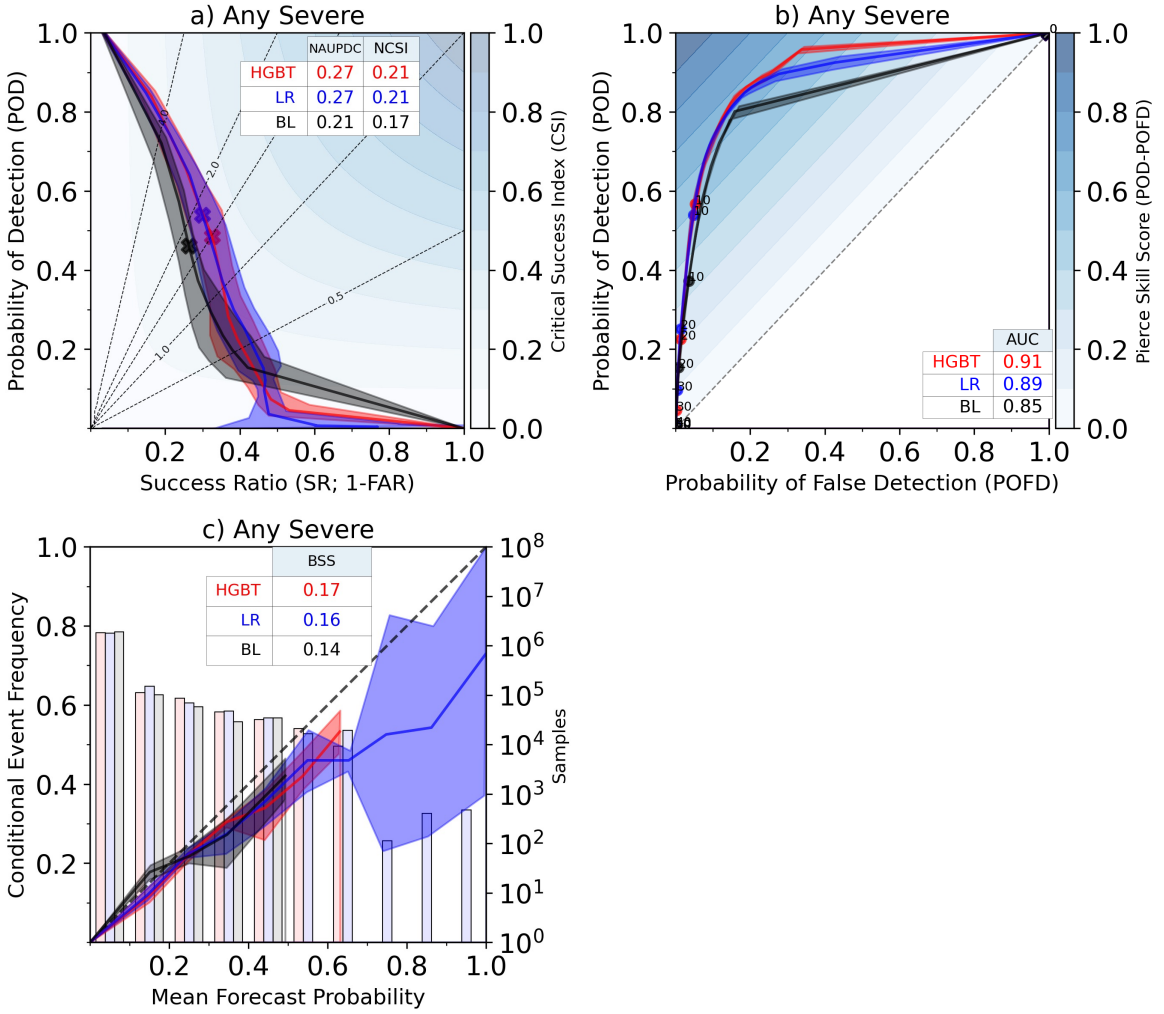

Figure 4.11: Performance diagrams (a), ROC curves (b), and reliability diagrams (c) for the any-severe HGBT (red), LR (blue), and baseline (black). These evaluate the performance of the guidance on the initial testing set after dropping five days where the ML had the highest BSS. Shading indicates the 2σ confidence interval. All models now show a larger bias towards overprediction and routinely lowered POD and SR.

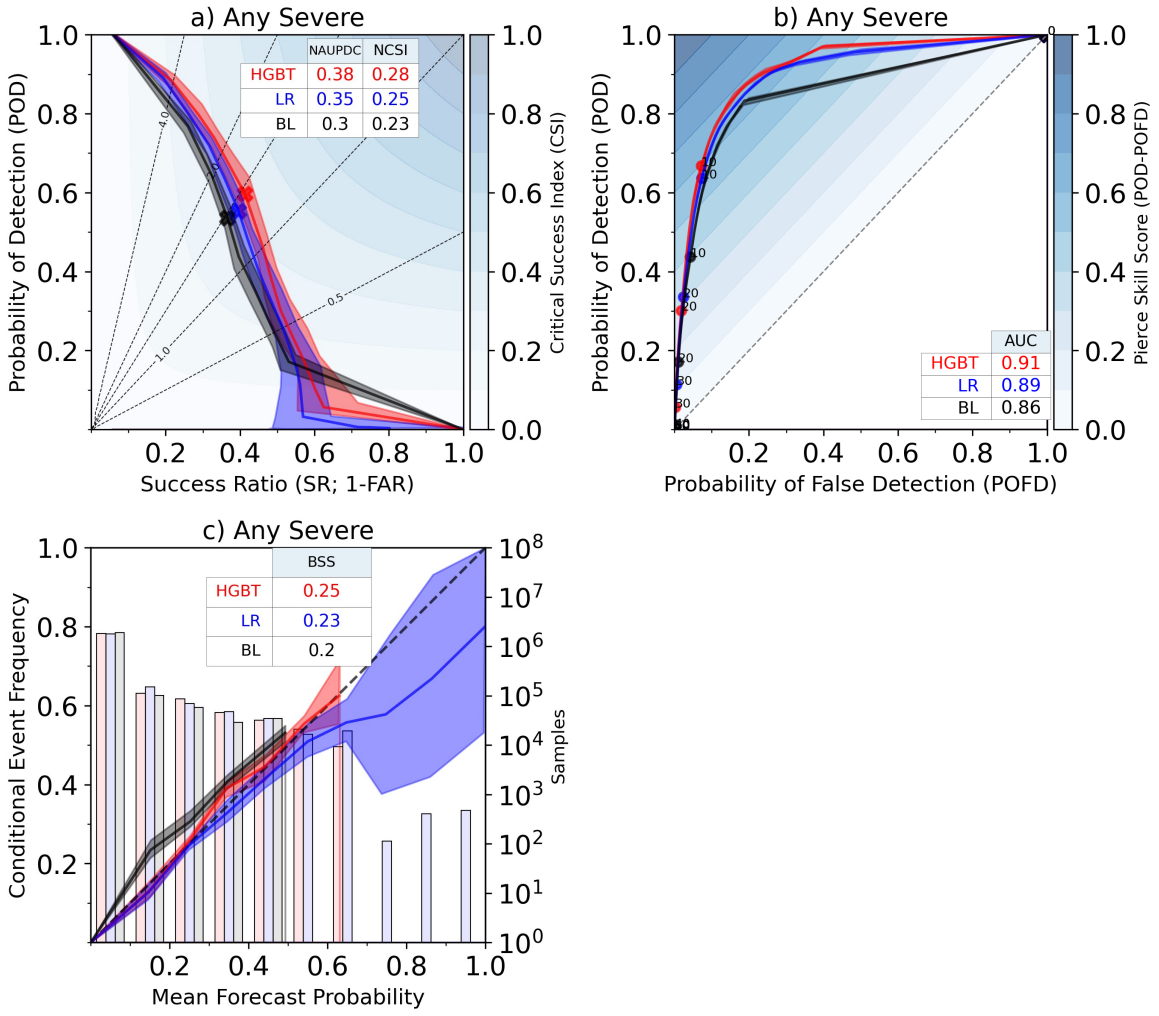

Figure 4.12: Performance diagrams (a), ROC curves (b), and reliability diagrams (c) for the any-severe HGBT (red), LR (blue), and baseline (black). These evaluate the performance of the guidance on the initial testing set after dropping five days where the BL had the lowest BSS. Shading indicates the 2σ confidence interval. While the discriminative ability of the ML guidance decreased, the performance and success ratio are slightly higher than on the full testing dataset.

4.3 Feature Ablation

To garner insight into which predictors are providing the largest benefit, we perform a feature ablation test with groups of predictors. Rather than running a full permutation of all 174 predictors, we categorize the predictors based on their field (intrastorm or environmental) and their degree of spatial smoothing (9, 27, or 45 km). We then train various instances of the ML architectures using combinations of these predictors.

The results of this are shown in Figures 4.13-4.16 with the figures corresponding to any-severe, severe wind, severe hail, and tornado respectively. Due to the volume of information, we show the results as scatter plots of the Brier Skill Score and AUPDC. While this reduces the quantity of information when compared to the verification diagrams utilized previously, it allows for a more succinct analysis of the bulk performance of these models. However, it no longer explicitly conveys information about how the products perform at various probabilities.

Based on this test, the bulk of the predictive skill comes from intrastorm predictors. Models incorporating these features, either alone or in combination with environmental features, generally perform higher than the baselines. However, models using only environmental features generally perform worse than the baseline. These results agree with the findings of Clark and Loken (2022). The notable exception is for severe wind; the HGBT environment-only severe wind guidance outperforms the baseline while the environment-only LR does not. Additionally, severe wind guidance sees an improvement in both reliability and AUPDC when both intrastorm and environmental fields are used in combination (Fig. 4.14). A similar trend is observed when utilizing both types of predictors for tornado guidance (Fig. 4.16); however, the inclusion of environmental features mainly acts to improve the reliability of the tornado guidance.

The inclusion of multi-scale features has a marginal impact on the skill of the guidance. Among the highest-performing models, there is little distinction between the various predictor scales. Additionally, there is no clear pattern in how the predictor scales impact the skill of the guidance using intrastorm or both types of predictors. However, multi-scale features do improve the skill of guidance produced with only environmental predictors. This is most apparent in the LR for any-severe (Fig. 4.13), both ML architectures for severe wind, and the HGBT for severe hail (Fig. 4.15). Regardless of the improved performance from multi-scale features, the guidance using only environmental predictors remains less skillful than other configurations with intrastorm predictors. It is possible that the predictor scales evaluated herein are too similar for substantial differences to appear; future work may consider incorporating predictors from a substantially larger spatial scale.

Finally, the results show that the HGBT guidance tends to be more skillful than the LR. Within each hazard, the top-performing guidance tends to come from the HGBT. This is mostly invariant of the predictor types used; the HGBT generally has the highest performance for intrastorm, environmental, and combined predictor sets. The primary exception to this is the tornado guidance. Multiple LR configurations exhibit an increase in BSS or AUPDC when compared to their HGBT counterparts. However, as the HGBT outperforms the LR for most hazards, the HGBT was selected as the primary architecture for continued evaluation on the final dataset.

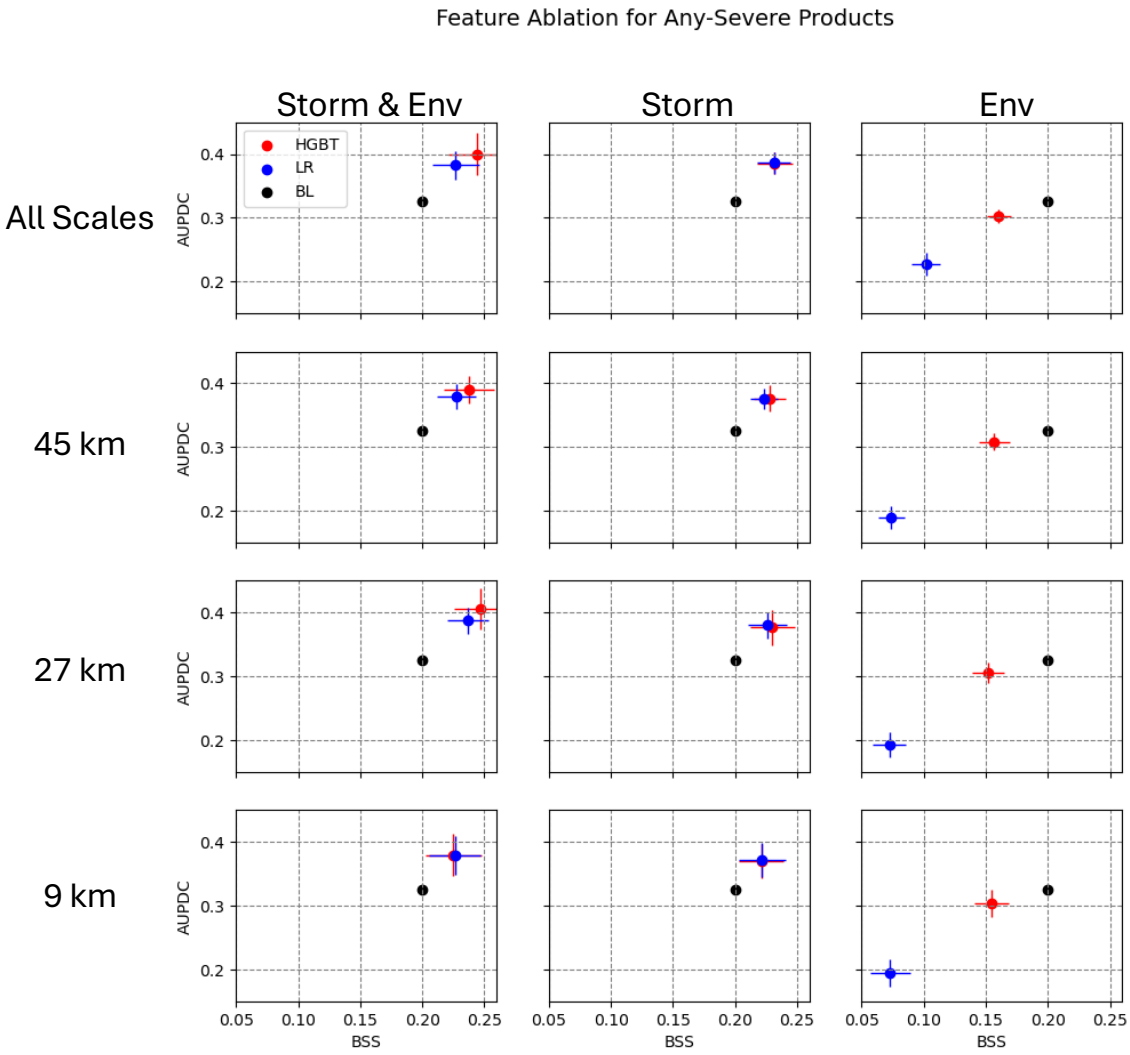

Figure 4.13: Results of feature ablation for any-severe on the initial testing set. Subplots represent various combinations of feature scales and types, with the feature scale (type) varying across the rows (columns). The HGBT (LR) performance using only these features is shown in red (blue). Baselines are retained in every panel as a point of comparison. Error bars indicate the 2σ confidence interval of the ML performance. Intrastorm features generally perform just as well as the configurations using intrastorm and environmental features. No substantial differences exist between the skill of the various scales of predictors.

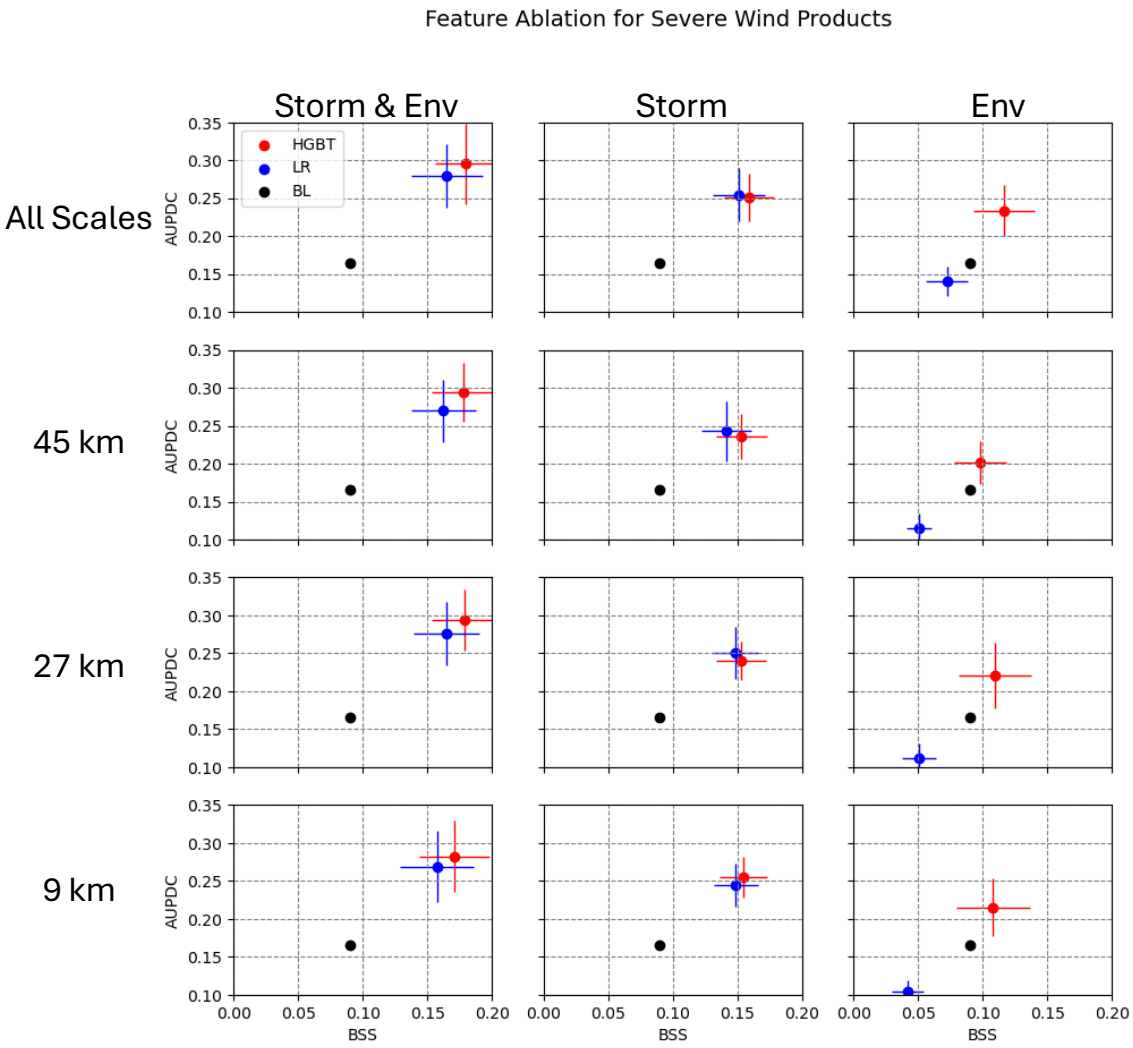

Figure 4.14: As in Fig. 4.13, but for severe wind. Marginal differences are observed between the various predictor scales when intrastorm features are included. The best-performing configuration utilizes both intrastorm and environmental features. The environment-only HGBT beat the baseline while the LR do not. Furthermore, the inclusion of multi-scale predictors does result in a skill increase for the environment-only HGBT. However, this configuration is not as skillful as the configurations using both intrastorm and environmental predictors.

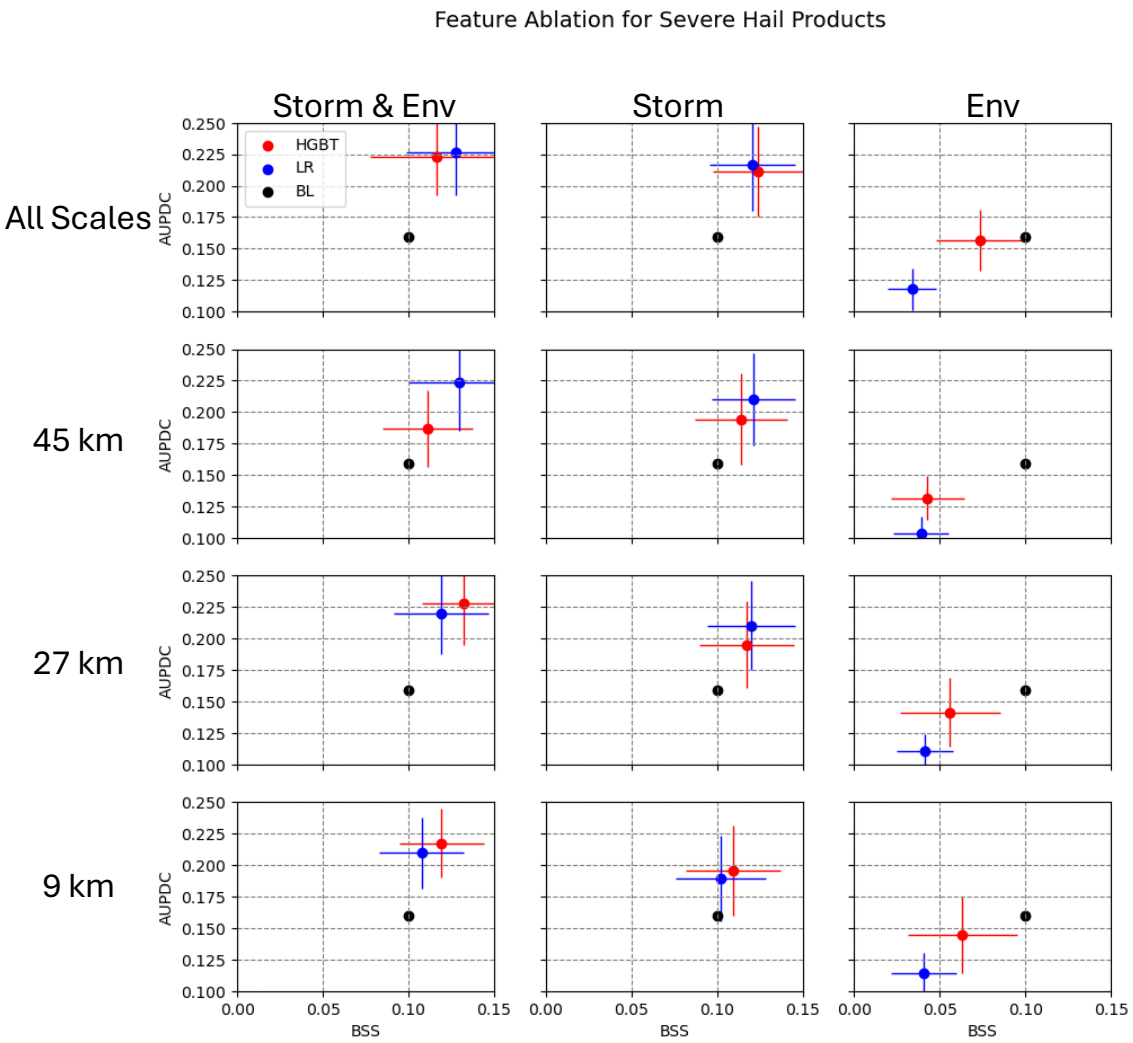

Figure 4.15: As in Fig. 4.13, but for severe hail. All environment-only configurations perform worse than the BL, while all configurations including intrastorm features perform better than the BL. The environment-only HGBT outperforms the environment-only LR. Conversely, the LR and HGBT are similar for most other configurations. The models using multi-scale features generally perform as well as the best-performing, single-scale configuration for a given predictor type.

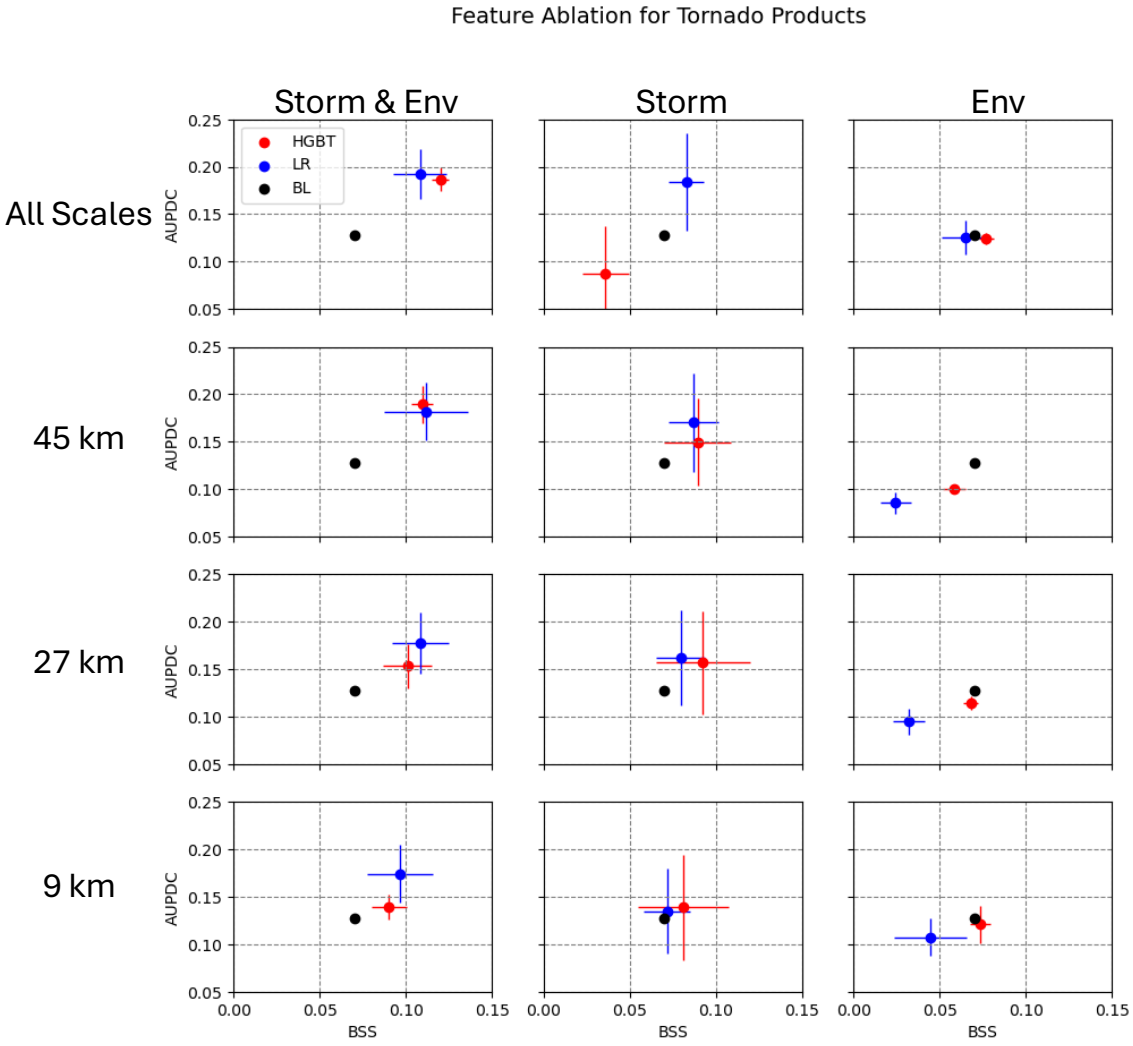

Figure 4.16: As in Fig. 4.13, but for tornadoes. The best-performing configurations again use both intrastorm and environmental predictors. However, little difference is observed between feature scales. The LR and HGBT are again fairly evenly matched, with neither architecture being the most consistently skillful. All environment-only configurations have lower AUPDC than the baseline, but some have a higher BSS.

4.4 Performance of Models on Final Dataset

4.4.1 Objective Skill

We now consider the skill of traditional ML architectures trained on the final dataset (2019-2023). All procedures follow the same process as those in Section 4.1.1. Given the results of the initial dataset, we only evaluate HGBT and the baselines and no longer consider LR.

Figure 4.17 displays the performance curves for HGBT and BL for each of the four hazard sets. The ML predictions achieve both higher POD and higher SR than the BL predictions. These results are not substantially different from those reviewed in Section 4.1.1; the ML is again shown to be more skillful than the BL across all considered hazards and probability thresholds with the largest improvement occurring for severe wind (Figure 4.17b). The HGBT for severe wind and tornadoes achieve their maximum CSI near PODs of 0.4 and 0.3 respectively. However, the any-severe and severe hail ML products maximize CSI when POD is near 0.5. Comparing Figures 4.1b and 4.17b, the initial dataset appears to provide an overly optimistic assessment of the BL performance for severe wind. Furthermore, both the ML and BL tornado products achieved lower skill on the final dataset than they experienced in the initial dataset. Conversely, both the BL and ML for severe hail achieve better performance on the final dataset than the initial dataset.

ROC curves for the ML and BL products are displayed in Figure 4.18. Again, the ranking of results has not changed much from the initial dataset. As measured by the AUC, the HGBT is a better discriminator than the BL across all hazards. It is notable that the discriminative ability of the BL is higher in the final dataset than the initial dataset for all hazards except severe wind. Conversely, all ML products except severe hail exhibit a lower discriminative ability on this dataset. As seen in Figure 4.18c,

the severe hail HGBT generally has a slightly lower POD than the baseline. However, the HGBT also achieves a lower POFD than the baseline. While the highest total discriminative ability belongs to the tornado HGBT product, both the tornado HGBT and BL products have substantially lower PODs than in the initial dataset. Consistent with previous results, the largest improvements are seen for the severe wind and tornado products. Comparing Figure 4.18d to Figure 4.17d, the HGBT's improvement over the baseline is much more apparent in ROC space than in performance space.

The reliability diagrams for the HGBT and BL predictions on the final testing set are shown in Figure 4.19. Most of the ML and BL products exhibit a systemic overprediction bias. As a result, the BSS is lower for all products except severe hail when compared to the initial dataset. This is not unique to the ML products; each of the baselines, except for the tornado baseline, also has an overprediction bias. The severe hail HGBT has the lowest overprediction bias of all four hazard types. While the initial severe hail HGBT learned to output much higher probabilities (4.3c), they were poorly calibrated. Thus, the final severe hail HGBT product is a more reliable but less confident system than the initial. Overall, the magnitude of overprediction remains fairly small for all four hazards. While not plotted, the HGBT products generally remain well above the no-skill line.

4.4.2 Case Study

As before, we now consider a case study. We utilize the guidance created from the 2200 UTC WoFS forecast produced on 10 May 2023. The SPC issued an enhanced risk for this day near the borders of Colorado, Kansas, and Nebraska. Again, the main risks were hail and tornadoes (SPC 2023). As we no longer consider the LR output,

Figure 4.20 shows the baselines and HGBT output for the any-severe and severe wind products while Figure 4.21 shows the baselines and HGBT output for severe hail and tornado.

Beginning with the any-severe guidance, both the baseline and HGBT captured that the hazards were split into two separate clusters. However, the any-severe HGBT correctly identifies the westernmost cluster as having a higher probability than the easternmost; comparatively, the baseline places equal importance on both clusters. As the westernmost cluster was associated with a larger quantity of hail reports and tornadoes, the former behavior is desired. While the baseline does not highlight the threat as well as the HGBT, the baseline does exhibit considerably fewer false alarms over the domain. Notably, the HGBT highlights a region in the southeast of the domain with no associated reports. The baseline still highlights this region but with a lower amplitude.

The severe wind guidance is again primarily dominated by a large false alarm signal. The HGBT correctly highlights the regions where severe wind reports occurred but also exhibits a considerable false alarm over nearly half of the domain. The wind baseline outputs similar probabilities to the HGBT in regions where wind reports occurred but also reduces the spurious signals towards the northern and southern edges of the domain.

The severe hail products have fairly similar performance. Both products correctly identify the clusters of hail reports but have substantial false alarms. While both products capture the cluster of reports in the center of the domain, the HGBT has a more focused maximum. The baseline also extends the elevated probabilities too far north. While the HGBT exhibits a similar coverage of false alarms, the probabilities are generally lower across the region. The HGBT guidance also features higher probabilities near the eastern reports; however, they are disjoint from both clusters of reports. As

such, the HGBT seems to represent the amplitude of the hazard well but is not able to perfect the location.

While both the HGBT and baseline tornado guidance capture the tornadoes that occur in northeast Colorado, the predictions have rather low confidence. The BL has higher amplitude probabilities over the central region, at the cost of a slightly higher false alarm rate. Comparatively, the HGBT guidance has a lower amplitude but represents the central region of interest better. Both the HGBT and baseline have similar false alarms near the eastern edge of the domain, but the HGBT erroneously extends these probabilities farther south. As such, the performance of the tornado guidance for this specific case favors the baseline.

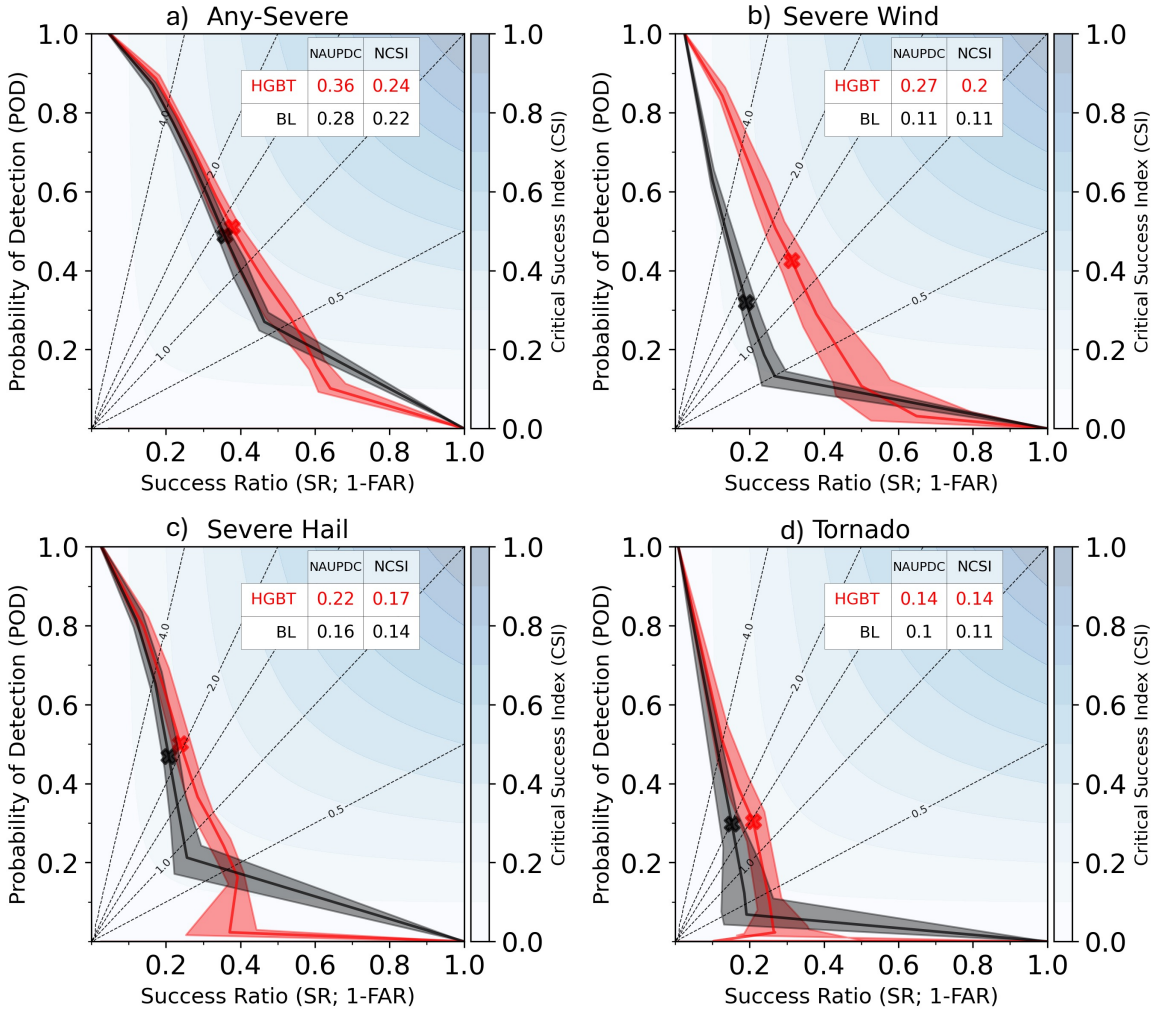

Figure 4.17: Performance diagrams evaluating HGBT (red) and baselines (black) on the final testing set for a) any-severe, b) severe wind, c) severe hail, and d) tornado. Shading indicates the 2σ confidence interval. HGBT again outperforms the BL for every hazard.

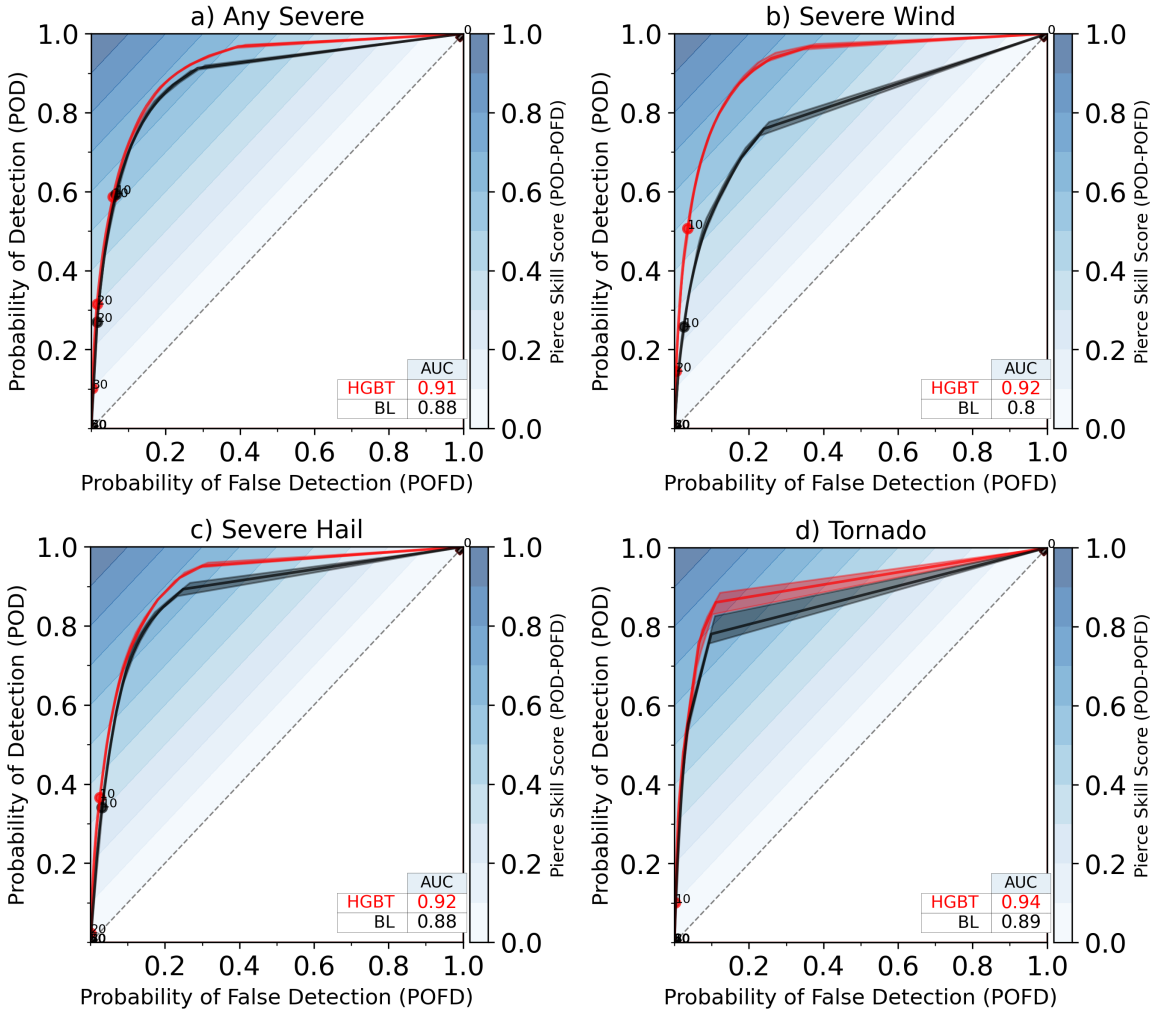

Figure 4.18: ROC diagrams evaluating HGBT (red) and baselines (black) on the final testing set for a) any-severe, b) severe wind, c) severe hail, and d) tornado. Shading indicates the 2σ confidence interval. The largest improvements in discrimination again come from severe wind and tornado; however, the HGBT exhibits lower POD on the final dataset than the initial.

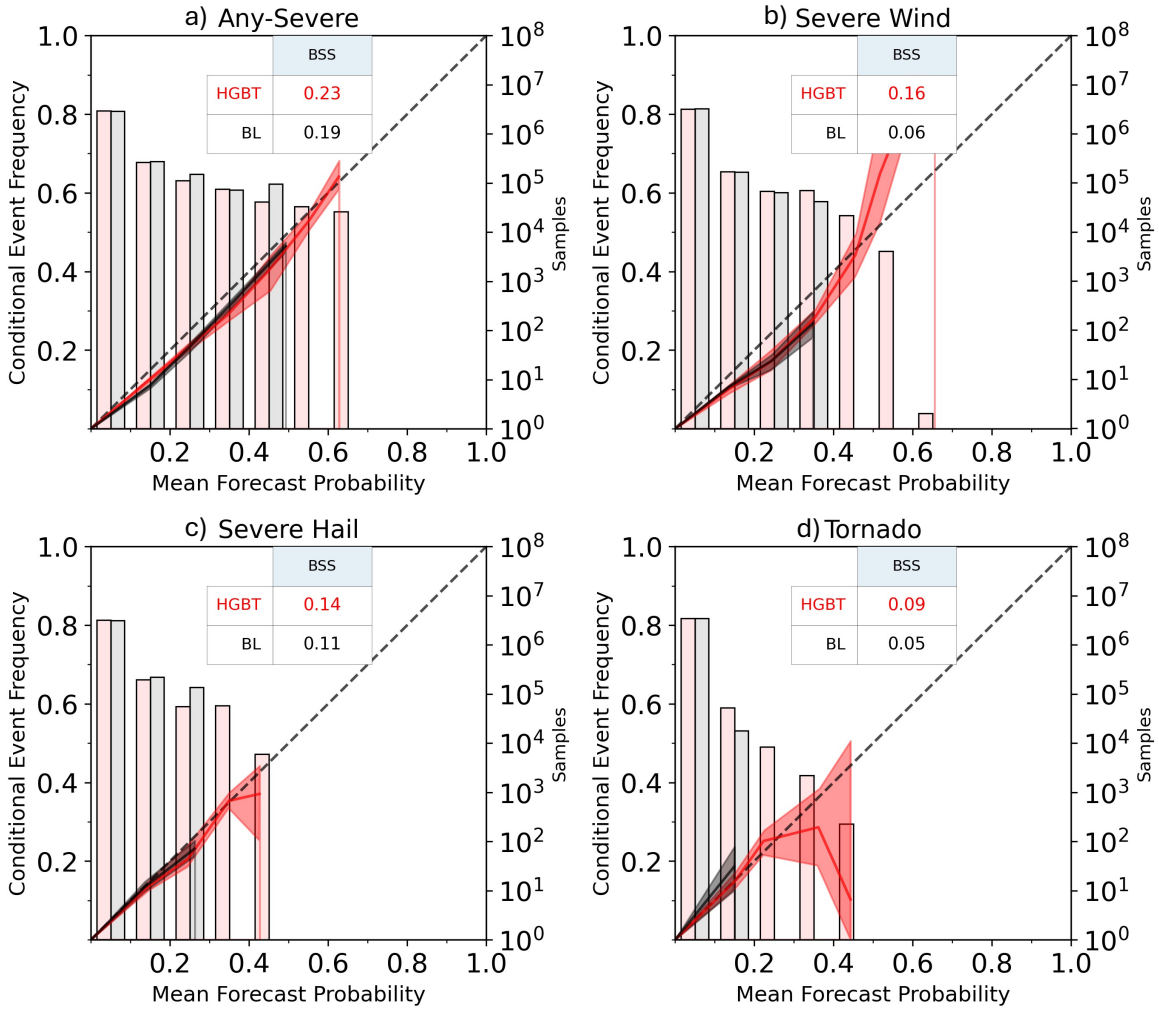

Figure 4.19: Reliability diagrams evaluating HGBT (red) and baselines (black) on the final testing set for a) any-severe, b) severe wind, c) severe hail, and d) tornado. Shading indicates the 2σ confidence interval. With the exception of tornado, both the BL and HGBT have a bias toward overprediction. While the HGBT and BL have similar levels of reliability, the HGBT learns to output higher probabilities. As such, the HGBT BSS exceeds the BL for all hazards.

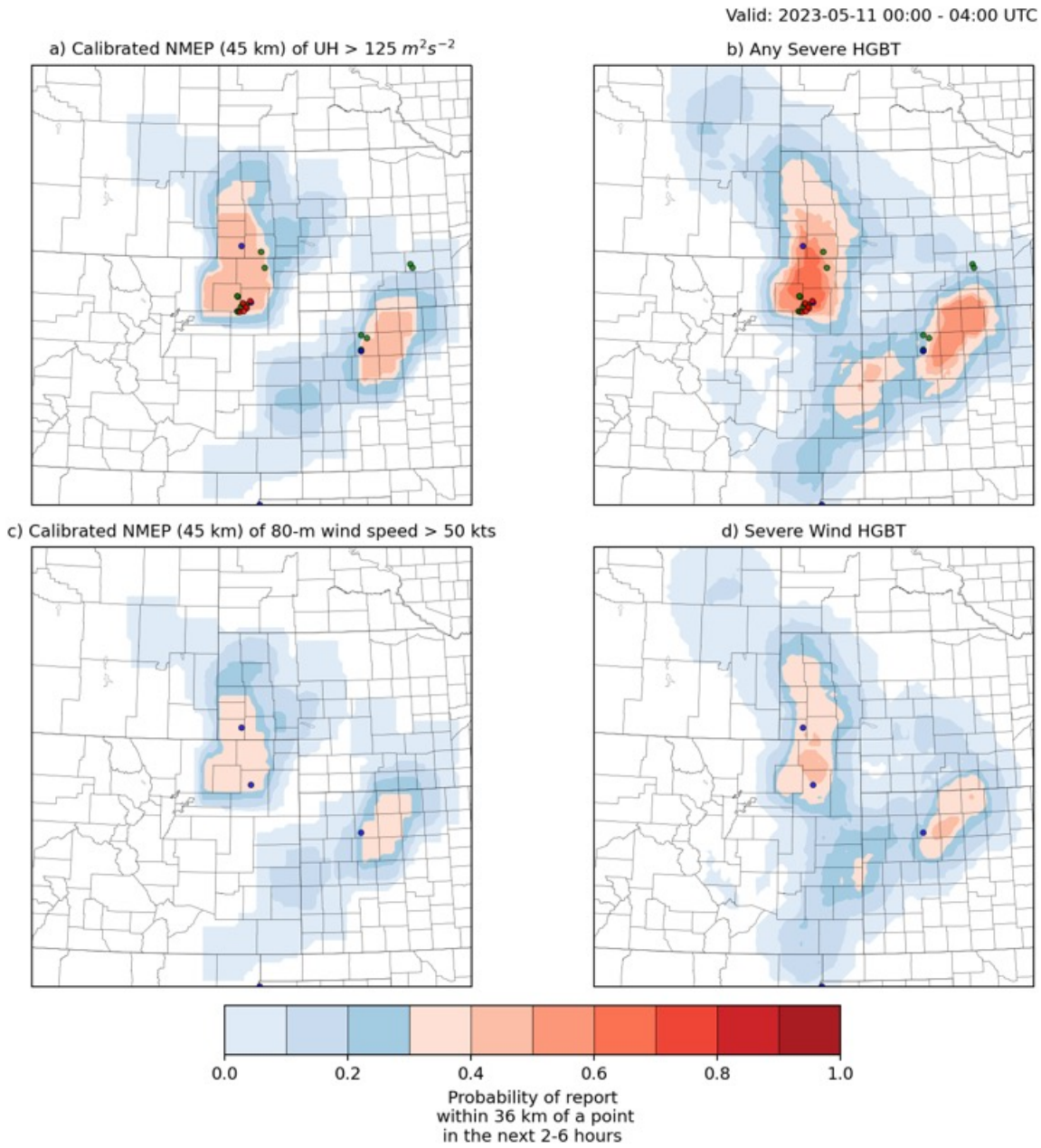

Figure 4.20: Watch-to-warning guidance produced using the WoFS forecast initialized at 2200 UTC on 10 May 2023. The panels correspond to the a) any-severe baseline, b) any-severe HGBT, c) severe wind baseline, and d) severe wind HGBT. The guidance is valid from 00-04 UTC on 11 May 2019. Reports are plotted as in previous figures. While the any-severe HGBT has elevated probabilities near the reports, it also highlights a tertiary false alarm to the south. False alarms again plague both severe wind products.

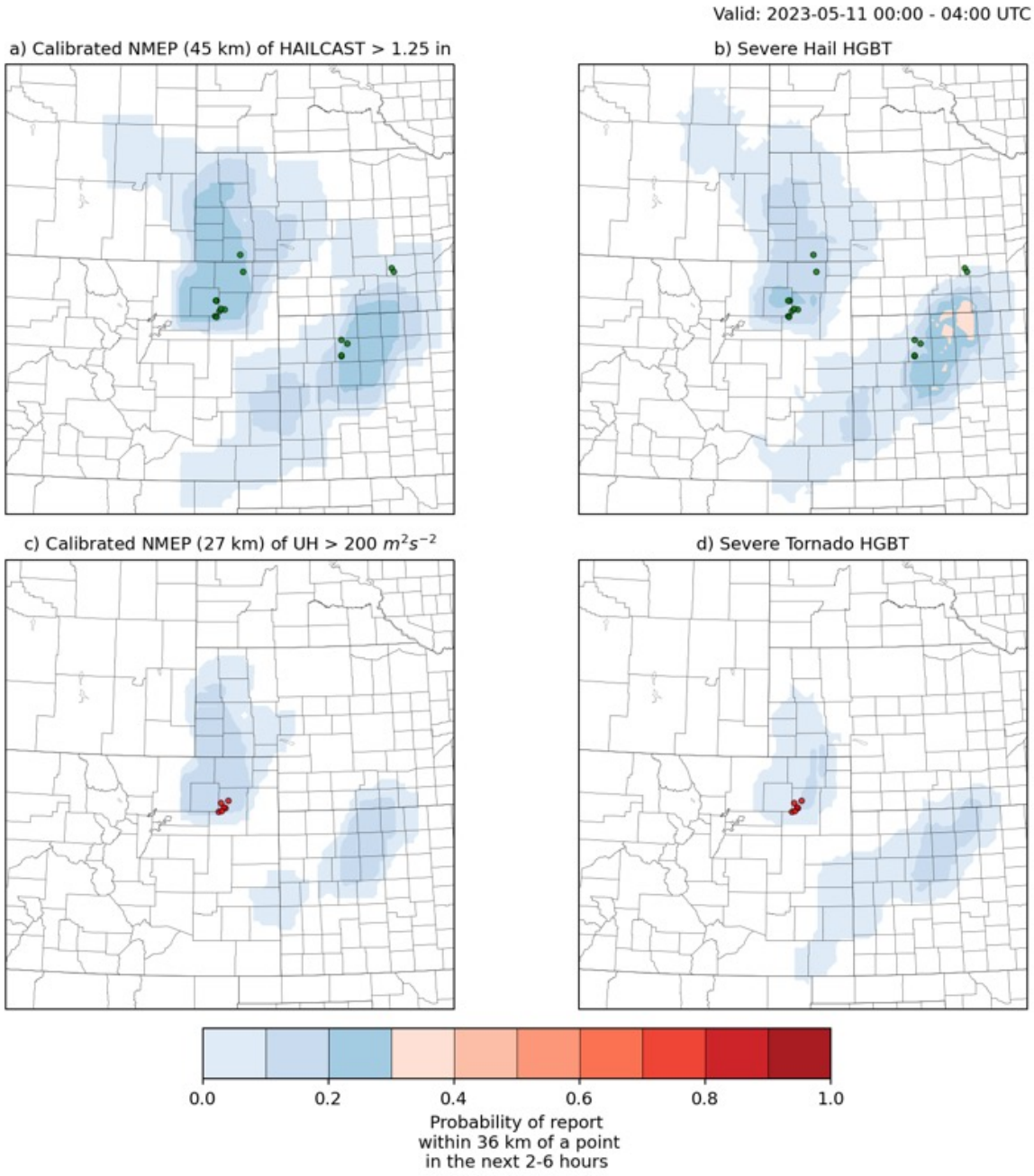

Figure 4.21: As in Fig. 4.20, but for a) severe hail baseline, b) severe hail HGBT, c) tornado baseline, and d) tornado HGBT. While the hail HGBT correctly raises probabilities, they are disjoint from the nearby reports. The tornado HGBT slightly reduces the false alarm in the center of the domain, but increases the false alarm to the south.

4.5 Performance of Deep Learning Techniques

4.5.1 Objective Skill

We now consider the skill of deep learning methods trained on the final dataset. We evaluate these products using two frameworks. Figure 4.22 shows the skill of two U-nets as evaluated on targets consisting of storm data and MESH. Conversely, Figure 4.23 shows the U-nets evaluated on targets only consisting of storm data. The results from Section 4.4 are most comparable to Figure 4.23 as the datasets have similar base rates. The inclusion of MESH in the verification set for Figure 4.22 results in a higher base rate of 6.3% compared to 5.1% when only using reports. The U-net systems share the same architecture and hyperparameters. The first U-net, U-net$_r$ is denoted U-Net (R) in figure captions and is trained to predict only storm data. The second U-net, U-Net$_{R+M}$ is denoted U-Net (R+M) in captions and is trained to predict storm data and MESH. U-nets were not trained for individual hazards; as such, we only discuss the performance of any-severe U-nets.

As expected, U-Net$_{R+M}$ is more skillful than U-Net$_R$ when evaluated on MESH and storm data. U-Net$_{R+M}$ achieves a higher POD and a marginally higher SR when evaluated on MESH and reports (4.22a). However, U-Net$_{R+M}$ is less reliable and more prone to overprediction; this is exacerbated when evaluated without MESH (4.23b). This implies that the most confident predictions from U-Net$_{R+M}$ are based on the MESH targets. This is supported by Figure 4.23a, which shows that U-Net$_{R+M}$ achieves a lower success ratio than U-Net$_R$ at high probabilities when MESH is not included.

U-Net$_R$ has a similar performance to U-Net$_{R+M}$, but with better reliability in both frameworks. The predictions of U-Net$_R$ overlap with MESH at some points, resulting in a higher NAUPDC and BSS when MESH is included. Overall, the U-net is not as confident when trained only using reports. This is indicated by U-Net$_{R+M}$ outputting

high probabilities more often than U-Net$_R$ as seen in Figure 4.23c. As these high probabilities from U-Net$_{R+M}$ are mostly linked to MESH, U-Net$_R$ is not penalized as heavily as U-Net$_{R+M}$ when MESH is removed.

In general, the U-nets are prone to overprediction similar to the traditional ML architectures evaluated. The addition of MESH as a target field drastically increases this tendency, likely due to the model being trained on a dataset with a higher base rate. However, there are two important factors to keep in mind with regards to the U-nets. First, they do not utilize isotonic regression like the other ML and BL products discussed within this work. If isotonic regression was applied in a similar manner to that discussed previously, the reliability of the U-nets would likely improve substantially. This is supported by the U-nets achieving high AUC values; they have the discriminative ability, but the calibration could be improved. Second, U-Net$_R$ achieves similar skill to the HGBT discussed previously in Section 4.4. However, the deep learning framework was substantially faster to implement than the traditional ML framework. This is primarily for two reasons: first, the deep learning dataset is substantially easier to produce as only 63 predictors are created manually. This is a substantial reduction from the 174 predictors of the traditional framework, mainly due to the removal of the smoothing step. Second, the U-nets were substantially faster to train than the HGBT. With access to a graphical processing unit (GPU), the U-nets take on the order of one minute to train. Comparatively, the HGBT takes multiple hours.

4.5.2 Case Study

To assess the deep learning, we consider the same case study discussed previously in Section 4.4.2. As discussed previously, this is intended to show how each architecture performs on a typical case. Figure 4.24 displays the baseline and HGBT guidance

discussed previously, as well as the output of U-net$_R$ and U-net$_{R+M}$. U-net$_R$ correctly reduces the probabilities in the northern region of the domain when compared to the HGBT guidance. Additionally, the central maximum produced by U-net$_R$ is more closely limited to the reports than the HGBT and BL guidance. However, the false alarm in the southern central region of the domain is substantially amplified compared to the baseline and HGBT.

Comparatively, the guidance from U-net$_{R+M}$ substantially raises the false alarm rate in the south compared to the other products. As expected given the reliability shown in Figure 4.23, U-net$_{R+M}$ exhibits a severe overprediction bias. Additionally, while U-net$_R$ correctly lowered the probabilities in the north of the domain, U-net$_{R+M}$ has the highest amplitude false alarm of all the guidance evaluated. Rather than two disparate regions of elevated probabilities, U-net$_{R+M}$ joins the regions into a continuous stretch of probabilities. Even with the inclusion of MESH targets in the evaluation (indicated by the black contours in 4.24), the signal from U-net$_{R+M}$ still results in a substantial false alarm in the central region.

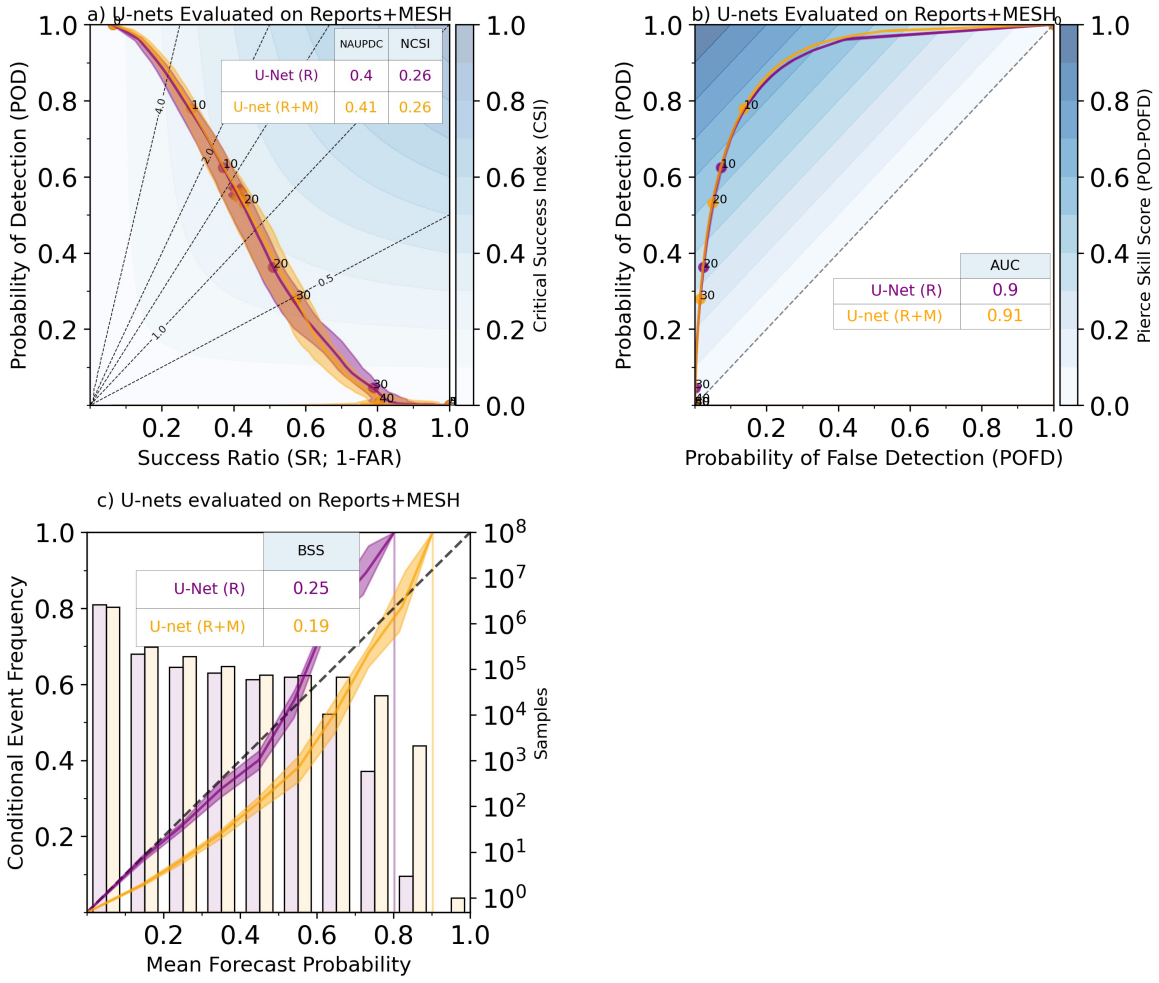

Figure 4.22: A performance diagram (a), reliability diagram (b), and ROC diagram (c) showing the skill of U-nets trained to predict storm reports (purple) and storm reports and MESH (yellow) evaluated using storm reports and MESH as targets. Shading indicates the 2σ confidence interval. While the U-net trained with MESH and reports exhibits a higher POD, it is substantially less reliable than the U-net trained only with reports.

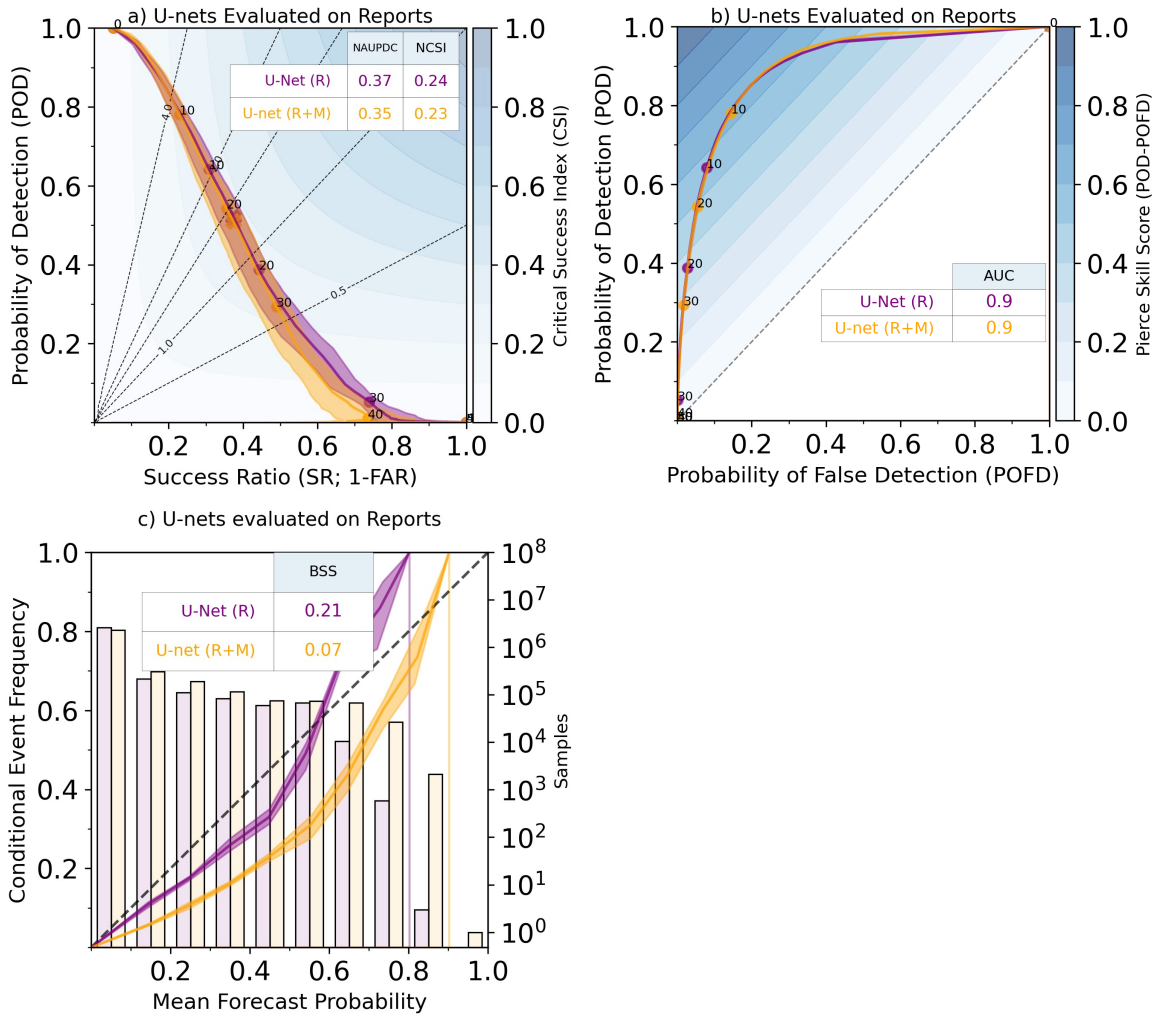

Figure 4.23: A performance diagram (a), reliability diagram (b), and ROC diagram (c) showing the skill of U-nets trained to predict storm reports (purple) and storm reports and MESH (yellow) evaluated using only storm reports as targets. Shading indicates the 2σ confidence interval. While the U-net trained with reports and MESH retains a higher POD, it suffers from a lower success ratio and worse reliability.

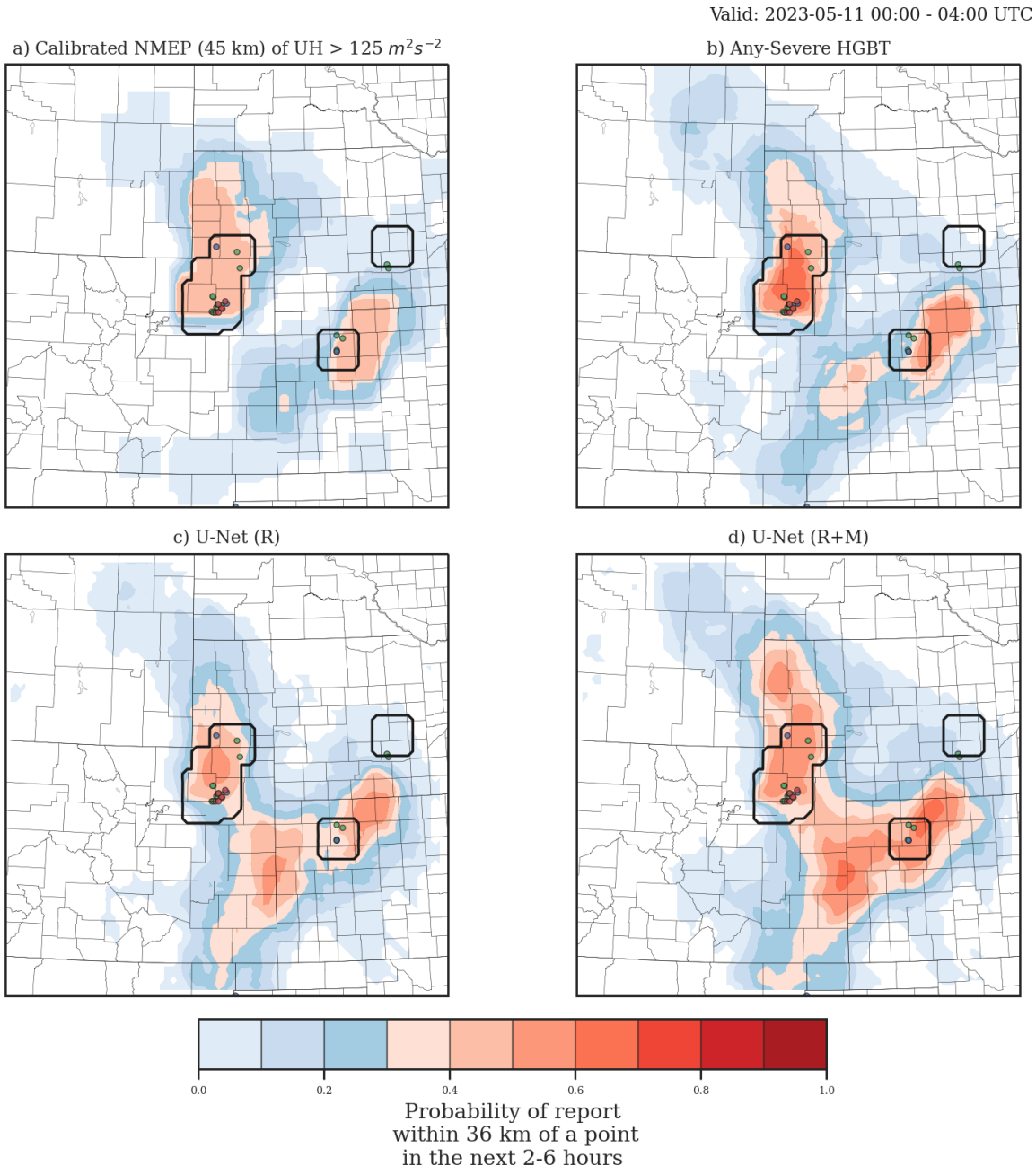

Figure 4.24: Watch-to-warning guidance produced using the WoFS forecast initialized at 2200 UTC on 10 May 2023. The panels correspond to the any-severe baseline (top left), any-severe HGBT (top right), U-Net$_R$ (bottom left), and U-Net$_{R+M}$. The guidance is valid from 00-04 UTC on 11 May 2019. Reports are plotted as in previous figures. For this case, both U-nets substantially increase the false alarms in the southern central region of the domain. U-Net$_R$ slightly reduces the false alarms in the northern half of the domain.

Chapter 5

Conclusions and Summary

5.1 Discussion

Following the results discussed previously, it is apparent that value is gained by utilizing machine learning to produce guidance during the watch-to-warning period. Each of the machine learning systems discussed within were capable of providing some benefit over the NMEP baselines. The logistic regression model is skillful and rather straightforward. The model is reasonably easy to interpret, as the feature weights give an indication of feature importance. However, the training process slows down for large datasets compared to the HGBT and deep learning architectures. The HGBT generally had the highest performance metrics and was second only to the deep learning in speed. As such, we chose to implement the HGBT as the architecture of choice in the real-time pipeline that runs alongside the WoFS. While the U-net had similar performance, we chose the HGBT as the features are created manually rather than implicitly.

The performance of the deep learning architectures is both inspiring and disheartening. The U-net can achieve the same skill as the traditional ML architectures but takes a fraction of the time to train when utilizing GPUs. Additionally, this system was substantially easier to implement than the explicit multi-scale predictors utilized by the traditional ML. As such, these types of networks could greatly expedite any future work. However, despite having a much more complex architecture, leveraging

explicit spatial information, and additional predictors, the U-nets did not surpass the performance of the traditional ML architectures. This suggests one primary possibility. Namely, the architectures evaluated within this work likely all exist within the context of a larger Rashomon Set. A Rashomon Set is a collection of models that all complete the same task at approximately equal levels of skill regardless of the various methods employed (Fisher et al. 2019). While the performance of the various models evaluated within this work asymptotically approaches a single level of skill, it remains unclear whether this is a local predictability limit or the true predictability limit. If the limit is local, it likely arises from imperfections in the dataset. Spurious convection and phase errors within the forecasts, as well as any other underlying limitations of WoFS, limit the machine learning's ability to correlate severe storms within the forecast to reports. Additionally, missing reports will limit the ability to discriminate which forecast storms are likely to become severe. In the case of a local bottleneck, substantial future effort may result in some level of increased skill. However, if this is the true predictability limit, then the guidance will be unable to surpass this bottleneck.

One of the primary issues with all of the guidance types discussed within this work is the false alarm rate. False alarms in severe weather guidance are correlated with an increase in weather-based casualties (Simmons and Sutter 2009). Cases such as the severe wind guidance on 23 May 2019 and 10 May 2023 depict the elevated false alarm threat well (e.g., Figures 4.6 and 4.20 respectively). While some false alarms are inevitable due to the nature of rare event predictions, it is not uncommon for large regions of the domain to have slightly elevated probabilities that do not verify. Additionally, if WoFS forecasts have spurious or displaced storms, the guidance will erroneously provide high probabilities for these regions (as seen in the south of the domain in Figure 4.20). These issues degrade the quality of the guidance, as well as confidence in the product, and as such, future efforts should be made to reduce them.

However, a decrease in false alarms should not be traded for a decrease in detection (see Simmons and Sutter 2009). Additionally, changes to the guidance must be made with caution as favoring one aspect of the predictions may degrade others (Murphy 1993).

Furthermore, it is currently unclear how many of the false alarms are correct (e.g., no severe weather occurred) versus incorrect (e.g., severe weather occurred but was not observed). Section 4.5 shows how the verification of the models' performance can be volatile based on what data are included in the verification set. Systems that learn meteorological signals may be penalized unfairly when verified against data that contain implicit social signals (e.g., the observed decrease in reliability of U-net$_{R+M}$ when MESH is not included in the evaluation). As such, any future work or verification must carefully consider the potential flaws of any data used within the verification process.

The current WoFS ML products operate at a shorter lead time and are shown to have higher skill, as would be expected (Flora et al. 2021). As such, the 2-6 hour guidance can serve as an initial first guess for regions where severe weather may develop. As the event grows closer, users can transition to alternative shorter-range, higher-skilled products available within the WoFS suite. Given the extra lead time of the watch-to-warning products and the associated risk of false alarms, further consideration of the end user is necessary. The impact of extended lead times on the general public's risk response is still unclear (Simmons and Sutter 2008; Hoekstra et al. 2011). Thus, it may be preferable to target this product toward an audience of decision-makers (e.g., Weather Forecast Offices) rather than the general public. Toward that end, the HGBT guidance discussed within this work was included for evaluation within the 2024 Hazardous Weather Testbed Spring Forecasting Experiment. As the results of

this experiment are not yet available, it remains to be seen what value users find in the ML-based watch-to-warning guidance.

5.2 Summary and Future Work

While current WoFS ML products are focused on lead times of 0-3 hours, the work discussed within this study represents the first WoFS ML product that targets lead times later in the watch-to-warning period (2-6 hours). It is also the first WoFS ML product that evaluates deep learning for producing severe weather guidance. Outside of the scope of WoFS products, this study is likely the first ML-based 2-6 hour product for severe weather guidance and is the first to be evaluated within the HWT SFE.

The data used to create predictors primarily come from WoFS ensemble forecasts from the Hazardous Weather Testbed Spring Forecasting Experiments. Initial experiments were conducted on a dataset consisting of forecasts from 2018-2021. However, a transition was made to a larger dataset consisting of data from 2019-2023 when it became available.

Predictors are created using both intrastorm and environmental fields from WoFS forecasts regridded to 9 km. Data from forecast hours 2-6 are used to produce a time composite of each field. Ensemble statistics are then calculated per field after smoothing the time composite at various resolutions. Thus, the predictors incorporate both spatial and ensemble information. The target field is binary, with all grid points within 36 km of a storm data report being mapped to the positive class. We then evaluate the skill of various architectures at predicting this class. Four categories of guidance are considered: any-severe, severe wind, severe hail, and tornado.

Three different machine learning architectures are evaluated within this work: logistic regression (LR), histogram-based gradient-boosting trees (HGBT), and U-nets.

The LR and HGBT make predictions in a tabular framework, while the U-net retains the spatial structure of the original fields when making predictions. Additionally, data from the MRMS suite is used to augment the WoFS data used by the U-nets. The skill of ML-based guidance is evaluated against rigorous non-ML baselines tuned for each hazard.

The findings of this work indicate that all ML architectures evaluated are capable of outperforming the non-ML baselines. This holds for each of the hazards considered. The ML generally provides improved skill and reliability when compared to the baseline. Of the traditional ML architectures evaluated, HGBT is both the highest-performing and most efficient. For this reason, HGBT is selected as the primary architecture of choice. While LR outperformed the baseline, it did not match the skill of the HGBT or the U-nets. U-nets had similar levels of skill to the HGBT and were substantially faster to train with access to a GPU. However, they also exhibited lower reliability. Additionally, despite incorporating additional predictors and target fields from the MRMS suite, U-nets were not capable of outperforming the HGBT when evaluated on storm reports.

The any-severe guidance generally had the highest performance, followed by the severe wind, severe hail, and tornado products. Consequently, the ML saw the largest improvement over the baselines for severe wind, followed by severe hail, tornado, and any-severe guidance. These improvements were robust to changes in the testing dataset and resulted in the ML generally outperforming the baseline for all initialization times.

Feature ablation reveals that the inclusion of multi-scale predictors had little impact on the skill of the best-performing models. Guidance created with access to only one scale of information was often just as skillful as guidance created with access to all considered scales. The largest factor in the skill of the models was the inclusion of intrastorm fields as predictors. The primary exception to this is severe wind, where

the most skillful guidance resulted from the use of both intrastorm and environmental predictors. Analysis of the U-net's performance shows that the evaluation process is highly sensitive to the data used as verification of severe weather. Models that have similar performance when MESH is included in the verification may have substantial differences in performance metrics when MESH is removed from the verification. As such, any future work should carefully consider the implications and innate biases of the verification dataset.

The immediate future of this project will largely be driven by the results of the 2024 Hazardous Weather Testbed Spring Forecasting Experiment as the 2024 SFE was the first exposure of these ML products to a broader user base. A long-term goal will likely focus on reducing the false alarm ratio of the guidance. However, the most efficient way to do that is currently unclear. Additionally, in the near future, the product must undergo more rigorous peer review to help ensure the guidance is fit for availability to the general public.

One alternative pathway is to shift away from predicting reports as a purely meteorological function and to instead predict reports as a function of both social and meteorological information. The current WoFS ML watch-to-warning guidance uses solely meteorological data. However, other systems, such as Nadocast, have shown promising results by including both NWP forecast data and data related to the climatology of reports (Hempel 2022). Adopting this approach would require further discussion as to the purpose of the ML guidance; namely, whether it intends to predict reports as a proxy for severe weather, or as a proxy for the observance of severe weather. Alternatively, future work may consider a complete redesign of the framework discussed herein. Specifically, the deep learning approach was fairly rudimentary and could likely be optimized better. However, given the results of this study, it is likely

that a substantial undertaking would be required to pass the bottleneck of skill and break out of the Rashomon set.

Finally, given the recent development of AI-based NWP systems, such as Google's GraphCast (Lam et al. 2022), it is unclear what role similar studies will have in the future. While published AI NWP models are not yet run at a convection-permitting scale, it is feasible that AI-based convection-allowing models will be developed in the near future. This could potentially lead to a regime shift when it comes to products created by post-processing NWP output, such as the inclusion of these products directly within the forecast system.

In conclusion, recent developments within the field may yield substantial changes in the near future. While other products may be developed to fill a similar niche, the work within this study represents the first WoFS-based ML product that targets extended lead times (2-6 hours). As such, only this product makes use of the unique capabilities of the WoFS at those lead times. As the WoFS continues to evolve and improve, the ML guidance will continue to advance alongside it.

Reference List

Bentéjac, C., A. Csörgő, and G. Martínez-Muñoz, 2021: A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, **54**, 1937–1967.

Breiman, L., 2017: *Classification and regression trees*. Routledge.

Brooks, H. E., 2004: Tornado-warning performance in the past and future: A perspective from signal detection theory. *Bulletin of the American Meteorological Society*, **85 (6)**, 837–844.

Brooks, H. E., and J. Correia Jr, 2018: Long-term performance metrics for national weather service tornado warnings. *Weather and Forecasting*, **33 (6)**, 1501–1511.

Brooks, H. E., M. L. Flora, and M. E. Baldwin, 2024: A rose by any other name: On basic scores from the 2× 2 table and the plethora of names attached to them. *Artificial Intelligence for the Earth Systems*, **3 (2)**, e230 104.

Brotzge, J. A., S. E. Nelson, R. L. Thompson, and B. T. Smith, 2013: Tornado probability of detection and lead time as a function of convective mode and environmental parameters. *Weather and forecasting*, **28 (5)**, 1261–1276.

Burke, A., N. Snook, D. J. Gagne II, S. McCorkle, and A. McGovern, 2020: Calibration of machine learning–based probabilistic hail predictions for operational forecasting. *Weather and Forecasting*, **35 (1)**, 149–168.

Calhoun, K. M., P. A. Campbell, R. B. Steeves, C. N. Satrio, T. Sandmael, E. D. Loken, M. K. Silcott, and P. C. Burke, 2024: Testing the future of storm-based probabilistic hazard creation and communication across the watch to warning paradigm. *51st Conference on Broadcast Meteorology/Seventh Conference on Weather Warnings and Communication*, AMS.

Cintineo, J. L., M. J. Pavolonis, and J. M. Sieglaff, 2022: Probsevere lightningcast: A deep-learning model for satellite-based lightning nowcasting. *Weather and Forecasting*, **37 (7)**, 1239–1257.

Cintineo, J. L., M. J. Pavolonis, J. M. Sieglaff, L. Cronce, and J. Brunner, 2020: Noaa probsevere v2. 0—probhail, probwind, and probtor. *Weather and Forecasting*, **35 (4)**, 1523–1543.

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. *Weather and Forecasting*, **29 (3)**, 639–653.

Cintineo, J. L., and Coauthors, 2018: The noaa/cimss probsevere model: Incorporation of total lightning and validation. *Weather and Forecasting*, **33 (1)**, 331–345.

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

Clark, A. J., and Coauthors, 2023: The first hybrid noaa hazardous weather testbed spring forecasting experiment for advancing severe weather prediction. *Bulletin of the American Meteorological Society*, **104 (12)**, E2305–E2307.

Farney, T. J., and P. G. Dixon, 2015: Variability of tornado climatology across the continental united states. *International Journal of Climatology*, **35 (10)**.

Fisher, A., C. Rudin, and F. Dominici, 2019: All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, **20 (177)**, 1–81.

Flora, M. L., B. Gallo, C. K. Potvin, A. J. Clark, and K. Wilson, 2024: Exploring the usefulness of machine learning severe weather guidance in the warn-on-forecast system: Results from the 2022 noaa hazardous weather testbed spring forecasting experiment. *Weather and Forecasting*.

Flora, M. L., C. K. Potvin, P. S. Skinner, S. Handler, and A. McGovern, 2021: Using machine learning to generate storm-scale probabilistic guidance of severe weather hazards in the warn-on-forecast system. *Monthly Weather Review*, **149 (5)**, 1535–1557.

Friedman, J. H., 2001: Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189–1232.

Gagne II, D. J., S. E. Haupt, D. W. Nychka, and G. Thompson, 2019: Interpretable deep learning for spatial analysis of severe hailstorms. *Monthly Weather Review*, **147 (8)**, 2827–2845.

Gallo, B. T., A. J. Clark, I. Jirak, D. Imy, B. Roberts, J. Vancil, K. Knopfmeier, and P. Burke, 2024: Wofs and the wisdom of the crowd: The impact of the warn-on-forecast system on hourly forecasts during the 2021 noaa hazardous weather testbed spring forecasting experiment. *Weather and Forecasting*, **39 (3)**, 485–500.

Gallo, B. T., and Coauthors, 2022: Exploring the watch-to-warning space: Experimental outlook performance during the 2019 spring forecasting experiment in noaa's hazardous weather testbed. *Weather and Forecasting*, **37 (5)**, 617–637.

Géron, A., 2022: *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow.* " O'Reilly Media, Inc.".

Heinselman, P., and Coauthors, 2024a: Forecasters' use of warn-on-forecast system probabilistic hazard information (wofs-phi) during the 2023 hwt watch-to-warning experiment. *104th AMS Annual Meeting*, AMS.

Heinselman, P. L., and Coauthors, 2024b: Warn-on-forecast system: From vision to reality. *Weather and Forecasting*, **39 (1)**, 75–95.

Hempel, B., 2022: Nadocast—conus severe weather probabilities via feature engineering and gradient boosted decision trees. *GitHub*.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting severe weather with random forests. *Monthly Weather Review*, **148 (5)**, 2135–2161.

Hill, A. J., and R. S. Schumacher, 2021: Forecasting excessive rainfall with random forests and a deterministic convection-allowing model. *Weather and Forecasting*, **36 (5)**, 1693–1711.

Hill, A. J., R. S. Schumacher, and I. L. Jirak, 2023: A new paradigm for medium-range severe weather forecasts: Probabilistic random forest–based predictions. *Weather and Forecasting*, **38 (2)**, 251–272.

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 public's perceptions of weather risks. *Weather, Climate, and Society*, **3 (2)**, 128–140.

Huang, H., and Coauthors, 2020: Unet 3+: A full-scale connected unet for medical image segmentation. *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, 1055–1059.

James, P. M., B. K. Reichert, and D. Heizenreder, 2018: Nowcastmix: Automatic integrated warnings for severe convection on nowcasting time scales at the german weather service. *Weather and Forecasting*, **33 (5)**, 1413–1433.

Jirak, I. L., C. J. Melick, and S. J. Weiss, 2016: Comparison of the spc storm-scale ensemble of opportunity to other convection-allowing ensembles for severe weather forecasting. Portland, OR, Amer. Meteor. Soc., 102" pp.

Jirak, I. L., S. J. Weiss, and C. J. Melick, 2012: *The SPC storm-scale ensemble of opportunity: Overview and results from the 2012 Hazardous Weather Testbed spring forecasting experiment*. Nashville, TN, Amer. Meteor. Soc., 137" pp.

Kramer, M. A., 1991: Nonlinear principal component analysis using autoassociative neural networks. *AIChE journal*, **37 (2)**, 233–243.

Lagerquist, R., A. McGovern, and T. Smith, 2017: Machine learning for real-time prediction of damaging straight-line convective wind. *Weather and Forecasting*, **32 (6)**, 2175–2193.

Lam, R., and Coauthors, 2022: Graphcast: Learning skillful medium-range global weather forecasting. *arXiv preprint arXiv:2212.12794*.

LeCun, Y., Y. Bengio, and G. Hinton, 2015: Deep learning. *nature*, **521 (7553)**, 436–444.

LeCun, Y., Y. Bengio, and Coauthors, 1995: Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, **3361 (10)**, 1995.

LeCun, Y., B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, 1989: Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, **2**.

Loken, E. D., A. J. Clark, and C. D. Karstens, 2020: Generating probabilistic next-day severe weather forecasts from convection-allowing ensembles using random forests. *Weather and Forecasting*, **35 (4)**, 1605–1631.

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

McGovern, A., R. J. Chase, M. Flora, D. J. Gagne, R. Lagerquist, C. K. Potvin, N. Snook, and E. Loken, 2023: A review of machine learning for convective weather. *Artificial Intelligence for the Earth Systems*, **2 (3)**, e220 077.

Miller, W. J., and Coauthors, 2022: Exploring the usefulness of downscaling free forecasts from the warn-on-forecast system. *Weather and Forecasting*, **37 (2)**, 181–203.

Murphy, A. H., 1993: What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8 (2)**, 281 – 293, https://doi.org/10.1175/1520-0434(1993)008⟨0281:WIAGFA⟩2.0.CO;2.

Obermeier, H. B., K. L. Berry, and J. E. Trujillo-Falcón, 2023: Understanding broadcast meteorologists' current and future use of severe weather watches, warnings, and probabilistic hazard information. *Weather, climate, and society*, **15 (4)**, 893–907.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, **12**, 2825–2830.

Platt, J., and Coauthors, 1999: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, **10 (3)**, 61–74.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, Springer, 234–241.

Sha, Y., R. A. Sobash, and D. J. Gagne, 2024: Generative ensemble deep learning severe weather prediction from a deterministic convection-allowing model. *Artificial Intelligence for the Earth Systems*, **3 (2)**, e230 094.

Simmons, K. M., and D. Sutter, 2006: Improvements in tornado warnings and tornado casualties. *International Journal of Mass Emergencies & Disasters*, **24 (3)**, 351–369.

Simmons, K. M., and D. Sutter, 2008: Tornado warnings, lead times, and tornado casualties: An empirical investigation. *Weather and Forecasting*, **23 (2)**, 246–258.

Simmons, K. M., and D. Sutter, 2009: False alarms, tornado warnings, and tornado casualties. *Weather, Climate, and Society*, **1 (1)**, 38–53.

Skinner, P. S., and Coauthors, 2018: Object-based verification of a prototype warn-on-forecast system. *Weather and forecasting*, **33 (5)**, 1225–1250.

Smith, T. M., and Coauthors, 2016: Multi-radar multi-sensor (mrms) severe weather and aviation products: Initial operating capabilities. *Bulletin of the American Meteorological Society*, **97 (9)**, 1617–1630.

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. *Weather and Forecasting*, **35 (5)**, 1981–2000.

Sobash, R. A., C. S. Schwartz, G. S. Romine, K. R. Fossell, and M. L. Weisman, 2016: Severe weather prediction using storm surrogates from an ensemble forecasting system. *Weather and Forecasting*, **31 (1)**, 255–271.

SPC, 2019: Day 1 outlook and prelim. reports valid: 1630 utc 05/23/2019 to 1200 utc 05/24/2019. NOAA, accessed 29 June 2024, https://www.spc.noaa.gov/products/outlook/archive/2019/day1otlk_v_20190523_1630.gif.

SPC, 2023: Day 1 outlook and prelim. reports valid: 1630 utc 05/10/2023 to 1200 utc 05/11/2023. NOAA, accessed 29 June 2024, https://www.spc.noaa.gov/products/outlook/archive/2023/day1otlk_v_20230510_1630.gif.

Stensrud, D. J., and Coauthors, 2009: Convective-scale warn-on-forecast system: A vision for 2020. *Bulletin of the American Meteorological Society*, **90 (10)**, 1487–1500.

Stensrud, D. J., and Coauthors, 2013: Progress and challenges with warn-on-forecast. *Atmospheric Research*, **123**, 2–16.

Wendt, N. A., and I. L. Jirak, 2021: An hourly climatology of operational mrms mesh-diagnosed severe and significant hail with comparisons to storm data hail reports. *Weather and Forecasting*, **36 (2)**, 645–659.

Wilson, K. A., 2023: The noaa weather prediction center's use and evaluation of experimental warn-on-forecast system guidance. *J. Oper. Meteor.*, **11**, 82–94.

Wilson, K. A., and Coauthors, 2024: Collaborative exploration of storm-scale probabilistic guidance for nws forecast operations. *Weather and Forecasting*, **39 (2)**, 387–402.

Witt, A., M. D. Eilts, G. J. Stumpf, J. Johnson, E. D. W. Mitchell, and K. W. Thomas, 1998: An enhanced hail detection algorithm for the wsr-88d. *Weather and Forecasting*, **13 (2)**, 286–303.

Zhou, Z., M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, 2018: Unet++: A nested u-net architecture for medical image segmentation. *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4*, Springer, 3–11.

Zinner, T., H. Mannstein, and A. Tafferner, 2008: Cb-tram: Tracking and monitoring severe convection from onset over rapid development to mature phase using multi-channel meteosat-8 seviri data. *Meteorology and Atmospheric Physics*, **101**, 191–210.